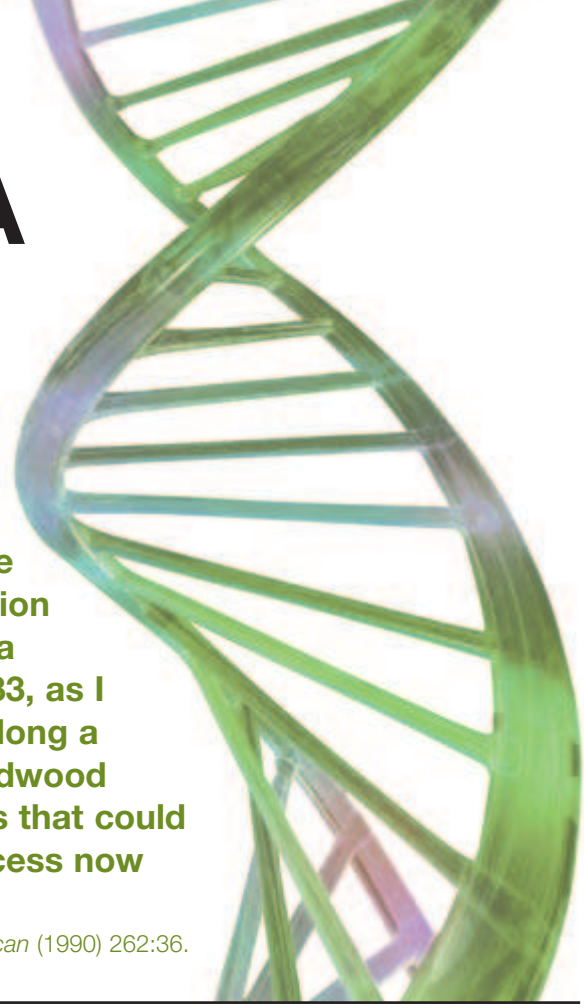
Chapter 8

Recombinant DNA technology and molecular cloning



Sometimes a good idea comes to you when you are not looking for it. Through an improbable combination of coincidences, naiveté and lucky mistakes, such a revelation came to me one Friday night in April, 1983, as I gripped the steering wheel of my car and snaked along a moonlit mountain road into northern California's redwood country. That was how I stumbled across a process that could make unlimited numbers of copies of genes, a process now known as the polymerase chain reaction (PCR).

Kary B. Mullis, *Scientific American* (1990) 262:36.

Outline

8.1 Introduction

The cornerstone of most molecular biology technologies is the gene. To facilitate the study of genes, they can be isolated and amplified. One method of isolation and amplification of a gene of interest is to clone the gene by inserting it into another DNA molecule that serves as a vehicle or vector that can be replicated in living cells. When these two DNAs of different origin are combined, the result is a recombinant DNA molecule. Although genetic processes such as crossing-over technically produce recombinant DNA, the term is generally reserved for DNA molecules produced by joining segments derived from different biological sources. The recombinant DNA molecule is placed in a host cell, either prokaryotic or eukaryotic. The host cell then replicates (producing a clone), and the vector with its foreign piece of DNA also replicates. The foreign DNA thus becomes amplified in number, and following its amplification can be purified for further analysis.

8.2 Historical perspective

In the early 1960s, before the advent of gene cloning, studies of genes often relied on indirect or fortuitous discoveries, such as the ability of bacteriophages to incorporate bacterial genes into their genomes. For example, a strain of phage phi 80 with the *lac* operator incorporated into its genome was used to demonstrate that the Lac repressor binds specifically to this DNA sequence (see Fig. 10.12). The synthesis of many disparate experimental observations into recombinant DNA technology occurred between 1972 and 1975, through the efforts of several research groups working primarily on bacteriophage lambda (λ).

Insights from bacteriophage lambda (λ) cohesive sites

In 1962, Allan Campbell noted that the linear genome of bacteriophage λ forms a circle upon entering the host bacterial cell, and a recombination (breaking and rejoining) event inserts the phage DNA into the host chromosome. Reversal of the recombination event leads to normal excision of the phage DNA. Rare excision events at different places can result in the incorporation of nearby bacterial DNA sequences (Fig. 8.1). Further analysis revealed that phage λ had short regions of single-stranded DNA whose base sequences were complementary to each other at each end of its linear genome. These single-stranded regions were called "cohesive" (*cos*) sites. Complementary base pairing of the *cos* sites allowed the linear genome to become a circle within the host bacterium. The idea of joining DNA segments by "cohesive sites" became the guiding principle for the development of genetic engineering. With the molecular characterization of restriction and modification systems in bacteria, it soon became apparent that the ideal engineering tools for making cohesive sites on specific DNA pieces were already available in the form of restriction endonucleases.

Insights from bacterial restriction and modification systems

Early on, Salvador Luria and other phage workers were intrigued by a phenomenon termed "restriction and modification." Phages grown in one bacterial host often failed to grow in different bacterial strains ("restriction"). However, some rare progeny phages were able to escape this restriction. Once produced in the restrictive host they had become "modified" in some way so that they now grew normally in this host. The entire cycle could be repeated, indicating that the modification was not an irreversible change. For example, phage λ grown on the C strain of *Escherichia coli* (λ·C) were restricted in the K-12 strain (the standard strain for most molecular work) (Fig. 8.2). However, the rare phage λ that managed to grow in the K-12 strain now had "K" modification (λ·K). These phages grew normally on both C and K-12; however, after growth on C, the phage λ with "C" modification (λ·C) was again restricted in K-12. Thus, the K-12 strain was able to mark its own resident DNA for preservation, but could eliminate invading DNA from another distantly related strain. In 1962, the molecular basis of restriction and modification was defined by Werner Arber and co-workers.

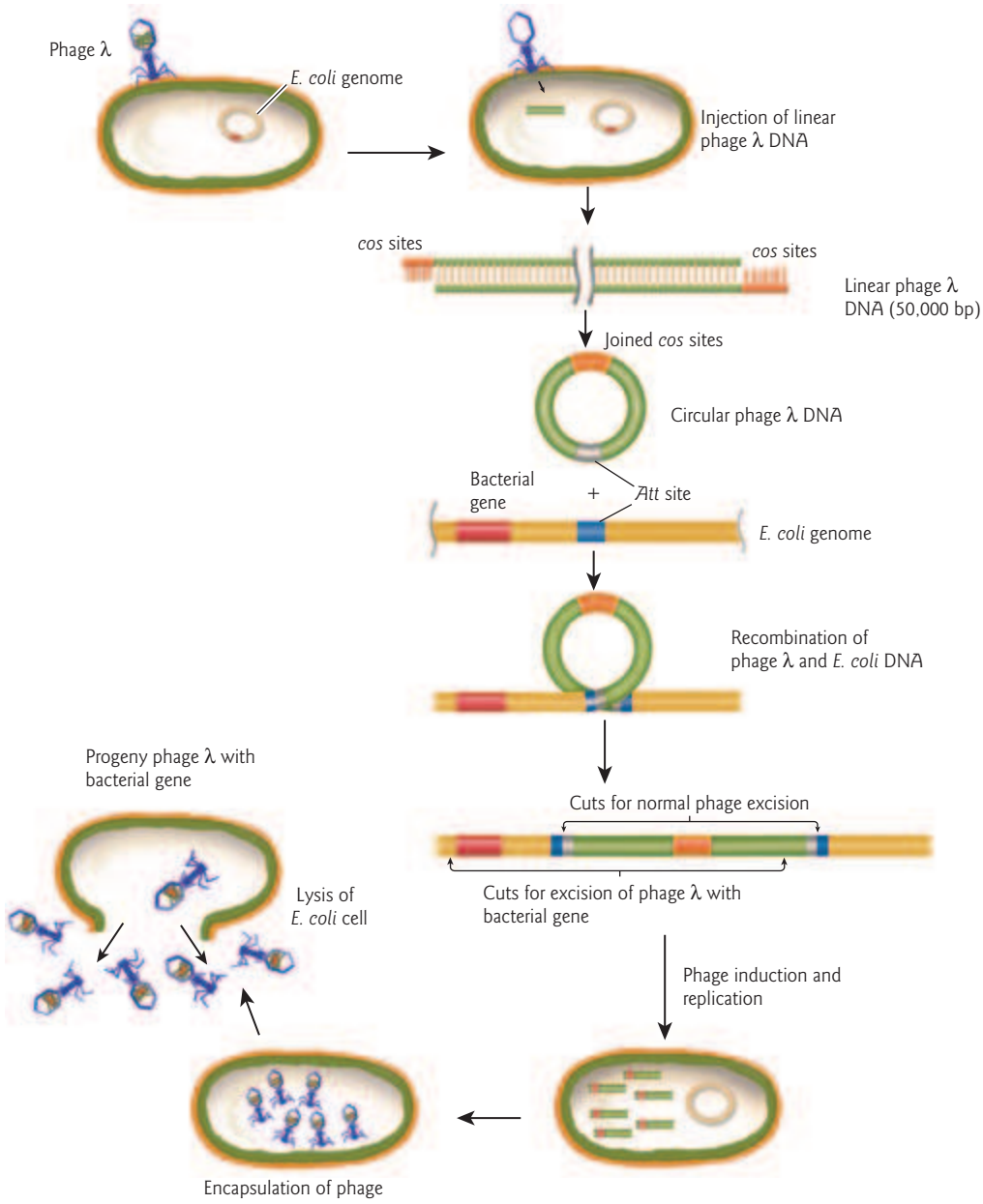


Figure 8.1 Bacteriophage lambda (λ) cohesive sites. Following the injection of a linear phage λ DNA into *E. coli* host cells, the phage λ genome circularizes by joining of the cohesive (*cos*) sites. In the lysogenic mode of replication, phage DNA is incorporated into the host genome by recombination at attachment (*Att*) sites on the phage and bacterial chromosome, and replicated as part of the host DNA. Under certain conditions, such as when the host encounters mutagenic chemicals or UV radiation, reversal of this recombination event leads to excision of the phage DNA. Rare excision events at different places allow phage λ to pick up bacterial genes. In the lytic mode of the phage life cycle, phage λ progeny with bacterial genes incorporated in their genomes are released from the lysed *E. coli*.

Restriction system

After demonstrating that phage λ DNA was degraded in a restricting host bacterium, Arber and co-workers hypothesized that the restrictive agent was a nuclease with the ability to distinguish whether DNA was resident or foreign. Six years later, such a nuclease was biochemically characterized in *E. coli* K-12 by Matt

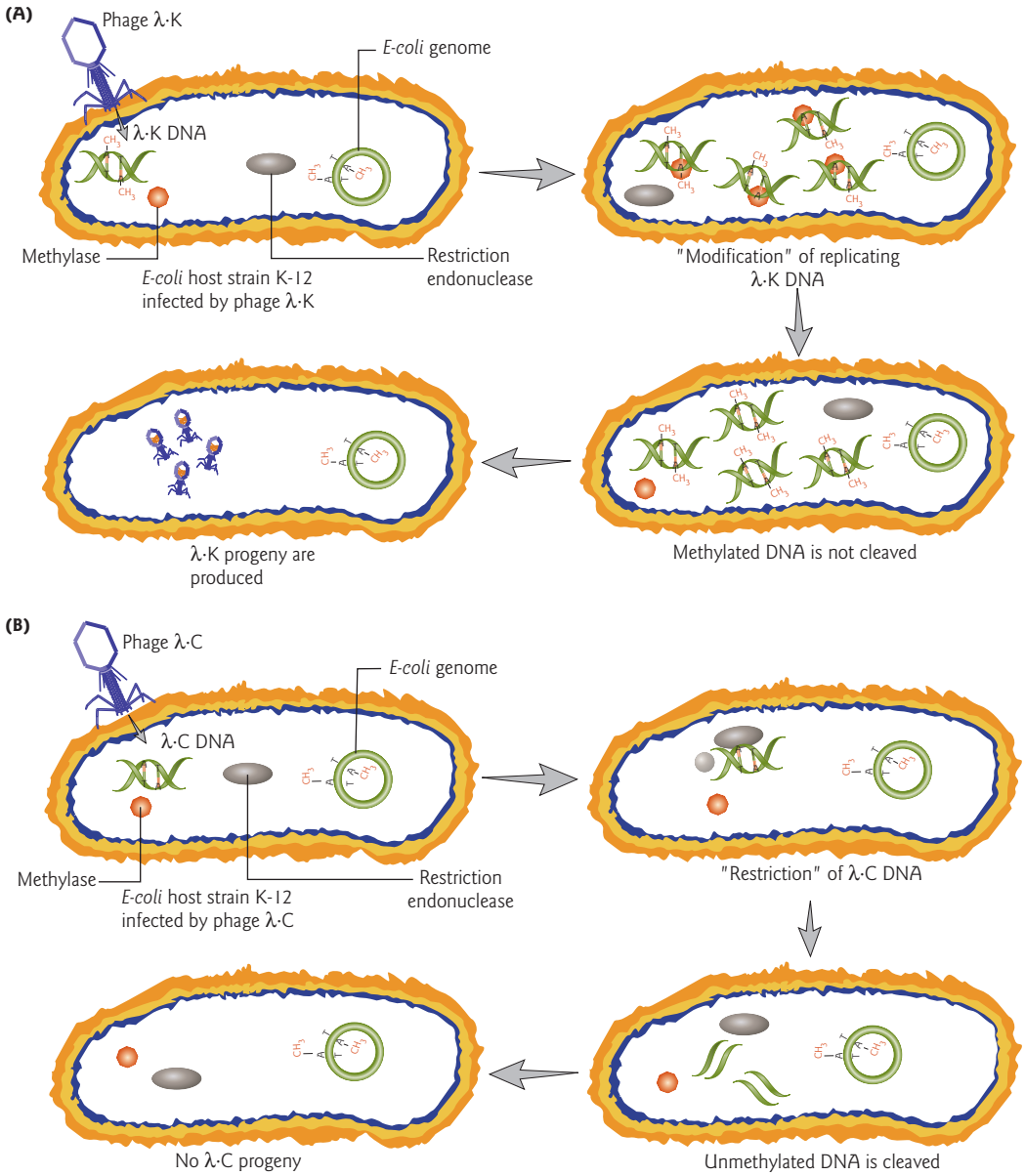


Figure 8.2 Restriction and modification systems in bacteria. Restriction endonucleases and their corresponding methylases function in bacteria to protect against bacteriophage infection. (A) Modification. When *E. coli* host strain K-12 is infected by phage λ·K, the phage DNA is not recognized as foreign because it has the same methylation pattern as the *E. coli* host genome. When the phage DNA replicates, the newly replicated DNA is modified by a specific methylase to maintain the pattern. Methylated DNA is not cleaved by restriction endonucleases, so progeny phage λ·K are produced. (B) When *E. coli* host strain K-12 is infected by phage λ·C, the phage DNA is recognized as foreign, because it does not have the same methylation pattern as the host genome. The phage DNA is cleaved by a specific restriction endonuclease, and no progeny phage λ·C are produced.

Meselson and Bob Yuan. The purified enzyme cleaved λ·C-modified DNA into about five pieces but did not attack λ·K-modified DNA (Fig. 8.2). Restriction endonucleases (also referred to simply as restriction enzymes) thus received their name because they restrict or prevent viral infection by degrading the invading nucleic acid.

Modification system

At the time, it was known that methyl groups were added to bacterial DNA at a limited number of sites. Most importantly, the location of methyl groups varied among bacterial species. Arber and colleagues were able to demonstrate that modification consisted of the addition of methyl groups to protect those sites in DNA sensitive to attack by a restriction endonuclease. In *E. coli*, adenine methylation (6-methyl adenine) is more common than cytosine methylation (5-methyl cytosine). Methyl-modified target sites are no longer recognized by restriction endonucleases and the DNA is no longer degraded. Once established, methylation patterns are maintained during replication. When resident DNA replicates, the old strand remains methylated and the new strand is unmethylated. In this hemimethylated state, the new strand is quickly methylated by specific methylases. In contrast, foreign DNA that is unmethylated or has a different pattern of methylation than the host cell DNA is degraded by restriction endonucleases.

The first cloning experiments

Hamilton Smith and co-workers demonstrated unequivocally that restriction endoncleases cleave a specific DNA sequence. Later, Daniel Nathans used restriction endonucleases to map the simian virus 40 (SV40) genome and to locate the origin of replication. These major breakthroughs underscored the great potential of restriction endonucleases for DNA work. Building on their discoveries, the cloning experiments of Herbert Boyer, Stanley Cohen, Paul Berg, and their colleagues in the early 1970s ushered in the era of recombinant DNA technology. One of the first recombinant DNA molecules to be engineered was a hybrid of phage λ and the SV40 mammalian DNA virus genome. In 1974 the first eukaryotic gene was cloned. Amplified ribosomal RNA (rRNA) genes or "ribosomal DNA" (rDNA) from the South African clawed frog *Xenopus laevis* were digested with a restriction endonuclease and linked to a bacterial plasmid. Amplified rDNA was used as the source of eukaryotic DNA since it was well characterized at the time and could be isolated in quantity by CsCl-gradient centrifugation. Within oocytes of the frog, rDNA is selectively amplified by a rolling circle mechanism from an extrachromosomal nucleolar circle (see Fig. 6.17). The number of rRNA genes in the oocyte is about 100- to 1000-fold greater than within somatic cells of the same organism. To the great excitement of the scientific community, the cloned frog genes were actively transcribed into rRNA in *E. coli*. This showed that recombinant plasmids containing both eukaryotic and prokaryotic DNA replicate stably in *E. coli*. Thus, genetic engineering could produce new combinations of genes that had never appeared in the natural environment, a feat which led to widespread concern about the safety of recombinant DNA work (Focus box 8.1).

8.3 Cutting and joining DNA

Two major categories of enzymes are important tools in the isolation of DNA and the preparation of recombinant DNA: restriction endonucleases and DNA ligases. Restriction endonucleases recognize a specific, rather short, nucleotide sequence on a double-stranded DNA molecule, called a restriction site, and cleave the DNA at this recognition site or elsewhere, depending on the type of enzyme. DNA ligase joins two pieces of DNA by forming phosphodiester bonds.

Major classes of restriction endonucleases

There are three major classes of restriction endonucleases. Their grouping is based on the types of sequences recognized, the nature of the cut made in the DNA, and the enzyme structure. Type I and III restriction endonucleases are not useful for gene cloning because they cleave DNA at sites other than the recognition sites and thus cause random cleavage patterns. In contrast, type II endonucleases are widely used for mapping and reconstructing DNA *in vitro* because they recognize specific sites and cleave just at these sites (Table 8.1). In addition, the type II endonuclease and methylase activities are usually separate, single subunit enzymes. Although the two enzymes recognize the same target sequence, they can be purified separately from each

Fear of recombinant DNA molecules

FOCUS BOX 8.1

In the wake of the first cloning experiments, there was immediate concern from both scientists and the general public about the possible dangers of recombinant DNA work. Concerns primarily focused on the ethics of "tampering with nature" and the potential for the escape of genetically engineered pathogenic bacteria from a controlled laboratory environment. One fear was that *E. coli* carrying cloned tumor virus DNA could be transferred to humans and trigger a global cancer epidemic. Not everyone shared these fears. James Watson wrote in his chapter in the book *Genetics and Society* (1993):

> I was tempted then to put together a book called the Whole Risk Catalogue. It would contain risks for old people and young people and so on. It would be a very popular book in our semi-paranoid society. Under "D" I would put dynamite, dogs, doctors, dieldrin [an insecticide] and DNA. I must confess to being more frightened of dogs. But everyone has their own things to worry about.

In 1975 a landmark meeting was held at the Asilomar Conference Center near San Francisco. The meeting was attended by over 100 molecular biologists. Recommendations arising from this meeting formed the basis for official guidelines developed by the National Institutes of Health (NIH) regarding containment. As time passed, there were no disasters that occurred as a result of recombinant DNA technology, and it was concluded by most scientists that under these guidelines the technology itself did not pose any risk to human health or the environment. Containment works very well and engineered bacteria and vectors do very poorly under natural conditions.

Currently, activities involving the handling of recombinant DNA molecules and organisms must be conducted in accordance with the *NIH Guidelines for Research Involving Recombinant DNA Molecules*. Four levels of risk are recognized, from minimal to high, for which four levels of containment (physical and biological barriers to the escape of dangerous organisms) are outlined. The highest risk level is for experiments dealing with highly infectious agents and toxins that are likely to cause serious or lethal human disease for which preventive or therapeutic interventions are not usually available. Precautions include negative-pressure air locks in laboratories and experiments done in laminar-flow hoods, with filtered or incinerated exhaust air. The bacteria used routinely in molecular biology, such as nonpathogenic strains of *E. coli*, are "Risk group I" agents, which are not associated with disease in healthy adult humans. Standard vectors for recombinant DNA are genetically designed to decrease, by many orders of magnitude, the probability of dissemination of recombinant DNA outside the laboratory.

Today, fears focus not so much on the technology *per se*, but on the application of recombinant DNA technology to agriculture, medicine, and bioterrorism. For example, there is concern about the safety of genetically engineered foods in the marketplace, the spread of herbicide-resistant genes from transgenic crop plants to weeds, the use of gene therapy for eugenics (artificial human selection), and the construction of recombinant DNA "designer weapons." The latter refers to engineering infectious microbes to be even more virulent, antibiotic-resistant, and environmentally stable. On December 13, 2002, new federal regulations were published to implement the US Public Health and Security and Bioterroism Preparedness and Response Act of 2002 (http://www.fda.gov/oc/bioterrorism/bioact.html). The regulations apply to the possession, use, and transfer of select agents that are considered potential bioterrorist agents, such as *Yersinia pesti* (plague), *Bacillus anthracis* (anthrax), and variola virus (smallpox).

other. Some type II restriction endonucleases do not conform to this narrow definition, making it necessary to define further subdivisions. The discussion here will focus on the "orthodox" type II restriction endonucleases that are commonly used in molecular biology research.

Restriction endonclease nomenclature

Restriction endonucleases are named for the organism in which they were discovered, using a system of letters and numbers. For example, *Hin*dIII (pronounced "hindee-three") was discovered in *Haemophilus*

Table 8.1 Major classes of restriction endonucleases.

Class	Abundance	Recognition site	Composition	Use in recombinant DNA research
Type I	Less common than type II	Cut both strands at a nonspecific location > 1000 bp away from recognition site	Three-subunit complex: individual recognition, endonuclease, and methylase activities	Not useful
Type II	Most common	Cut both strands at a specific, usually palindromic, recognition site (4–8 bp)	Endonuclease and methylase are separate, single-subunit enzymes	Very useful
Type III	Rare	Cleavage of one strand only, 24–26 bp downstream of the 3′ recognition site	Endonuclease and methylase are separate two-subunit complexes with one subunit in common	Not useful

influenza (strain d). The *Hin* comes from the first letter of the genus name and the first two letters of the species name; d is for the strain type; and III is for the third enzyme of that type. *Sma*I is from *Serratia marcescens* and is pronounced "smah-one," *Eco*RI (pronounced "echo-r-one") was discovered in *Escherichia coli* (strain R), and *Bam*HI is from *Bacillus amyloliquefaciens* (strain H). Over 3000 type II restriction endonucleases have been isolated and characterized to date. Approximately 240 are available commercially for use by molecular biologists.

Recognition sequences for type II restriction endonucleases

Each orthodox type II restriction endonuclease is composed of two identical polypeptide subunits that join together to form a homodimer. These homodimers recognize short symmetric DNA sequences of 4–8 bp. Six base pair cutters are the most commonly used in molecular biology research. Usually, the sequence read in the 5′ → 3′ direction on one strand is the same as the sequence read in the 5′ → 3′ direction on the complementary strand. Sequences that read the same in both directions are called palindromes (from the Greek word *palindromos* for "run back"). Figure 8.3 shows some common restriction endonucleases and their recognition sequences. Some enzymes, such as *Eco*R1, generate a staggered cut, in which the single-stranded complementary tails are called "sticky" or cohesive ends because they can hydrogen bond to the single-stranded complementary tails of other DNA fragments. If DNA molecules from different sources share the same palindromic recognition sites, both will contain complementary sticky ends (single-stranded tails) when digested with the same restriction endonuclease. Other type II enzymes, such as *Sma*I, cut both strands of the DNA at the same position and generate blunt ends with no unpaired nucleotides when they cleave the DNA.

Restriction endonucleases exhibit a much greater degree of sequence specificity in the enzymatic reaction than is exhibited in the binding of regulatory proteins, such as the Lac repressor to DNA (see Section 10.6). For example, a single base pair change in a critical operator sequence usually reduces the affinity of the Lac repressor by 10- to 100-fold, whereas a single base pair change in the recognition site of a restriction endonuclease essentially eliminates all enzymatic activity.

Like other DNA-binding proteins, the first contact of a restriction endonuclease with DNA is nonspecific (Fig. 8.4). Nonspecific binding usually does not involve interactions with the bases but only with the DNA sugar–phosphate backbone. The restriction endonuclease is loosely bound and its catalytic center is kept at a safe distance from the phosphodiester backbone. Nonspecific binding is a prerequisite for efficient target site location. For example, *Bam*HI moves along the DNA in a linear fashion by a process called "sliding." Sliding involves helical movement due to tracking along a groove of the DNA over short distances (< 30–50 bp). This reduces the volume of space through which the protein needs to search to one dimension. However, the "random walk" nature of linear diffusion gives equal probabilities for forward and reverse steps, so if the distances between the nonspecific binding site and the recognition site are large (> 30–50 bp), the protein

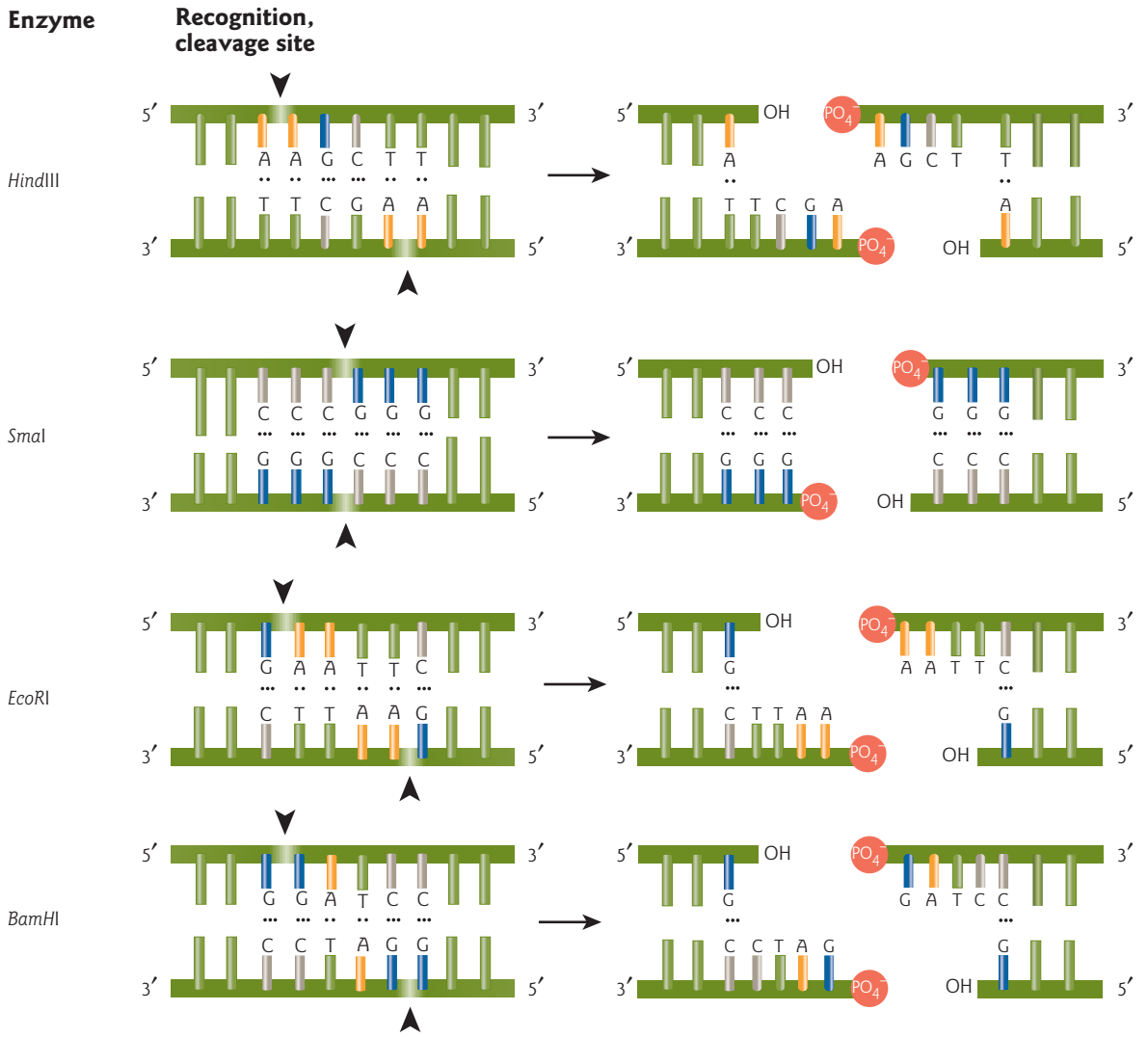


Figure 8.3 Cleavage patterns of some common restriction endonucleases. The recognition and cleavage sites, and cleavage patterns of *Hin*dIII, *Sma*I, *Eco*RI, and *Bam*HI are shown. Restriction endonucleases catalyze the hydrolysis of phosphodiester bonds in palindromic DNA sequences to produce double-strand breaks, resulting in the formation of 5′-PO_4^- and 3′-OH termini with "sticky" ends (*Hin*dIII, *Eco*RI, and *Bam*HI) or "blunt" ends (*Sma*I).

would return repeatedly to its start point. The main mode of translocation over long distances is thus by "hopping" or "jumping." In this process, the protein moves between binding sites through three-dimensional space, by dissociating from its initial site before reassociating elsewhere in the same DNA chain. Because of relatively small diffusion constants of proteins, most rebinding events will be short range "hops" back to or near the initial binding site. In the example of *Bam*HI, once the target restriction site is located, the recognition process triggers large conformational changes of the enzyme and the DNA (called coupling), which leads to the activation of the catalytic center (Fig. 8.4). In addition to indirect interaction with the DNA backbone, specific binding is characterized by direct interaction of the enzyme with the nitrogenous bases.

All structures of orthodox type II restriction endonucleases characterized by X-ray crystallography so far show a common structural core composed of four conserved β-strands and one α-helix (Focus box 8.2). In the presence of the essential cofactor Mg^{2+}, the enzyme cleaves the DNA on both strands at the same time within or in close proximity to the recognition sequence (restriction site). The enzyme cuts the DNA

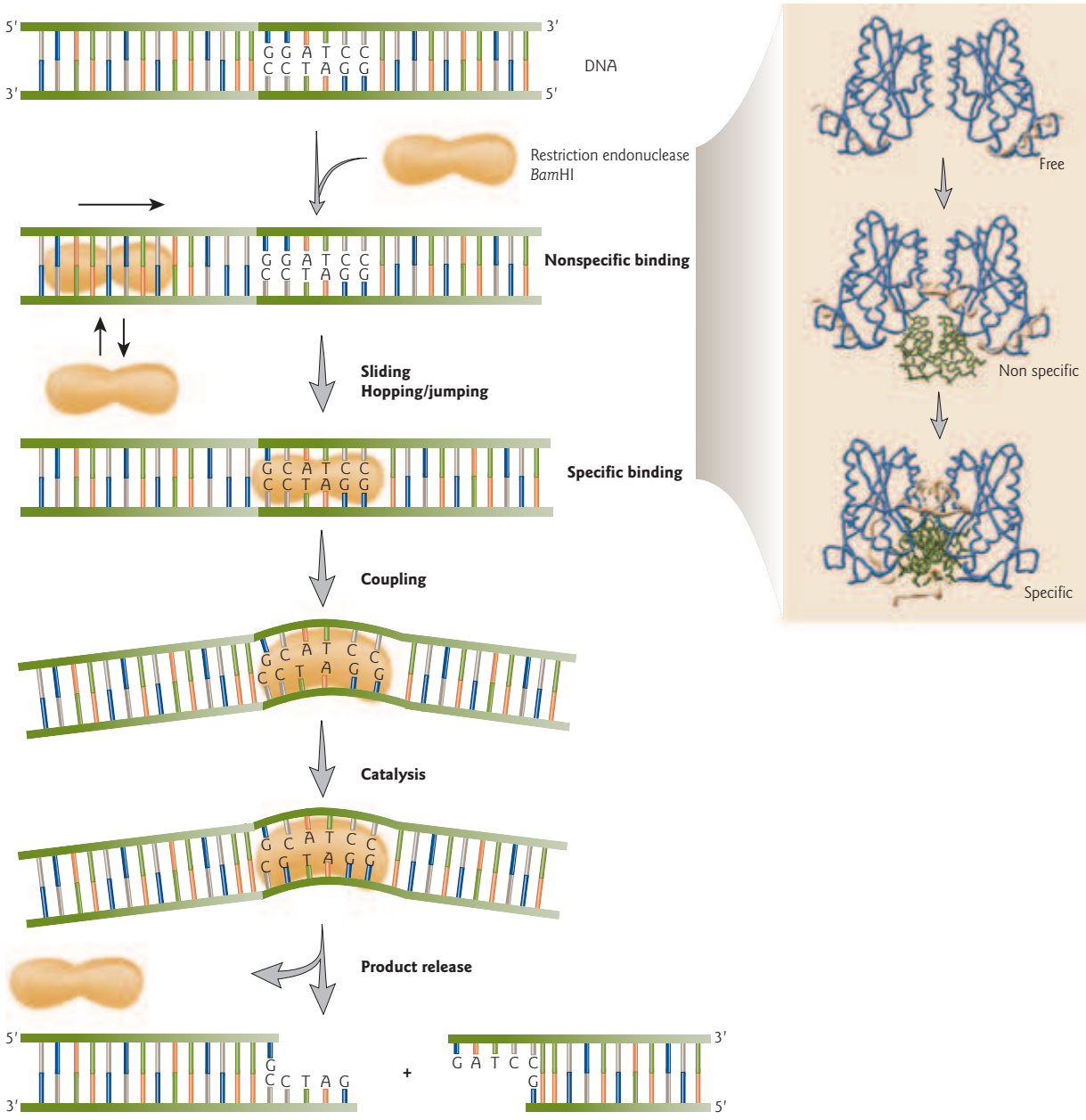


Figure 8.4 The steps involved in DNA binding and cleavage by a type II restriction endonuclease. Type II restriction endonucleases, like *Bam*HI, bind DNA as dimers. The first contact with DNA is nonspecific. The target site is then located by a combination of linear diffusion or "sliding" of the enzyme along the DNA over short distances, and hopping/jumping over longer distances. Once the target restriction site is located, the recognition process (coupling) triggers large conformational changes of the enzyme and the DNA, which leads to activation of the catalytic center. Catalysis results in product release. (Pingoud, A., Jeltsch, A. 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Research* 29:3705–3727; and Gowers, D.M., Wilson, G.G., Halford, S.E. 2005. Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proceedings of the National Academy of Sciences USA* 102:15883–15888.) (Inset) Structures of free, nonspecific, and specific DNA-bound forms of *Bam*HI. The two dimers are shown in brown, the DNA backbone is in green and the bases in gray. *Bam*HI becomes progressively more closed around the DNA as it goes from the nonspecific to specific DNA binding mode. (Protein Data Bank, PDB:1ESG. Adapted from Viadiu, H., Aggarwal, A.K. 2000. Structure of *Bam*HI bound to nonspecific DNA: a model for DNA sliding. *Molecular Cell* 5:889–895. Copyright © 2000, with permission from Elsevier.)

duplex by breaking the covalent, phosphodiester bond between the phosphate of one nucleotide and the sugar of an adjacent nucleotide, to give free 5′-phosphate and 3′-OH ends. Type II restriction endonucleases do not require ATP hydrolysis for their nucleolytic activity. Although there are a number of models for how this nucleophilic attack on the phosphodiester bond occurs (Focus box 8.2), the exact mechanism by which restriction endonucleases achieve DNA cleavage has not yet been proven experimentally for any type II restriction endonuclease.

DNA ligase

The study of DNA replication and repair processes led to the discovery of the DNA-joining enzyme called DNA ligase. DNA ligases catalyze formation of a phosphodiester bond between the 5′-phosphate of a nucleotide on one fragment of DNA and the 3′-hydroxyl of another (see Fig. 6.14). This joining of linear DNA fragments together with covalent bonds is called ligation. Unlike the type II restriction endonucleases, DNA ligase requires ATP as a cofactor.

Because it can join two pieces of DNA, DNA ligase became a key enzyme in genetic engineering. If restriction-digested fragments of DNA are placed together under appropriate conditions, the DNA fragments from two sources can anneal to form recombinant molecules by hydrogen bonding between the complementary base pairs of the sticky ends. However, the two strands are not covalently bonded by phosphodiester bonds. DNA ligase is required to seal the gaps, covalently bonding the two strands and regenerating a circular molecule. The DNA ligase most widely used in the lab is derived from the bacteriophage T4. T4 DNA ligase will also ligate fragments with blunt ends, but the reaction is less efficient and higher concentrations of the enzyme are usually required *in vitro*. To increase the efficiency of the reaction, researchers often use the enyzme terminal deoxynucleotidyl transferase to modify the blunt ends. For example, if a single-stranded poly(dA) tail is added to DNA fragments from one source, and a single-stranded poly(dT) tail is added to DNA from another source, the complementary tails can hydrogen bond (Fig. 8.5). Recombinant DNA molecules can then be created by ligation.

8.4 Molecular cloning

The basic procedure of molecular cloning involves a series of steps. First, the DNA fragments to be cloned are generated by using restriction endonucleases, as described in Section 8.3. Second, the fragments produced by digestion with restriction enzymes are ligated to other DNA molecules that serve as vectors. Vectors can replicate autonomously (independent of host genome replication) in host cells and facilitate the manipulation of the newly created recombinant DNA molecule. Third, the recombinant DNA molecule is transferred to a host cell. Within this cell, the recombinant DNA molecule replicates, producing dozens of identical copies known as clones. As the host cells replicate, the recombinant DNA is passed on to all progeny cells, creating a population of identical cells, all carrying the cloned sequence. Finally, the cloned DNA segments can be recovered from the host cell, purified, and analyzed in various ways.

Vector DNA

Cloning vectors are carrier DNA molecules. Four important features of all cloning vectors are that they: (i) can independently replicate themselves and the foreign DNA segments they carry; (ii) contain a number of unique restriction endonuclease cleavage sites that are present only once in the vector; (iii) carry a selectable marker (usually in the form of antibiotic resistance genes or genes for enzymes missing in the host cell) to distinguish host cells that carry vectors from host cells that do not contain a vector; and (iv) are relatively easy to recover from the host cell. There are many possible choices of vector depending on the purpose of cloning. The greatest variety of cloning vectors has been developed for use in the bacterial host *E. coli*. Thus, the first practical skill generally required by a molecular biologist is the ability to grow pure cultures of bacteria.

FOCUS BOX 8.2

*Eco*RI: kinking and cutting DNA

*Eco*RI functions as a homodimer of two identical 31,000 molecular weight subunits and catalyzes the cleavage of a double-stranded sequence d(GAATTC). The interaction of the restriction endonuclease *Eco*RI with DNA illustrates how subtle features of its shape and surface characteristics allow it to interact with complementary surfaces on the DNA.

The crystal structure of *Eco*RI complexed with a 12 bp DNA duplex was determined in 1986. One dimer contains a

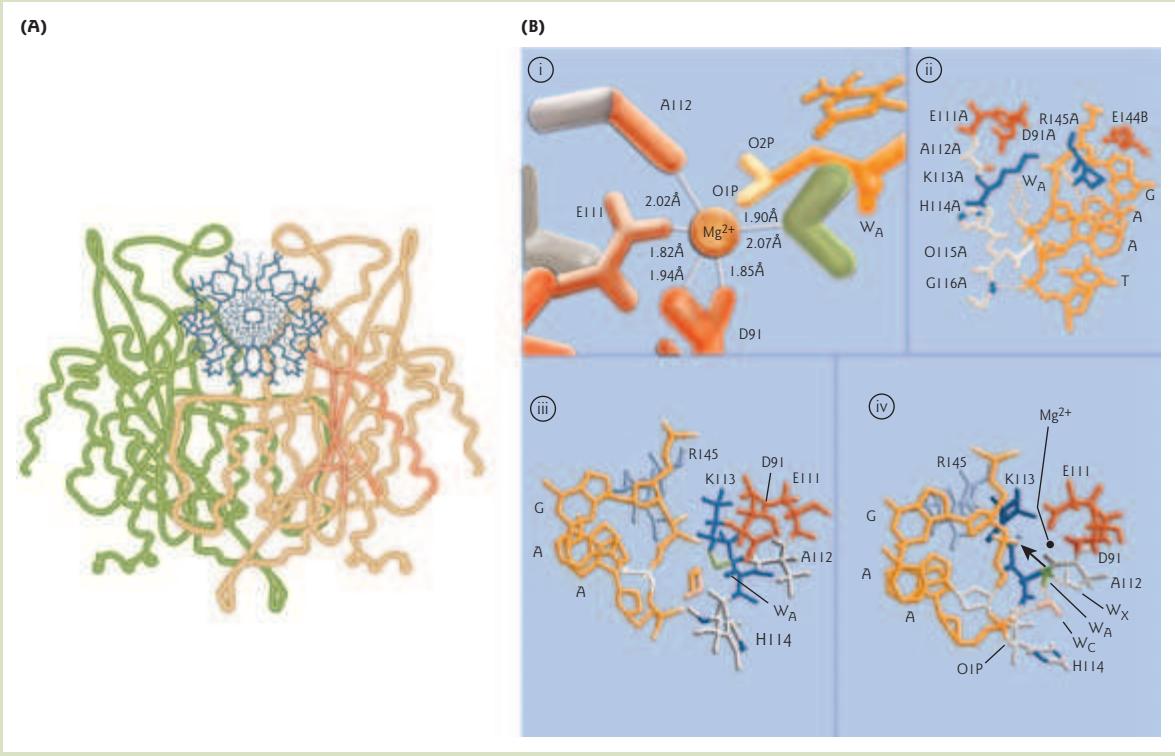


Figure 1 Structure of *Eco*RI. (A) Crystal structure of the two subunits (green and light orange) of *Eco*RI bound to DNA (blue). In one subunit the four strictly conserved β-strands and one α-helix of the common core are shown in red. (Protein Data Bank, PDB:1ERI. Adapted from Pingoud, A. and Jeltsch, A. 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Research* 29:3705–3727. Copyright © 2001, with permission of the Oxford University Press.). (B) Catalytic centers of the *Eco*RI–DNA complex. (i) Coordination of Mg^{2+} by six ligands in the catalytic center: one carboxylate oxygen of the glutamic acid at position 111 (E111); two carboxylate oxygens of asparagine 91 (D91); the main-chain carbonyl of alanine 112 (A112); the O1P oxygen of the scissile phosphate GpAA (to polarize the phosphate and facilitate nucleophilic attack); and a water molecule, W_A, that forms the attacking nucleophile. (ii) Catalytic and recognition elements of the crystal structure of the Mg^{2+}-free *Eco*RI–DNA complex. The letters following the side chain numbers denote protein subunits A and B. Only one DNA strand (orange) is shown for part of the recognition site. (iii and iv) The *Eco*RI–DNA complexes in the absence (iii) and presence (iv) of Mg^{2+}. The black arrow in (iv) shows the direction of nucleophilic attack on phosphorus. The presence of Mg^{2+} causes a number of structural changes, including alteration of the position and orientation of the water molecules W_A and W_C, movement of D91, and movement of lysine 113 (K113) away from its hydrogen-bonding partner E111. (Reproduced from Kurpiewski, M.R., Engler, L.E., Wozniak, L.A., Kobylanska, A., Koziolkiewicz, M., Stec, W.J., Jen-Jacobsen, L. 2004. Mechanisms of coupling between DNA recognition and specificity and catalysis in *Eco*RI endonuclease. *Structure* 12:1775–1788. Copyright © 2004, with permission from Elsevier.)

*Eco*RI: kinking and cutting DNA

FOCUS BOX 8.2

conserved four-stranded β-sheet surrounded on either side by α-helices (Fig. 1). The active site of the endonuclease lies at the C-terminus of this parallel β-sheet and forms a catalytic center, in which Mg^{2+} is bound by interaction with six amino acids (β2 and β3 contain the amino acid residues directly involved in catalysis). Upon specific DNA binding, about 150 water molecules are released; this expulsion of solvent molecules from the interface allows for close contact between the enzyme and the DNA. The N-terminus of the protein forms an arm that partially wraps around the DNA. A bundle of four parallel α-helices, two from each dimer, pushes into the major groove and directly recognizes the DNA base sequence. A major portion of the sequence specificity exhibited by this enzyme appears to be achieved through an array of 12 hydrogen bond donors and acceptors from protein side chains. These donors and acceptors are complementary to the donors and acceptors presented by the exposed edges of the base pairs in the hexanucleotide recognition sequence.

The binding of *Eco*RI to its recognition site induces a dramatic conformational change not only in the enzyme itself, but also in the DNA. A central kink (or bend) of about 20–40° in the DNA brings the critical phosphodiester bond between G and A deeper into the active site. The kink is accompanied by unwinding of the DNA. This unwinding of the top 6 bp relative to the bottom 6 bp results in a widening of the major groove by about 3.5 Å. The widening allows the two α-helices from each subunit of the dimer to fit (end on) into the major groove. Further, the realignment of base pairs produced by the kink creates sites for multiple hydrogen bonds with the protein not present in the undistorted DNA. Thus, the protein-induced distortions of the DNA are an intimate part of the recognition and catalysis process.

Choice of vector is dependent on insert size and application

The classic cloning vectors are plasmids, phages, and cosmids, which are limited to the size insert they can accommodate, taking up to 10, 20, and 45 kb, respectively (Table 8.2). The feature of plasmids and phages and their use as cloning vectors will be discussed in more detail in later sections. A cosmid is a plasmid carrying a

Table 8.2 Principal features and applications of different cloning vector systems.

Vector	Basis	Size limits of insert	Major application
Plasmid	Naturally occuring multicopy plasmids	≤ 10 kb	Subcloning and downstream manipulation, cDNA cloning and expression assays
Phage	Bacteriophage λ	5–20 kb	Genomic DNA cloning, cDNA cloning, and expression libraries
Cosmid	Plasmid containing a bacteriophage λ *cos* site	35–45 kb	Genomic library construction
BAC (bacterial artificial chromosome)	*Escherichia coli* F factor plasmid	75–300 kb	Analysis of large genomes
YAC (yeast artificial chromosome)	*Saccharomyces cerevisiae* centromere, telomere, and autonomously replicating sequence	100–1000 kb (1 Mb)	Analysis of large genomes, YAC transgenic mice
MAC (mammalian artificial chromosome)	Mammalian centromere, telomere, and origin of replication	100 kb to > 1 Mb	Under development for use in animal biotechnology and human gene therapy

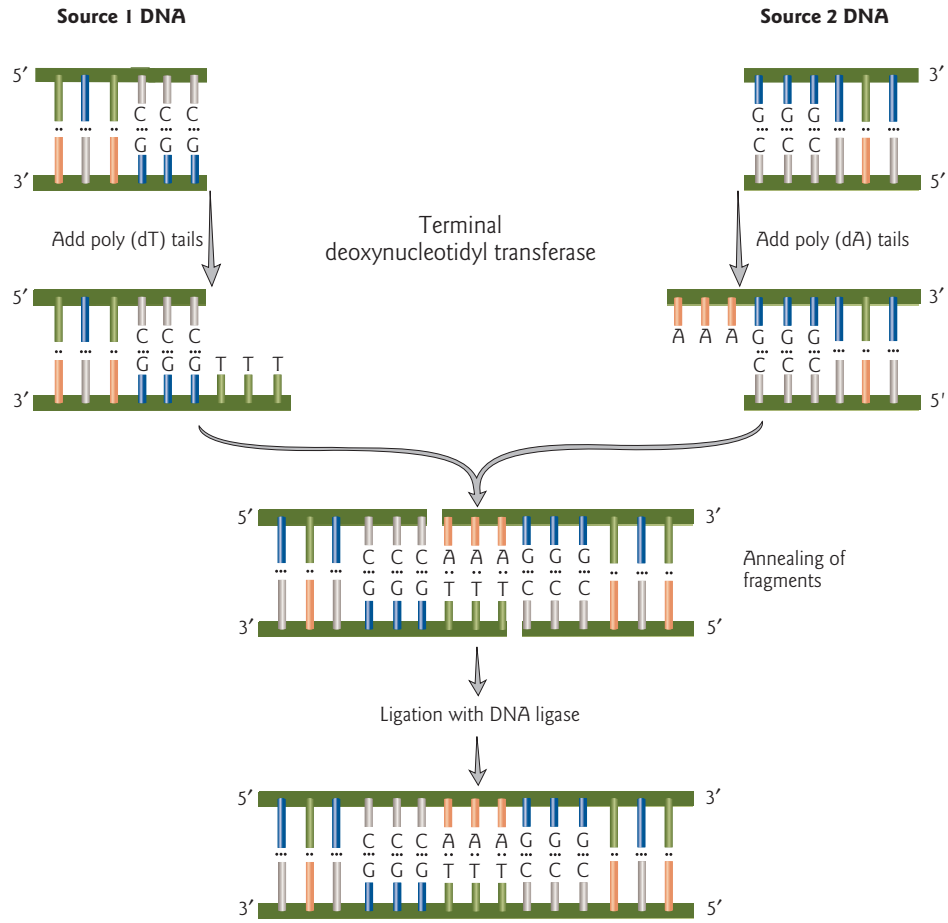


Figure 8.5 Modified blunt end ligation. Recombinant DNA molecules can be formed from DNA cut with restriction endonucleases that leave blunt ends, such as *Sma*I. Without end modification, blunt end ligation is of low efficiency. The efficiency is increased through using the enzyme terminal deoxynucleotidyl transferase to create complementary tails by the addition of poly(dA) and poly(dT) to the cleaved fragments. These tails allow DNA fragments from two different sources to anneal. "Source 1" DNA and "source 2" DNA are then covalently linked by treatment with DNA ligase to create a recombinant DNA molecule. Note that the *Sma*I site is destroyed in the process.

phage λ *cos* site, allowing it to be packaged into a phage head. Cosmids infect a host bacterium as do phages, but replicate like plasmids and the host cells are not lysed. Mammalian genes are often greater than 100 kb in size, so originally there were limitations in cloning complete gene sequences. Vectors engineered more recently have circumvented this problem by mimicking the properties of host cell chromosomes. This new generation of artificial chromosome vectors includes bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), and mammalian artificial chromosomes (MACs).

Plasmid DNA as a vector

Plasmids are naturally occurring extrachromosomal double-stranded circular DNA molecules that carry an origin of replication and replicate autonomously within bacterial cells (see Section 3.4). The plasmid vector pBR322, constructed in 1974, was one of the first genetically engineered plasmids to be used in

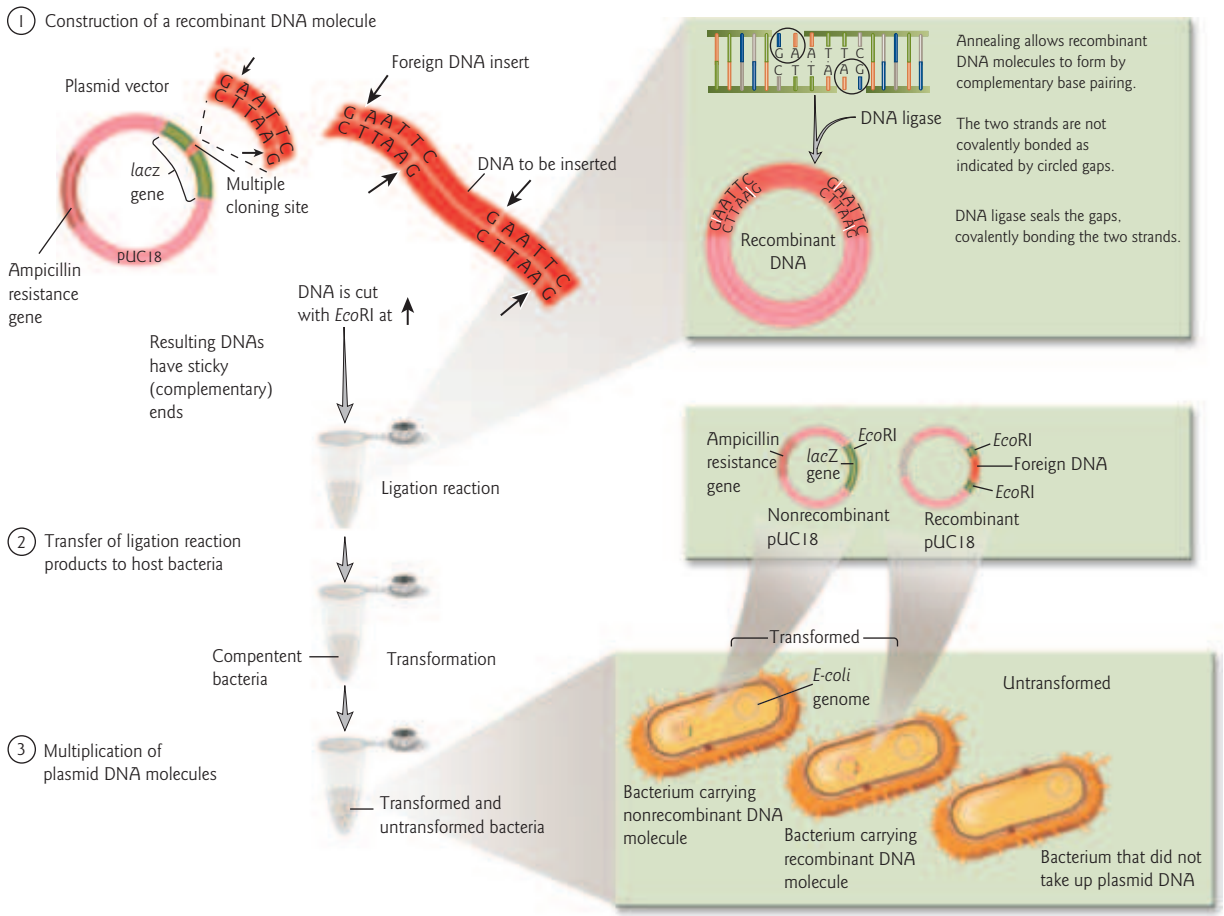


Figure 8.6 Molecular cloning using a plasmid vector. Molecular cloning using a plasmid vector involves five major steps. (1) Construction of a recombinant DNA molecule. In this example, vector DNA (the plasmid pUC18) and the foreign DNA insert are cleaved with *Eco*RI and mixed together in a ligation reaction containing DNA ligase. pUC18 carries the ampicillin resistance gene and has a large number of restriction sites comprising a multiple cloning site within a selectable marker gene. (2) Transfer of ligation reaction products to host bacteria. Competent *E. coli* are transformed with ligation reaction products. Any DNA that remains linear will not be taken up by the host bacteria. (3) Multiplication of plasmid DNA molecules. Within each transformed host bacterium, there is autonomous multiplication of plasmid DNA. Each bacterium may contain as many as 500 copies of pUC18. Some bacteria in the mixture will be untransformed (not carrying either recombinant or nonrecombinant plasmid DNA).

recombinant DNA. Plasmids are named with a system of uppercase letters and numbers, where the lowercase "p" stands for "plasmid." In the case of pBR322, the BR identifies the original constructors of the vector (Bolivar and Rodriquez), and 322 is the identification number of the specific plasmid. These early vectors were often of low copy number, meaning that they replicate to yield only one or two copies in each cell. pUC18, the vector shown in Fig. 8.6, is a derivative of pBR322. This is a "high copy number" plasmid (> 500 copies per bacterial cell).

Plasmid vectors are modified to contain a specific antibiotic resistance gene and a multiple cloning site (also called the polylinker region) which has a number of unique target sites for restriction endonucleases. Cutting the circular plasmid vector with one of these enzymes results in a single cut, creating a linear plasmid. A foreign DNA molecule, referred to as the "insert," cut with the same enzyme, can then be joined to the vector in a ligation reaction (Fig. 8.6). Ligations of the insert to vector are not 100% productive, because

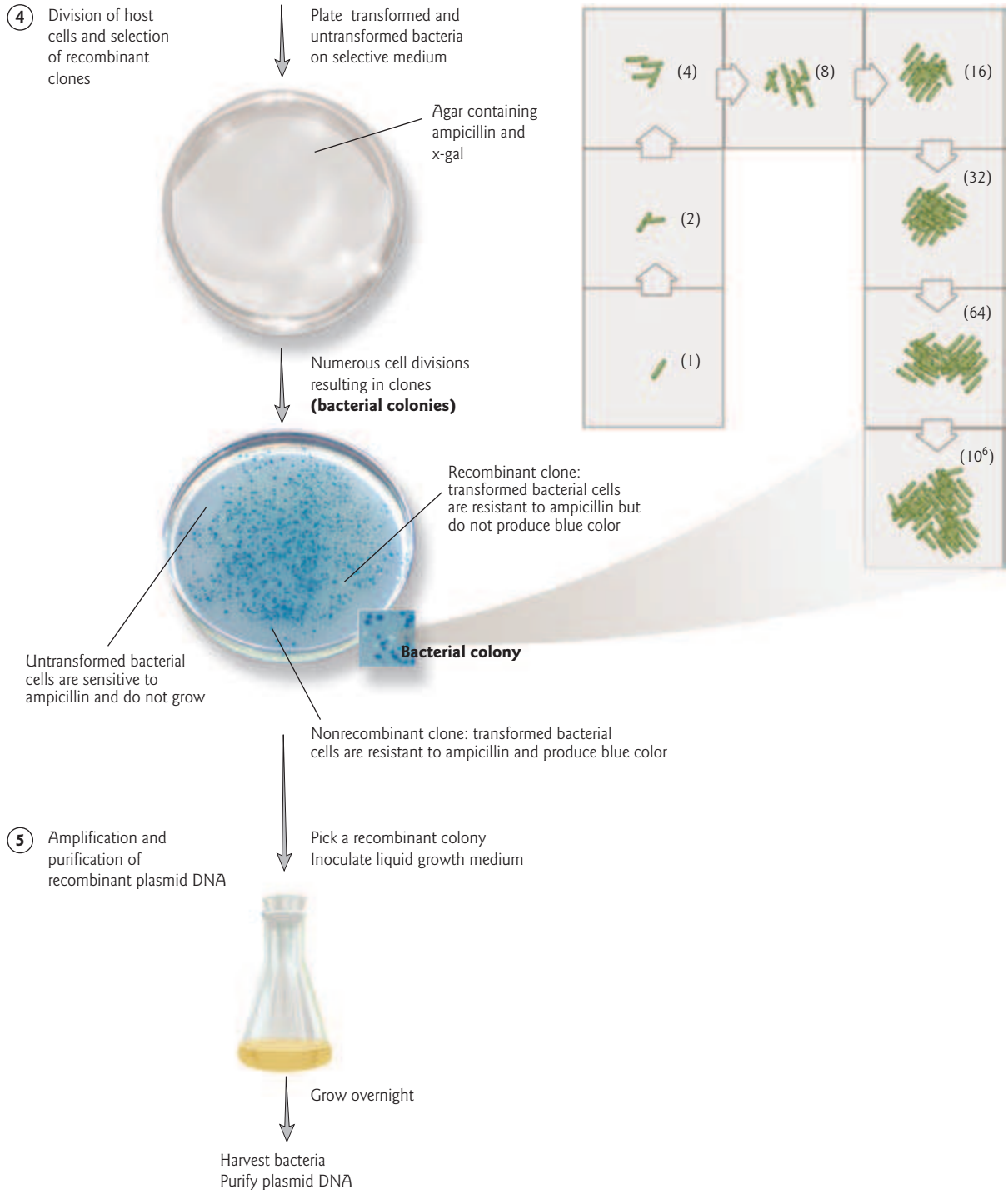


Figure 8.6 (*cont'd*) (4) Division of host cells and selection of recombinant clones by blue-white screening. Bacteria are plated on a selective agar medium containing the antibiotic ampicillin and X-gal (see Fig. 8.7). If foreign DNA is inserted into the multiple cloning site, then the *lac*Z′ coding region is disrupted and the N-terminal portion of β-galactosidase is not produced. Since there is no functional β-galactosidase in the bacteria, the substrate X-gal remains colorless, and the bacterial colony containing recombinant plasmid DNA appears white, thus allowing the direct identification of colonies carrying cloned DNA inserts. If there is no insertion of foreign DNA in the multiple cloning site, then the *lac*Z′ gene is intact and enzymatically active β-galactosidase is produced. The bacterial colonies containing nonrecombinant plasmid DNA thus appear blue. (Photograph courtesy of Vinny Roggero and the Spring 2006 Molecular Genetics Lab, College of William and Mary.) (5) Amplification and purification of recombinant plasmid DNA. A recombinant colony is used to inoculate liquid growth medium. After growing the bacteria overnight, the culture is harvested, bacterial cells are lysed, and the plasmid DNA is purified away from other cellular components.

the two ends of a plasmid vector can be readily ligated together, which is called self-ligation. The degree of self-ligation can be reduced by treatment of the vector with the enzyme phosphatase, which removes the terminal 5′-phosphate. When the 5′-phosphate is removed from the plasmid it cannot be recircularized by ligase, since there is nothing with which to make a phosphodiester bond. But, if the vector is joined with a foreign insert, the 5′-phosphate is provided by the foreign DNA. Another strategy involves using two different restriction endonuclease cutting sites with noncomplementary sticky ends. This inhibits self-ligation and promotes annealing of the foreign DNA in the desired orientation within the vector.

Transformation: transfer of recombinant plasmid DNA to a bacterial host

The ligation reaction mixture of recombinant and nonrecombinant DNA described in the preceding section is introduced into bacterial cells in a process called transformation (Fig. 8.6). The traditional method is to incubate the cells in a concentrated calcium salt solution to make their membranes leaky. The permeable "competent" cells are then mixed with DNA to allow entry of the DNA into the bacterial cell. Alternatively, a process called electroporation can be used that drives DNA into cells by a strong electric current.

Since bacterial species use a restriction-modification system to degrade foreign DNA lacking the appropriate methylation pattern, including plasmids, the question arises: why don't the transformed bacteria degrade the foreign DNA? The answer is that molecular biologists have cleverly circumvented this defense system by using mutant strains of bacteria, deficient for both restriction and modification, such as the common lab strain *E. coli* DH5α.

Successfully transformed bacteria will carry either recombinant or nonrecombinant plasmid DNA. Multiplication of the plasmid DNA occurs within each transformed bacterium. A single bacterial cell placed on a solid surface (agar plate) containing nutrients can multiply to form a visible colony made of millions of identical cells (Fig. 8.6). As the host cell divides, the plasmid vectors are passed on to progeny, where they continue to replicate. Numerous cell divisions of a single transfomed bacteria result in a clone of cells (visible as a bacterial colony) from a single parental cell. This step is where "cloning" got its name. The cloned DNA can then be isolated from the clone of bacterial cells.

Recombinant selection

What needs to be included in the medium for plating cells so that nontransformed bacterial cells are not able to grow at all? The answer depends on the particular vector, but in the case of pUC18, the vector carries a selectable marker gene for resistance to the antibiotic ampicillin. Ampicillin, a derivative of penicillin, blocks synthesis of the peptidoglycan layer that lies between the inner and outer cell membranes of *E. coli* (Table 8.3). Ampicillin does not affect existing cells with intact cell envelopes but kills dividing cells as they synthesize new peptidoglycan. The ampicillin resistance genes carried by the recombinant plasmids produce an enzyme, β-lactamase, that cleaves a specific bond in the four-membered ring (β-lactam ring) in the ampicillin molecule that is essential to its antibiotic action. If the plasmid vector is introduced into a plasmid-free antibiotic-sensitive bacterial cell, the cell becomes resistant to ampicillin. Nontransformed cells contain no pUC18 DNA, therefore they will not be antibiotic-resistant, and their growth will be inhibited on agar containing ampicillin. Transformed bacterial cells may contain either nonrecombinant pUC18 DNA (self-ligated vector only) or recombinant pUC18 DNA (vector containing foreign DNA insert). Both types of transformed bacterial cells will be ampicillin-resistant.

Blue-white screening

To distinguish nonrecombinant from recombinant transformants, blue-white screening or "*lac* selection" (also called α-complementation) can be used with this particular vector (Figs 8.6, 8.7). Bacterial colonies are grown on selective medium containing ampicillin and a colorless chromogenic compound called X-gal, for short (5-bromo-4-chloro-3-indolyl-β-D-galactoside). pUC18 carries a portion of the *lacZ* gene (called

Table 8.3 Some commonly used antibiotics and antibiotic resistance genes.

Antibiotic	Mode of action	Resistance gene
Ampicillin	Inhibits bacterial cell wall synthesis by disrupting peptidoglycan cross-linking	β-Lactamase (*amp*^{r}) gene product is secreted and hydrolyzes ampicillin
Tetracycline	Inhibits binding of aminoacyl tRNA to the 30S ribosomal subunit	*tet*^{r} gene product is membrane bound and prevents tetracycline accumulation by an efflux mechanism
Kanamycin	Inactivates translation by interfering with ribosome function	Neomycin or aminoglycoside phosphotransferase (*neo*^{r}) gene product inactivates kanamycin by phosphorylation

lacZ′) that encodes the first 146 amino acids for the enzyme β-galactosidase (see Section 10.5). The multiple cloning site resides in the coding region. If the *lacZ′* region is not interrupted by inserted DNA, the amino-terminal portion of β-galactosidase is synthesized. Importantly, an *E. coli* deletion mutant strain is used (e.g. DH5α) that harbors a mutant sequence of *lacZ* that encodes only the carboxyl end of β-galactosidase (*lacZ′* ΔM15). Both the plasmid and host *lacZ* fragments encode nonfunctional proteins. However, by α-complementation the two partial proteins can associate and form a functional enzyme. When present, the enzyme β-galactosidase catalyzes hydrolysis of X-gal, converting the colorless substrate into a blue-colored product (see Figs 8.6, 8.7).

Amplification and purification of recombinant plasmid DNA

Further screening of positive (white) colonies can be done by restriction endonuclease digest to confirm the presence and orientation of the insert (see Section 8.9). When a positive colony containing recombinant plasmid DNA is transferred aseptically to liquid growth medium, the cells will continue to multiply exponentially. Within a day or two, a culture containing trillions of identical cells can be harvested.

The final step in molecular cloning is the recovery of the cloned DNA. Plasmid DNA can be purified from crude cell lysates by chromatography (see Tool box 8.1) using silica gel or anion exchange resins that preferentially bind nucleic acids under appropriate conditions and allow for the removal of proteins and polysaccharides. The purified plasmid DNA can then be eluted and recovered by ethanol precipitation in the presence of monovalent cations. Ethanol precipitation of plasmid DNA from aqueous solutions yields a clear pellet that can be easily dissolved in an appropriate buffered solution.

Bacteriophage lambda (λ) as a vector

Bacteriophage lambda (λ) has been widely used in recombinant DNA since engineering of the first viral cloning vector in 1974. Phage λ vectors are particularly useful for preparing genomic libraries, because they can hold a larger piece of DNA than a plasmid vector (see Section 8.5). Today many variations of λ vectors exist. Insertion vectors have unique restriction endonuclease sites that allow the cloning of small DNA fragments in addition to the phage λ genome. These are often used for preparing cDNA expression libraries. Replacement vectors have paired cloning sites on either side of a central gene cluster. This central cluster contains genes for lysogeny and recombination, which are not essential for the lytic life cycle (see Fig. 8.1). The central gene cluster can be removed and foreign DNA inserted between the "arms." All phage vectors used as cloning vectors have been disarmed for safety and can only function in special laboratory conditions. A typical strategy for the use of a phage λ replacement vector is depicted in Fig. 8.8. The recombinant viral particle infects bacterial host cells, in a process called "transduction." The host cells lyse after phage

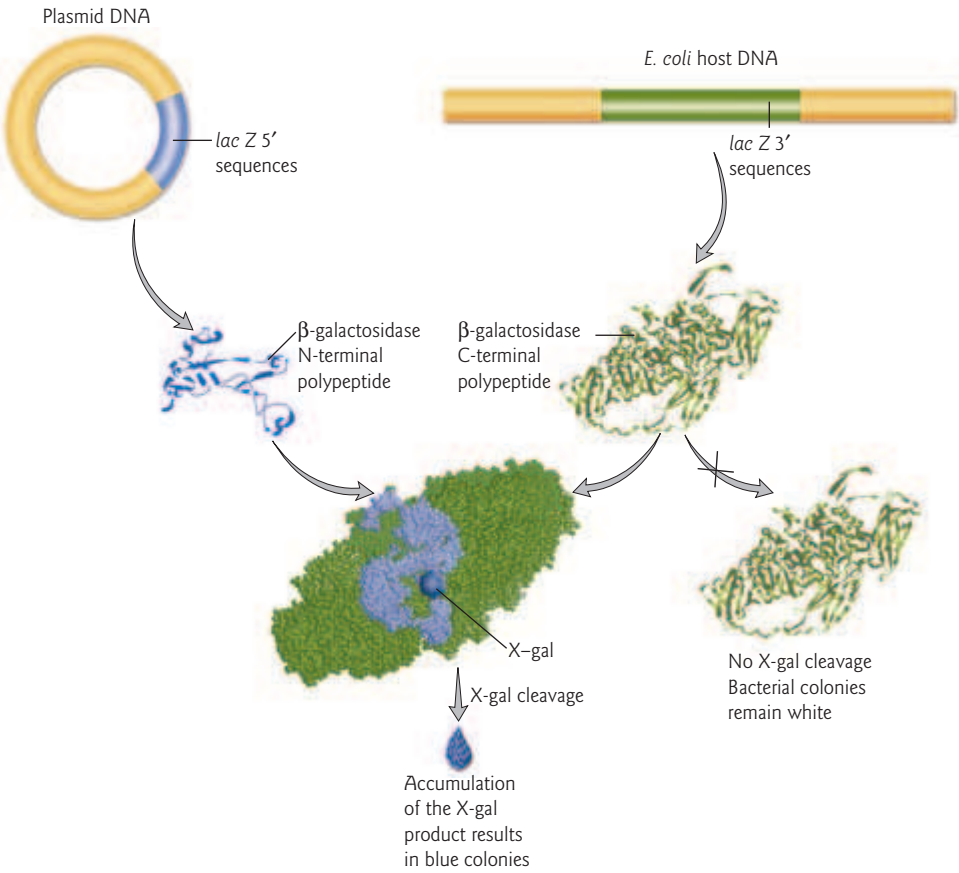


Figure 8.7 β-Galactosidase activity can be used as an indicator of the presence of a foreign DNA insert. Plasmids that express the N-terminal fragment of β-galactosidase (*lacZ* 5′) can be used in *E. coli* strains expressing the C-terminal fragment of the enzyme (*lacZ* 3′ sequences). The N-terminal and C-terminal fragments join and four subunits come together to form a functional tetrameric-enzyme. β-Galactosidase activity can be measured in live cells using a colorless chromogenic substrate called 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal). Cleavage of X-gal produces a blue-colored product that can be visualized as a blue colony on agar plates. If a foreign insert has disrupted the *lacZ* 5′ coding sequence, then only the C-terminal polypeptide will be produced in the bacterial cell. Thus, X-gal is not cleaved and bacterial colonies remain white.

reproduction, releasing progeny virus particles. The viral particles appear as a clear spot of lysed bacteria or "plaque" on an agar plate containing a lawn of bacteria. Each plaque represents progeny of a single recombinant phage and contains millions of recombinant phage particles. Most contemporary vectors carry a *lacZ′* gene allowing blue-white selection.

Artificial chromosome vectors

Bacterial artificial chromosomes (BACs) and yeast artificial chromosomes (YACs) are important tools for mapping and analysis of complex eukaryotic genomes. Much of the work on the Human Genome Project and other genome sequencing projects depends on the use of BACs and YACs, because they can hold greater than 300 kb of foreign DNA. BACs are constructed using the fertility factor plasmid (F factor) of *E. coli* as a starting point. The plasmid is naturally 100 kb in size and occurs at a very low copy number in

TOOL BOX 8.1

Liquid chromatography

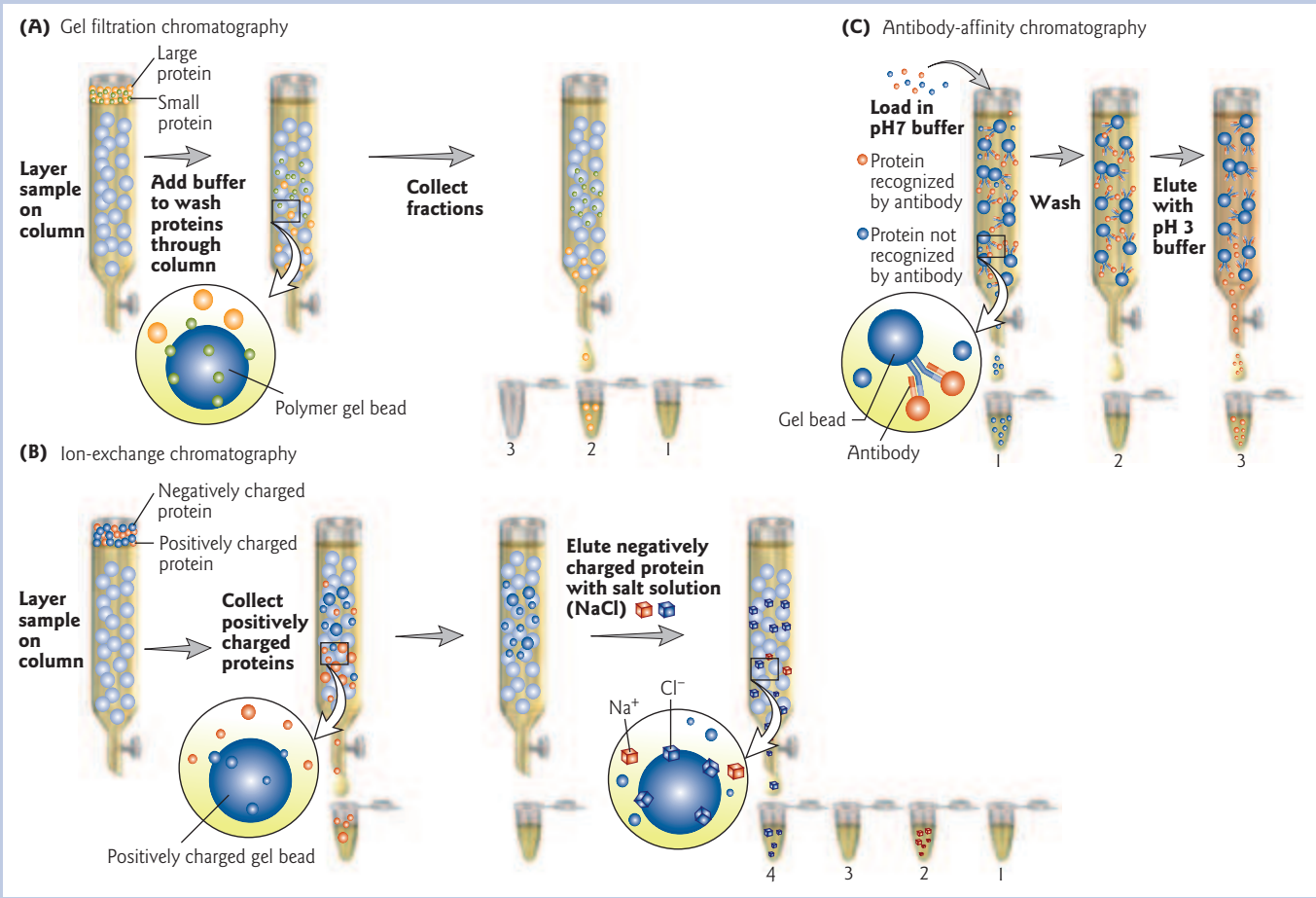


Figure 1 Liquid chromatography techniques. (A) Gel filtration chromatography is used to separate macromolecules that differ in size. For example, a protein mixture is layered on the top of a column packed with porous beads (agarose or polyacrylamide). Larger proteins flow around the beads. Because smaller proteins penetrate into the beads, they travel through the beads more slowly than larger proteins. Different proteins can be collected in separate liquid fractions. (B) Ion-exchange chromatography is used to separate macromolecules (such as proteins or nucleic acids) that differ in net charge. For example, proteins are added to a column packed with beads that are coated by amino (NH^{3+}) or carboxyl (COO^{-}) groups that carry either a positive charge (shown here) or a negative charge at neutral pH. Acidic proteins with the opposite charge (net negative charge) bind to the positively charged beads, while basic or neutral proteins with the same net charge flow through the column. Bound proteins, in this case negatively charged, are eluted by passing a salt gradient through the column. As the negatively charged salt ions bind to the beads, the protein is released. (C) Affinity chromatography relies on the ability of a protein or nucleic acid to bind specifically to another molecule. Columns are packed with beads to which ligand molecules are covalently attached that bind the protein or nucleic acid of interest. Ligands can be antibodies, enzyme substrates, or other small molecules that bind a specific macromolecule. For example, in antibody-affinity chromatography, the column contains a specific antibody covalently attached to beads. Only proteins with a high affinity for the antibody are retained by the column, regardless of mass or charge, while other proteins flow through. The bound protein can be eluted in an acidic solution, by adding an excess of ligand, or by changing the salt concentration.

Liquid chromatography

TOOL BOX 8.1

An important tool in molecular biology is chromatography. The technique of chromatography was first developed in the early 1900s by a botanist named Mikhail Semenovich Tswett. Tswett passed a leaf extract through a vertical tube packed with some absorbent resin. Through this procedure he was able to separate the main green and orange pigments from the leaves. The chlorophylls, xanthophylls, and carotenes appeared as distinct colored bands in the column. Based on these observations, Tswett name the technique "chromatography" (from the Greek word *khroma* for "color," and *graphein*, "to write").

Today, there are many variants of chromatography, but they all rely on the principles first observed by Tswett, that molecules dissolved in a solution will interact (bind and dissociate) with a solid surface. When the solution is allowed to flow across the surface, molecules that interact weakly with the solid surface will spend less time bound to the surface and will move more rapidly than molecules that interact strongly with the surface. Liquid chromatography is commonly used to separate mixtures of nucleic acids and proteins by passing them through a column packed tightly with spherical beads. The nature of these beads determines whether the separation of the nucleic acids or proteins depends on differences in mass (gel filtration chromatography), charge (ion-exchange chromatography), or binding affinity (affinity chromatography) (Fig. 1).

the host. The engineered BAC vector is 7.4 kb (including a replication origin, cloning sites, and selectable markers) and thus can accommodate a large insert of foreign DNA. The characteristics of YAC vectors are discussed below.

Immediately after the construction of the first YAC in 1983, efforts were undertaken to develop a mammalian artificial chromosome (MAC). From there on, it took 14 years until the first prototype MAC was described in 1997. Like YACs, MACs rely on the presence of centromeric sequences, sequences that can initiate DNA replication, and telomeric sequences. Their development is considered an important advance in animal biotechnology and human gene therapy for two main reasons. First, they involve autonomous replication and segregation in mammalian cells, as opposed to random integration into chromosomes (as for other vectors). Second, they can be modified for their use as expression systems of large genes, including not only the coding region but all control elements. A major drawback limiting application at this time, however, is that they are difficult to handle due to their large size and can be recovered only in small quantities. Two principal procedures exist for the generation of MACs. In one method, telomere-directed fragmentation of natural chromosomes is used. For example, a human artificial chromosome (HAC) has been derived from chromosome 21 using this method. Another method involves *de novo* assembly of cloned centromeric, telomeric, and replication origins *in vitro*.

Yeast artificial chromosome (YAC) vectors

Yeast, although a eukaryote, is a small single cell that can be manipulated and grown in the lab much like bacteria. YAC vectors are designed to act like chromosomes. Their design would not have been possible without a detailed knowledge of the requirements for chromosome stability and replication, and genetic analysis of yeast mutants and biochemical pathways. YAC vectors include an origin of replication (autonomously replicating sequence, ARS) (see Section 6.6), a centromere to ensure segregation into daughter cells, telomeres to seal the ends of the chromosomes and confer stability, and growth selectable markers in each arm (Fig. 8.9). These markers allow for selection of molecules in which the arms are joined and which contain a foreign insert. For example, the yeast genes *URA3* and *TRP1* are often used as markers. Positive selection is carried out by auxotrophic complementation of a *ura3-trp1* mutant yeast strain,

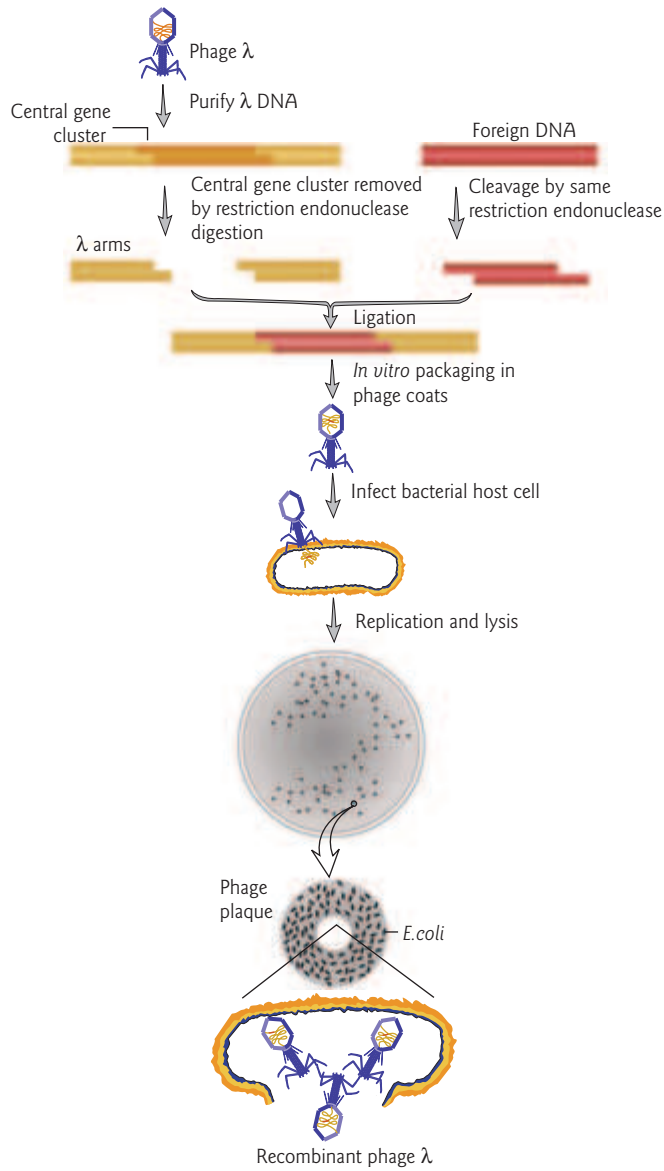


Figure 8.8 Use of bacteriophage lambda (λ) as a cloning vector. DNA is extracted from phage λ and the central gene cluster is removed by restriction endonuclease digestion. The foreign DNA to be cloned is cut with the same enzyme and ligated to the left and right "arms" of the phage λ DNA. The recombinant DNA is then mixed with phage proteins *in vitro*. The DNA is packaged into the phage head and tail fibers are attached via a self-assembly pathway. The recombinant viral particle is then able to infect bacterial cells on an agar plate. The phage replicates its genome, including the foreign DNA insert. Recombinant phage λ DNA directs the cell to make phage particles. The bacteria become filled with new phage particles, break open (lyse), and release millions of recombinant phages. The holes in the lawn of host bacteria, called plaques, are regions where phages have killed the bacteria. Each plaque represents progeny of a single recombinant phage.

which requires supplementation with uracil and tryptophan to grow. *URA3* encodes an enzyme that is required for the biosynthesis of the nitrogenous base uracil (orotidine-5′-phosphate decarboxylase). *TRP1* encodes an enzyme that is required for biosynthesis of the amino acid tryptophan (phosphoribosyl-anthranilate isomerase). YAC vectors are maintained as a circle prior to inserting foreign DNA. After cutting with restriction endonucleases *Bam*HI and *Eco*RI, the left arm and right arm become linear, with the end sequences forming the telomeres. Foreign DNA is cleaved with *Eco*RI and the YAC arms and foreign DNA are ligated and then transferred into yeast host cells. The yeast host cells are maintained as spheroplasts (lacking yeast cell wall). Yeast cells are grown on selective nutrient regeneration plates that lack uracil and tryptophan, to select for molecules in which the arms are joined bringing together the *URA3* and *TRP1* genes.

Red-white selection In the example shown in Fig. 8.9, recombinant YACs are screened for by a "red-white selection" process. Within the multiple cloning site of the YAC in this example, there is another marker,

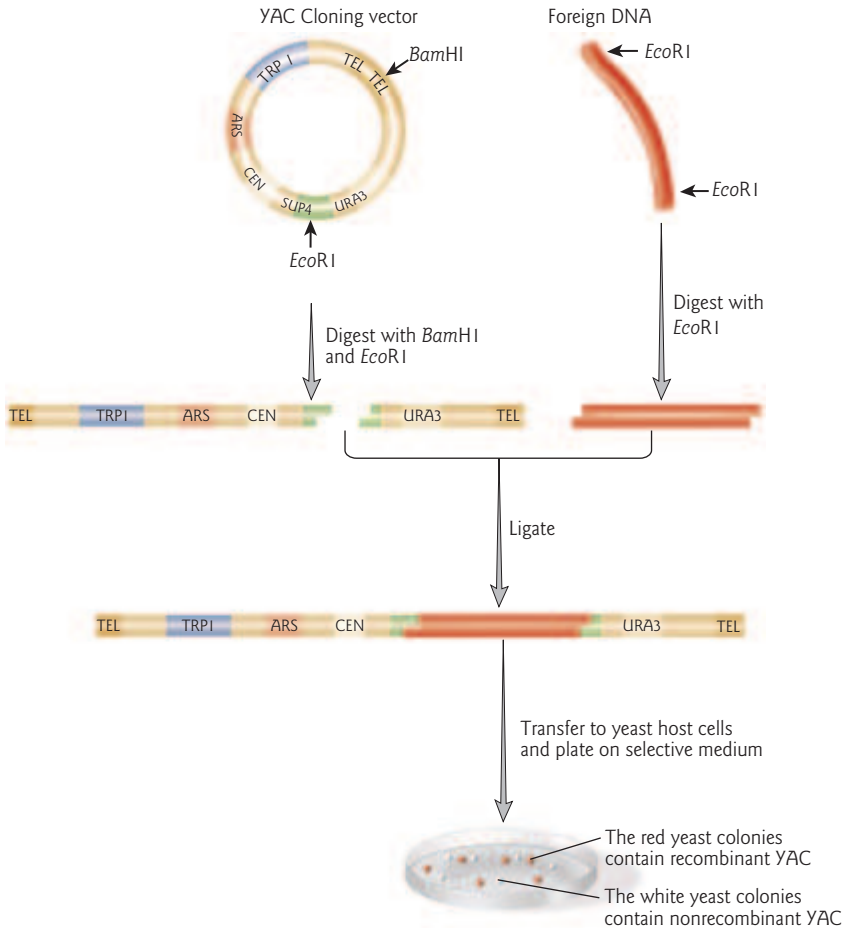


Figure 8.9 Use of yeast artificial chromosome (YAC) cloning vectors. YAC cloning vectors contain functional elements for chromosome maintenance in the yeast *Saccharomyces cerevisiae*. The YAC shown in this example contains an autonomously replicating sequence (ARS) to function as an origin of replication, centromere elements (CEN) for chromosome segregation during cell division, telomeric sequences (TEL) for chromosome stability, and growth selectable markers (URA3 and TRP1) to select positively for chromosome maintenance. Foreign DNA is partially digested with *Eco*RI and the material is then ligated to YAC vector DNA that has been digested with *Bam*HI to liberate telomeric ends and with *Eco*RI to create the insert cloning site. Yeast transformants containing recombinant YAC DNA can be identified by red-white color selection using a yeast strain that is Trp1$^-$ and Ura3$^-$ and contains the *Ade2-1* mutation, which is suppressed by the *SUP4* gene product. Inactivation of *SUP4* by DNA insertion into the *Eco*RI site results in the formation of a red colony.

SUP4. *SUP4* encodes a tRNA that suppresses the *Ade2-1* UAA mutation. *ADE1* and *ADE2* encode enzymes involved in the synthesis of adenine (phosphoribosylamino-imidazole-succinocarbozamide synthetase and phosphoribosylamino-imidazole carboxylase, respectively). In the absence of these critical enzymes, *Ade2-1* mutant cells produce a red pigment, derived from the polymerization of the intermediate phosphoribosylamino-imidazole. But *Ade2-1* mutant cells expressing *SUP4* are white (the color of wild-type yeast strains), because the *Ade2-1* mutation is suppressed. When foreign DNA is inserted in the multiple cloning site, *SUP4* expression is interrupted. In the absence of *SUP4* expression the red pigment reappears because the *Ade2-1* mutation is no longer suppressed. In contrast, the nonrecombinant YAC vectors retain the active *SUP4* suppressor. Thus, red colonies contain recombinant YAC vector DNA, whereas the white colonies contain nonrecombinant YAC vector DNA.

Sources of DNA for cloning

The cloning that has been described so far will work for any random piece of DNA. But since the goal of many cloning experiments is to obtain a sequence of DNA that directs the production of a specific protein, we need to first consider where to obtain such DNA. Sources of DNA for cloning into vectors may be DNA fragments representing a specific gene or portion of a gene, or may be sequences of the entire genome of an organism, depending on the end goal of the researcher. Typical "inserts" include genomic DNA, cDNA (Tool box 8.2), polymerase chain reaction (PCR) products (Tool box 8.3), and chemically synthesized oligonucleotides. When previously isolated clones are transferred into a different vector for other applications, this is called "subcloning."

8.5 Constructing DNA libraries

Vectors are used to compile a library of DNA fragments that have been isolated from the genomes of a variety of organisms. This collection of fragments can then be used to isolate specific genes and other DNA sequences of interest. DNA fragments are generated by cutting the DNA with a specific restriction endonuclease. These fragments are ligated into vector molecules, and the collection of recombinant molecules is transferred into host cells, one molecule in each cell. The total number of all DNA molecules makes up the library. This library is searched, that is screened, with a molecular probe that specifically identifies the target DNA. Once prepared the library can be perpetuated indefinitely in the host cells and is readily retrieved whenever a new probe is available to seek out a particular fragment. Two main types of libraries can be used to isolate specific DNAs: genomic and cDNA libraries.

Genomic library

A genomic library contains DNA fragments that represent the entire genome of an organism. The first step in creating a genomic library is to break the DNA into manageable size pieces (e.g. 15–20 kb for phage λ vectors), usually by partial restriction endonuclease digest. Under limiting conditions, any particular restriction site is cleaved only occasionally, so not all sites are cleaved in any particular DNA molecule. This generates a continuum of overlapping fragments. The second step is to purify fragments of optimal size by gel electrophoresis or centrifugation techniques. The final step is to insert the DNA fragments into a suitable vector. In humans, the genome size is approximately 3×10^9 bp. With an average insert size of 20 kb, the number of random fragments to ensure with high probability (95–99%) that every sequence is represented is approximately 10^6 clones for humans. The maths actually works out to 1.5×10^5 (i.e. $(3 \times 10^9$ bp$)/(2 \times 10^4$ bp)) but more clones are needed in practice, since insertion is random. Bacteriophage λ or cosmid vectors are typically used for genomic libraries. Since a larger insert size can be accommodated by these vectors compared with plasmids, there is a greater chance of cloning a gene sequence with both the coding sequence and the regulatory elements in a single clone.

cDNA library

The principle behind cDNA cloning is that an mRNA population isolated from a specific tissue, cell type, or developmental stage (e.g. embryo mRNA) should contain mRNAs specific for any protein expressed in that cell type or during that stage, along with "housekeeping" mRNAs that encode essential proteins such as the ribosomal proteins, and other mRNAs common to many cell types or stages of development. Thus, if mRNA can be isolated, a small subset of all the genes in a genome can be studied. mRNA cannot be cloned directly, but a cDNA copy of the mRNA can be cloned (see Tool box 8.2). Because a cDNA library is derived from mRNA, the library contains the coding region of expressed genes only, with no introns or regulatory regions. This latter point becomes important for applications of recombinant DNA technology to the production of transgenic animals and for human gene therapy (see Chapters 15 and 17).

8.6 Probes

Searching for a specific cloned DNA sequence in a library is called library screening. One of the key elements required to identify a gene during library screening is the probe. The term probe generally refers to a nucleic acid (usually DNA) that has the same or a similar sequence to that of a specific gene or DNA sequence of interest, such that the denatured probe and target DNA can hybridize when they are renatured together. The probe not only must have the same or a similar sequence to the gene of interest but the researcher must also be able to detect its hybridization. Thus, the probe is labeled; that is, it is chemically modified in some way which allows it, and hence anything it hybridizes to, to be detected. Specific enzymes are used that can add labeled nucleotides in a variety of ways. Typically the probe is made radioactive and added to a solution (Tool boxes 8.4, 8.5). Filters containing immobilized clones are then bathed in the solution. The principle behind this step is that the probe will bind to any clone containing sequences similar to those found on the probe. This binding step is called hybridization. In some cases a library is screened with a protein. For example, when a cDNA library is being screened an antibody can be used to identify the protein that is being expressed by the insert of the clone. In this case, the library is said to be "incubated" with the antibody probe, not hybridized. The use of antibodies in molecular biology research is discussed in more detail in Chapter 9 (Tool box 9.4).

Hybridization can occur between DNA and DNA, DNA and RNA, and RNA and RNA. There are three major types of probe: (i) oligonucleotide probes, which are synthesized chemically and end-labeled; (ii) DNA probes, which are cloned DNAs and may either be end-labeled or internally labeled during *in vitro* replication; and (iii) RNA probes (riboprobes), which are internally labeled during *in vitro* transcription from cloned DNA templates. RNA probes and oligonucleotide probes are generally single-stranded. DNA may be labeled as a double-stranded or single-stranded molecule, but it is only useful as a probe when single-stranded and therefore must be denatured before use. Oligonucleotide, cloned DNA, and RNA probes are of two major types: heterologous and homologous.

Heterologous probes

A heterologous probe is a probe that is similar to, but not exactly the same as, the nucleic acid sequence of interest. If the gene being sought is known to have a similar nucleotide sequence to a second gene that has already been cloned, then it is possible to use this known sequence as a probe. For example, a mouse probe could be used to search a human genomic library.

Homologous probes

A homologous probe is a probe that is exactly complementary to the nucleic acid sequence of interest. Homologous probes can be designed and constructed in a number of different ways. Examples include degenerate probes, expressed sequence tag (EST) based probes, and cDNA probes that are used to locate a genomic clone.

Use of degenerate probes: historical perspective

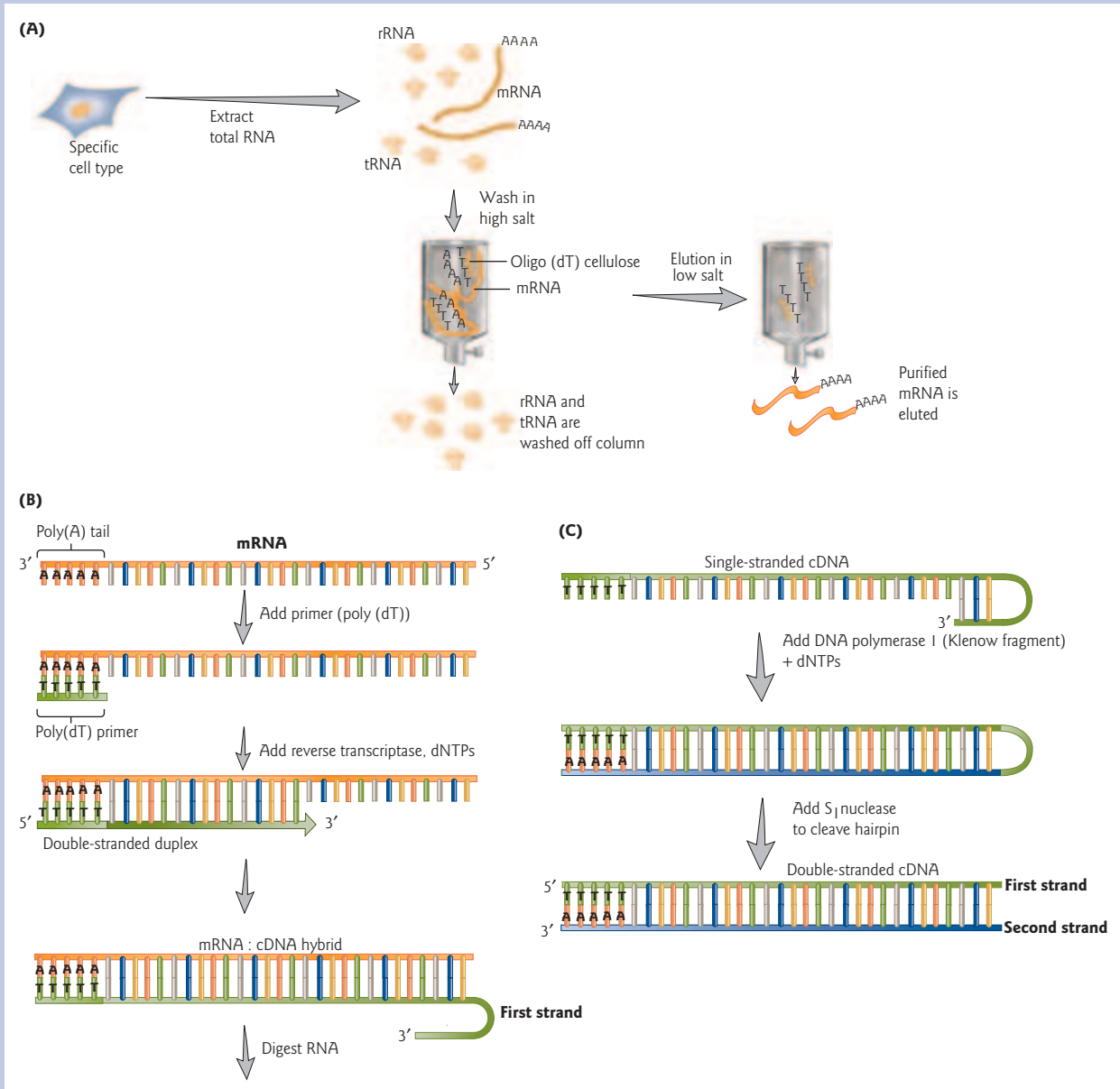
Before the advent of genome sequence databases, the classic method for designing a probe and screening a library relied on having a partial amino acid sequence of a purified protein. To generate an 18–21 nt oligonucleotide probe, all that was required was to know the sequence of about six to seven amino acids. Long before the advent of DNA sequencing, amino acid sequencing was a routine procedure for biochemists. In fact, in 1953, the same year that Watson and Crick proposed the double helix structure of DNA, Frederick Sanger – also at Cambridge University – worked out the sequence of amino acids in the polypeptide chains of the hormone insulin. This was a most important achievement, since it had long been thought that protein sequencing would be a nearly impossible task. Traditionally, protein sequencing was performed using the Edman degradation method (Fig. 8.10). Today, protein sequencing is more often

Complementary DNA (cDNA) synthesis

Most eukaryotic mRNAs are polyadenylated at the 3′ end to form a poly(A) tail (see Section 13.5). This has an important practical consequence that has been exploited by molecular biologists. The poly(A) region can be used to selectively isolate mRNA from total RNA by affinity chromatography (Fig. 1A). The purified mRNA can then be used as a template for synthesis of a complementary DNA (cDNA) (Fig. 1B,C).

Purification of mRNA

Total RNA is extracted from a specific cell type that expresses a specific set of genes. Of this total cellular RNA, 80–90% is rRNA, tRNA, and histone mRNA, not all of which have a poly(A) tail. These RNAs can be separated from the poly(A) mRNA by passing the total RNA through an affinity column of oligo(dT) or oligo(U) bound to resin beads. Under conditions of relatively high salt the poly(A) RNA is retained



Complementary DNA (cDNA) synthesis

TOOL BOX 8.2

by formation of hydrogen bonds with the complementary bases, and the RNA lacking a poly(A) tail flows through. The salt conditions for hybridization are similar to the ion concentration in cells (e.g. 0.3–0.6 M NaCl). The poly(A) mRNA is then eluted from the column in low salt elution buffer (e.g. 0.01 M NaCl), which promotes denaturation of the hybrid (see Section 2.6).

First strand synthesis of cDNA

A number of strategies can be used to synthesize cDNA from purified mRNA. One strategy is as follows. In brief, cDNA is synthesized by the action of reverse transcriptase and DNA polymerase (Fig. 1B). The reverse transcriptase catalyzes the synthesis of a single-stranded DNA from the mRNA template. Like a regular DNA polymerase, reverse transcriptase also needs a primer to get started. A poly(dT) primer is added to provide a free 3′-OH end that can be used for extension by reverse transcriptase in the presence of deoxynucleoside triphosphates (dNTPs). Usually a viral reverse transcriptase is employed such as one from avian myeloblastosis virus (AMV). The reverse transcriptase adds dNTPs from 5′ to 3′ by complementary base pairing. This is called first strand synthesis. The mRNA is then degraded with a ribonuclease or an alkaline solution.

Second strand synthesis of cDNA

For most applications, including cloning of cDNAs, double-stranded DNA is required. The second DNA strand is generated by the Klenow fragment of DNA polymerase I from *E. coli* (Fig. 1C). The 5′ → 3′ exonuclease activity of DNA polymerase I from *E. coli* (see Focus box 6.1) makes it unsuitable for many applications. However, this enzymatic activity can be readily removed from the holoenzyme by exposure to a protease. The large or Klenow fragment of DNA polymerase I generated by proteolysis has 5′ → 3′ polymerase and 3′ → 5′ exonuclease (proofreading) activity, and is widely used in molecular biology. Commercially available Klenow fragments are usually produced by expression in bacteria from a truncated form of the DNA polymerase I gene.

There is a tendency for the reverse transcriptase enzyme used in first strand synthesis to loop back on itself and start to make another complementary strand. This hairpin forms a natural primer for DNA polymerase and a second strand of DNA is generated. S1 nuclease (from *Aspergillus oryzae*) is then added to cleave the single-stranded DNA hairpin. Double-strand DNA linkers with ends that are complementary to an appropriate cloning vector are added to the double-strand DNA molecule before ligation into the cloning vector. The end result is a double-stranded cDNA in which the second strand corresponds to the sequence of the mRNA, thus representing the coding strand of the gene. The sequences that appear in the literature are the 5′ → 3′ sequences of the second strand cDNA (Fig. 1). Sequences corresponding to introns and to promoters and all regions upstream of the transcriptional start site are not represented in cDNAs. The library created from all the cDNAs derived from the mRNAs in the specific cell type forms the cDNA library of cDNA clones.

Figure 1 (*opposite*) Traditional cDNA synthesis. (A) Purification of mRNA. Total RNA is extracted from a specific cell type and loaded on an oligo(dT) affinity chromatography column under conditions (high salt buffer) that promote hybridization between the 3′ poly(A) tails of the mRNA and the oligo(dT) covalently coupled to the column matrix. After hybridization, the rRNAs and tRNAs are washed out of the column. The mRNA is eluted with a low salt buffer. The resulting purified mRNA contains many different mRNAs encoding different proteins. (B) First strand synthesis. Synthesis of the first strand of cDNA is carried out using the enzyme reverse transcriptase and a poly(dT) primer in the presence of dNTPs. An mRNA–cDNA hybrid is produced and the mRNA is then digested with an alkaline solution or the enzyme ribonuclease. (C) Second strand synthesis. Synthesis of double-stranded cDNA uses a self-priming method. The Klenow fragment of DNA polymerase I catalyzes synthesis of the second strand, using the natural hairpin of the first strand as a primer. The hairpin is cleaved with a single-strand DNA nuclease (S1 nuclease). The end result is a collection of double-strand cDNAs that correspond to the sequences of the many different mRNAs extracted from the cell.

TOOL BOX 8.3

Polymerase chain reaction (PCR)

The polymerase chain reaction (PCR) is the one of the most powerful techniques that has been developed recently in the area of recombinant DNA research. PCR has had a major impact on many areas of molecular cloning and genetics. With this technique, a target sequence of DNA can be amplified a billion-fold in several hours. Amplification of particular segments of DNA by PCR is distinct from the amplification of DNA during cloning and propagation within a host cell. The procedure is carried out entirely *in vitro*. In addition to its use in many molecular cloning strategies, PCR is also used in the analysis of gene expression (see Section 9.5), forensic analysis where minute samples of DNA are isolated from a crime scene (see Section 16.2), and diagnostic tests for genetic diseases (see Disease box 8.1).

PCR is a DNA polymerase reaction. As with any DNA polymerase reaction it requires a DNA template and a free 3′-OH to get the polymerase started. The template is provided by the DNA sample to be amplified and the free 3′-OH groups are provided by site-specific oligonucleotide primers. The primers are complementary to each of the ends of the sequence that is to be amplified. Note that *in vivo* DNA polymerase would use an RNA primer (see Section 6.4), but a more stable, more easily synthesized DNA primer is used *in vitro*. The three steps of the reaction are denaturation, annealing of primers, and primer extension (Fig. 1):

1 **Denaturation**. In the first step, the target sequence of DNA is heated to denature the template strands and render the DNA single-stranded.
2 **Annealing**. The DNA is then cooled to allow the primers to anneal, that is, to bind the appropriate complementary strand. The temperature for this step varies depending on the size of the primer, the GC content, and its homology to the target DNA. Primers are generally DNA oligonucleotides of approximately 20 bases each.
3 **Primer extension**. In the presence of Mg^{2+}, DNA polymerase extends the primers on both strands from 5′ to 3′ by its polymerase activity. Primer extension is performed at a temperature optimal for the particular polymerase that is used. Currently, the most popular enzyme for this step is *Taq* polymerase, the DNA polymerase from the thermophilic (heat-loving) bacteria *Thermus aquaticus*. This organism lives in hot springs that can be near boiling and thus requires a thermostable polymerase.

These three steps are repeated from 28 to 35 times. With each cycle, more and more fragments are generated with just the region between the primers amplified. These accumulate exponentially. The contribution of strands with extension beyond the target sequence becomes negligible since these accumulate in a linear manner. After 25 cycles in an automated thermocycler machine, there is a 2^{25} amplification of the target sequence. PCR products can be visualized on a gel stained with nucleic acid-specific fluorescent compounds such as ethidium bromide or SYBR green. The error rate of *Taq* is 2×10^{-4}. If an error occurs early on in the cycles, it could become prominent. Other polymerases, such as *Pfu*, have greater fidelity. *Pfu* DNA polymerase is from *Pyrococcus furiosus*. Base misinsertions that may occur infrequently during polymerization are rapidly excised by the $3' \rightarrow 5'$ exonuclease (proofreading) activity of this enzyme.

When Kary Mullis first developed the PCR method in 1985, his experiments used *E. coli* DNA polymerase. Because *E. coli* DNA polymerase is heat-sensitive, its activity was destroyed during the denaturation step at 95°C. Therefore, a new aliquot of the enzyme had to be added in each cycle. The purification, and ultimately the cloning, of the DNA polymerase from *T. aquaticus* made the reaction much simpler. In his first experiments, Mullis had to move the reaction manually between the different temperatures. Fortunately, this procedure has been automated by the development of thermal cyclers. These instruments have the capability of rapidly switching between the different temperatures that are required for the PCR reaction. Thus the reactions can be set up and placed in the thermal cycler, and the researcher can return several hours later (or the next morning) to obtain the products.

Polymerase chain reaction (PCR)

TOOL BOX 8.3

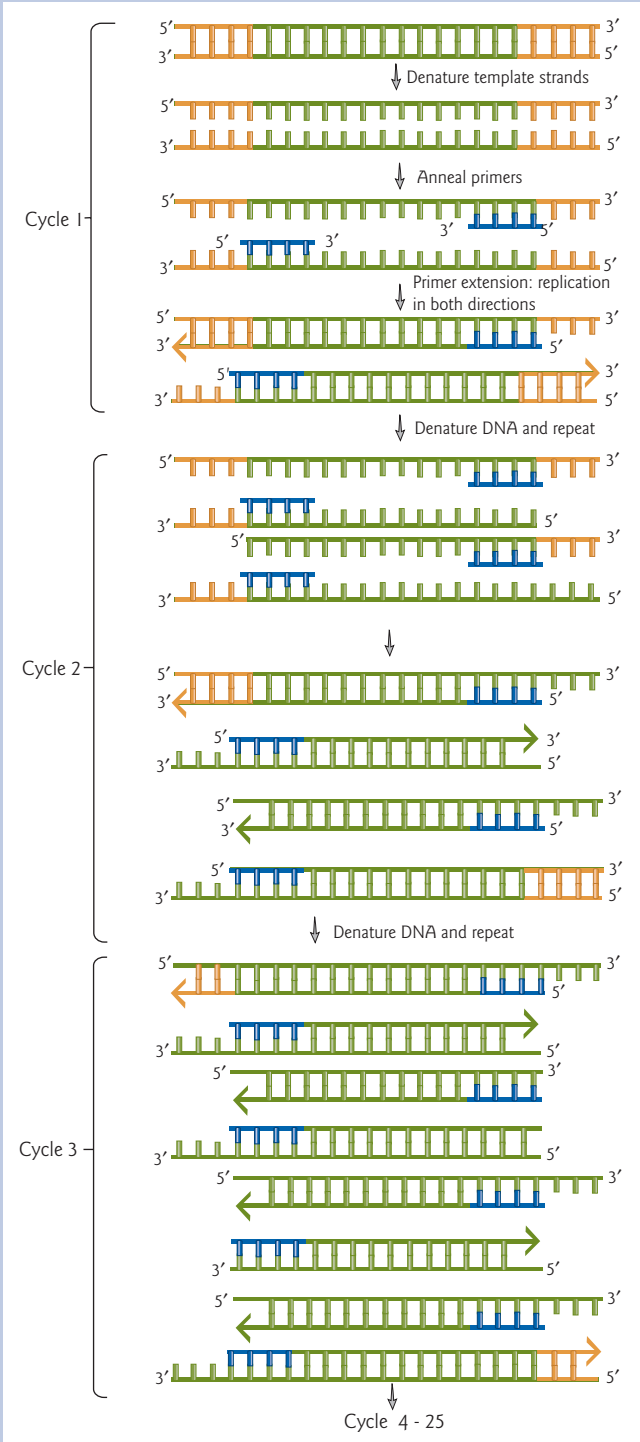


Figure 1 Polymerase chain reaction (PCR). PCR is an *in vitro* DNA replication method. The starting material is a double-stranded DNA. The target sequence to be amplified is indicated in green. Large numbers of two primers (blue) are added, each with a sequence complementary to that found in one strand at the end of the region to be amplified. A thermostable DNA polymerase (e.g. *Taq* polymerase) and dNTPs are also added. In the first cycle, heating to 95°C denatures the double-stranded DNA and subsequent cooling to 55–65°C then allows the primers to anneal to their complementary sequences in the target DNA. *Taq* polymerase extends each primer from 5′ to 3′, generating newly synthesized strands in both directions, which extend to the end of the template strands. The extension is performed at 72°C. In the second cycle, the original and newly made DNA strands are denatured at 95°C and primers are annealed to their complementary sequences at 55–65°C. Each annealed primer again is extended by *Taq* polymerase. In the third cycle, two double-strand DNA molecules are generated exactly equal to the target sequence. These two are doubled in the fourth cycle and are doubled again with each successive cycle.

TOOL BOX 8.4

Radioactive and nonradioactive labeling methods

The world of cells and macromolecules is invisible to the unaided eye. Development of the tools of molecular biology has allowed researchers to make this world visible. Since World War II, when radioactive materials first became widely available as byproducts of work in nuclear physics, they have become indispensible tools for detecting biological molecules. Hundreds of biological compounds (e.g. nucleotides, amino acids, and numerous metabolic intermediates) are commercially available. The presence of a radioisotope does not change the chemical properties of a radioactively labeled precursor of a macromolecule. Enzymes, both *in vitro* and *in vivo*, catalyze reactions involving labeled substrates just as readily as those involving nonlabeled substrates. Because radioisotopes emit easily detected particles, the fate of radiolabeled molecules can be traced in cells and cellular extracts. For example, labeling nucleic acids is important for tracking their localization, for defining synthetic processes, and for labeling of hybridization probes.

Commonly used radioisotopes in molecular biology

Radioisotopes are unstable isotopes of an element. Isotopes of a given element contain the same number of protons but a different number of neutrons. During radioactive decay there is a change in the number of neutrons and protons from an unstable combination to a more stable combination. The nuclide has less mass after decay (mass converted to energy). For the radioisotopes used in molecular biology research, this energy is emitted as beta (β) particles (small, electrically charged particles that are identical to electrons) or gamma (γ) rays. For example, the amino acids methionine and cysteine labeled with sulfur-35 (^{35}S) are widely used to label cellular proteins (Table 1). Phosphorus-32 (^{32}P) labeled nucleotides are routinely used to label both RNA and DNA in cell-free systems (*in vitro*). For metabolic labeling (labeling *in vivo*), compounds labeled with hydrogen-3 (^{3}H, tritium) are more commonly used. For example, to identify the site of RNA synthesis, cells can be incubated for a short period with ^{3}H-uridine and then subjected to a fractionation procedure to separate the various organelles or to autoradiography. The radioisotope iodine-125 (^{125}I) is often covalently linked to antibodies for the detection of specific proteins.

Detection techniques

The technique of autoradiography makes use of the fact that radioactive isotopes expose photographic film. The visible silver grains on the film can be counted to provide an estimate of the quantity of radioactive material present. Quantitative measurements of radioactivity in a labeled material can also be performed with several different instruments. A Geiger counter measures ions produced in a gas by β-particles or γ-rays emitted from a radioisotope. In a scintillation counter, a radiolabeled sample is mixed with a liquid containing a fluorescent compound that emits a flash of light when it absorbs the energy of the β-particles

Table 1 Some radioisotopes commonly used in molecular biology research.

Radioisotope	Symbol	Labeled macromolecule	Half-life*	Application
Tritium (hydrogen-3)	^{3}H	Nitrogenous bases (e.g. ^{3}H-uridine, ^{3}H-thymidine) ^{3}H-dNTPs, ^{3}H-NTPs	12.28 years	RNA and DNA labeling *in vivo* Probes for *in situ* hybridization
Carbon-14	^{14}C	^{14}C-chloramphenicol	5730 years	CAT assays
Phosphorus-32	^{32}P	^{32}P-NTPs ^{32}P-dNTPs	14.29 days	Hybridization probes
Sulfur-35	^{35}S	Amino acids (^{35}S-cysteine, ^{35}S-methionine)	87.4 days	Protein labeling (*in vivo* and *in vitro*)
Iodine-125	^{125}I	N/A	60.14 days	Antibody labeling

* The half-life is a means of classifying the rate of decay of radioisotopes according to the time it takes them to lose half their strength (intensity).

Radioactive and nonradioactive labeling methods

TOOL BOX 8.4

or γ-rays released during decay of the radioisotope; a phototube in the instrument detects and counts these light flashes. Phosphorimagers are used to detect radiolabeled compounds on a surface, storing digital data on the number of decays in disintegrations per minute (dpm) per small pixel of surface area. These instruments are commonly used to quantitate radioactive molecules separated by gel electrophoresis and are replacing photographic film for this purpose.

Nonradioactive labeling

As noted above, traditionally, nucleic acids have been labeled with radioisotopes. These radiolabeled probes are very sensitive, but their handling is subject to stringent safety precautions regulated in the US by the federal Nuclear Regulatory Commission, and, in the case of ^{32}P and ^{35}S, the signal decays relatively quickly. More recently, a series of nonradioactive labeling methods have been developed that generate colorimetric or chemiluminescent signals. A widely used label is digoxygenin, a plant steroid isolated from foxglove, *Digitalis*. This can be conjugated to nucleotides and incorporated into DNA, RNA, or oligonucleotide probes and then detected using an antibody to digoxygenin. The antibody can be attached covalently (conjugated) to fluorescent dyes or enzymes that facilitate signal detection. For example, often anti-digoxygenin antibodies are conjugated to the enzyme alkaline phosphate. When a specific substrate is added, the attached enzyme catalyzes a chemical reaction producing light which exposes an X-ray film. Another system uses biotin, a vitamin, and the bacterial protein streptavidin, which binds to biotin with extremely high affinity. Biotin-conjugated nucleotides are incorporated as a label and detected using enzyme-conjugated streptavidin (see Focus box 10.1 for an application).

performed using mass spectrometry technology, such as matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) (see Fig. 16.15).

In the example shown in Fig. 8.11, all possible oligonucleotide combinations are synthesized as probes. This is based on the degeneracy of the genetic code (see Section 5.3). Some amino acids are coded for by more than one triplet combination. This can be optimized by choosing a region of the protein that has a high percentage of single or two codon amino acids. The oligonucleotides are made synthetically in the lab and then used to screen a library to identify the gene (or cDNA) encoding the purified protein. One of the oligonucleotides will be exactly the same as the cloned gene. For organisms with sequenced genomes (see Sections 16.4 and 16.5), the protein sequence of their gene products can be simply deduced from the DNA sequence of the respective genes. As a result of this, and the development of EST-based probes, degenerate probes are rarely used today.

Unique EST-based probes

The use of EST-based probes is a newer method than making degenerate probes. Although EST-based probes are no longer frequently used for organisms in which the whole genome sequence is available, they are still useful for organisms where only limited sequence information is available. ESTs are partial cDNA sequences of about 200–400 bp (because they represent just a short portion of the cDNA, they are called "tags"). This method uses cDNA sequence data and identifies a single oligonucleotide rather than a degenerate mixture. A computer program applies the genetic code to translate an EST into a partial amino acid sequence. If a match is found with the protein under study, the EST provides the unique DNA sequence of that portion of cDNA. A probe can then be synthesized and used to screen a library for the entire cDNA (or genomic) clone.

Nucleic acid labeling

There are a variety of methods for labeling RNA and DNA. The choice of method depends on the application. Probes of the highest specific activity (proportion of incorporated label per mass of probes) are generated using internal labeling, where many labeled nucleotides are incorporated uniformly during DNA or RNA synthesis *in vitro*. End-labeling involves either adding a labeled nucleotide to the 3′-hydroxyl end of a DNA strand or exchanging the unlabeled 5′-phosphate group for a labeled phosphate. When deciding on a labeling method, one of the first questions to ask is whether internal labeling or end-labeling of the nucleic acid is desirable (Table 1). Internal labeling provides maximal labeling and is often used for probes to screen libraries or for Southern blots (see Tool box 8.7). End-labeling is used when precise definition of one end of the DNA (5′ or 3′) is required. An example of this is labeling a DNA fragment for use in DNase I footprinting (see Fig. 9.15), where the researcher wants to know the orientation of the fragments; that is, is a protein-binding site near the 5′ or 3′ end of the DNA sequence? Common methods of uniform labeling involve DNA or RNA synthesis reactions; for example, random primed labeling and synthesis of riboprobes. Some methods for end-labeling DNA fragments involve DNA synthesis reactions (e.g. Klenow fill-in), while oligonucleotides are generally end-labeled by other enzyme-mediated reactions.

Random primed labeling

Random primed labeling is a method of incorporating radioactive nucleotides along the length of a fragment of DNA. The DNA is denatured and random oligonucleotides are annealed to both strands. Each batch of random oligonucleotides contains all possible sequences (for hexamers, which are most commonly employed, this would be 4096 different oligonucleotides), so any DNA template can be used with this method. The odds are that some of these primers will anneal to the DNA of interest. The oligonucleotide primers provide the required free 3′-hydroxyl group for the initiation of DNA synthesis. A Klenow fragment of *E. coli* DNA polymerase I is then used to extend the oligonucleotides, using three unlabeled ("cold") nucleotides and one radioactively labeled ("hot") nucleotide (or a nonradioactively labeled nucleotide) provided in the reaction mixture, to produce a uniformly labeled double-stranded probe. Subsequently, the double-stranded probe is denatured and added to a hybridization solution (Fig. 1A).

In vitro transcription

RNA can be labeled by *in vitro* transcription from a DNA template. The DNA template (often a cDNA) is cloned into a special plasmid vector so that it can be transcribed under the control of a promoter specific for recognition by a bacteriophage RNA polymerase, typically SP6 RNA polymerase, T7 RNA polymerase, or T3 RNA polymerase. The transcription reaction is carried out *in vitro* by the addition of all four NTPs, with one or more labeled, and the appropriate phage RNA polymerase (Fig. 1B). Labeled RNA can be used for tracking the movement and localization of RNA transcripts in cells, analyzing RNA processing pathways, and as hybridization probes. RNA probes (riboprobes) may be either complementary to the sense or antisense strand of DNA depending on the purpose.

Klenow fill-in

A "fill-in" reaction is used to generate blunt ends on fragments created by cleavage with restriction endonucleases that leave 5′ single-stranded overhangs. The Klenow fragment of *E. coli* DNA polymerase I is used to fill in the gaps from 5′ to 3′, in the presence of dNTPs, including one labeled dNTP (Fig. 1C). The result is a double-stranded DNA with the 3′ ends labeled.

Table 1 Labeling of nucleic acids.

Labeling method	Type of labeling	Enzyme	Example of application
Random priming	Uniform	Klenow DNA polymerase	Hybridization probes
***In vitro* transcription**	Uniform	Phage SP6, T7, or T3 RNA polymerase	Hybridization probes, tracking RNA localization
Klenow fill-in	3′ end-labeling	Klenow DNA polymerase	DNase I footprinting
Oligonucleotides	5′ end-labeling 3′ end-labeling	T4 polynucleotide kinase Terminal transferase	Hybridization probes, EMSA

Oligonucleotide labeling

Methods for end-labeling oligonucleotides do not involve DNA synthesis reactions. Instead, they make use of other enzymes. In 5′ end-labeling, a γ-phosphate from ATP is added to the 5′ end of an oligonucleotide by the enzyme T4 polynucleotide kinase. In 3′ end-labeling, a labeled dNTP is added to the 3′ end by the enzyme terminal transferase.

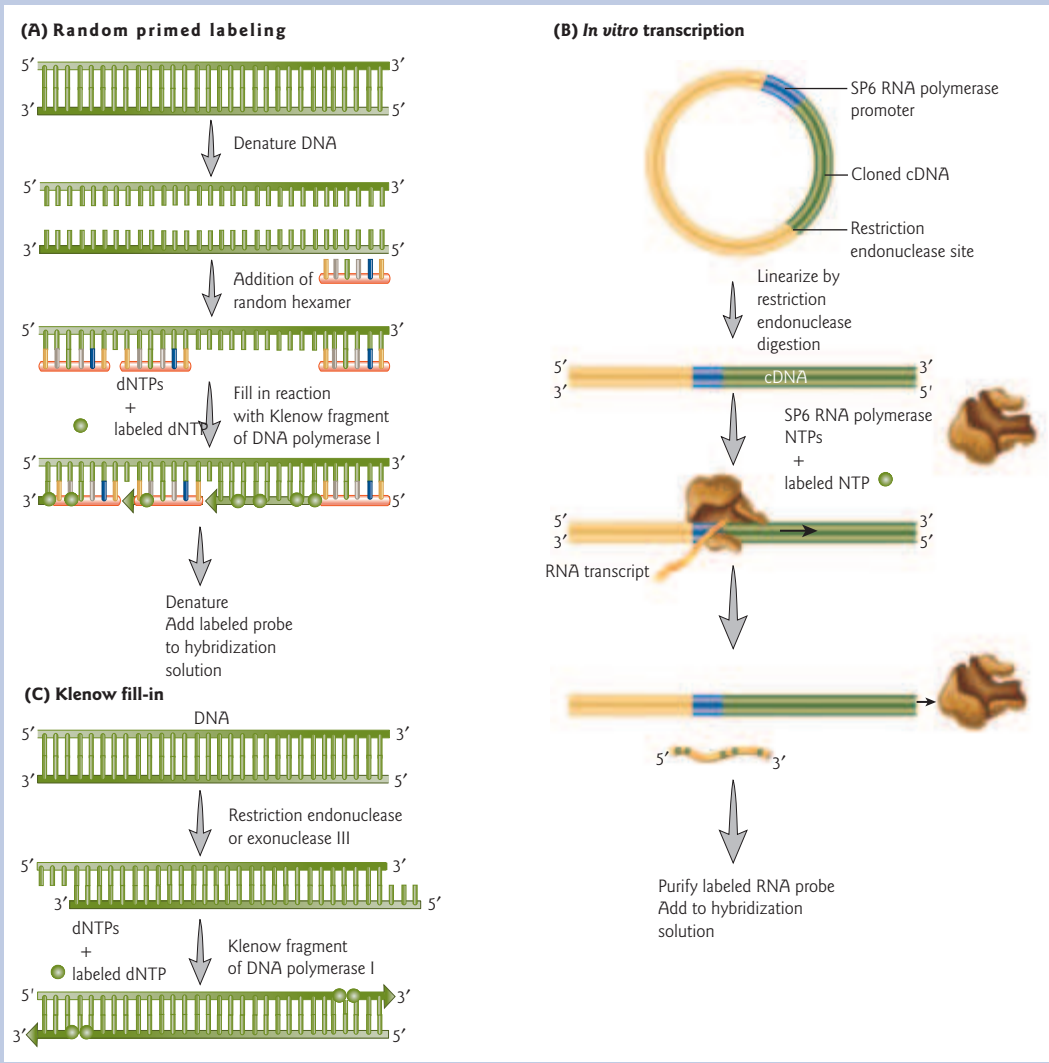


Figure 1 Some methods for labeling nucleic acids. Two methods of uniform labeling and one method of end-labeling are depicted that involve nucleic acid synthesis reactions. (A) In random primed labeling, the template DNA is denatured and random hexamer primers are annealed to both strands. Only one strand is shown for simplicity. The primers provide the 3′-OH for the initiation of DNA synthesis upon addition of the Klenow fragment of DNA polymerase I, and unlabeled and labeled dNTPs. The resulting labeled double-stranded DNA probe is denatured and added to a hybridization solution. (B) *In vitro* transcription generates a labeled RNA from a DNA template. The DNA template (cDNA) is cloned in a plasmid vector containing a promoter for bacteriophage SP6 RNA polymerase. The transcription reaction is carried out by the addition of all labeled and unlabeled NTPs and SP6 RNA polymerase. The labeled RNA probe can then be purified and added to a hybridization solution. (C) Klenow fill-in is used to create labeled blunt ends on fragments created by cleavage with a restriction endonuclease that leaves a 5′ overhang. The reaction is conceptually identical to the one described in (A), but with a long primer and a very short segment of single-stranded template.

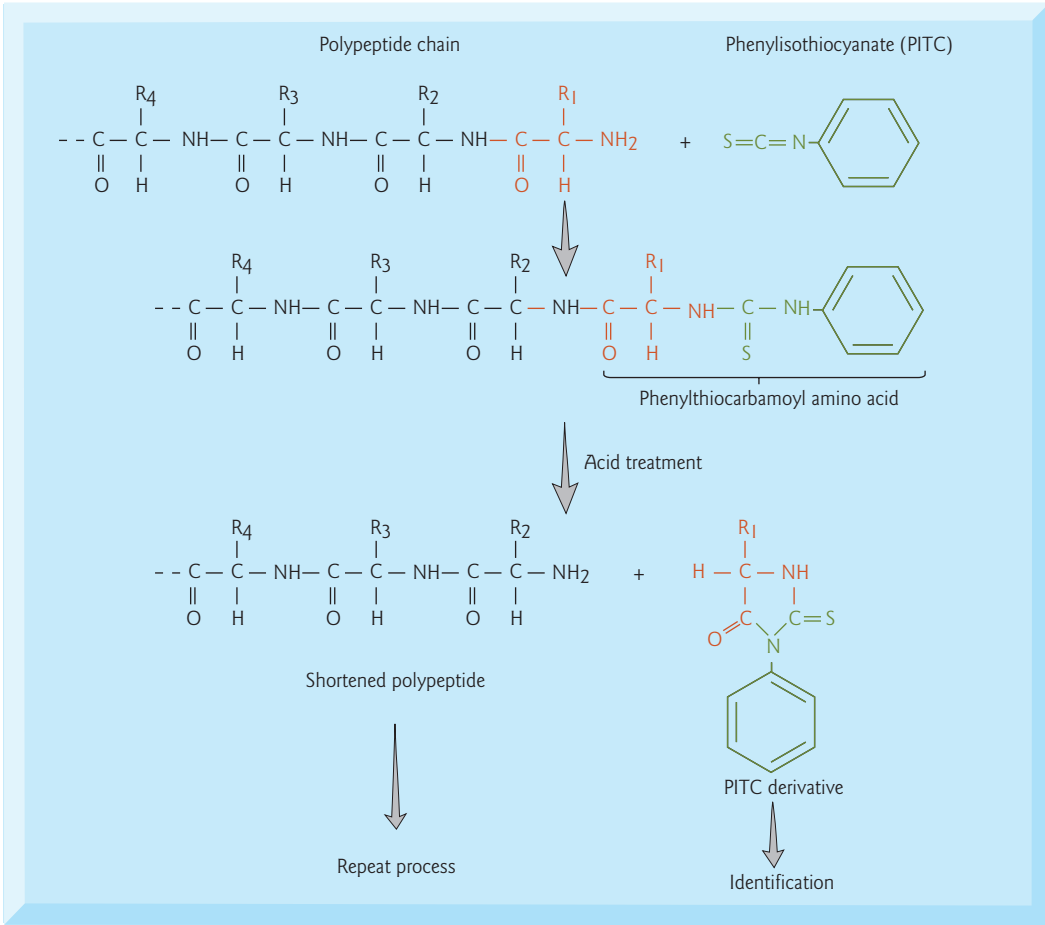


Figure 8.10 Protein sequencing by the Edman degradation method. The NH_2-terminal end of the polypeptide to be sequenced is chemically modified by bonding to phenylisothiocyanate (PITC). Acid treatment results in a PITC derivative of the NH_2-terminal amino acid (phenylthiocarbamoyl amino acid) and a polypeptide that is one amino acid shorter. The PITC derivative can then be identified by analytical methods. The process is repeated for each amino acid at the NH_2-terminus until the polypeptide is sequenced.

Using an identified cDNA to locate a genomic clone

Since a cDNA contains only the coding region of a gene, researchers often need to isolate a genomic clone for analysis of regulatory regions, introns, etc. Use of an identified cDNA to locate a genomic clone provides a highly specific probe for the gene of interest.

8.7 Library screening

Nowadays, because of the wealth of genomic sequence data available for many organisms, a DNA sequence of interest is more likely to be isolated by polymerase chain reaction (PCR) (see Tool box 8.3) than by a library screen. DNA cloning and PCR both amplify tiny samples of DNA into large quantities by repeated rounds of DNA duplication, either carried out by cycles of cell division in a host or cycles of DNA synthesis *in vitro*. In PCR, the pair of oligonucleotide primers limits the amplification process to the particular DNA sequence of interest from the beginning. In contrast, once prepared, a DNA library can be

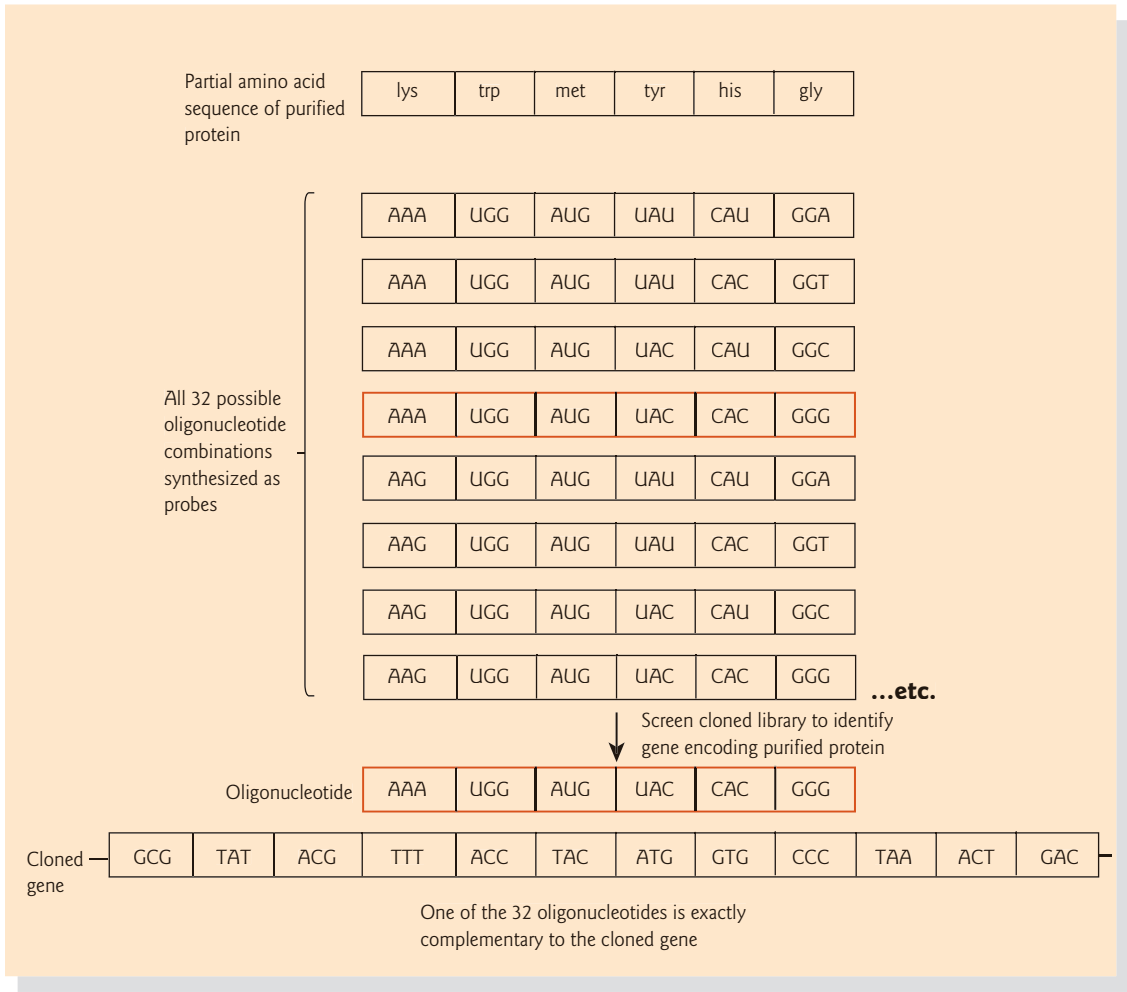


Figure 8.11 Generating degenerate oligonucleotide probes. The coding DNA sequence can be deduced from the partial amino acid sequence of a protein. In this example, two amino acids, tryptophan (Trp) and methionine (Met) have only one codon each. Three others, lysine (Lys), histidine (His), and tyrosine (Tyr) are encoded by two codons, while glycine (Gly) is encoded by codons that include all base combinations at the third position. For this amino acid sequence, 32 oligonucleotide sequences encompass all the possible combinations of codons. For simplicity, only eight of these are depicted in the diagram. The exact coding sequence of the gene of interest must be one of the base combinations. Using a mixture of radioactively labeled oligonucleotides as probes, a cloned library can be screened to isolate the gene for the complete protein.

perpetuated indefinitely in host cells, and can be readily retrieved whenever the researcher wants to seek out some particular fragment. Assuming a probe is available for the cloned sequence of interest, the library can be screened using the principles of hybridization. Complementary base pairing of single-stranded nucleic acids underlies some of the most important biological processes: replication, recombination, DNA repair, transcription, and translation. Library screening exploits this fundamental ability of double-stranded nucleic acids to undergo denaturation or melting (separation into single strands) and for complementary single strands to spontaneously anneal to a labeled nucleic acid probe to form a hybrid duplex (heteroduplex). The power of the technique is that the labeled nucleic acid probe can detect a complementary molecule in a complex mixture with exquisite sensitivity and specificity.

Like many procedures in molecular biology, the process of library screening sounds simple, but in practice it can be labor intensive, and may result in "false positives" if the hybridization conditions are not stringent enough. The example shown in Fig. 8.12 is for screening a cDNA library cloned into plasmid vectors. A similar protocol would be employed for screening a phage library, but would involve plaque hybridization instead of colony hybridization.

Transfer of colonies to a DNA-binding membrane

Bacterial colonies with recombinant vectors containing inserts representing the entire library are grown on nutrient agar plates, forming hundreds or thousands of colonies (Fig. 8.12). The colonies (members of the library) are transferred to nitrocellulose or nylon membranes by gently pressing to make a replica. Bacteria attach to the membrane and the cells are then lysed and the DNA purified by treatment with alkali and proteases. The DNA is denatured to make it single stranded, and fixed to the membrane either by heat treatment or ultraviolet irradiation. The DNA is covalently bound by its sugar–phosphate backbone and the unpaired bases are exposed for complementary base pairing.

Colony hybridization

In the next step of library screening, a radioactively labeled, single-stranded DNA probe is applied (Fig. 8.12). The hybridization step is performed at a nonstringent temperature that ensures the probe will bind to any clone containing a similar sequence. At the same time, some nonspecific hybridization will occur because some of the clones will contain limited, but not significant, similarity to the probe. A series of washes are performed at a stringent temperature that is high enough to remove the probe from all clones to which it has bound in a nonspecific manner. Heteroduplex stability is influenced by the number of hydrogen bonds between the bases and base stacking by hydrophobic interactions that hold the two single strands together. The number of hydrogen bonds is determined by various properties of the heteroduplex, including the length of the duplex, its GC content, and the degree of mismatch between the probe and the complementary target DNA sequence. The shorter the duplex, the lower the GC content, and the more mismatches there are, the lower the melting temperature (T_m) will be because there are fewer hydrogen bonds and base-stacking interactions to disrupt (see Section 2.6). The appropriate hybridization temperature is calculated according to the GC content and the percent homology of probe to target according to the following equation:

$$T_m = 49.82 + 0.41(\%\ G + C) - (600/l)$$

where l is the length of the hybrid in base pairs.

It is important that the temperature is not so high that it removes the probe from clones that contain sequences that are similar (only a few mismatches) or identical to the probe itself (exact complementarity). Therefore, consideration about the source of the probe (homologous or heterologous) determines the temperature at which the high stringency washes are performed.

Detection of positive colonies

In the final phase of library screening, an X-ray film is applied to the membrane and is exposed by any remaining specifically bound radioactive probe. The resulting autoradiogram has a dark spot on the developed film where DNA–DNA hybrids have formed, by virtue of sequence complementarity to the radiolabeled probe. If the gene is large, it may be fragmented. Various fragments in different clones may need to be identified by finding overlapping fragments and reconstructing the order. The original plate is used to pick bacterial cells with recombinant plasmids that hybridized to the probe. Cells are transferred to medium for growth and further analysis.

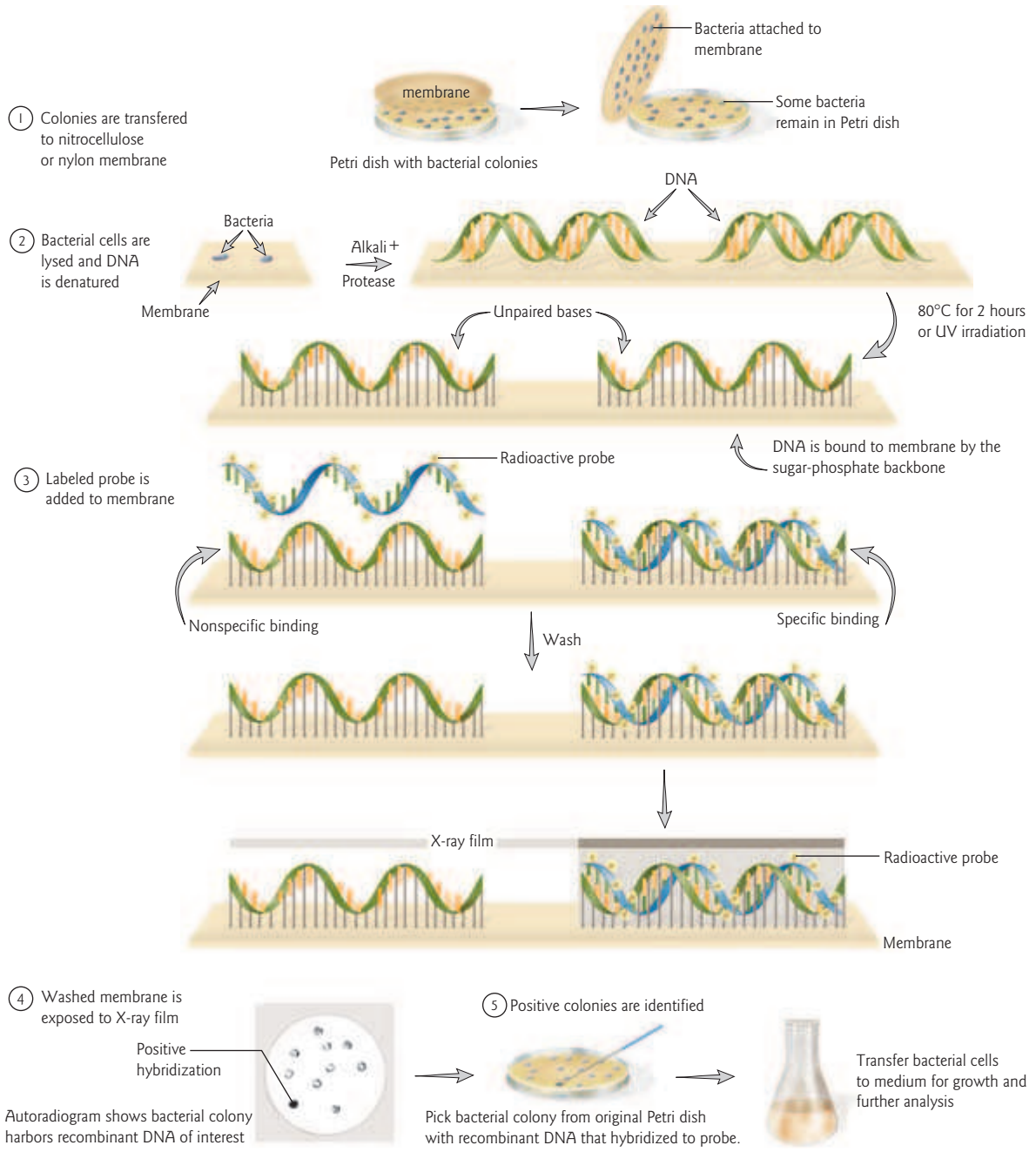


Figure 8.12 Screening a library by nucleic acid hybridization. The example depicts a method for screening a cDNA library. Colony hybridization is used to identify bacterial cells that harbor a specific recombinant plasmid. (1) A sample of each bacterial colony in the library is transferred to a membrane. (2) The bacterial cells on the membrane are lysed and the DNA is denatured. Using heat or UV treatment, the DNA is covalently bound to the membrane. (3) The membrane is placed in a hybridization bag along with a labeled single-stranded DNA probe. After hybridization, the membrane is removed from the bag and washed to remove excess probe and nonspecifically bound probe. (4) Hybrids are detected by placing a piece of X-ray film over the membrane and exposing for a short time. The film is developed and the hybridization events are visualized as dark spots on the autoradiogram. (5) From the orientation of the film, a positive colony containing the insert that hybridized to the probe can be identified. Bacterial cells are picked from this colony for growth and further analysis.

8.8 Expression libraries

Expression libraries are made with a cloning vector that contains the required regulatory elements for gene expression, such as the promoter region (see Section 10.3). In an *E. coli* expression vector (plasmid or bacteriophage), an *E. coli* promoter is placed next to a unique restriction site where DNA can be inserted. When a foreign cDNA is cloned into an expression vector in the correct reading frame, the coding region is transcribed and translated in the *E. coli* host. Expression libraries are useful for identifying a clone containing the cDNA of interest when an antibody to the protein encoded by that gene or cDNA is available. Binding of a radioactively labeled antibody (see Tool box 9.4), using a technique similar to nucleic acid hybridization, can be used to identify a specific protein made by one of the clones of the expression library.

8.9 Restriction mapping

Once the clone of interest has been isolated, the first stage of analysis is often the creation of a restriction map. Restriction mapping provides a compilation of the number, order, and distance between restriction endonuclease cutting sites along a cloned DNA fragment. In addition, restriction mapping plays an important role in characterizing DNA, mapping genes, and diagnostic tests for genetic diseases (see Section 8.10).

To make a restriction map, a cloned DNA fragment is cut with restriction endonucleases and loaded on to an agarose gel for electrophoresis (Tool box 8.6). The lengths of the DNA fragments can be determined by comparing their position in the gel to reference DNAs of known lengths in the gel. A DNA fragment migrates a distance that is inversely proportional to the logarithm of the fragment length in base pairs over a limited range in the gel. Thus, agarose gel electrophoresis allows the restriction fragment lengths to be determined. The pattern of cutting in single and double digests indicates what the relationship is between the two sites. Assume, for example, that you have attempted to subclone a cDNA into a plasmid vector. You have obtained a positive clone by blue-white screening. Because you want to use the recombinant plasmid DNA for the preparation of an antisense RNA probe, you need to check that the orientation of the insert in the plasmid vector is correct. According to the plasmid map shown in Fig. 8.13, if the insert is in the desired orientation, the fragment sizes generated by a double digest with *Eco*RI and *Hin*dIII will be 4.5 and 1.3 kb, respectively. If the insert is in the opposite orientation, the order of restriction sites will be reversed and the fragment sizes will be 3.5 and 2.3 kb. The results of restriction endonuclease digests show that the insert is in the correct orientation, and that you have successfully subcloned a template for *in vitro* riboprobe preparation.

8.10 Restriction fragment length polymorphism (RFLP)

In 1980, Mark Skolnick, Ray White, David Botstein, and Ronald Davis created a restriction fragment length polymorphism (RFLP, pronounced "rif-lip") marker map of the human genome. A RFLP is defined by the existence of alternative alleles associated with restriction fragments that differ in size from each other. RFLPs are visualized by digesting DNA from different individuals with restriction endonucleases, followed by gel electrophoresis to separate fragments according to size, then Southern blotting (Tool box 8.7), and hybridization to a labeled probe that identifies the locus under investigation. A RFLP is demonstrated whenever the Southern blot pattern obtained with one individual is different from the one obtained with another individual. These variable regions do not necessarily occur in genes, and the function of most of those in the human genome is unknown. An exception is a RFLP that can be used to diagnose sickle cell anemia (Fig. 8.14). In individuals with sickle cell anemia, a point mutation in the β-globin gene has destroyed the recognition site for the restriction endonuclease *Mst*II. This mutation can be distinguished by the presence of a larger restriction fragment on a Southern blot in an affected individual, compared with a shorter fragment in a normal individual.

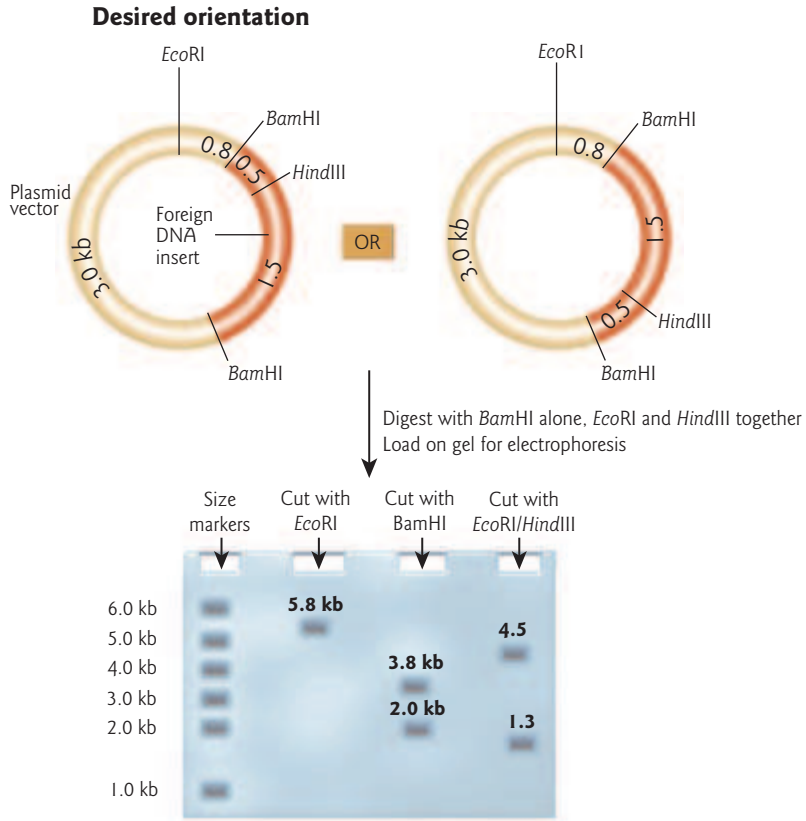


Figure 8.13 Analysis of recombinant DNA by restriction endonuclease digestion. Assume a 2.0 kb foreign DNA insert has been successfully ligated into the *Bam*HI site of a 3.8 kb plasmid vector. However, the orientation of the insert is unknown. Samples of the recombinant plasmid are digested with restriction endonucleases: one sample is digested with *Eco*RI, one with *Bam*HI, and one with both *Eco*RI and *Hin*dIII. The resulting fragments are separated by agarose gel electrophoresis. The sizes of the separated fragments can be measured by comparison with molecular weight standards in an adjacent lane. *Eco*RI linearizes the 5.8 kb plasmid which appears as a single band on the gel. *Bam*HI generates two fragments of 3.8 and 2.0 kb in size, representing the vector and the insert, respectively. Digestion with *Eco*RI and *Hin*dIII generates two fragments of 4.5 and 1.3 kb. These data indicate that the foreign DNA has been inserted in the desired orientation. If the DNA had been inserted in the opposite orientation, a double digest with *Eco*RI and *Hin*dIII would have generated fragment sizes of 3.5 and 2.3 kb.

RFLPs can serve as markers of genetic diseases

By carefully examining the DNA of members of families that carry genetic diseases, it has been possible to find forms of particular RFLPs that tend to be inherited with particular diseases. The simplest RFLPs are those caused by single base pair substitutions. However, RFLPs can also be generated by the insertion of genetic material such as transposable elements, or by tandem duplications, deletions, translocations, or other chromosomal rearrangements. In linkage analysis, families in which individuals are at risk for a genetic disease are identified (i.e. both parents are heterozygous for an autosomal recessive mutation associated with a particular disease). DNA samples from various family members are then analyzed to determine the frequency with which specific RFLP markers segregate with the mutant allele causing the disease. This frequency is a measure of the distance between the markers and the mutation-defined locus.

Generally, the fragment size differences occur not because a restriction site was created or disrupted by the diseased state itself, but rather because the nucleotide sequence differences just happen to be near the gene

TOOL BOX 8.6

Electrophoresis

Electrophoresis is the standard method for analyzing, identifying, and purifying fragments of DNA or RNA that differ in size, charge, or conformation. It is one of the most widely used techniques in molecular biology. Walk into any molecular biology lab, and the odds are you will see at least one gel apparatus in operation. When charged molecules are placed in an electric field, they migrate toward the positive (anode, red) or negative (cathode, black) pole according to their charge. In contrast to proteins, which can have either a net positive or net negative charge, nucleic acids have a consistent negative change due to their phosphate backbone, and they migrate toward the anode. Proteins and nucleic acids are separated by electrophoresis within a matrix or "gel." Most commonly, the gel is cast in the shape of a thin slab, with wells for loading the sample. The gel is immersed within an electrophoresis buffer that provides ions to carry a current and some type of buffer to maintain the pH at a relatively constant value.

The gels used for eletrophoresis are composed either of agarose or polyacrylamide. Agarose gels are used in a horizontal gel apparatus, while polyacrylamide gels are used in a vertical gel apparatus. These two differ in resolving power. Agarose gels are used for the analysis and preparation of fragments between 100 and 50,000 bp in size with moderate resolution, and polyacrylamide gels are used for the analysis and preparation of small molecules with single nucleotide resolution. This high resolution is required for applications such as for DNA sequencing (see Section 8.11), DNase I footprinting, and electrophoretic mobility shift assays (see Fig. 9.15). In contrast to agarose, polyacrylamide gels are also widely used for the electrophoresis of proteins (see Tool box 9.3).

Agarose gel electrophoresis

Agarose is a polysaccharide extracted from seaweed. Agarose gels are easily prepared by mixing agarose powder with buffer solution, boiling in a microwave to melt, and pouring the gel into a mold where the agarose (generally 0.5–2.0%) solidifies into a slab (Fig. 1). Agarose gels have a large range of separation, but relatively low resolving power. By varying the concentration of agarose, fragments of DNA from about 100 to 50,000 bp can be separated using standard electrophoretic techniques. A toothed comb forms wells in the agarose. The agarose slab is submerged in a buffer solution and an electric current is passed through the gel, with the negatively charged DNA (due to the phosphate in the sugar–phosphate backbone) moving through the gel from the negatively charged electrode (cathode) towards the positive electrode (anode). Pores between the agarose molecules act like a sieve that separates the molecules by size. In gel filtration chromatography (see Tool box 8.1), nucleic acids flow around the spherical agarose beads. In contrast, in an electrophoretic gel, nucleic acids migrate through the pores; thus fragments separate by size with the smallest pieces moving the fastest and farthest through the gel. Because DNA by itself is not visible in the gel, the DNA is stained with a fluorescent dye such as ethidium bromide (EtBr). EtBr intercalates between the bases causing DNA to fluoresce orange when the dye is illuminated by ultraviolet light.

Pulsed field gel electrophoresis (PFGE)

DNA changes conformation as it moves through gels, alternating between extended and compact forms. The mobility of a DNA molecule depends upon the relationship between the pore size of the gel and the globular size of the DNA in its compact form, with larger molecules moving more slowly. Once DNA reaches a critical size, the molecule is too large to fit through any of the pores in an agarose gel and can move only as an extended molecule. The mobility of the DNA thus becomes independent of size, resulting in co-migration of all large molecules. To fractionate large DNA molecules such as YACs (see Section 8.4), agarose gel electrophoresis is carried out with a pulsed electric field. The periodic field causes the DNA molecule to reorient; longer molecules take longer to realign than shorter ones, thus delaying their progress through the gel and allowing them to be resolved. DNA molecules up to 200–400 Mb in size have been separated by various pulsed field-based methods.

Polyacrylamide gel electrophoresis (PAGE)

Polyacrylamide is a cross-linked polymer of acrylamide. The length of the polymer chains is dictated by the concentration of acrylamide used, which is typically between 3.5 and 20%. Polyacrylamide gels are significantly more cumbersome to prepare than agarose gels. Because oxygen inhibits the polymerization process, they must be

Electrophoresis

TOOL BOX 8.6

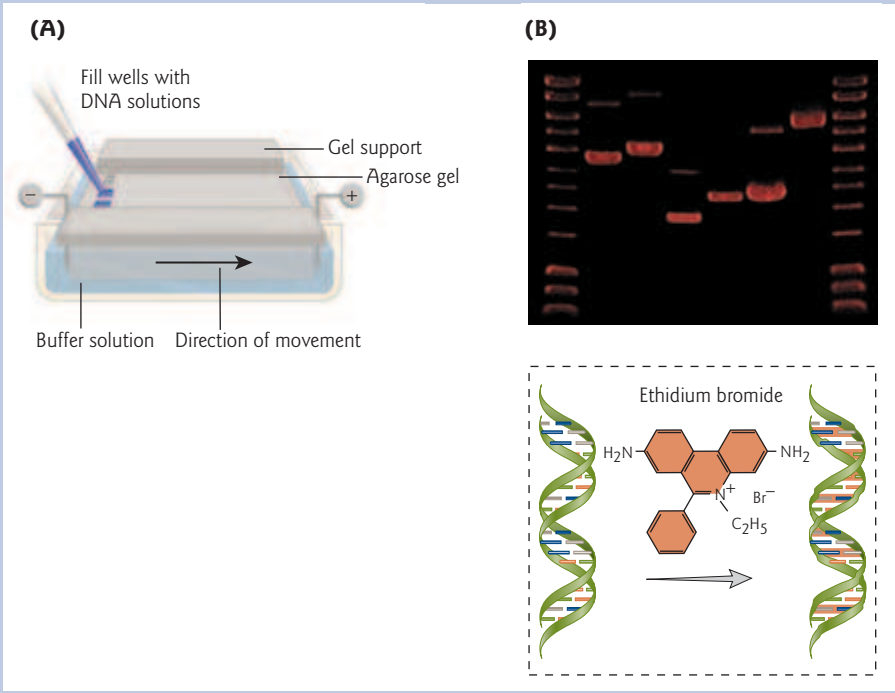


Figure 1 Agarose gel electrophoresis is used to separate DNA (and RNA) molecules according to size. (A) A pipet is used to load DNA samples on an agarose gel in a horizontal gel apparatus. The negatively charged nucleic acids move toward the positive electrode. Larger molecules move more slowly than smaller molecules, so the DNA (or RNA) molecules are separated according to size. (B) Photograph of an agarose gel stained with ethidium bromide (EtBr) to make the DNA bands visible. EtBr molecules intercalate between the bases (see inset) causing the DNA to fluoresce orange when the gel is illuminated with UV light. (Photograph courtesy of Vinny Roggero, College of William and Mary.)

poured between glass plates. Gels can be either nondenaturing or denaturing (e.g. contain 8 M urea). Denaturing gels are used, for example, when single-stranded DNA is being analyzed, as in DNA sequencing (see Fig. 8.15). Polyacrylamide gels have a rather small range of separation, but very high resolving power. In the case of DNA, polyacrylamide is used for separating fragments of less than about 500 bp. However, under appropriate conditions, fragments of DNA differing in length by a single base pair are easily resolved. Bands in polyacrylamide gels are usually detected by autoradiography, although silver staining can also be used.

involved. A particular form of a polymorphism that is close to a diseased gene tends to stay with that gene during crossing-over (recombination) in meiosis. Relatively large segments of chromosomes are involved in crossing-over, so markers close together on a given chromosome are more likely to be transmitted together (not separated during recombination) than those that are far apart. Linkage thus refers to the likelihood of having one marker transmitted with another through meiosis. Markers that are transmitted together frequently are said to be closely linked. Thus RFLPs can serve as markers of disease, even when the RFLP is

TOOL BOX 8.7

Southern blot

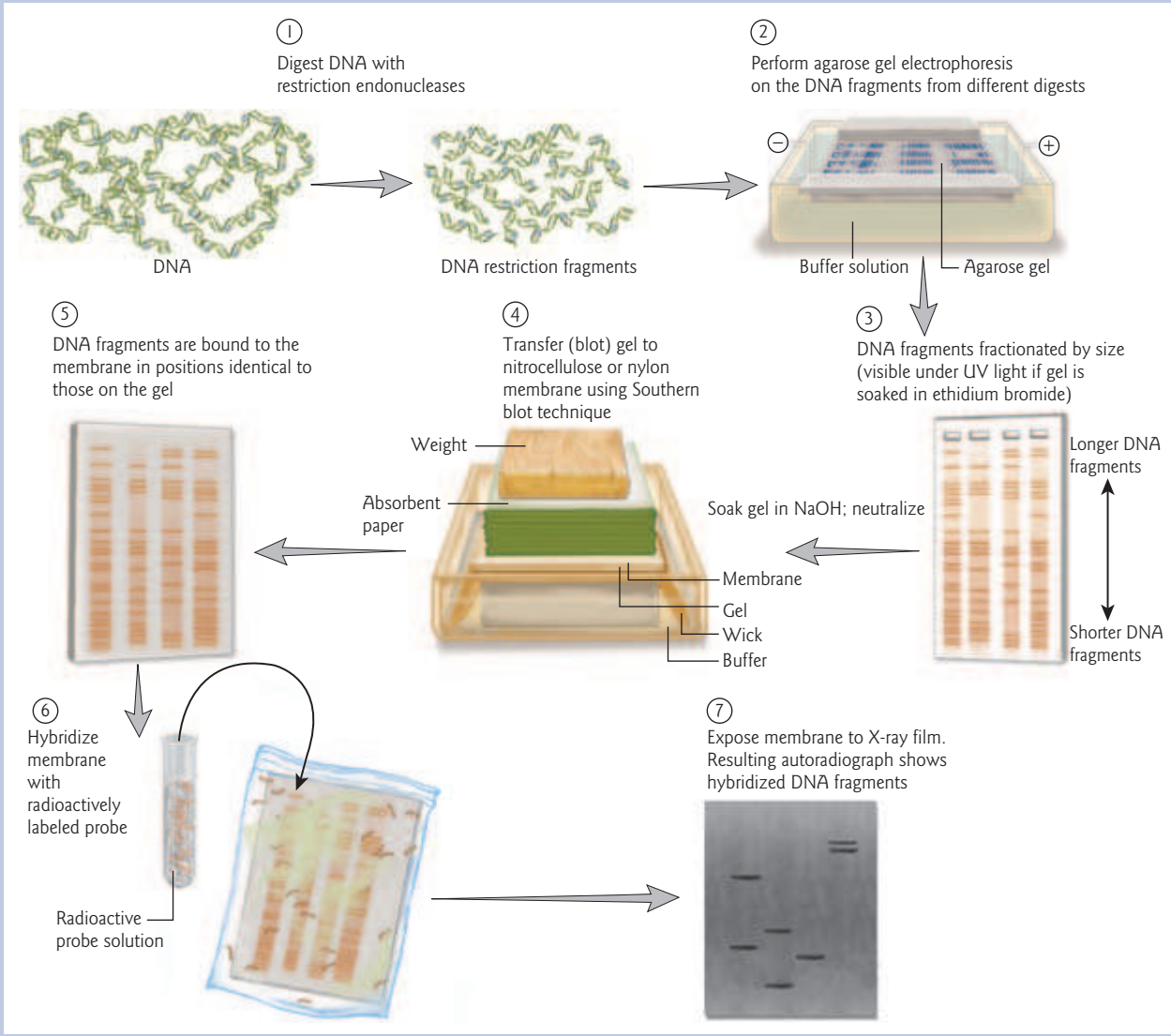


Figure 1 Southern blot. The steps involved in performing a Southern blot hybridization are depicted. (1–3) Samples of the DNA to be probed are cut with restriction endonucleases and the fragments are separated by gel electrophoresis. (4) After soaking the gel in an alkaline solution to denature the DNA, the single-stranded DNA is transferred to a DNA-binding membrane for hybridization by making a sandwich of the gel, membrane, filter paper, and absorbent paper. The blot is held in place with a weight. (5) Capillary action draws the buffer through the gel, transferring the pattern of DNA fragments from the gel to the membrane. (6) The membrane is hybridized with a radioactively labeled probe and then washed to remove any nonspecifically bound probe. (7) The membrane is overlaid with a piece of X-ray film for autoradiography. The hybridized fragments show up as bands on the X-ray film.

Southern blot

TOOL BOX 8.7

The capillary transfer of fragments of DNA separated on an agarose gel from the gel to a solid support was first carried out by Edward Southern. This technique thus bears his name and is called a Southern blot. Southern blots have many applications, but their primary purpose is to identify a specific gene fragment from the often many DNA bands on a gel.

Southern blot method

In the classic Southern blot method, DNA samples are first digested with one or more restriction endonucleases to reduce the size of the DNA molecules (Fig. 1). An endonuclease with a six base pair recognition sequence, statistically cuts once every 4^6 base pairs – producing many thousands of restriction fragments in the genomes of higher eukaryotes. For example, 732,422 restriction fragments will be generated in mouse genomic DNA by *Eco*RI. The digested DNA is then separated by agarose gel electrophoresis. The DNA fragments are denatured by an alkaline (high pH) buffer, and the single strands are transferred by capillary action to a nylon or nitrocellulose membrane. The membrane is a replica of the agarose gel. To identify the DNA fragment that contains a gene of interest, a specific DNA probe, such as a small region of the DNA of interest, can be used to hybridize to the membrane, as described earlier for colony hybridization (see Section 8.7). If the probe hybridizes to fragments on the membrane, then photographic film applied to the membrane will be exposed where the probe has hybridized to a specific DNA band or bands (Fig. 1).

Applications

Southern blots can be used to complement restriction mapping of cloned DNA and to identify overlapping fragments. For example, the exact location of a 2.0 kb cDNA within a 10 kb insert of foreign DNA can be identified so that it can be isolated for further analysis or sequencing. Alternatively, Southern blots can used to identify structural differences between genomes, through analysis of RFLPs, or to study families of related DNA sequences. To illustrate this point, assume that a gene of interest does not contain any internal *Hin*dIII sites. If a clone (e.g. a cDNA clone for some gene) is hybridized to a *Hin*dIII digestion of the DNA sample and two fragments are seen, then it can be concluded that the organism being analyzed has two copies of the gene. Normally, it is not known at the start of analysis what restriction endonuclease sites are located in the gene, so a series of digestions and hybridizations are performed. If four out of the five restriction endonucleases reveal two bands, and the probe hybridizes to three fragments of DNA digested with a fifth restriction endonuclease (e.g. *Eco*RI), then it can be concluded that two genes exist and one of these genes contains a restriction endonuclease site for *Eco*RI. In the case of transgenic animals and plants, Southern blots can be used to determine whether the foreign gene is present and is now part of the host chromosome (see Sections 15.2 and 15.6). The integration of foreign DNA can be visualized by an increase in restriction fragment sizes.

not within the disease gene. RFLPs have been useful for detecting such genetic diseases as cystic fibrosis, Huntington's disease, and hemophilia.

RFLPs were the predominant form of DNA variation used for linkage analysis until the advent of PCR. The main advantage of RFLP analysis over PCR-based protocols is that no prior sequence information or oligonucleotide synthesis is required. However, when a PCR assay for typing a particular locus is developed, it is generally preferable to RFLP analysis. In some cases, a combined approach of PCR and RFLP is used for analysis (Disease box 8.1).

8.11 DNA sequencing

Until 1977, determining the sequence of bases in DNA was a labor-intensive process that could be applied only to very short sequences, such as the template region for tRNA. With the development of techniques for rapid, large-scale DNA sequencing, today molecular biologists determine the order of bases in DNA as a matter of course. DNA sequencing is used to provide the ultimate characterization once a gene has

DISEASE BOX 8.1

PCR-RFLP assay for maple syrup urine disease

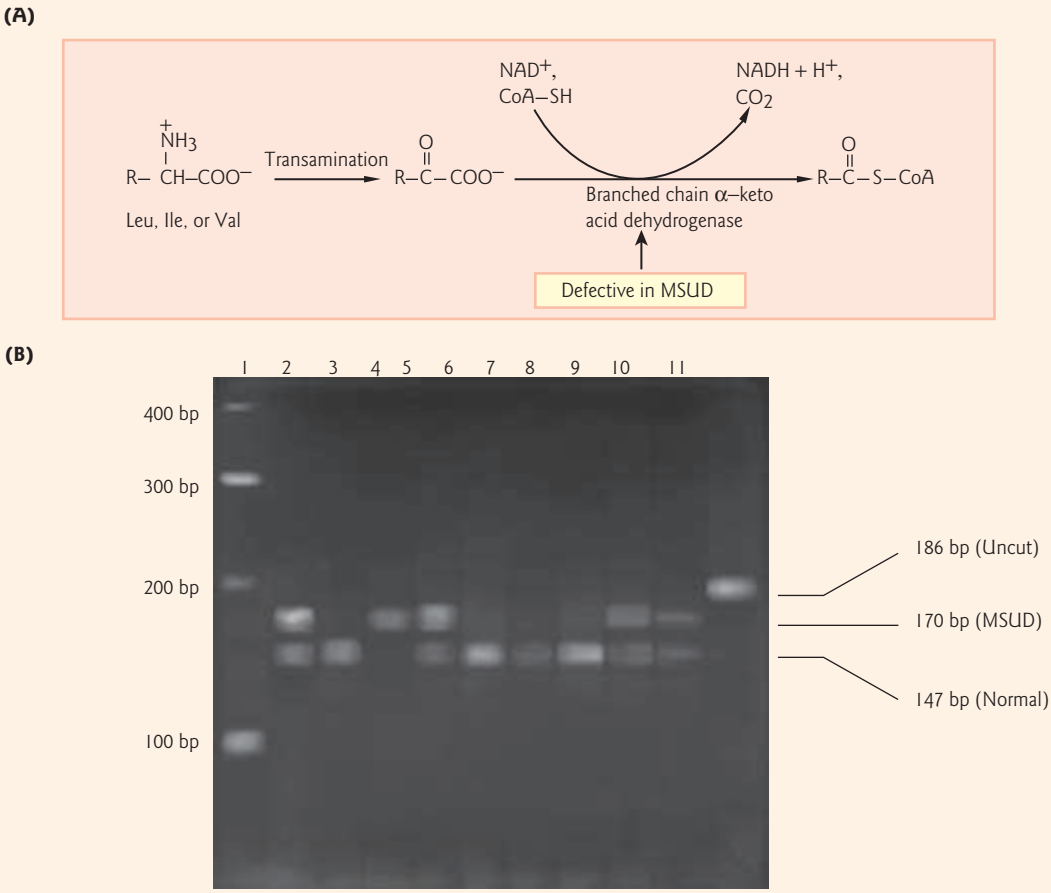


Figure 1 Diagnosis of maple syrup urine disease (MSUD). (A) Degradation of the amino acids leucine, isoleucine, and valine starts with transamination followed by oxidative decarboxylation of the respective keto acids. The latter reaction is carried out by a multienzyme complex, called the branched-chain α-keto acid dehydrogenase complex, which is defective in MSUD patients. (B) Genotype analysis of MSUD by PCR-RFLP assay. DNA samples were amplified by PCR with primers specific for the MSUD allele. Following digestion with restriction endonuclease *Sca*I, samples were visualized by agarose gel electrophoresis and staining with ethidium bromide. Heterozygous individuals are indicated by the presence of a 170 and 147 bp DNA fragment: lane 2 (father), lane 5 (sibling), lane 9 (mother), and lane 10 (maternal grandmother). Individuals homozygous for the normal allele are indicated by the presence of the 147 bp DNA fragment only: lane 3 (paternal grandmother) and lanes 6–8 (maternal great grandfather, maternal grandfather, maternal great grandmother, respectively). Lane 4 indicates the resulting 170 bp fragment from the family member homozygous for the MSUD allele. Lane 11 shows the undigested 186 bp PCR product. (Love-Gregory, L.D., Dyer, J.A., Grasela, J., Hillman, R.E., Philips, C.L. 2001. Carrier detection and rapid newborn diagnostic test for the common Y393N maple syrup urine disease allele by PCR-RFLP: culturally permissible testing in the Mennonite community. *Journal of Inherited and Metabolic Diseases* 24:393–403, Fig. 2. Copyright © 2001 SSIEM and Kluwer Academic Publishers. Reprinted with kind permission of Springer Science and Business Media.)

PCR-RFLP assay for maple syrup urine disease

DISEASE BOX 8.1

Maple syrup urine disease (MSUD) was first described in 1954. It is a metabolic disorder inherited as an autosomal recessive that affects the metabolism of the three branched-chain amino acids, leucine, isoleucine, and valine. In a normal individual, when more protein is consumed than is needed for growth, the branched-chain amino acids are degraded to generate energy. Breakdown of these amino acids involves a series of chemical reactions. The second step of the degradation pathway is mediated by an enzyme system consisting of six components, called the multienzyme branched-chain α-keto acid dehydrogenase complex (Fig. 1). In MSUD, one or several of the genes encoding components of this complex are mutated. Because the degradation pathway is blocked, the α-keto acid derivatives of isoleucine, leucine, and valine accumulate in the blood and urine. This accumulation of keto acids gives the urine of affected children a sweet odor resembling maple syrup, and causes a toxic effect that interferes with brain function.

Infants with classic MSUD appear normal at birth and become symptomatic within 4–7 days after birth. They follow a progressive course of neurological deterioration, exhibiting lethargy, seizures, coma, and death within the first 2–3 weeks of life if untreated. Early diagnosis is essential for the child with MSUD to develop normally. Treatment involves a special, carefully controlled diet that requires detailed monitoring of protein intake. The diet centers around a synthetic formula that provides nutrients and all the amino acids except for leucine, isoleucine, and valine. These three amino acids are then added to the diet in carefully controlled amounts to provide enough of these essential amino acids for normal growth and development without exceeding the level of tolerance. In older patients, this special metabolic product provides basic nutrients and takes the place of cow's milk in a normal diet. The remainder of the diet is essentially a vegetarian diet.

Worldwide, the incidence of MSUD is one in 185,000–225,000; however, in certain Old Order Mennonite communities, the incidence of classic MSUD is estimated to be one in 176 live births. The defect in Mennonite MSUD patients is caused by a single nucleotide change in one of the genes encoding a component of the enzyme complex. The missense mutation results in a tyrosine (Y) to asparagine (N) substitution, called the Y393N allele. MSUD in the non-Mennonite population is clinically and genetically heterogeneous; however the Y393N allele is present in a significant portion of the non-Mennonite MSUD population. A mismatch PCR-RFLP assay has been designed to identify the Y393N allele. Owing to religious and cultural preferences, prenatal testing is not permitted in the Mennonite community. Hence, neonatal testing is vital. Various tests are available to monitor the levels of the amino acids and their α-keto acid derivatives in the blood and urine. Traditional serum-based assays may not give results quickly enough, and there have been reports of infants dying before newborn screening results were reported. The mismatch PCR-RFLP assay provides rapid turnaround times. Since the assay is DNA-based it does not require time for the levels of keto acid derivatives to increase in the blood serum, which may take up to 72 hours. Buccal swabs or blood samples are used and results are available in a minimum of 8 hours. Identification of the normal allele (147 bp) results from the cleavage of a PCR product at a *Sca*I site. Identification of the Y393N allele (170 bp) results in the absence of this second *Sca*I site (Fig. 1).

been cloned or amplified by PCR. Although a sequence on its own is of limited value, it is the necessary stepping-stone to more informative analyses of the cloned gene. DNA sequencing is used to identify genes, determine the sequence of promoters and other regulatory DNA elements that control expression, reveal the fine structure of genes and other DNA sequences, confirm the DNA sequence of cDNA and other DNA synthesized *in vitro* (for example, after *in vitro* mutagenesis to confirm the mutation), and help deduce the amino acid sequence of a gene or cDNA from the DNA sequence. With the advent of automated DNA sequencing technology, large genome sequencing projects are yielding information about the evolution of genomes, the location of coding regions, regulatory elements, and other sequences, and the presence of mutations that give rise to genetic diseases (see Section 16.3).

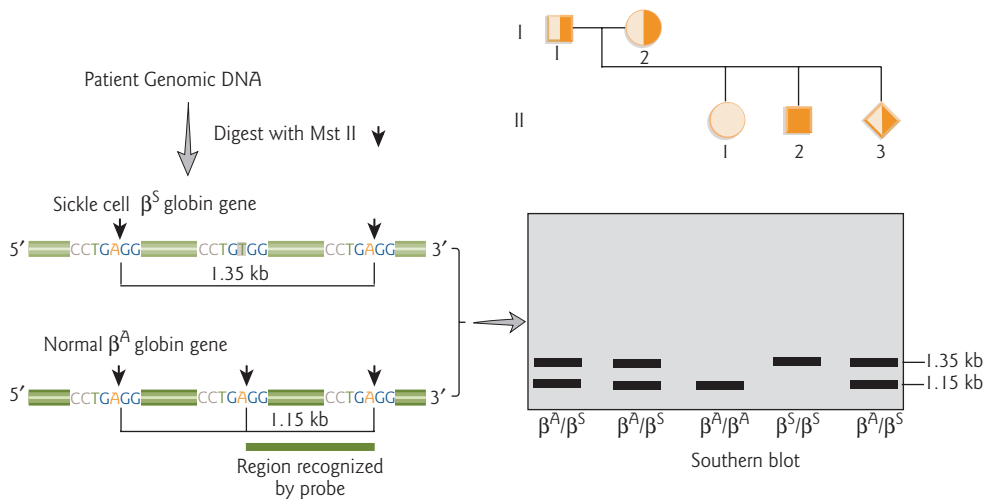


Figure 8.14 Diagnosis of sickle cell anemia by restriction fragment length polymorphism (RFLP) and Southern blot. Black arrows represent the location of recognition sites for the restriction endonuclease *Mst*II in the β-globin gene. In the mutant β-globin gene (β^S), a point mutation (GAG → GTG) has destroyed one *Mst*II recognition site. Digestion of patient genomic DNA with *Mst*II results in a 1.35 kb DNA fragment, compared with a 1.15 kb DNA fragment in normal individuals. For diagnosis, the restriction-digested DNA fragments are separated by gel electrophoresis and transferred to a nylon or nitrocellulose membrane (see Tool box 8.7). The fragments are visualized by hybridization using a probe that spans a portion of the β-globin gene where the 1.15 kb *Mst*II restriction fragment resides. In the pedigree, the family has one unaffected homozygous normal daughter (II-1), an affected homozygous son (II-2), and an unaffected heterozygous fetus (II-3). The genotypes of each family member can be read directly from the Southern blot. (Wilson, J.T., Milner, P.F., Summer, M.E., Nallaseth, F.S., Fadel, H.E., Reindollar, R.H., McDonough, P.G., Wilson, L.B. 1982. Use of restriction endonucleases for mapping the allele for β^S-globin. *Proceedings of the National Academy of Sciences USA* 79:3628–3631; and Chang, J.C., Alberti, A., Kan, Y.W. 1983. A β-thalassemia lesion abolishes the same *Mst*II site as the sickle mutation. *Nucleic Acids Research* 11:7789–7794.)

Manual DNA sequencing by the Sanger "dideoxy" DNA method

In 1977, Frederick Sanger, Allan Maxam, and Walter Gilbert pioneered DNA sequencing. The Maxam and Gilbert sequencing method uses a chemical method that involves selective degradation of bases. The most widely used method for DNA sequencing is the Sanger or "dideoxy" method, which is, in essence, a DNA synthesis reaction (Fig. 8.15). In this method, single-stranded DNA is mixed with a radioactively labeled primer to provide the 3′-OH required for DNA polymerase to initiate DNA synthesis. The primer is usually complementary to a region of the vector just outside the multiple cloning site. The sample is then split into four aliquots, each containing DNA polymerase, four dNTPs (at high concentration), and a low concentration of a replication terminator. The replication terminators are dideoxynucleoside triphosphates (ddNTPs) that are missing the 3′-OH. Because they lack the 3′-OH, they cannot form a phosphodiester bond with another nucleotide. Thus, each reaction proceeds until a replication-terminating nucleotide is added, and each of the four sequencing reactions produces a series of single-stranded DNA molecules, each one base longer than the last. The polymerase of choice for DNA sequencing is phage T7 DNA polymerase (called "Sequenase"). The sequencing mixtures are loaded into separate lanes of a denaturing polyacrylamide gel and electrophoresis is used to separate the DNA fragments. Autoradiography is used to detect a ladder of radioactive bands. The radioactive label (primer) is at the 5′ end of each newly synthesized DNA molecule. Thus, the smallest fragment at the bottom of the gel represents the 5′ end of the DNA. Reading the sequence of bases from the bottom up (5′ → 3′) gives the sequence of the DNA molecule synthesized in the sequencing reaction. The sequence of the original strand of DNA is complementary to the sequence read from the gel (3′ → 5′).

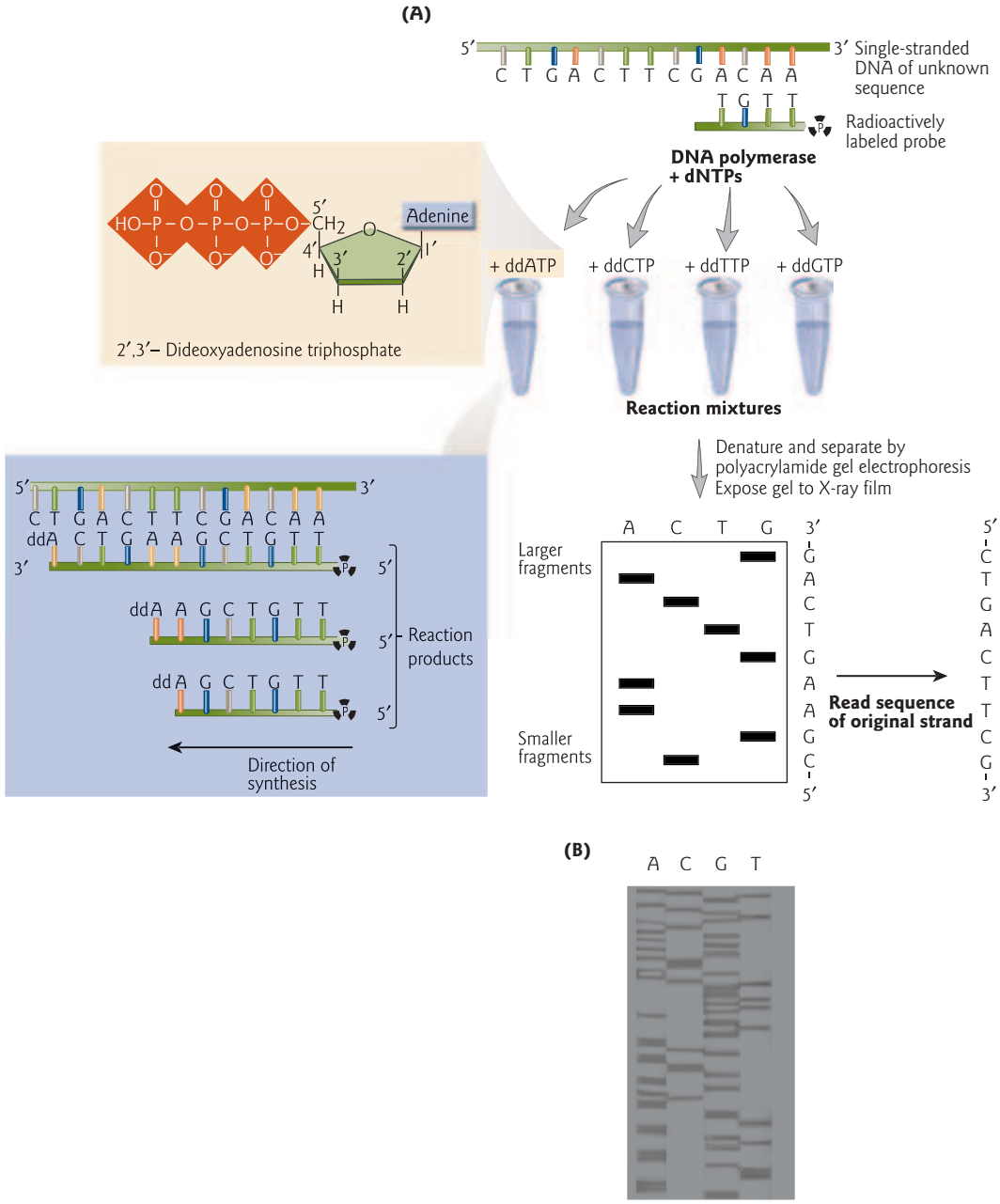


Figure 8.15 Sanger "dideoxy" DNA sequencing. (A) Four DNA synthesis reaction mixes are set up using template DNA, a labeled primer, DNA polymerase, and a mixture of dNTPs and one each of the four dideoxy NTPs (ddNTPs). The direction of synthesis is from 5′ to 3′. The radioactive products of each reaction mixture are separated by polyacrylamide gel electrophoresis and located by exposing the gel to X-ray film. The nucleotide sequence of the newly synthesized DNA is read directly from the autoradiogram in the 5′ → 3′ direction, beginning at the bottom of autoradiogram. The sequence in the original template strand is its complement (3′ → 5′). (Inset) The random incorporation of dideoxy ATP into the growing chain generates a series of smaller DNA fragments ending at all possible positions where adenine is found in the newly synthesized fragments. These correspond to positions where thymine occurs in the original template strand. (B) An exposed X-ray film of a DNA sequencing gel. The four lanes represent A, C, G, and T dideoxy reaction mixes, respectively. (Photograph courtesy of Jim Nicoll, College of William and Mary.)

This method can separate DNA of approximately 500 nt. For longer sequences, overlapping fragments are sequenced. The technique is very laborious and the sequences have to be read by hand.

Automated DNA sequencing

In 1986, Leroy Hood and Lloyd Smith automated Sanger's method. In this new sequencing technology, radioactive markers are replaced with fluorescent ones. Each ddNTP terminator is tagged with a different color of fluorophore: red, green, blue, or yellow. Thus, instead of having to run four separate sequencing reactions, the reactions can be combined into one tube. The first automated sequencer made use of a polyacrylamide gel to resolve the samples, a laser to excite the dye molecules as they reached a detector near the end of the gel, and a computer to read the results as a DNA sequence. In this system each automated sequencer was able to produce 4800 bases of sequence per day. The current automated systems replace the old-style gel with arrays of tiny capilliaries, each of which acts as a "lane." A pump loads special capillaries with a polymer that serves as the separation matrix. DNA samples in a 96-well plate are loaded into the capillary array by a short burst of electrophoresis, called "electrokinetic injection." The capillary array is immersed in running buffer and the DNA fragments then migrate through the capillary matrix by size, smallest to largest. As the DNA fragments reach the detection window, a laser beam excites the dye molecules causing them to fluoresce. Emitted light from 96 capillaries is collected at once, spectrally separated into the four colors and focused onto a CCD camera. Computer software interprets the pattern of peaks to produce a graph of fluorescence intensity versus time (electropherogram), which is then converted to the DNA sequence (Fig. 8.16). With this system, as many as 2 million bases can be sequenced per day.

Chapter summary

Insights from bacteriophage λ cohesive sites and bacterial restriction and modification systems led to the development of genetic engineering, and the characterization of restriction endonucleases. Type II restriction endonucleases are widely used for mapping and reconstructing DNA *in vitro* because they recognize specific 4–8 bp sequences in double-stranded DNA and make cuts in both strands just at these sites. Restriction endonucleases function as homodimers. The first contact with DNA is nonspecific. By linear diffusion (sliding) along the DNA in combination with repeated dissociation/reassociation (hopping/jumping), the enzyme locates the target restriction site. The recognition process triggers a large conformational change in the enzyme and DNA, which leads to catalysis. Because the cuts in the two strands are frequently staggered, restriction endonucleases can create sticky ends that help link together two DNA molecules from different sources *in vitro*. DNA ligase is used to join the two pieces of DNA, forming a recombinant DNA molecule.

Cloning vectors are carrier DNA molecules that can independently replicate themselves and the foreign DNA segments they carry in host cells. Sources of foreign DNA include genomic DNA, cDNA, polymerase chain reaction (PCR) products, and chemically synthesized oligonucleotides. There are many possible choices of vector depending on the purpose of cloning and the size of the foreign insert. The classic cloning vectors are plasmids, phages, and cosmids, which are limited to inserts of up to 10, 20, or 45 kb, respectively. A new generation of artificial chromosome vectors that can carry much larger inserts include bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), and mammalian artificial chromosomes (MACs).

Among the first generations of plasmid cloning vectors were pUC plasmids that replicate autonomously in bacterial cells after transformation of the bacteria. These have an ampicillin resistance gene and a multiple cloning site that interrupts a partial *lacZ* (β-galactosidase) gene. The multiple cloning sites make it convenient to carry out directional cloning into two different restriction sites. Ampicillin-resistant clones are screened for those that do not make active β-galactosidase and therefore do not turn the indicator substrate, X-gal, blue. Positive bacterial colonies can be amplified in liquid growth medium, followed by purification of the amplified recombinant plasmid DNA by liquid chromatography methods. Ion-exchange chromatography can be used to separate substances according to their charges. Gel filtration chromatography

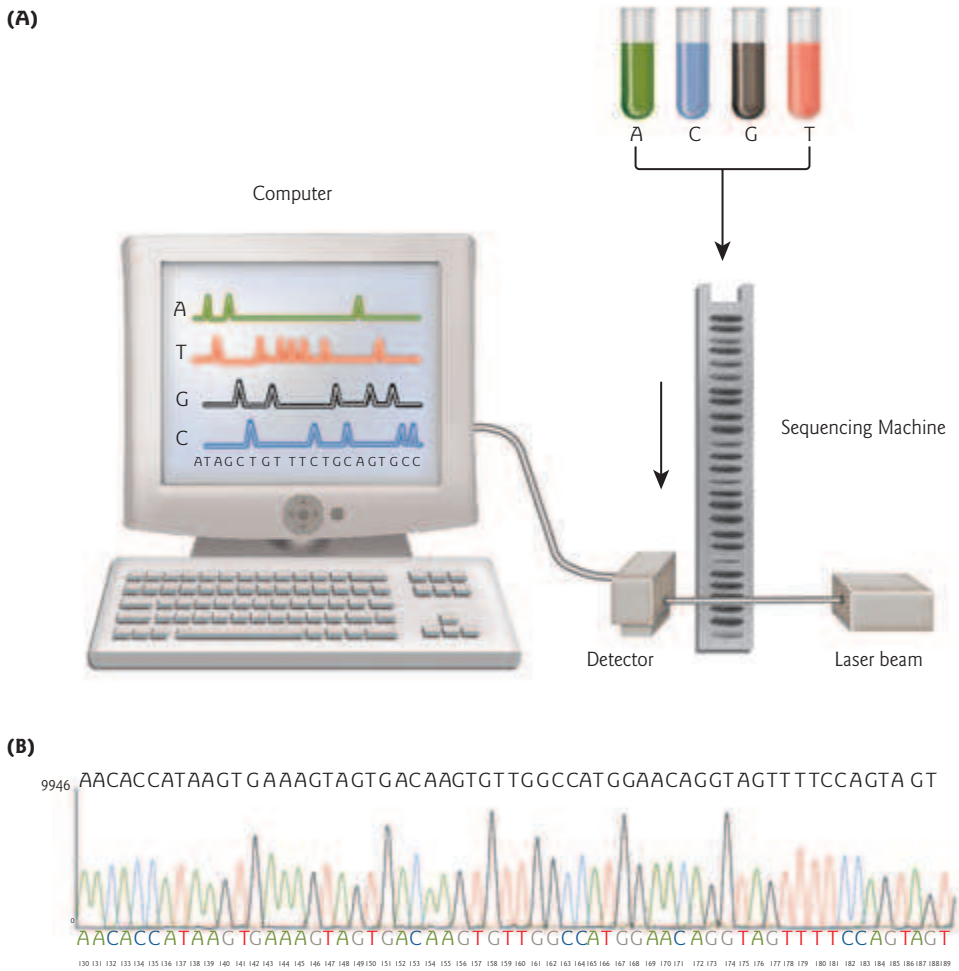


Figure 8.16 Automated DNA sequencing. (A) The diagram depicts the steps involved in automated DNA sequencing. This technique uses dideoxynucleotides, just as in the manual technique shown in Fig. 8.15, but the primers used in each of the four reactions are tagged with different fluorescent molecules. The products from each tube will emit a different color fluorescence when excited by light. (B) Sample computer printout of an automated DNA sequencing experiment. (Electropherogram courtesy of Ghislain Bonamy, College of William and Mary.)

uses columns filled with porous resins that let in smaller substances, but exclude larger ones. Thus, the smaller substances are slowed in their journey through the column, but larger substances travel relatively rapidly through the column.

Engineered phage λ from which certain nonessential genes have been removed to make room for inserts are useful for preparing genomic libraries, in which it is important to have large pieces of genomic DNA in each clone. Even more useful are YAC vectors that are designed to act like chromosomes in host yeast cells, and can accommodate up to 1 Mb of foreign DNA. YAC vectors included an origin of replication, a centromere, telomeres, and growth selectable markers in each arm. Positive selection is carried out by auxotrophic complementation for molecules in which the arms are joined together. Recombinant YACs are often screened for by a "red-white" selection process, in which insertion of foreign DNA leads to the expression of a red pigment in particular mutant strains of yeast.

A genomic library contains DNA fragments that represent the entire genome of an organism. The library created from all the cDNAs derived from the expressed mRNAs in a specific cell type forms a cDNA library

of cDNA clones. To make a cDNA library, one can synthesize cDNAs one strand at a time, using mRNAs from a cell as templates for the first strands, and these first strands as templates for the second strands. Reverse transcriptase generates the first strands and the Klenow subunit of *E. coli* DNA polymerase I generates the second strands. Double-stranded DNA linkers with ends that are complementary to an appropriate cloning vector are added to the cDNA before ligation into the cloning vector. Gene coding sequences of expressed genes are represented in the library. Sequences corresponding to introns and regulatory regions are not present.

The polymerase chain reaction (PCR) amplifies a region of DNA between two predetermined sites. Oligonucleotides complementary to these sites serve as primers for the synthesis of copies of the DNA between the sites. Each cycle of PCR doubles the number of copies of the amplified DNA until a large quantity has been made. PCR is used extensively in many areas of molecular cloning, analysis of gene expression, and diagnosis of genetic diseases.

Particular recombinant DNA clones in a library can be detected by colony or plaque hybridization with labeled probes, or with antibodies if an expression vector is used. Specific clones can be identified using heterologous or homologous probes that bind to the gene itself. Knowing the amino acid sequence of a gene product, one can design a set of degenerate or expressed sequence tag (EST) based oligonucleotides that encode part of this amino acid sequence. cDNA probes are often used to screen genomic libraries.

There are a variety of methods for labeling RNA and DNA. Probes of the highest specific activity are generated using internal labeling, where many labeled nucleotides are incorporated uniformly during DNA or RNA synthesis *in vitro*. End-labeling involves either adding a labeled nucleotide to the 3′-OH end of a DNA strand or exchanging the unlabeled 5′-phosphate group for a labeled phosphate, and is used when precise definition of one end of the DNA is required. If the probe is radiolabeled one can detect it by autoradiography, using X-ray film or a phosphorimager, or by liquid scintillation counting. Some very sensitive nonradioactive labeling methods are now available. Those that employ chemiluminescence can be detected by autoradiography or by phosphorimaging. Those that produce colored products can be detected directly.

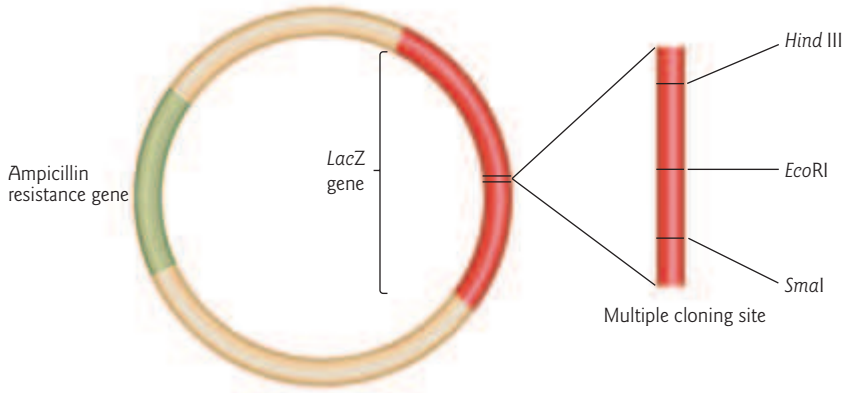
Restriction mapping determines the number, order, and distance between restriction endonuclease cutting sites along a cloned DNA fragment. To make a restriction map, a cloned DNA fragment is cut with restriction endonucleases and loaded on an agarose gel for electrophoresis. Both DNA and RNA fragments can be separated by size using gel electrophoresis. The most common gel used in nucleic acid electrophoresis is agarose, but polyacrylamide is used for the separation of smaller fragments such as in DNA sequencing. Some restriction fragment length polymorphisms (RFLPs) are used as markers for genetic diseases.

Labeled DNA (or RNA) probes can be used to hybridize to DNAs of the same, or very similar, sequence on a Southern blot. The number of bands that hybridize to a short probe gives an estimate of the number of closely related genes in an organism. Southern blots can be used to complement restriction mapping of cloned DNA and to identify overlapping fragments, or to identify structural differences between genomes, through analysis of RFLPs.

The Sanger DNA sequencing method uses a radiolabeled primer to initiate synthesis by DNA polymerase and dideoxynucleotides (ddNTPs) to terminate DNA synthesis, yielding a series of labeled DNA fragments, each one base longer than the last. These fragments can be separated according to size by electrophoresis. The last base in each of these fragments is known, because we know which ddNTP was used to teminate each of four reactions. Ordering these fragments by size tells us the base sequence of the DNA. The sequence of the original strand of DNA is complementary to the sequence read from the autoradiograph of the gel. In automated DNA sequencers, radioactive markers are replaced with fluorescent ones. An electropherogram – a graph of fluorescence intensity versus time – is converted to the DNA sequence.

Analytical questions

1 You have attempted to ligate a 1.5 kb fragment of foreign DNA into the *Eco*RI site in the multiple cloning site of the 4.0 kb plasmid vector shown below:



(a) After ligation you use the DNA in the ligation mixture to transform host bacteria. Why is it important to use host bacteria that are deficient for restriction modification?

(b) You screen the bacteria that supposedly have been transformed with recombinant plasmid DNA. Some of the bacterial colonies growing on the nutrient agar plate that contains ampicillin and X-gal are white and some are blue. Explain these results.

(c) To confirm the presence of the foreign DNA insert, you perform *Eco*RI restriction endonuclease digests on DNA extracted from bacterial colonies. Draw a diagram of an agarose gel showing the orientation of the positive and negative electrodes and the pattern of bands (label their size in kilobases) you would expect to see for *Eco*RI-digested recombinant plasmid and *Eco*RI-digested nonrecombinant plasmid vector, after electrophoresis and staining of the gel with ethidium bromide.

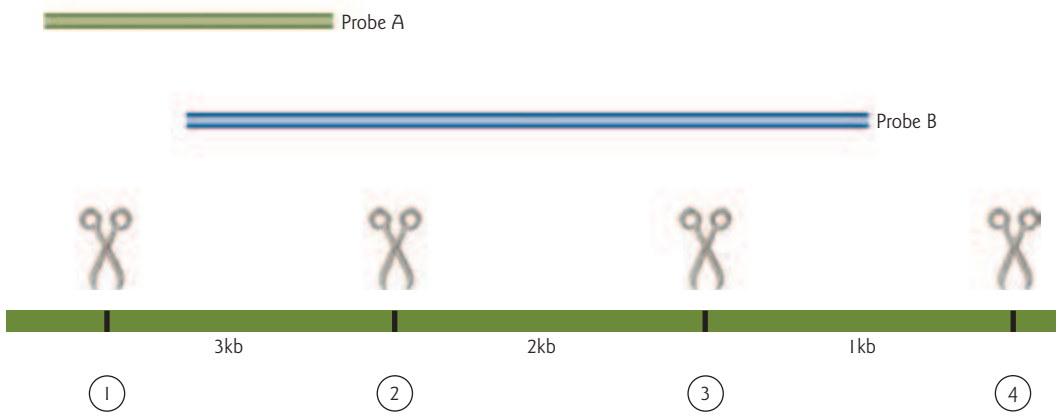
2 A chromatographic column in which oligo-dT is linked to an inert substance is useful in separating eukaryotic mRNA from other RNA molecules. On what principle does this column operate?

3 Starting with the nucleotide sequence of the human DNA ligase I gene, describe how you would search for a homologous gene in another organism whose genome has been sequenced, such as the pufferfish *Tetraodon nigroviridis*. Then, describe how you would obtain the protein and test it for ligase activity.

4 You plan to use the polymerase chain reaction to amplify part of the DNA sequence shown below, using oligonucleotide primers that are hexamers matching the regions shown in bold. (In practice, hexamers are too short for most purposes.) State the sequence of the primer oligonucleotides that should be used, including their polarity (5′ → 3′), and give the sequence of the DNA molecule that results from amplification.

5′-TAGGCAT**GCAATG**GTAATTTTTCAGGAACCAGGGCCCTT**AAGCCG**TAGGCAT-3′
3′-ATCGGTA**CGTTAC**CATTAAAAACTCCTTGGTCCCGGGAA**TTCGGC**ATCGGTA-5′

5 The following is a physical map of a region you are mapping for RFLP analysis:



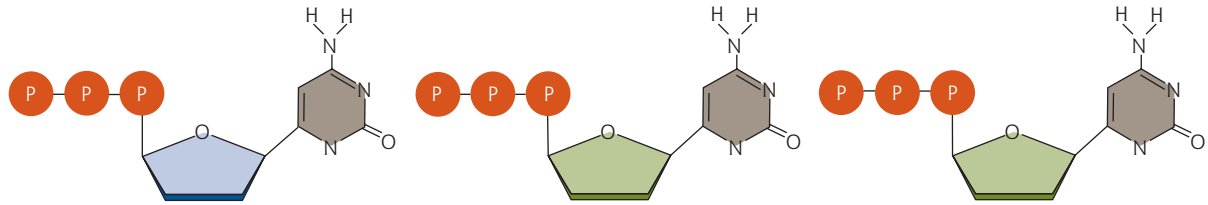
The numbered vertical lines represent restriction sites recognized by *Sma*I. Sites 2 and 3 are polymorphic, the others are not. You cut the DNA with *Sma*I, electrophorese the fragments, and Southern blot them to a membrane. You have a choice of two probes that recognize the DNA regions shown above: probe A, green line; probe B, blue line.

(a) Explain which probe you would use for analysis and why the other choice would be unsuitable.

(b) Give the sizes of bands you will detect in individuals homozygous for the following genotypes with respect to sites 2 and 3:

Haplotype	Site 2	Site 3
A	Present	Present
B	Present	Absent
C	Absent	Present
D	Absent	Absent

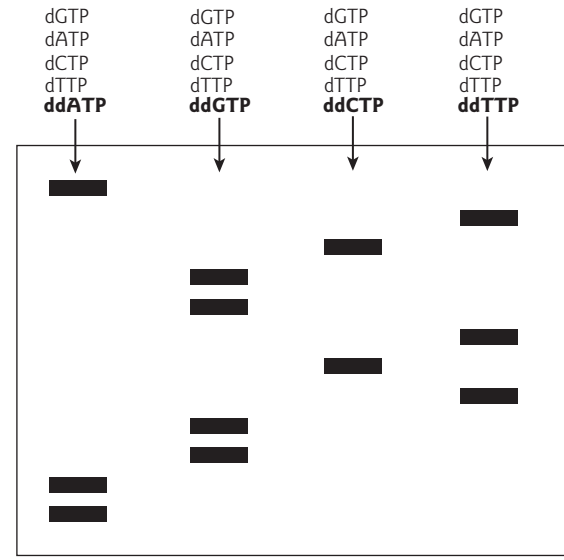
6 Complete the incomplete diagrams below to show the key structural difference between an NTP, dNTP, and ddNTP (e.g. CTP, dCTP, ddCTP):



(a) What do "d" and "dd" stand for?

(b) Explain why ddNTPs are called "chain terminators" in DNA sequencing reactions.

7 The nucleotide sequence of a DNA fragment was determined by the Sanger (dideoxy) DNA sequencing method. The data are shown below. What is the 5′ → 3′ sequence of the nucleotides in the original DNA fragment?



Suggestions for further reading

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K., eds (2002) *Short Protocols in Molecular Biology*, 5th edn. John Wiley & Sons, New York.

Echols, H. (2001) *Operators and Promoters. The Story of Molecular Biology and Its Creators*. University of California Press, Berkeley, CA.

Gowers, D.M., Wilson, G.G., Halford, S.E. (2005) Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proceedings of the National Academy of Sciences USA* 102:15883–15888.

Harrington, J.J., van Bokkelen, G., Mays, R.W., Gutashaw, K., Willard, H.F. (1997) Formation of *de novo* centromeres and construction of first generation human artificial microchromosomes. *Nature Genetics* 125:345–355.

Katoh, M., Ayabe, F., Norikane, S. et al. (2004) Construction of a novel human artificial chromosome vector for gene delivery. *Biochemical and Biophysical Research Communications* 321:280–290.

Kurpiewski, M.R., Engler, L.E., Wozniak, L.A., Kobylanska, A., Koziolkiewicz, M., Stec, W.J., Jen-Jacobsen, L. (2004) Mechanisms of coupling between DNA recognition and specificity and catalysis in *Eco*RI endonuclease. *Structure* 12:1775–1788.

Lipps, H.J., Jenke, A.C.W., Nehlsen, K., Scinteie, M.F., Stehle, I.M., Bode, J. (2003) Chromosome-based vectors for gene therapy. *Gene* 304:23–33.

Love-Gregory, L.D., Dyer, J.A., Grasela, J., Hillman, R.E., Philips, C.L. (2001) Carrier detection and rapid newborn diagnostic test for the common Y393N maple syrup urine disease allele by PCR-RFLP: culturally permissible testing in the Mennonite community. *Journal of Inherited and Metabolic Diseases* 24:393–403.

Morrow, J.F., Cohen, S.N., Chang, A.C.Y., Boyer, H.W., Goodman, H.M., Helling, R.B. (1974) Replication and transcription of eukaryotic DNA in *Escherichia coli*. *Proceedings of the National Academy of Science USA* 71:1743–1747.

Mullis, K.B. (1990) The unusual origin of the polymerase chain reaction. *Scientific American* 262:36–43.

Pingoud, A., Jeltsch, A. (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Research* 29:3705–3727.

Sambrook, J., Russell, D.W. (2001) *Molecular Cloning: a Laboratory Manual*, 3rd edn. Cold Spring Harbor Laboratory Press, New York.

Viadiu, H., Aggarwal, A.K. (2000) Structure of *Bam*HI bound to nonspecific DNA: a model for DNA sliding. *Molecular Cell* 5:889–895.

Watson, J. (1993) The human genome initiative. In: *Genetics and Society* (eds B. Holland, C. Kyriacou), pp. 13–26. Addison-Wesley Publishing Co., New York.