

spectrometer then separates these fragments and displays their masses. The amino acid sequence can be deduced from the differences in mass between the peptides (Figure 8–20B). Post-translational modifications are identified when the amino acid to which they are attached show a characteristically increased mass.

To learn more about the structure and function of a protein, one must obtain large amounts of the protein for analysis. This is most often accomplished by using the powerful recombinant DNA technologies discussed next.

Summary

Populations of cells can be analyzed biochemically by disrupting them and fractionating their contents by ultracentrifugation. Further fractionations allow functional cell-free systems to be developed; such systems are required to determine the molecular details of complex cellular processes. Protein synthesis, DNA replication, RNA splicing, the cell cycle, mitosis, and various types of intracellular transport can all be studied in this way. The molecular weight and subunit composition of even very small amounts of a protein can be determined by SDS polyacrylamide-gel electrophoresis. In two-dimensional gel electrophoresis, proteins are resolved as separate spots by isoelectric focusing in one dimension, followed by SDS polyacrylamide-gel electrophoresis in a second dimension. These electrophoretic separations can be applied even to proteins that are normally insoluble in water.

The major proteins in soluble cell extracts can be purified by column chromatography; depending on the type of column matrix, biologically active proteins can be separated on the basis of their molecular weight, hydrophobicity, charge characteristics, or affinity for other molecules. In a typical purification the sample is passed through several different columns in turn—the enriched fractions obtained from one column are applied to the next. Once a protein has been purified to homogeneity, its biological activities can be examined in detail. Using mass spectrometry, the masses of proteins and peptides derived from them can be rapidly determined. With this information one can refer to genome databases to deduce the remaining amino acid sequence of the protein from the nucleotide sequence of its gene.

ISOLATING, CLONING, AND SEQUENCING DNA

Until the early 1970s DNA was the most difficult cellular molecule for the biochemist to analyze. Enormously long and chemically monotonous, the string of nucleotides that forms the genetic material of an organism could be examined only indirectly, by protein or RNA sequencing or by genetic analysis. Today the situation has changed entirely. From being the most difficult macromolecule of the cell to analyze, DNA has become the easiest. It is now possible to isolate a specific region of a genome, to produce a virtually unlimited number of copies of it, and to determine the sequence of its nucleotides overnight. At the height of the Human Genome Project, large facilities with automated machines were generating DNA sequences at the rate of 1000 nucleotides per second, around the clock. By related techniques, an isolated gene can be altered (engineered) at will and transferred back into the germ line of an animal or plant, so as to become a functional and heritable part of the organism's genome.

These technical breakthroughs in genetic engineering—the ability to manipulate DNA with precision in a test tube or an organism—have had a dramatic impact on all aspects of cell biology by facilitating the study of cells and their macromolecules in previously unimagined ways. They have led to the discovery of whole new classes of genes and proteins, while revealing that many proteins have been much more highly conserved in evolution than had been suspected. They have provided new tools for determining the functions of proteins and of individual domains within proteins, revealing a host of unexpected relationships between them. By making available large amounts of any protein, they have shown the way to efficient mass production of protein hormones and vaccines. Finally, by allowing the regulatory regions of genes to be dissected,

they provide biologists with an important tool for unraveling the complex regulatory networks by which eucaryotic gene expression is controlled.

Recombinant DNA technology comprises a mixture of techniques, some new and some borrowed from other fields such as microbial genetics (Table 8–7). Central to the technology are the following key techniques:

1. Cleavage of DNA at specific sites by restriction nucleases, which greatly facilitates the isolation and manipulation of individual genes.
2. DNA cloning either through the use of cloning vectors or the polymerase chain reaction, whereby a single DNA molecule can be copied to generate many billions of identical molecules.
3. Nucleic acid hybridization, which makes it possible to find a specific sequence of DNA or RNA with great accuracy and sensitivity on the basis of its ability to bind a complementary nucleic acid sequence.
4. Rapid sequencing of all the nucleotides in a purified DNA fragment, which makes it possible to identify genes and to deduce the amino acid sequence of the proteins they encode.

TABLE 8–7 Some Major Steps in the Development of Recombinant DNA and Transgenic Technology

1869	Miescher first isolates DNA from white blood cells harvested from pus-soaked bandages obtained from a nearby hospital.
1944	Avery provides evidence that DNA, rather than protein, carries the genetic information during bacterial transformation.
1953	Watson and Crick propose the double-helix model for DNA structure based on x-ray results of Franklin and Wilkins .
1955	Kornberg discovers DNA polymerase, the enzyme now used to produce labeled DNA probes.
1961	Marmur and Doty discover DNA renaturation, establishing the specificity and feasibility of nucleic acid hybridization reactions.
1962	Arber provides the first evidence for the existence of DNA restriction nucleases, leading to their purification and use in DNA sequence characterization by Nathans and H. Smith .
1966	Nirenberg, Ochoa, and Khorana elucidate the genetic code.
1967	Gellert discovers DNA ligase, the enzyme used to join DNA fragments together.
1972–1973	DNA cloning techniques are developed by the laboratories of Boyer, Cohen, Berg , and their colleagues at Stanford University and the University of California at San Francisco.
1975	Southern develops gel-transfer hybridization for the detection of specific DNA sequences.
1975–1977	Sanger and Barrell and Maxam and Gilbert develop rapid DNA-sequencing methods.
1981–1982	Palmiter and Brinster produce transgenic mice; Spradling and Rubin produce transgenic fruit flies.
1982	GenBank , NIH's public genetic sequence database, is established at Los Alamos National Laboratory.
1985	Mullis and co-workers invent the polymerase chain reaction (PCR).
1987	Capecchi and Smithies introduce methods for performing targeted gene replacement in mouse embryonic stem cells.
1989	Fields and Song develop the yeast two-hybrid system for identifying and studying protein interactions
1989	Olson and colleagues describe sequence-tagged sites, unique stretches of DNA that are used to make physical maps of human chromosomes.
1990	Lipman and colleagues release BLAST, an algorithm used to search for homology between DNA and protein sequences.
1990	Simon and colleagues study how to efficiently use bacterial artificial chromosomes, BACs, to carry large pieces of cloned human DNA for sequencing.
1991	Hood and Hunkapillar introduce new automated DNA sequence technology.
1995	Venter and colleagues sequence the first complete genome, that of the bacterium <i>Haemophilus influenzae</i> .
1996	Goffeau and an international consortium of researchers announce the completion of the first genome sequence of a eucaryote, the yeast <i>Saccharomyces cerevisiae</i> .
1996–1997	Lockhart and colleagues and Brown and DeRisi produce DNA microarrays, which allow the simultaneous monitoring of thousands of genes.
1998	Sulston and Waterston and colleagues produce the first complete sequence of a multicellular organism, the nematode worm <i>Caenorhabditis elegans</i> .
2001	Consortia of researchers announce the completion of the draft human genome sequence.

Figure 8–21 The DNA nucleotide sequences recognized by four widely used restriction nucleases. As in the examples shown, such sequences are often six base pairs long and “palindromic” (that is, the nucleotide sequence is the same if the helix is turned by 180 degrees around the center of the short region of helix that is recognized). The enzymes cut the two strands of DNA at or near the recognition sequence. For some enzymes, such as *HpaI*, the cleavage leaves blunt ends; for others, such as *EcoRI*, *HindIII*, and *PstI*, the cleavage is staggered and creates cohesive ends. Restriction nucleases are obtained from various species of bacteria: *HpaI* is from *Hemophilus parainfluenzae*, *EcoRI* is from *Escherichia coli*, *HindIII* is from *Hemophilus influenzae*, and *PstI* is from *Providencia stuartii*.

5. Simultaneous monitoring of the expression level of each gene in a cell, using nucleic acid microarrays that allow tens of thousands of hybridization reactions to be performed simultaneously.

In this chapter we describe each of these basic techniques, which together have revolutionized the study of cell biology.

Large DNA Molecules Are Cut into Fragments by Restriction Nucleases

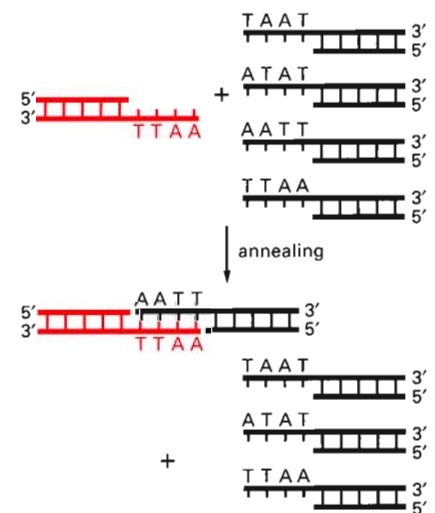
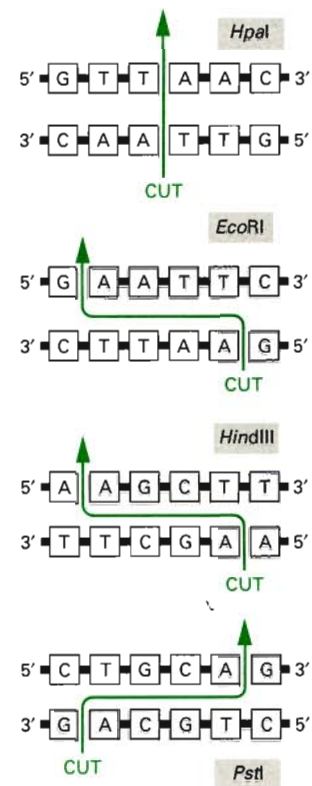
Unlike a protein, a gene does not exist as a discrete entity in cells, but rather as a small region of a much longer DNA molecule. Although the DNA molecules in a cell can be randomly broken into small pieces by mechanical force, a fragment containing a single gene in a mammalian genome would still be only one among a hundred thousand or more DNA fragments, indistinguishable in their average size. How could such a gene be purified? Because all DNA molecules consist of an approximately equal mixture of the same four nucleotides, they cannot be readily separated, as proteins can, on the basis of their different charges and binding properties. Moreover, even if a purification scheme could be devised, vast amounts of DNA would be needed to yield enough of any particular gene to be useful for further experiments.

The solution to all of these problems began to emerge with the discovery of **restriction nucleases**. These enzymes, which can be purified from bacteria, cut the DNA double helix at specific sites defined by the local nucleotide sequence, thereby cleaving a long double-stranded DNA molecule into fragments of strictly defined sizes. Different restriction nucleases have different sequence specificities, and it is relatively simple to find an enzyme that can create a DNA fragment that includes a particular gene. The size of the DNA fragment can then be used as a basis for partial purification of the gene from a mixture.

Different species of bacteria make different restriction nucleases, which protect them from viruses by degrading incoming viral DNA. Each nuclease recognizes a specific sequence of four to eight nucleotides in DNA. These sequences, where they occur in the genome of the bacterium itself, are protected from cleavage by methylation at an A or a C residue; the sequences in foreign DNA are generally not methylated and so are cleaved by the restriction nucleases. Large numbers of restriction nucleases have been purified from various species of bacteria; several hundred, most of which recognize different nucleotide sequences, are now available commercially.

Some restriction nucleases produce staggered cuts, which leave short single-stranded tails at the two ends of each fragment (Figure 8–21). Ends of this type are known as *cohesive ends*, as each tail can form complementary base pairs with the tail at any other end produced by the same enzyme (Figure 8–22). The

Figure 8–22 Restriction nucleases produce DNA fragments that can be easily joined together. Fragments with the same cohesive ends can readily join by complementary base-pairing between their cohesive ends, as illustrated. The two DNA fragments that join in this example were both produced by the *EcoRI* restriction nuclease, whereas the three other fragments were produced by different restriction nucleases that generated different cohesive ends (see Figure 8–21). Blunt-ended fragments, like those generated by *HpaI* (see Figure 8–21), can be spliced together with more difficulty.



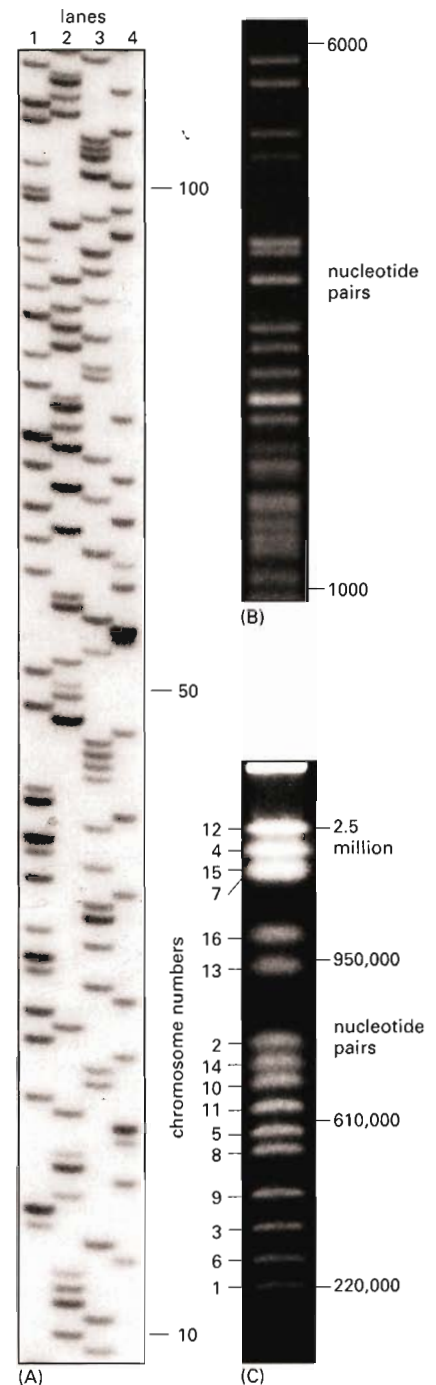
cohesive ends generated by restriction enzymes allow any two DNA fragments to be easily joined together, as long as the fragments were generated with the same restriction nuclease (or with another nuclease that produces the same cohesive ends). DNA molecules produced by splicing together two or more DNA fragments are called **recombinant DNA** molecules; they have made possible many new types of cell-biological studies.

Gel Electrophoresis Separates DNA Molecules of Different Sizes

The length and purity of DNA molecules can be accurately determined by the same types of gel electrophoresis methods that have proved so useful in the analysis of proteins. The procedure is actually simpler than for proteins: because each nucleotide in a nucleic acid molecule already carries a single negative charge, there is no need to add the negatively charged detergent SDS that is required to make protein molecules move uniformly toward the positive electrode. For DNA fragments less than 500 nucleotides long, specially designed polyacrylamide gels allow separation of molecules that differ in length by as little as a single nucleotide (Figure 8–23A). The pores in polyacrylamide gels, however, are too small to permit very large DNA molecules to pass; to separate these by size, the much more porous gels formed by dilute solutions of agarose (a polysaccharide isolated from seaweed) are used (Figure 8–23B). These DNA separation methods are widely used for both analytical and preparative purposes.

A variation of agarose gel electrophoresis, called *pulsed-field gel electrophoresis*, makes it possible to separate even extremely long DNA molecules. Ordinary gel electrophoresis fails to separate such molecules because the steady electric field stretches them out so that they travel end-first through the gel in snakelike configurations at a rate that is independent of their length. In pulsed-field gel electrophoresis, by contrast, the direction of the electric field is changed periodically, which forces the molecules to reorient before continuing to move snakelike through the gel. This reorientation takes much more time for larger molecules, so that longer molecules move more slowly than shorter ones. As a consequence, even entire bacterial or yeast chromosomes separate into discrete

Figure 8–23 Gel electrophoresis techniques for separating DNA molecules by size. In the three examples shown, electrophoresis is from top to bottom, so that the largest—and thus slowest-moving—DNA molecules are near the top of the gel. In (A) a polyacrylamide gel with small pores is used to fractionate single-stranded DNA. In the size range 10 to 500 nucleotides, DNA molecules that differ in size by only a single nucleotide can be separated from each other. In the example, the four lanes represent sets of DNA molecules synthesized in the course of a DNA-sequencing procedure. The DNA to be sequenced has been artificially replicated from a fixed start site up to a variable stopping point, producing a set of partial replicas of differing lengths. (Figure 8–36 explains how such sets of partial replicas are synthesized.) Lane 1 shows all the partial replicas that terminate in a G, lane 2 all those that terminate in an A, lane 3 all those that terminate in a T, and lane 4 all those that terminate in a C. Since the DNA molecules used in these reactions are radiolabeled, their positions can be determined by autoradiography, as shown. In (B) an agarose gel with medium-sized pores is used to separate double-stranded DNA molecules. This method is most useful in the size range 300 to 10,000 nucleotide pairs. These DNA molecules are fragments produced by cleaving the genome of a bacterial virus with a restriction nuclease, and they have been detected by their fluorescence when stained with the dye ethidium bromide. In (C) the technique of pulsed-field agarose gel electrophoresis has been used to separate 16 different yeast (*Saccharomyces cerevisiae*) chromosomes, which range in size from 220,000 to 2.5 million nucleotide pairs (see Figure 4–13). The DNA was stained as in (B). DNA molecules as large as 10^7 nucleotide pairs can be separated in this way. (A, courtesy of Leander Lauffer and Peter Walter; B, courtesy of Ken Kreuzer; C, from D. Vollrath and R.W. Davis, *Nucleic Acids Res.* 15:7865–7876, 1987. © Oxford University Press.)



bands in pulsed-field gels and so can be sorted and identified on the basis of their size (Figure 8–23C). Although a typical mammalian chromosome of 10^8 base pairs is too large to be sorted even in this way, large segments of these chromosomes are readily separated and identified if the chromosomal DNA is first cut with a restriction nuclease selected to recognize sequences that occur only rarely (once every 10,000 or more nucleotide pairs).

The DNA bands on agarose or polyacrylamide gels are invisible unless the DNA is labeled or stained in some way. One sensitive method of staining DNA is to expose it to the dye *ethidium bromide*, which fluoresces under ultraviolet light when it is bound to DNA (see Figures 8–23B,C). An even more sensitive detection method incorporates a radioisotope into the DNA molecules before electrophoresis; ^{32}P is often used as it can be incorporated into DNA phosphates and emits an energetic β particle that is easily detected by autoradiography (as in Figure 8–23A).

Purified DNA Molecules Can Be Specifically Labeled with Radioisotopes or Chemical Markers *in vitro*

Two procedures are widely used to label isolated DNA molecules. In the first method a DNA polymerase copies the DNA in the presence of nucleotides that are either radioactive (usually labeled with ^{32}P) or chemically tagged (Figure 8–24A). In this way “DNA probes” containing many labeled nucleotides can be produced for nucleic acid hybridization reactions (discussed below). The second procedure uses the bacteriophage enzyme polynucleotide kinase to transfer a single ^{32}P -labeled phosphate from ATP to the 5′ end of each DNA chain (Figure 8–24B). Because only one ^{32}P atom is incorporated by the kinase into each DNA strand, the DNA molecules labeled in this way are often not radioactive enough to be used as DNA probes; because they are labeled at only one end, however, they have been invaluable for other applications including DNA footprinting, as we see shortly.

Today, radioactive labeling methods are being replaced by labeling with molecules that can be detected chemically or through fluorescence. To produce such nonradioactive DNA molecules, specially modified nucleotide precursors are used (Figure 8–24C). A DNA molecule made in this way is allowed to bind to its complementary DNA sequence by hybridization, as discussed in the next section, and is then detected with an antibody (or other ligand) that specifically recognizes its modified side chain (see Figure 8–28).

Nucleic Acid Hybridization Reactions Provide a Sensitive Way of Detecting Specific Nucleotide Sequences

When an aqueous solution of DNA is heated at 100°C or exposed to a very high pH ($\text{pH} \geq 13$), the complementary base pairs that normally hold the two strands of the double helix together are disrupted and the double helix rapidly dissociates into two single strands. This process, called *DNA denaturation*, was for many years thought to be irreversible. In 1961, however, it was discovered that complementary single strands of DNA readily re-form double helices by a process called **hybridization** (also called *DNA renaturation*) if they are kept for a prolonged period at 65°C . Similar hybridization reactions can occur between any two single-stranded nucleic acid chains (DNA/DNA, RNA/RNA, or RNA/DNA), provided that they have complementary nucleotide sequences. These specific hybridization reactions are widely used to detect and characterize specific nucleotide sequences in both RNA and DNA molecules.

Single-stranded DNA molecules used to detect complementary sequences are known as **probes**; these molecules, which carry radioactive or chemical markers to facilitate their detection, can be anywhere from fifteen to thousands of nucleotides long. Hybridization reactions using DNA probes are so sensitive and selective that they can detect complementary sequences present at a concentration as low as one molecule per cell. It is thus possible to determine how many copies of any DNA sequence are present in a particular DNA sample. The

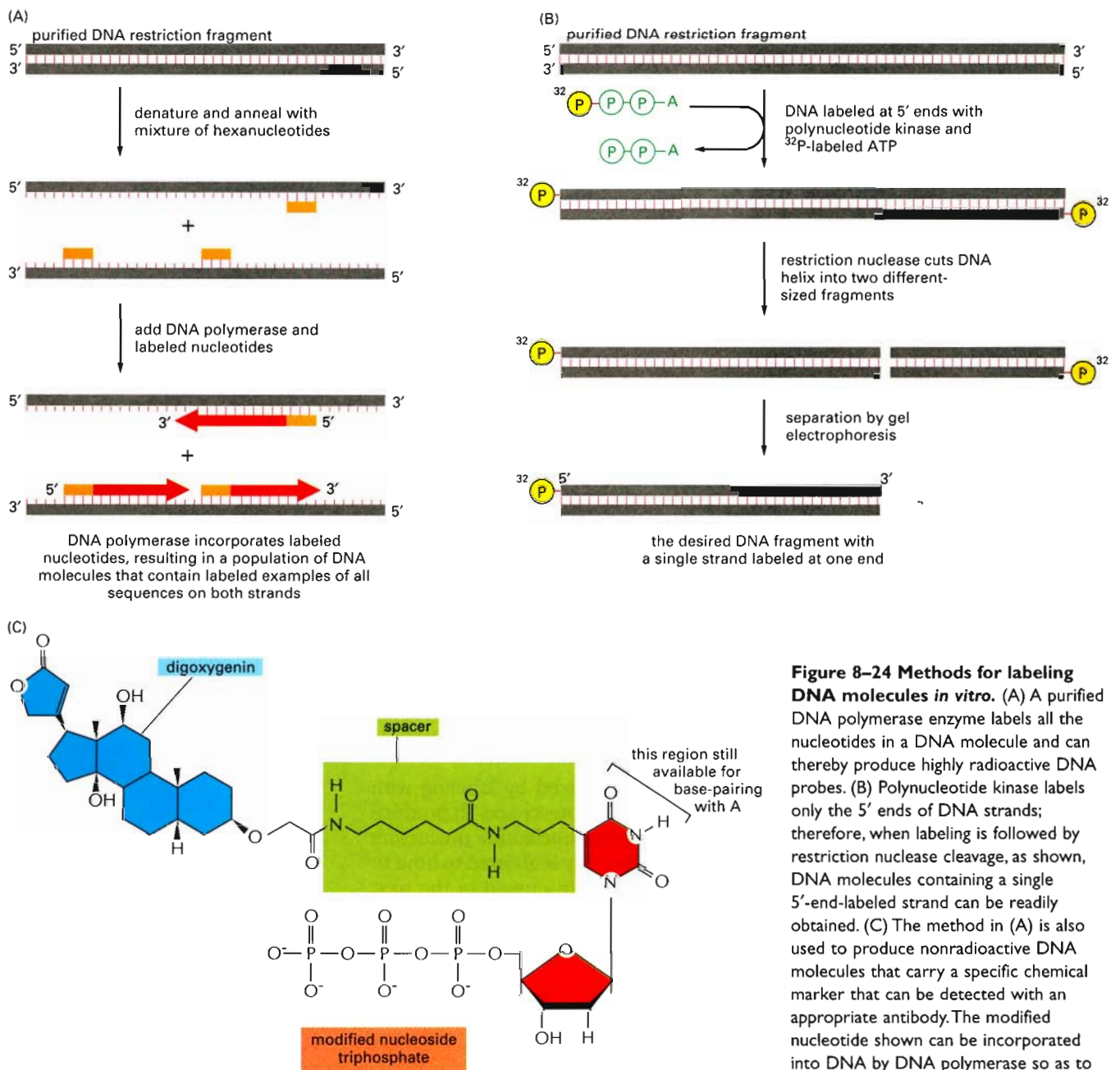


Figure 8-24 Methods for labeling DNA molecules *in vitro*. (A) A purified DNA polymerase enzyme labels all the nucleotides in a DNA molecule and can thereby produce highly radioactive DNA probes. (B) Polynucleotide kinase labels only the 5' ends of DNA strands; therefore, when labeling is followed by restriction nuclease cleavage, as shown, DNA molecules containing a single 5'-end-labeled strand can be readily obtained. (C) The method in (A) is also used to produce nonradioactive DNA molecules that carry a specific chemical marker that can be detected with an appropriate antibody. The modified nucleotide shown can be incorporated into DNA by DNA polymerase so as to allow the DNA molecule to serve as a probe that can be readily detected. The base on the nucleoside triphosphate shown is an analog of thymine in which the methyl group on T has been replaced by a spacer arm linked to the plant steroid digoxigenin. To visualize the probe, the digoxigenin is detected by a specific antibody coupled to a visible marker such as a fluorescent dye. Other chemical labels such as biotin can be attached to nucleotides and used in essentially the same way.

same technique can be used to search for related but nonidentical genes. To find a gene of interest in an organism whose genome has not yet been sequenced, for example, a portion of a known gene can be used as a probe (Figure 8-25).

Alternatively, DNA probes can be used in hybridization reactions with RNA rather than DNA to find out whether a cell is expressing a given gene. In this case a DNA probe that contains part of the gene's sequence is hybridized with RNA purified from the cell in question to see whether the RNA includes molecules matching the probe DNA and, if so, in what quantities. In somewhat more elaborate procedures the DNA probe is treated with specific nucleases after the hybridization is complete, to determine the exact regions of the DNA probe that have paired with cellular RNA molecules. One can thereby determine the start and stop sites for RNA transcription, as well as the precise boundaries of the intron and exon sequences in a gene (Figure 8-26).

Today, the positions of intron/exon boundaries are usually determined by sequencing the cDNA sequences that represent the mRNAs expressed in a cell. Comparing this expressed sequence with the sequence of the whole gene reveals where the introns lie. We review later how cDNAs are prepared from mRNAs.

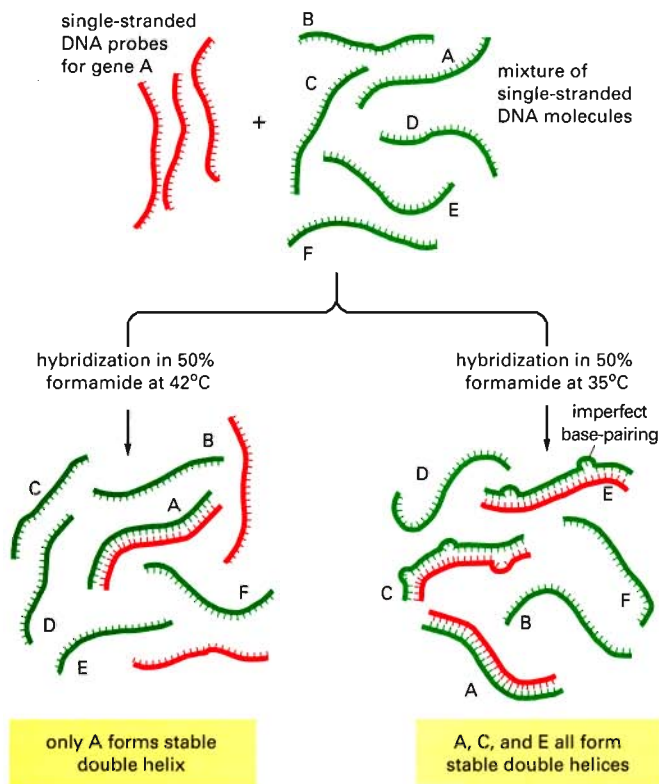


Figure 8–25 Different hybridization conditions allow less than perfect DNA matching. When only an identical match with a DNA probe is desired, the hybridization reaction is kept just a few degrees below the temperature at which a perfect DNA helix denatures in the solvent used (its *melting temperature*), so that all imperfect helices formed are unstable. When a DNA probe is being used to find DNAs that are related, but not identical, in sequence, hybridization is performed at a lower temperature. This allows even imperfectly paired double helices to form. Only the lower-temperature hybridization conditions can be used to search for genes (C and E in this example) that are nonidentical but related to gene A (see Figure 10–18).

We have seen that genes are switched on and off as a cell encounters new signals in its environment. The hybridization of DNA probes to cellular RNAs allows one to determine whether or not a particular gene is being transcribed; moreover, when the expression of a gene changes, one can determine whether the change is due to transcriptional or posttranscriptional controls (see Figure 7–87). These tests of gene expression were initially performed with one DNA probe at a time. DNA microarrays now allow the simultaneous monitoring of

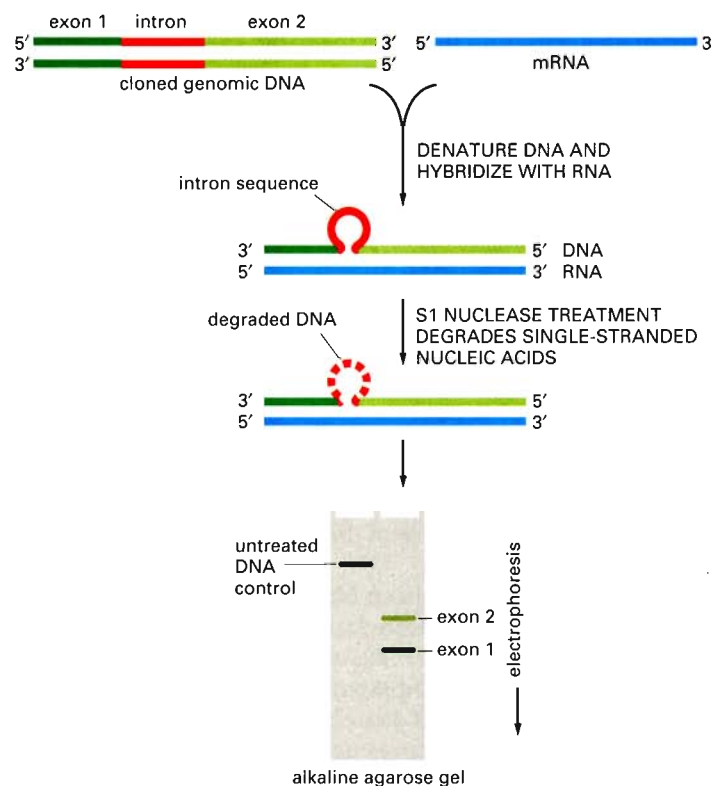


Figure 8–26 The use of nucleic acid hybridization to determine the region of a cloned DNA fragment that is present in an mRNA molecule. The method shown requires a nuclease that cuts the DNA chain only where it is not base-paired to a complementary RNA chain. The positions of the introns in eucaryotic genes are mapped by the method shown; the beginning and the end of an RNA molecule can be determined in the same way. For this type of analysis the DNA is electrophoresed through a denaturing agarose gel, which causes it to migrate as single-stranded molecules.

hundreds or thousands of genes at a time, as we discuss later. Hybridization methods are in such wide use in cell biology today that it is difficult to imagine how we could study gene structure and expression without them.

Northern and Southern Blotting Facilitate Hybridization with Electrophoretically Separated Nucleic Acid Molecules

DNA probes are often used to detect, in a complex mixture of nucleic acids, only those molecules with sequences that are complementary to all or part of the probe. Gel electrophoresis can be used to fractionate the many different RNA or DNA molecules in a crude mixture according to their size before the hybridization reaction is performed; if molecules of only one or a few sizes become labeled with the probe, one can be certain that the hybridization was indeed specific. Moreover, the size information obtained can be invaluable in itself. An example illustrates this point.

Suppose that one wishes to determine the nature of the defect in a mutant mouse that produces abnormally low amounts of albumin, a protein that liver cells normally secrete into the blood in large amounts. First, one collects identical samples of liver tissue from mutant and normal mice (the latter serving as controls) and disrupts the cells in a strong detergent to inactivate cellular nucleases that might otherwise degrade the nucleic acids. Next, one separates the RNA and DNA from all of the other cell components: the proteins present are completely denatured and removed by repeated extractions with phenol—a potent organic solvent that is partly miscible with water; the nucleic acids, which remain in the aqueous phase, are then precipitated with alcohol to separate them from the small molecules of the cell. Then one separates the DNA from the RNA by their different solubilities in alcohols and degrades any contaminating nucleic acid of the unwanted type by treatment with a highly specific enzyme—either an RNase or a DNase. The mRNAs are typically separated from bulk RNA by retention on a chromatography column that specifically binds the poly-A tails of mRNAs.

To analyze the albumin-encoding mRNAs with a DNA probe, a technique called **Northern blotting** is used. First, the intact mRNA molecules purified from mutant and control liver cells are fractionated on the basis of their sizes into a series of bands by gel electrophoresis. Then, to make the RNA molecules accessible to DNA probes, a replica of the pattern of RNA bands on the gel is made by transferring (“blotting”) the fractionated RNA molecules onto a sheet of nitrocellulose or nylon paper. The paper is then incubated in a solution containing a labeled DNA probe whose sequence corresponds to part of the template strand that produces albumin mRNA. The RNA molecules that hybridize to the labeled DNA probe on the paper (because they are complementary to part of the normal albumin gene sequence) are then located by detecting the bound probe by autoradiography or by chemical means (Figure 8–27). The size of the RNA molecules in each band that binds the probe can be determined by reference to bands of RNA molecules of known sizes (RNA standards) that are electrophoresed side by side with the experimental sample. In this way one might discover that liver cells from the mutant mice make albumin RNA in normal amounts and of normal size; alternatively, albumin RNA of normal size might be detected in greatly reduced amounts. Another possibility is that the mutant albumin RNA molecules might be abnormally short and therefore move unusually quickly through the gel; in this case the gel blot could be retested with a series of shorter DNA probes, each corresponding to small portions of the gene, to reveal which part of the normal RNA is missing.

An analogous gel-transfer hybridization method, called **Southern blotting**, analyzes DNA rather than RNA. Isolated DNA is first cut into readily separable fragments with restriction nucleases. The double-stranded fragments are then separated on the basis of size by gel electrophoresis, and those complementary to a DNA probe are identified by blotting and hybridization, as just described for RNA (see Figure 8–27). To characterize the structure of the albumin gene in the mutant mice, an albumin-specific DNA probe would be used to construct a

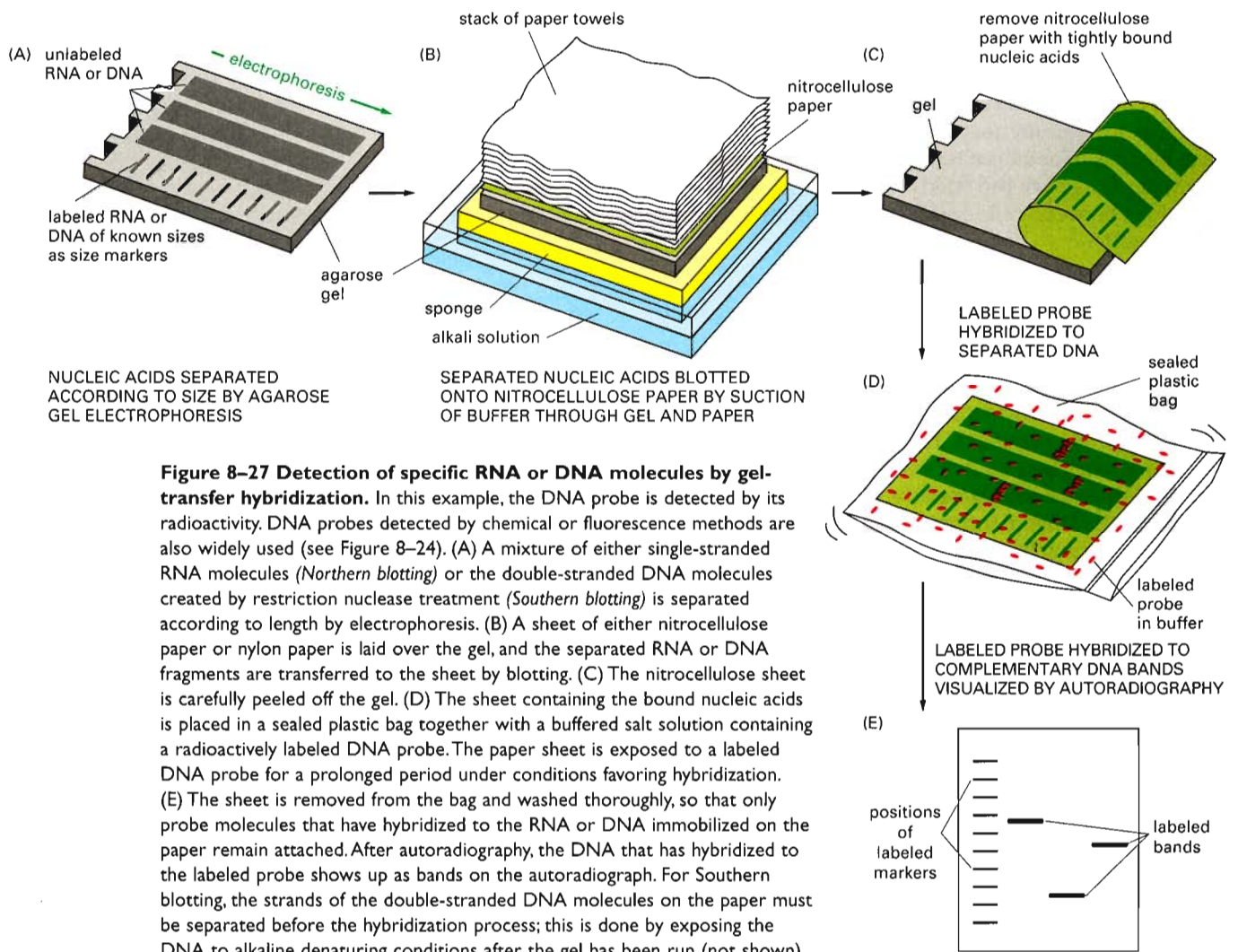


Figure 8-27 Detection of specific RNA or DNA molecules by gel-transfer hybridization. In this example, the DNA probe is detected by its radioactivity. DNA probes detected by chemical or fluorescence methods are also widely used (see Figure 8-24). (A) A mixture of either single-stranded RNA molecules (*Northern blotting*) or the double-stranded DNA molecules created by restriction nuclease treatment (*Southern blotting*) is separated according to length by electrophoresis. (B) A sheet of either nitrocellulose paper or nylon paper is laid over the gel, and the separated RNA or DNA fragments are transferred to the sheet by blotting. (C) The nitrocellulose sheet is carefully peeled off the gel. (D) The sheet containing the bound nucleic acids is placed in a sealed plastic bag together with a buffered salt solution containing a radioactively labeled DNA probe. The paper sheet is exposed to a labeled DNA probe for a prolonged period under conditions favoring hybridization. (E) The sheet is removed from the bag and washed thoroughly, so that only probe molecules that have hybridized to the RNA or DNA immobilized on the paper remain attached. After autoradiography, the DNA that has hybridized to the labeled probe shows up as bands on the autoradiograph. For Southern blotting, the strands of the double-stranded DNA molecules on the paper must be separated before the hybridization process; this is done by exposing the DNA to alkaline denaturing conditions after the gel has been run (not shown).

detailed *restriction map* of the genome in the region of the albumin gene. From this map one could determine if the albumin gene has been rearranged in the defective animals—for example, by the deletion or the insertion of a short DNA sequence; most single base changes, however, could not be detected in this way.

Hybridization Techniques Locate Specific Nucleic Acid Sequences in Cells or on Chromosomes

Nucleic acids, no less than other macromolecules, occupy precise positions in cells and tissues, and a great deal of potential information is lost when these molecules are extracted by homogenization. For this reason, techniques have been developed in which nucleic acid probes are used in much the same way as labeled antibodies to locate specific nucleic acid sequences *in situ*, a procedure called *in situ* hybridization. This procedure can now be done both for DNA in chromosomes and for RNA in cells. Labeled nucleic acid probes can be hybridized to chromosomes that have been exposed briefly to a very high pH to disrupt their DNA base pairs. The chromosomal regions that bind the probe during the hybridization step are then visualized. Originally, this technique was developed with highly radioactive DNA probes, which were detected by autoradiography. The spatial resolution of the technique, however, can be greatly improved by labeling the DNA probes chemically (Figure 8-28) instead of radioactively, as described earlier.

In situ hybridization methods have also been developed that reveal the distribution of specific RNA molecules in cells in tissues. In this case the tissues are

not exposed to a high pH, so the chromosomal DNA remains double-stranded and cannot bind the probe. Instead the tissue is gently fixed so that its RNA is retained in an exposed form that can hybridize when the tissue is incubated with a complementary DNA or RNA probe. In this way the patterns of differential gene expression can be observed in tissues, and the location of specific RNAs can be determined in cells (Figure 8–29). In the *Drosophila* embryo, for example, such patterns have provided new insights into the mechanisms that create distinctions between cells in different positions during development (described in Chapter 21).

Genes Can Be Cloned from a DNA Library

Any DNA fragment that contains a gene of interest can be cloned. In cell biology, the term **DNA cloning** is used in two senses. In one sense it literally refers to the act of making many identical copies of a DNA molecule—the amplification of a particular DNA sequence. However, the term is also used to describe the isolation of a particular stretch of DNA (often a particular gene) from the rest of a cell's DNA, because this isolation is greatly facilitated by making many identical copies of the DNA of interest.

DNA cloning in its most general sense can be accomplished in several ways. The simplest involves inserting a particular fragment of DNA into the purified DNA genome of a self-replicating genetic element—generally a virus or a plasmid. A DNA fragment containing a human gene, for example, can be joined in a test tube to the chromosome of a bacterial virus, and the new recombinant DNA molecule can then be introduced into a bacterial cell. Starting with only one such recombinant DNA molecule that infects a single cell, the normal replication mechanisms of the virus can produce more than 10^{12} identical virus DNA molecules in less than a day, thereby amplifying the amount of the inserted human DNA fragment by the same factor. A virus or plasmid used in this way is known as a *cloning vector*, and the DNA propagated by insertion into it is said to have been *cloned*.

To isolate a specific gene, one often begins by constructing a *DNA library*—a comprehensive collection of cloned DNA fragments from a cell, tissue, or organism. This library includes (one hopes) at least one fragment that contains the gene of interest. Libraries can be constructed with either a virus or a plasmid vector and are generally housed in a population of bacterial cells. The principles underlying the methods used for cloning genes are the same for either type of cloning vector, although the details may differ. Today most cloning is performed with plasmid vectors.

The **plasmid vectors** most widely used for gene cloning are small circular molecules of double-stranded DNA derived from larger plasmids that occur naturally in bacterial cells. They generally account for only a minor fraction of the total host bacterial cell DNA, but they can easily be separated owing to their

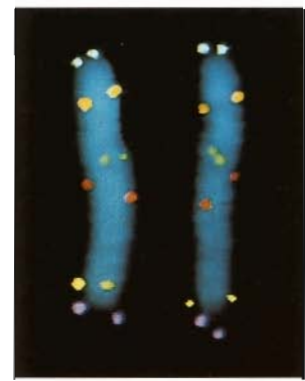


Figure 8–28 *In situ* hybridization to locate specific genes on chromosomes. Here, six different DNA probes have been used to mark the location of their respective nucleotide sequences on human chromosome 5 at metaphase. The probes have been chemically labeled and detected with fluorescent antibodies. Both copies of chromosome 5 are shown, aligned side by side. Each probe produces two dots on each chromosome, since a metaphase chromosome has replicated its DNA and therefore contains two identical DNA helices. (Courtesy of David C. Ward.)

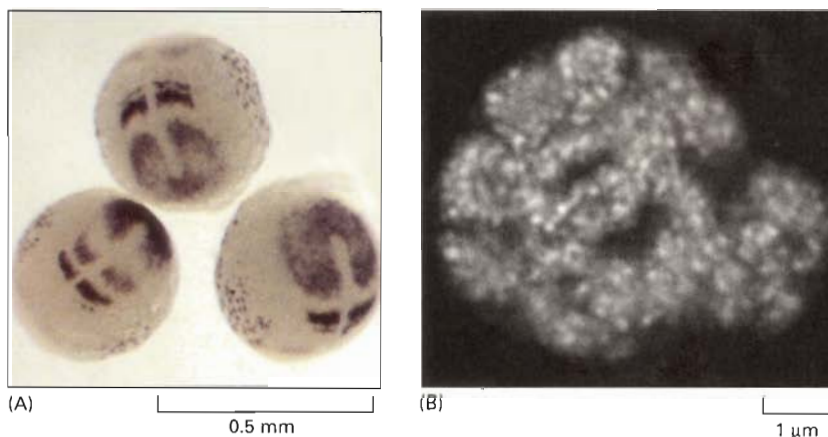


Figure 8–29 *In situ* hybridization for RNA localization. (A) Expression pattern of *deltaC* in the early zebrafish embryo. This gene codes for a ligand in the Notch signaling pathway (discussed in Chapter 15), and the pattern shown here reflects its role in the development of somites—the future segments of the vertebrate trunk and tail. (B) High-resolution RNA *in situ* localization reveals the sites within the nucleolus of a pea cell where ribosomal RNA is synthesized. The sausage-like structures, 0.5–1 μm in diameter, correspond to the loops of chromosomal DNA that contain the genes encoding rRNA. Each small, white spot represents transcription of a single rRNA gene. (A, courtesy of Yun-jin Jiang; B, courtesy of Peter Shaw.)

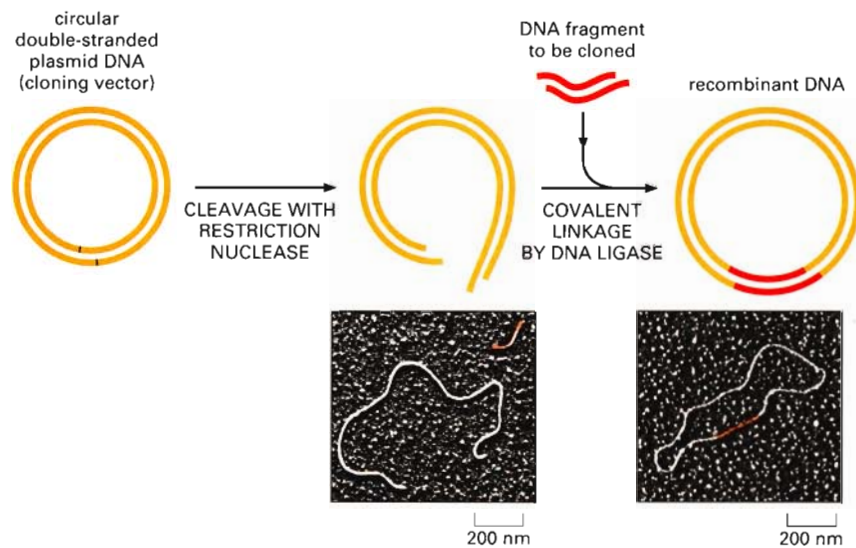


Figure 8–30 The insertion of a DNA fragment into a bacterial plasmid with the enzyme DNA ligase. The plasmid is cut open with a restriction nuclease (in this case one that produces cohesive ends) and is mixed with the DNA fragment to be cloned (which has been prepared with the same restriction nuclease), DNA ligase, and ATP. The cohesive ends base-pair, and DNA ligase seals the nicks in the DNA backbone, producing a complete recombinant DNA molecule. (Micrographs courtesy of Huntington Potter and David Dressler.)

small size from chromosomal DNA molecules, which are large and precipitate as a pellet upon centrifugation. For use as cloning vectors, the purified plasmid DNA circles are first cut with a restriction nuclease to create linear DNA molecules. The cellular DNA to be used in constructing the library is cut with the same restriction nuclease, and the resulting restriction fragments (including those containing the gene to be cloned) are then added to the cut plasmids and annealed via their cohesive ends to form recombinant DNA circles. These recombinant molecules containing foreign DNA inserts are then covalently sealed with the enzyme DNA ligase (Figure 8–30).

In the next step in preparing the library, the recombinant DNA circles are introduced into bacterial cells that have been made transiently permeable to DNA; such cells are said to be *transfected* with the plasmids. As these cells grow and divide, doubling in number every 30 minutes, the recombinant plasmids also replicate to produce an enormous number of copies of DNA circles containing the foreign DNA (Figure 8–31). Many bacterial plasmids carry genes for antibiotic resistance, a property that can be exploited to select those cells that have been successfully transfected; if the bacteria are grown in the presence of the antibiotic, only cells containing plasmids will survive. Each original bacterial cell that was initially transfected contains, in general, a different foreign DNA insert; this insert is inherited by all of the progeny cells of that bacterium, which together form a small colony in a culture dish.

For many years, plasmids were used to clone fragments of DNA of 1,000 to 30,000 nucleotide pairs. Larger DNA fragments are more difficult to handle and were harder to clone. Then researchers began to use yeast artificial chromosomes (YACs), which could handle very large pieces of DNA (Figure 8–32). Today, new plasmid vectors based on the naturally occurring F plasmid of *E. coli* are used to clone DNA fragments of 300,000 to 1 million nucleotide pairs. Unlike

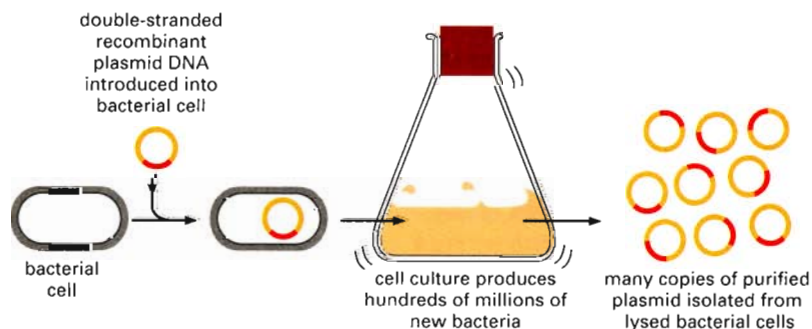


Figure 8–31 Purification and amplification of a specific DNA sequence by DNA cloning in a bacterium. To produce many copies of a particular DNA sequence, the fragment is first inserted into a plasmid vector, as shown in Figure 8–30. The resulting recombinant plasmid DNA is then introduced into a bacterium, where it can be replicated many millions of times as the bacterium multiplies.

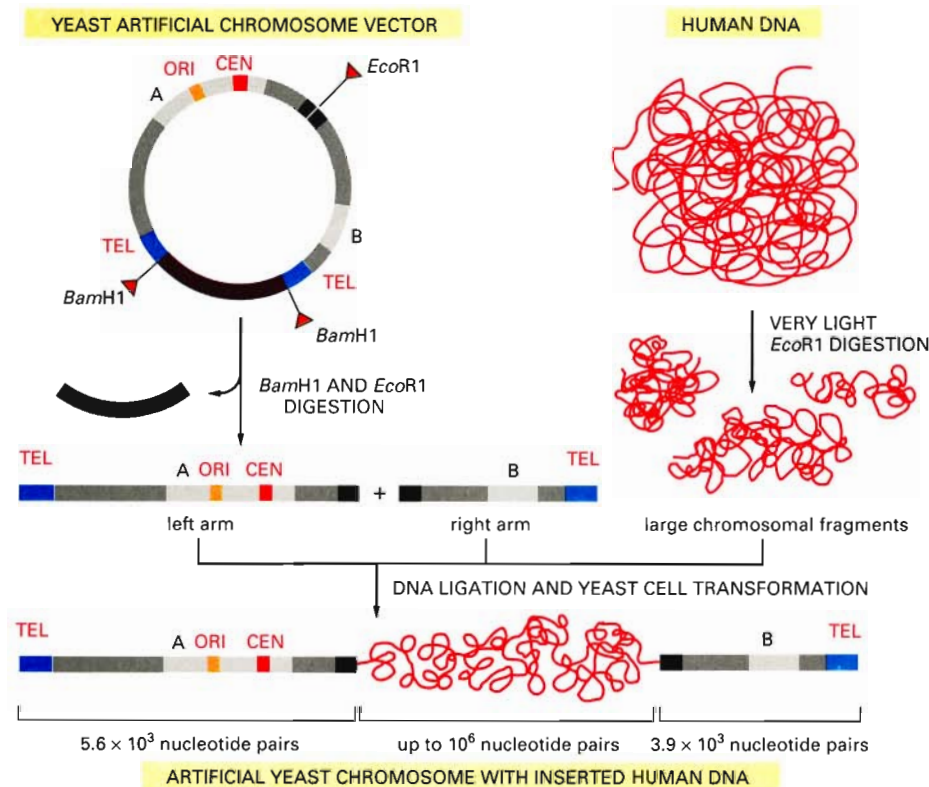


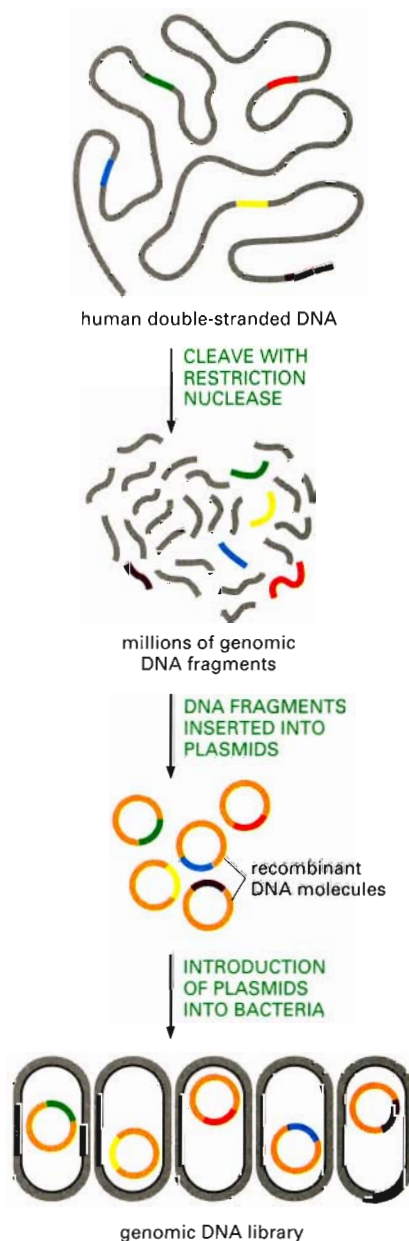
Figure 8–32 The making of a yeast artificial chromosome (YAC). A YAC vector allows the cloning of very large DNA molecules. TEL, CEN, and ORI are the telomere, centromere, and origin of replication sequences, respectively, for the yeast *Saccharomyces cerevisiae*. BamHI and EcoRI are sites where the corresponding restriction nucleases cut the DNA double helix. The sequences denoted A and B encode enzymes that serve as selectable markers to allow the easy isolation of yeast cells that have taken up the artificial chromosome. (Adapted from D.T. Burke, G.F. Carle, and M.V. Olson, *Science* 236:806–812, 1987.)

smaller bacterial plasmids, the F plasmid—and its derivative, the **bacterial artificial chromosome (BAC)**—is present in only one or two copies per *E. coli* cell. The fact that BACs are kept in such low numbers in bacterial cells may contribute to their ability to maintain large cloned DNA sequences stably: with only a few BACs present, it is less likely that the cloned DNA fragments will become scrambled due to recombination with sequences carried on other copies of the plasmid. Because of their stability, ability to accept large DNA inserts, and ease of handling, BACs are now the preferred vector for building DNA libraries of complex organisms—including those representing the human and mouse genomes.

Two Types of DNA Libraries Serve Different Purposes

Cleaving the entire genome of a cell with a specific restriction nuclease and cloning each fragment as just described is sometimes called the “shotgun” approach to gene cloning. This technique can produce a very large number of DNA fragments—on the order of a million for a mammalian genome—which will generate millions of different colonies of transfected bacterial cells. (When working with BACs rather than typical plasmids, larger fragments can be inserted, so fewer transfected bacterial cells are required to cover the genome.) Each of these colonies is composed of a clone of cells derived from a single ancestor cell, and therefore harbors many copies of a particular stretch of the fragmented genome (Figure 8–33). Such a plasmid is said to contain a **genomic DNA clone**, and the entire collection of plasmids is called a **genomic DNA library**. But because the genomic DNA is cut into fragments at random, only some fragments contain genes. Many of the genomic DNA clones obtained from the DNA of a higher eucaryotic cell contain only noncoding DNA, which, as we discussed in Chapter 4, makes up most of the DNA in such genomes.

Figure 8–33 Construction of a human genomic DNA library. A genomic library is usually stored as a set of bacteria, each carrying a different fragment of human DNA. For simplicity, cloning of just a few representative fragments (colored) is shown. In reality, all of the gray DNA fragments would also be cloned.



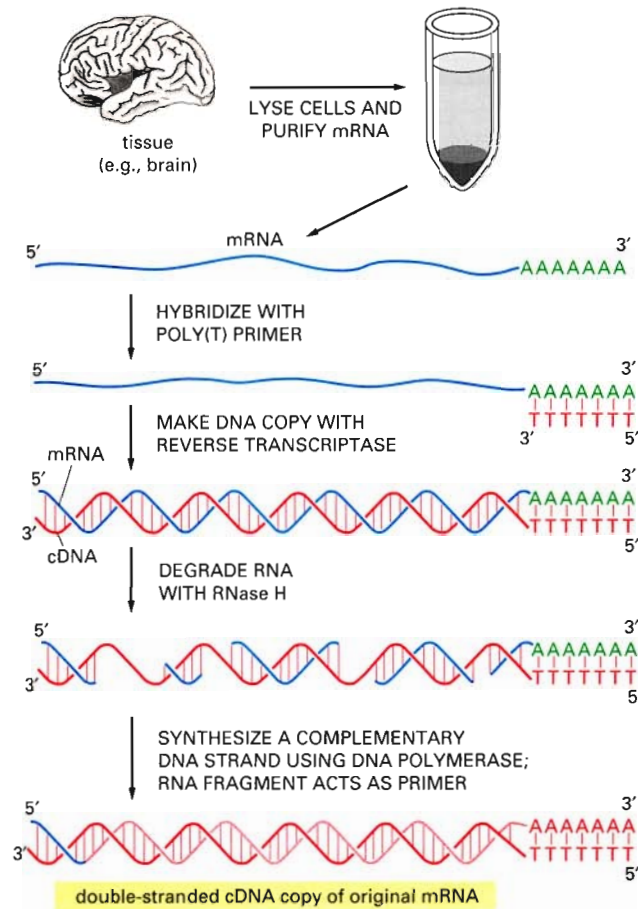


Figure 8–34 The synthesis of cDNA. Total mRNA is extracted from a particular tissue, and DNA copies (cDNA) of the mRNA molecules are produced by the enzyme reverse transcriptase (see p. 289). For simplicity, the copying of just one of these mRNAs into cDNA is illustrated. A short oligonucleotide complementary to the poly-A tail at the 3' end of the mRNA (discussed in Chapter 6) is first hybridized to the RNA to act as a primer for the reverse transcriptase, which then copies the RNA into a complementary DNA chain, thereby forming a DNA/RNA hybrid. Treating the DNA/RNA hybrid with RNase H (see Figure 5–13) creates nicks and gaps in the RNA strand. The remaining single-stranded cDNA is then copied into double-stranded cDNA by the enzyme DNA polymerase. The primer for this synthesis reaction is provided by a fragment of the original mRNA, as shown. Because the DNA polymerase used to synthesize the second DNA strand can synthesize through the bound RNA molecules, the RNA fragment that is base-paired to the 3' end of the first DNA strand usually acts as the primer for the final product of the second strand synthesis. This RNA is eventually degraded during subsequent cloning steps. As a result, the nucleotide sequences at the extreme 5' ends of the original mRNA molecules are often absent from cDNA libraries.

An alternative strategy is to begin the cloning process by selecting only those DNA sequences that are transcribed into mRNA and thus are presumed to correspond to protein-encoding genes. This is done by extracting the mRNA (or a purified subfraction of the mRNA) from cells and then making a complementary DNA (cDNA) copy of each mRNA molecule present; this reaction is catalyzed by the reverse transcriptase enzyme of retroviruses, which synthesizes a DNA chain on an RNA template. The single-stranded DNA molecules synthesized by the reverse transcriptase are converted into double-stranded DNA molecules by DNA polymerase, and these molecules are inserted into a plasmid or virus vector and cloned (Figure 8–34). Each clone obtained in this way is called a **cDNA clone**, and the entire collection of clones derived from one mRNA preparation constitutes a **cDNA library**.

There are important differences between genomic DNA clones and cDNA clones, as illustrated in Figure 8–35. Genomic clones represent a random sample of all of the DNA sequences in an organism and, with very rare exceptions, are the same regardless of the cell type used to prepare them. By contrast, cDNA clones contain only those regions of the genome that have been transcribed into mRNA. Because the cells of different tissues produce distinct sets of mRNA molecules, a distinct cDNA library is obtained for each type of cell used to prepare the library.

cDNA Clones Contain Uninterrupted Coding Sequences

The use of a cDNA library for gene cloning has several advantages. First, some proteins are produced in very large quantities by specialized cells. In this case, the mRNA encoding the protein is likely to be produced in such large quantities that a cDNA library prepared from the cells is highly enriched for the cDNA molecules encoding the protein, greatly reducing the problem of identifying the desired clone in the library (see Figure 8–35). Hemoglobin, for example, is made

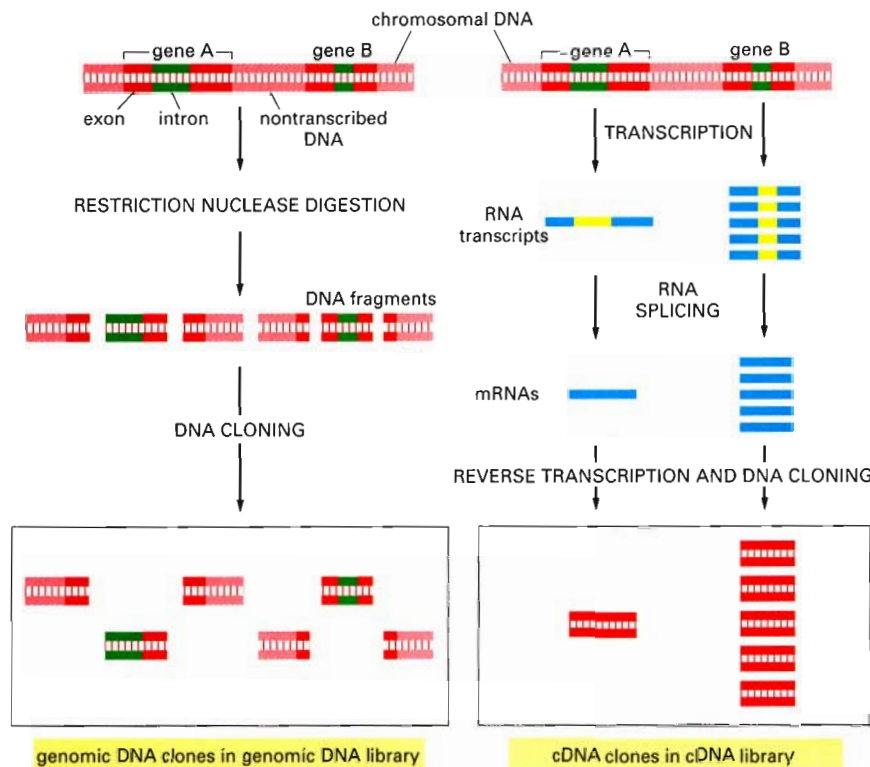


Figure 8-35 The differences between cDNA clones and genomic DNA clones derived from the same region of DNA. In this example gene A is infrequently transcribed, whereas gene B is frequently transcribed, and both genes contain introns (green). In the genomic DNA library, both the introns and the nontranscribed DNA (pink) are included in the clones, and most clones contain, at most, only part of the coding sequence of a gene (red). In the cDNA clones the intron sequences (yellow) have been removed by RNA splicing during the formation of the mRNA (blue), and a continuous coding sequence is therefore present in each clone. Because gene B is transcribed more abundantly than gene A in the cells from which the cDNA library was made, it is represented much more frequently than A in the cDNA library. In contrast, A and B are in principle represented equally in the genomic DNA library.

in large amounts by developing erythrocytes (red blood cells); for this reason the globin genes were among the first to be cloned.

By far the most important advantage of cDNA clones is that they contain the uninterrupted coding sequence of a gene. As we have seen, eucaryotic genes usually consist of short coding sequences of DNA (exons) separated by much longer noncoding sequences (introns); the production of mRNA entails the removal of the noncoding sequences from the initial RNA transcript and the splicing together of the coding sequences. Neither bacterial nor yeast cells will make these modifications to the RNA produced from a gene of a higher eucaryotic cell. Thus, when the aim of the cloning is either to deduce the amino acid sequence of the protein from the DNA sequence or to produce the protein in bulk by expressing the cloned gene in a bacterial or yeast cell, it is much preferable to start with cDNA.

Genomic and cDNA libraries are inexhaustible resources that are widely shared among investigators. Today, many such libraries are also available from commercial sources.

Isolated DNA Fragments Can Be Rapidly Sequenced

In the late 1970s methods were developed that allowed the nucleotide sequence of any purified DNA fragment to be determined simply and quickly. They have made it possible to determine the complete DNA sequences of tens of thousands of genes, and many organisms have had their DNA genomes fully sequenced (see Table 1-1, p. 20). The volume of DNA sequence information is now so large (many tens of billions of nucleotides) that powerful computers must be used to store and analyze it.

Large volume DNA sequencing was made possible through the development in the mid-1970s of the **dideoxy method** for sequencing DNA, which is based on *in vitro* DNA synthesis performed in the presence of chain-terminating dideoxyribonucleoside triphosphates (Figure 8-36).

Although the same basic method is still used today, many improvements have been made. DNA sequencing is now completely automated: robotic devices mix the reagents and then load, run, and read the order of the nucleotide

bases from the gel. This is facilitated by using chain-terminating nucleotides that are each labeled with a different colored fluorescent dye; in this case, all four synthesis reactions can be performed in the same tube, and the products can be separated in a single lane of a gel. A detector positioned near the bottom of the

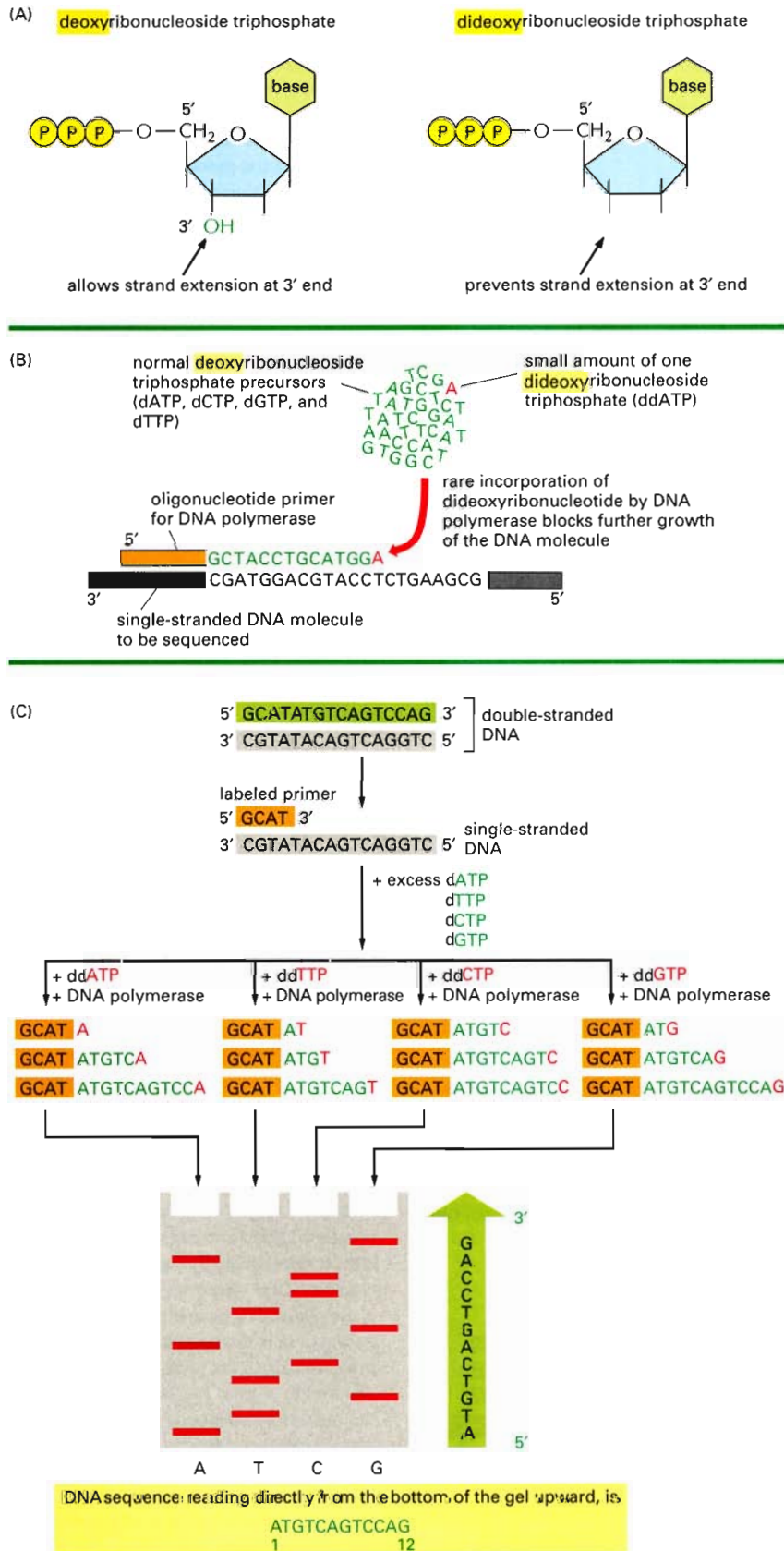


Figure 8–36 The enzymatic—or dideoxy—method of sequencing DNA. (A) This method relies on the use of dideoxyribonucleoside triphosphates, derivatives of the normal deoxyribonucleoside triphosphates that lack the 3' hydroxyl group. (B) Purified DNA is synthesized *in vitro* in a mixture that contains single-stranded molecules of the DNA to be sequenced (*gray*), the enzyme DNA polymerase, a short primer DNA (*orange*) to enable the polymerase to start DNA synthesis, and the four deoxyribonucleoside triphosphates (dATP, dCTP, dGTP, dTTP; *green* A, C, G, and T). If a dideoxyribonucleotide analog (*red*) of one of these nucleotides is also present in the nucleotide mixture, it can become incorporated into a growing DNA chain. Because this chain now lacks a 3' OH group, the addition of the next nucleotide is blocked, and the DNA chain terminates at that point. In the example illustrated, a small amount of dideoxyATP (ddATP, symbolized here as a *red* A) has been included in the nucleotide mixture. It competes with an excess of the normal deoxyATP (dATP, *green* A), so that ddATP is occasionally incorporated, at random, into a growing DNA strand. This reaction mixture will eventually produce a set of DNAs of different lengths complementary to the template DNA that is being sequenced and terminating at each of the different A's. The exact lengths of the DNA synthesis products can then be used to determine the position of each A in the growing chain. (C) To determine the complete sequence of a DNA fragment, the double-stranded DNA is first separated into its single strands and one of the strands is used as the template for sequencing. Four different chain-terminating dideoxyribonucleoside triphosphates (ddATP, ddCTP, ddGTP, ddTTP, again shown in *red*) are used in four separate DNA synthesis reactions on copies of the same single-stranded DNA template (*gray*). Each reaction produces a set of DNA copies that terminate at different points in the sequence. The products of these four reactions are separated by electrophoresis in four parallel lanes of a polyacrylamide gel (labeled here A, T, C, and G). The newly synthesized fragments are detected by a label (either radioactive or fluorescent) that has been incorporated either into the primer or into one of the deoxyribonucleoside triphosphates used to extend the DNA chain. In each lane, the bands represent fragments that have terminated at a given nucleotide (e.g., A in the leftmost lane) but at different positions in the DNA. By reading off the bands in order, starting at the bottom of the gel and working across all lanes, the DNA sequence of the newly synthesized strand can be determined. The sequence is given in the *green* arrow to the right of the gel. This sequence is identical to that of the 5' → 3' strand (*green*) of the original double-stranded DNA molecule.

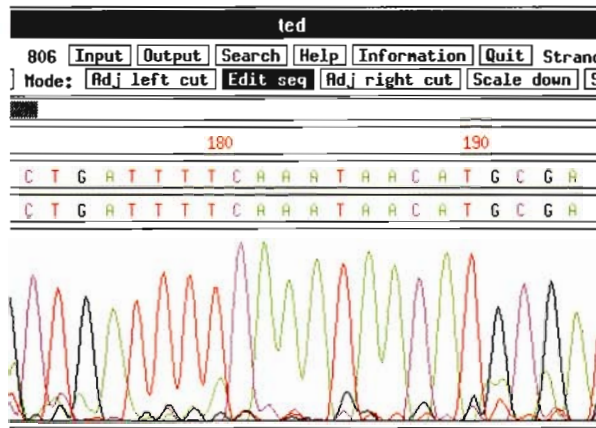


Figure 8-37 Automated DNA sequencing. Shown here is a tiny part of the data from an automated DNA-sequencing run as it appears on the computer screen. Each colored peak represents a nucleotide in the DNA sequence—a clear stretch of nucleotide sequence can be read here between positions 173 and 194 from the start of the sequence. This particular example is taken from the international project that determined the complete nucleotide sequence of the genome of the plant *Arabidopsis*. (Courtesy of George Murphy.)

gel reads and records the color of the fluorescent label on each band as it passes through a laser beam (Figure 8-37). A computer then reads and stores this nucleotide sequence.

Nucleotide Sequences Are Used to Predict the Amino Acid Sequences of Proteins

Now that DNA sequencing is so rapid and reliable, it has become the preferred method for determining, indirectly, the amino acid sequences of most proteins. Given a nucleotide sequence that encodes a protein, the procedure is quite straightforward. Although in principle there are six different reading frames in which a DNA sequence can be translated into protein (three on each strand), the correct one is generally recognizable as the only one lacking frequent stop codons (Figure 8-38). As we saw when we discussed the genetic code in Chapter 6, a random sequence of nucleotides, read in frame, will encode a stop signal for protein synthesis about once every 20 amino acids. Those nucleotide sequences that encode a stretch of amino acids much longer than this are candidates for presumptive exons, and they can be translated (by computer) into amino acid sequences and checked against databases for similarities to known proteins from other organisms. If necessary, a limited amount of amino acid sequence can then be determined from the purified protein to confirm the sequence predicted from the DNA.

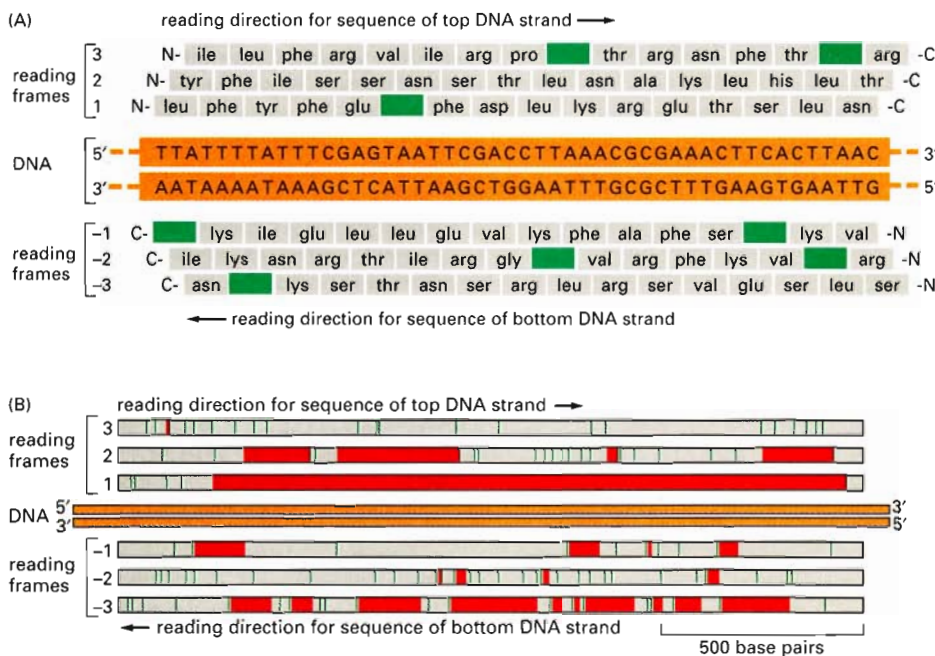


Figure 8-38 Finding the regions in a DNA sequence that encode a protein. (A) Any region of the DNA sequence can, in principle, code for six different amino acid sequences, because any one of three different reading frames can be used to interpret the nucleotide sequence on each strand. Note that a nucleotide sequence is always read in the 5'-to-3' chain direction and encodes a polypeptide from the amino (N) to the carboxyl (C) terminus. For a random nucleotide sequence read in a particular frame, a stop signal for protein synthesis is encountered, on average, about once every 21 amino acids (once every 63 nucleotides). In this sample sequence of 48 base pairs, each such signal (*stop codon*) is colored green, and only reading frame 2 lacks a stop signal. (B) Search of a 1700 base-pair DNA sequence for a possible protein-encoding sequence. The information is displayed as in (A), with each stop signal for protein synthesis denoted by a green line. In addition, all of the regions between possible start and stop signals for protein synthesis (see pp. 348-350) are displayed as red bars. Only reading frame 1 actually encodes a protein, which is 475 amino acid residues long.

The problem comes, however, in determining which nucleotide sequences—within a whole genome sequence—represent genes that encode proteins. Identifying genes is easiest when the DNA sequence is from a bacterial or archeal chromosome, which lacks introns, or from a cDNA clone. The location of genes in these nucleotide sequences can be predicted by examining the DNA for certain distinctive features (discussed in Chapter 6). Briefly these genes that encode proteins are identified by searching the nucleotide sequence for open reading frames (ORFs) that begin with an initiation codon, usually ATG, and end with a termination codon, TAA, TAG, or TGA. To minimize errors, computers used to search for ORFs are often directed to count as genes only those sequences that are longer than, say, 100 codons in length.

For more complex genomes, such as those of eucaryotes, the process is complicated by the presence of large introns embedded within the coding portion of genes. In many multicellular organisms, including humans, the average exon is only 150 nucleotides long. Thus in eucaryotes, one must also search for other features that signal the presence of a gene, for example, sequences that signal an intron/exon boundary or distinctive upstream regulatory regions.

A second major approach to identifying the coding regions in chromosomes is through the characterization of the nucleotide sequences of the detectable mRNAs (in the form of cDNAs). The mRNAs (and the cDNAs produced from them) lack introns, regulatory DNA sequences, and the nonessential “spacer” DNA that lies between genes. It is therefore useful to sequence large numbers of cDNAs to produce a very large collection (called a database) of the coding sequences of an organism. These sequences are then readily used to distinguish the exons from the introns in the long chromosomal DNA sequences that correspond to genes.

Finally, nucleotide sequences that are conserved between closely related organisms usually encode proteins. Comparison of these conserved sequences in different species can also provide insight into the function of a particular protein or gene, as we see later in the chapter.

The Genomes of Many Organisms Have Been Fully Sequenced

Owing in large part to the automation of DNA sequencing, the genomes of many organisms have been fully sequenced; these include plant chloroplasts and animal mitochondria, large numbers of bacteria and archaea, and many of the model organisms that are studied routinely in the laboratory, including several yeasts, a nematode worm, the fruit fly *Drosophila*, the model plant *Arabidopsis*, the mouse, and, last but not least, humans. Researchers have also deduced the complete DNA sequences for a wide variety of human pathogens. These include the bacteria that cause cholera, tuberculosis, syphilis, gonorrhea, Lyme disease, and stomach ulcers, as well as hundreds of viruses—including smallpox virus and Epstein–Barr virus (which causes infectious mononucleosis). Examination of the genomes of these pathogens should provide clues about what makes them virulent, and will also point the way to new and more effective treatments.

Haemophilus influenzae (a bacterium that can cause ear infections or meningitis in children) was the first organism to have its complete genome sequence—all 1.8 million nucleotides—determined by the shotgun sequencing method, the most common strategy used today. In the shotgun method, long sequences of DNA are broken apart randomly into many shorter fragments. Each fragment is then sequenced and a computer is used to order these pieces into a whole chromosome or genome, using sequence overlap to guide the assembly. The shotgun method is the technique of choice for sequencing small genomes. Although larger, more repetitive genome sequences are more tricky to assemble, the shotgun method has been useful for sequencing the genomes of *Drosophila melanogaster*, mouse, and human.

With new sequences appearing at a steadily accelerating pace in the scientific literature, comparison of the complete genome sequences of different organisms allows us to trace the evolutionary relationships among genes and

organisms, and to discover genes and predict their functions. Assigning functions to genes often involves comparing their sequences with related sequences from model organisms that have been well characterized in the laboratory, such as the bacterium *E. coli*, the yeasts *S. cerevisiae* and *S. pombe*, the nematode worm *C. elegans*, and the fruit fly *Drosophila* (discussed in Chapter 1).

Although the organisms whose genomes have been sequenced share many cellular pathways and possess many proteins that are homologous in their amino acid sequences or structure, the functions of a very large number of newly identified proteins remain unknown. Some 15–40% of the proteins encoded by these sequenced genomes do not resemble any other protein that has been characterized functionally. This observation underscores one of the limitations of the emerging field of genomics: although comparative analysis of genomes reveals a great deal of information about the relationships between genes and organisms, it often does not provide immediate information about how these genes function, or what roles they have in the physiology of an organism. Comparison of the full gene complement of several thermophilic bacteria, for example, does not reveal why these bacteria thrive at temperatures exceeding 70°C. And examination of the genome of the incredibly radioresistant bacterium *Deinococcus radiodurans* does not explain how this organism can survive a blast of radiation that can shatter glass. Further biochemical and genetic studies, like those described in the final sections of this chapter, are required to determine how genes function in the context of living organisms.

Selected DNA Segments Can Be Cloned in a Test Tube by a Polymerase Chain Reaction

Now that so many genome sequences are available, genes can be cloned directly without the need to construct DNA libraries first. A technique called the **polymerase chain reaction (PCR)** makes this rapid cloning possible. PCR allows the DNA from a selected region of a genome to be amplified a billionfold, effectively “purifying” this DNA away from the remainder of the genome.

Two sets of DNA oligonucleotides, chosen to flank the desired nucleotide sequence of the gene, are synthesized by chemical methods. These oligonucleotides are then used to prime DNA synthesis on single strands generated by heating the DNA from the entire genome. The newly synthesized DNA is produced in a reaction catalyzed *in vitro* by a purified DNA polymerase, and the primers remain at the 5′ ends of the final DNA fragments that are made (Figure 8–39A).

Nothing special is produced in the first cycle of DNA synthesis; the power of the PCR method is revealed only after repeated rounds of DNA synthesis. Every cycle doubles the amount of DNA synthesized in the previous cycle. Because each cycle requires a brief heat treatment to separate the two strands of the template DNA double helix, the technique requires the use of a special DNA polymerase, isolated from a thermophilic bacterium, that is stable at much higher temperatures than normal, so that it is not denatured by the repeated heat treatments. With each round of DNA synthesis, the newly generated fragments serve as templates in their turn, and within a few cycles the predominant product is a single species of DNA fragment whose length corresponds to the distance between the two original primers (see Figure 8–39B).

In practice, 20–30 cycles of reaction are required for effective DNA amplification, with the products of each cycle serving as the DNA templates for the next—hence the term polymerase “chain reaction.” A single cycle requires only about 5 minutes, and the entire procedure can be easily automated. PCR thereby makes possible the “cell-free molecular cloning” of a DNA fragment in a few hours, compared with the several days required for standard cloning procedures. This technique is now used routinely to clone DNA from genes of interest directly—starting either from genomic DNA or from mRNA isolated from cells (Figure 8–40).

The PCR method is extremely sensitive; it can detect a single DNA molecule in a sample. Trace amounts of RNA can be analyzed in the same way by first

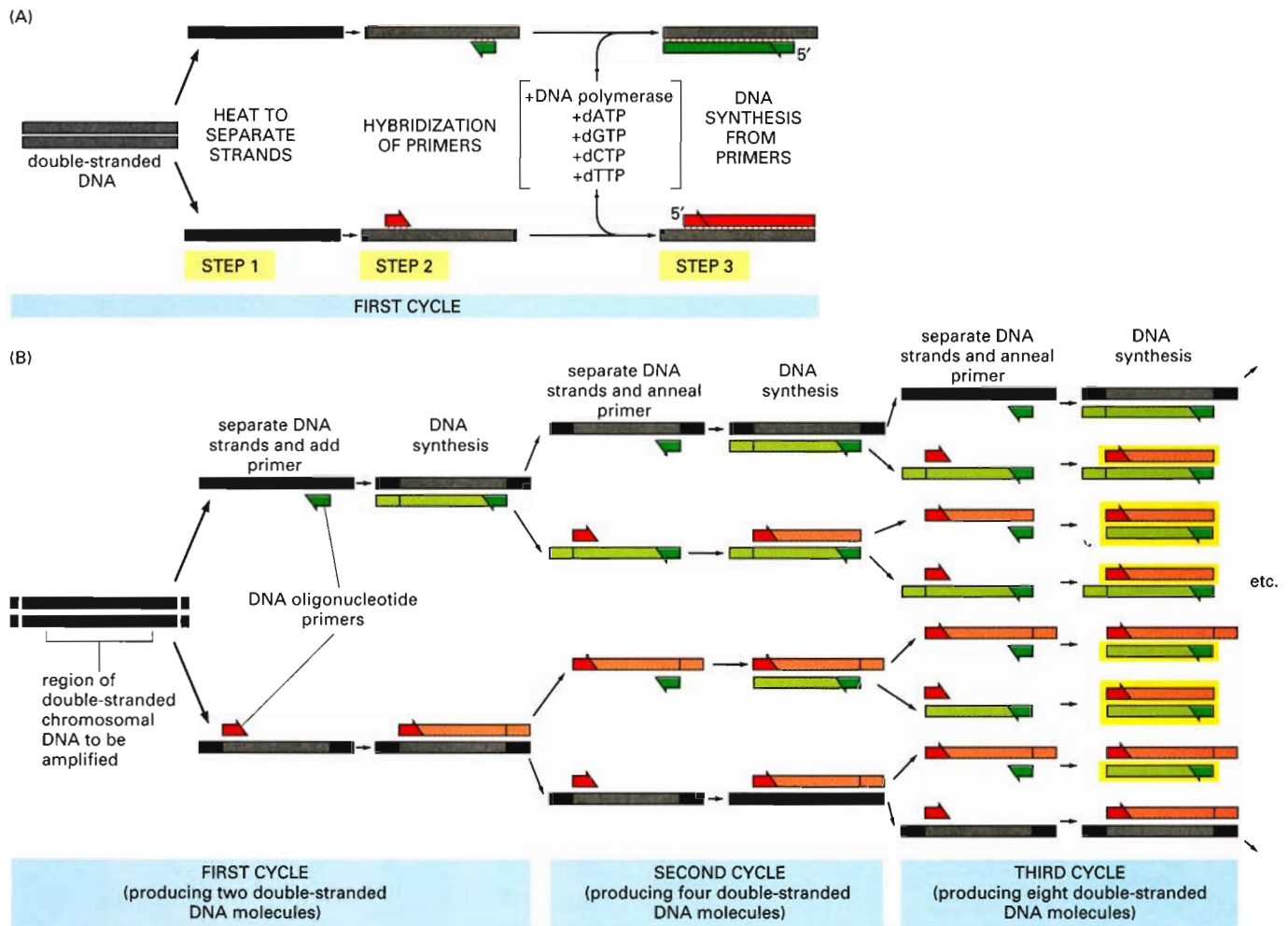


Figure 8–39 Amplification of DNA using the PCR technique. Knowledge of the DNA sequence to be amplified is used to design two synthetic DNA oligonucleotides, each complementary to the sequence on one strand of the DNA double helix at opposite ends of the region to be amplified. These oligonucleotides serve as primers for *in vitro* DNA synthesis, which is performed by a DNA polymerase, and they determine the segment of the DNA that is amplified. (A) PCR starts with a double-stranded DNA, and each cycle of the reaction begins with a brief heat treatment to separate the two strands (step 1). After strand separation, cooling of the DNA in the presence of a large excess of the two primer DNA oligonucleotides allows these primers to hybridize to complementary sequences in the two DNA strands (step 2). This mixture is then incubated with DNA polymerase and the four deoxyribonucleoside triphosphates so that DNA is synthesized, starting from the two primers (step 3). The entire cycle is then begun again by a heat treatment to separate the newly synthesized DNA strands. (B) As the procedure is performed over and over again, the newly synthesized fragments serve as templates in their turn, and within a few cycles the predominant DNA is identical to the sequence bracketed by and including the two primers in the original template. Of the DNA put into the original reaction, only the sequence bracketed by the two primers is amplified because there are no primers attached anywhere else. In the example illustrated in (B), three cycles of reaction produce 16 DNA chains, 8 of which (boxed in yellow) are the same length as and correspond exactly to one or the other strand of the original bracketed sequence shown at the far left; the other strands contain extra DNA downstream of the original sequence, which is replicated in the first few cycles. After three more cycles, 240 of the 256 DNA chains correspond exactly to the original bracketed sequence, and after several more cycles, essentially all of the DNA strands have this unique length.

transcribing them into DNA with reverse transcriptase. The PCR cloning technique has largely replaced Southern blotting for the diagnosis of genetic diseases and for the detection of low levels of viral infection. It also has great promise in forensic medicine as a means of analyzing minute traces of blood or other tissues—even as little as a single cell—and identifying the person from whom they came by his or her genetic “fingerprint” (Figure 8–41).

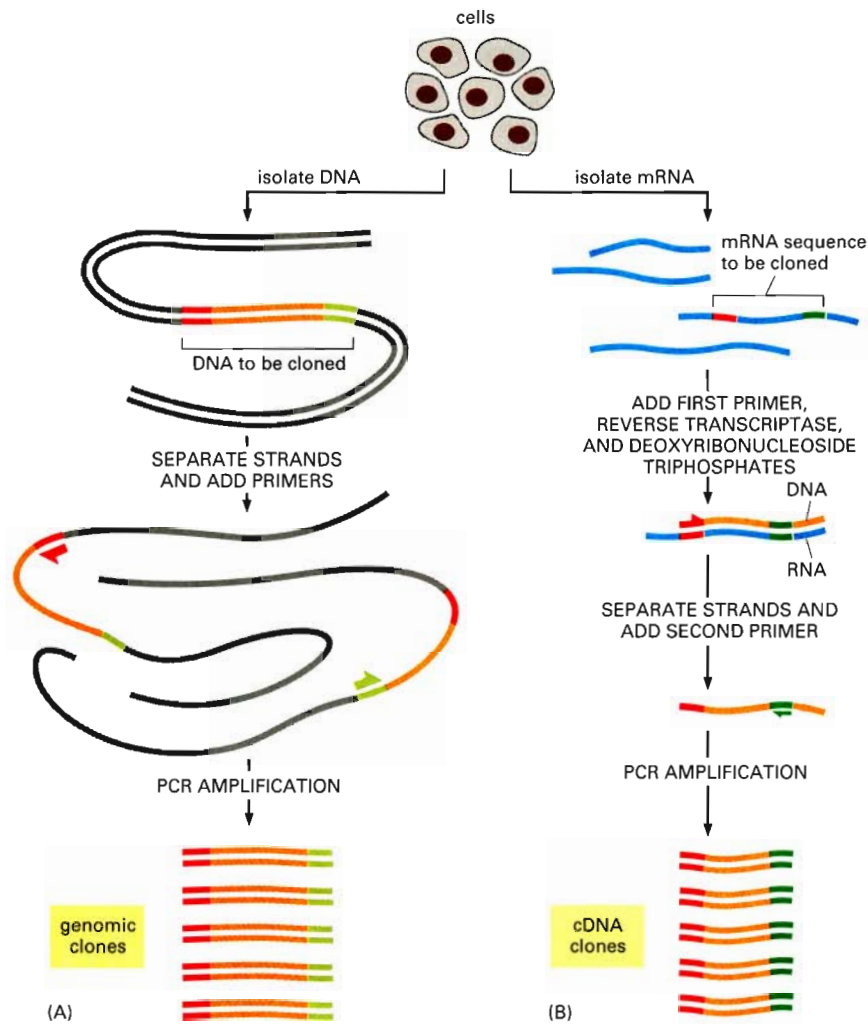


Figure 8–40 Use of PCR to obtain a genomic or cDNA clone.

(A) To obtain a genomic clone by using PCR, chromosomal DNA is first purified from cells. PCR primers that flank the stretch of DNA to be cloned are added, and many cycles of the reaction are completed (see Figure 8–39). Since only the DNA between (and including) the primers is amplified, PCR provides a way to obtain a short stretch of chromosomal DNA selectively in a pure form. (B) To use PCR to obtain a cDNA clone of a gene, mRNA is first purified from cells. The first primer is then added to the population of mRNAs, and reverse transcriptase is used to make a complementary DNA strand. The second primer is then added, and the single-stranded DNA molecule is amplified through many cycles of PCR, as shown in Figure 8–39. For both types of cloning, the nucleotide sequence of at least part of the region to be cloned must be known beforehand.

Cellular Proteins Can Be Made in Large Amounts Through the Use of Expression Vectors

Fifteen years ago, the only proteins in a cell that could be studied easily were the relatively abundant ones. Starting with several hundred grams of cells, a major protein—one that constitutes 1% or more of the total cellular protein—can be purified by sequential chromatography steps to yield perhaps 0.1 g (100 mg) of pure protein. This amount was sufficient for conventional amino acid sequencing, for detailed analysis of biochemical activities, and for the production of antibodies, which could then be used to localize the protein in the cell. Moreover, if suitable crystals could be grown (often a difficult task), the three-dimensional structure of the protein could be determined by x-ray diffraction techniques, as we will discuss later. The structure and function of many abundant

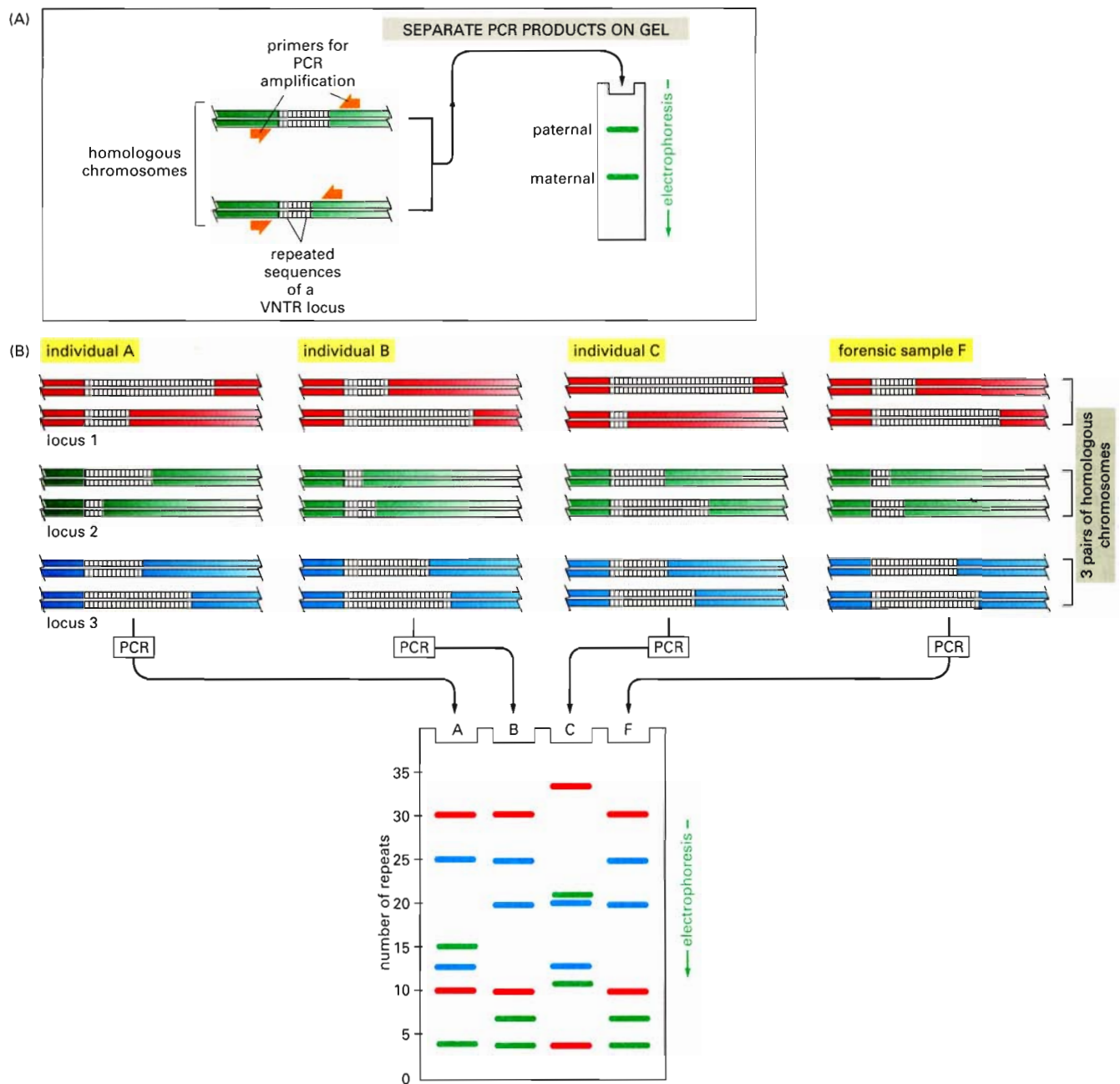
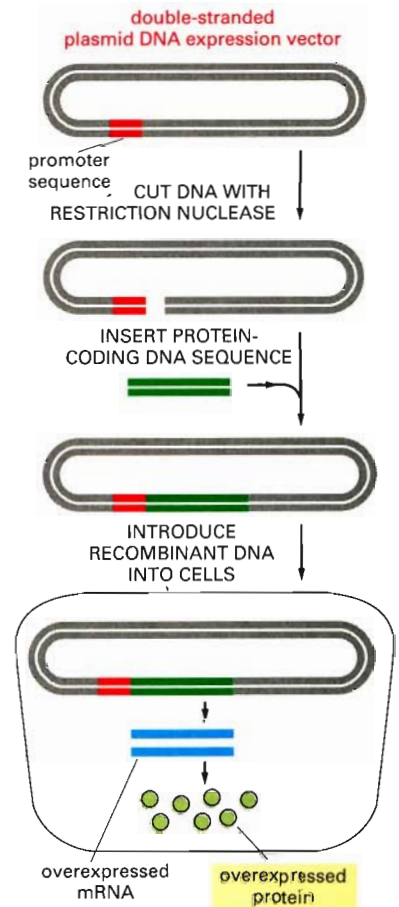


Figure 8–41 How PCR is used in forensic science. (A) The DNA sequences that create the variability used in this analysis contain runs of short, repeated sequences, such as CACACA . . . , which are found in various positions (loci) in the human genome. The number of repeats in each run can be highly variable in the population, ranging from 4 to 40 in different individuals. A run of repeated nucleotides of this type is commonly referred to as a *hypervariable microsatellite* sequence—also known as a VNTR (*variable number of tandem repeat*) sequence. Because of the variability in these sequences at each locus, individuals usually inherit a different variant from their mother and from their father; two unrelated individuals therefore do not usually contain the same pair of sequences. A PCR analysis using primers that bracket the locus produces a pair of bands of amplified DNA from each individual, one band representing the maternal variant and the other representing the paternal variant. The length of the amplified DNA, and thus the position of the band it produces after electrophoresis, depends on the exact number of repeats at the locus. (B) In the schematic example shown here, the same three VNTR loci are analyzed (requiring three different pairs of specially selected oligonucleotide primers) from three suspects (individuals A, B, and C), producing six DNA bands for each person after polyacrylamide gel electrophoresis. Although some individuals have several bands in common, the overall pattern is quite distinctive for each. The band pattern can therefore serve as a “fingerprint” to identify an individual nearly uniquely. The fourth lane (F) contains the products of the same reactions carried out on a forensic sample. The starting material for such a PCR can be a single hair or a tiny sample of blood that was left at the crime scene. When examining the variability at 5 to 10 different VNTR loci, the odds that two random individuals would share the same genetic pattern by chance can be approximately one in 10 billion. In the case shown here, individuals A and C can be eliminated from further enquiries, whereas individual B remains a clear suspect for committing the crime. A similar approach is now routinely used for paternity testing.

Figure 8–42 Production of large amounts of a protein from a protein-coding DNA sequence cloned into an expression vector and introduced into cells. A plasmid vector has been engineered to contain a highly active promoter, which causes unusually large amounts of mRNA to be produced from an adjacent protein-coding gene inserted into the plasmid vector. Depending on the characteristics of the cloning vector, the plasmid is introduced into bacterial, yeast, insect, or mammalian cells, where the inserted gene is efficiently transcribed and translated into protein.



proteins—including hemoglobin, trypsin, immunoglobulin, and lysozyme—were analyzed in this way.

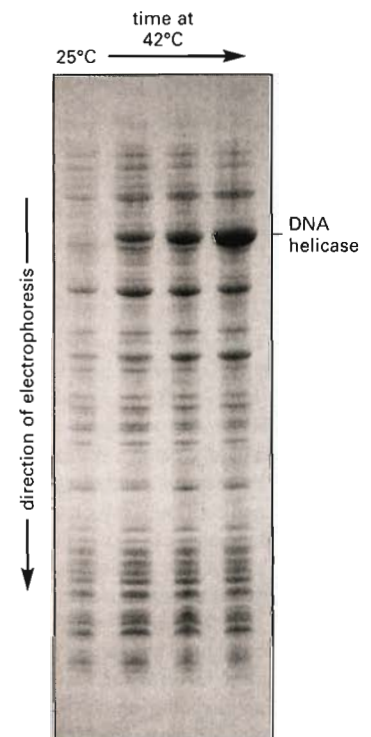
The vast majority of the thousands of different proteins in a eucaryotic cell, however, including many with crucially important functions, are present in very small amounts. For most of them it is extremely difficult, if not impossible, to obtain more than a few micrograms of pure material. One of the most important contributions of DNA cloning and genetic engineering to cell biology is that they have made it possible to produce any of the cell's proteins in nearly unlimited amounts.

Large amounts of a desired protein are produced in living cells by using **expression vectors** (Figure 8–42). These are generally plasmids that have been designed to produce a large amount of a stable mRNA that can be efficiently translated into protein in the transfected bacterial, yeast, insect, or mammalian cell. To prevent the high level of the foreign protein from interfering with the transfected cell's growth, the expression vector is often designed so that the synthesis of the foreign mRNA and protein can be delayed until shortly before the cells are harvested (Figure 8–43).

Because the desired protein made from an expression vector is produced inside a cell, it must be purified away from the host cell proteins by chromatography following cell lysis; but because it is such a plentiful species in the cell lysate (often 1–10% of the total cell protein), the purification is usually easy to accomplish in only a few steps. Many expression vectors have been designed to add a molecular tag—a cluster of histidine residues or a small marker protein—to the expressed protein to make possible easy purification by affinity chromatography, as discussed previously (see pp. 483–484). A variety of expression vectors are available, each engineered to function in the type of cell in which the protein is to be made. In this way cells can be induced to make vast quantities of medically useful proteins—such as human insulin and growth hormone, interferon, and viral antigens for vaccines. More generally, these methods make it possible to produce every protein—even those that may be present in only a few copies per cell—in large enough amounts to be used in the kinds of detailed structural and functional studies that we discuss in the next section (Figure 8–44).

DNA technology can also be used to produce large amounts of any RNA molecule whose gene has been isolated. Studies of RNA splicing, protein synthesis, and RNA-based enzymes, for example, are greatly facilitated by the availability of pure RNA molecules. Most RNAs are present in only tiny quantities in cells, and they are very difficult to purify away from other cellular components—especially from the many thousands of other RNAs present in the cell. But any RNA of interest can be synthesized efficiently *in vitro* by transcription of its DNA sequence with a highly efficient viral RNA polymerase. The single species of RNA produced is then easily purified away from the DNA template and the RNA polymerase.

Figure 8–43 Production of large amounts of a protein by using a plasmid expression vector. In this example, bacterial cells have been transfected with the coding sequence for an enzyme, DNA helicase; transcription from this coding sequence is under the control of a viral promoter that becomes active only at temperatures of 37°C or higher. The total cell protein has been analyzed by SDS-polyacrylamide gel electrophoresis, either from bacteria grown at 25°C (no helicase protein made), or after a shift of the same bacteria to 42°C for up to 2 hours (helicase protein has become the most abundant protein species in the lysate). (Courtesy of Jack Barry.)



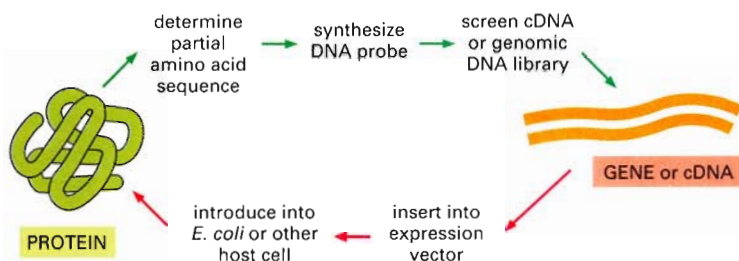


Figure 8–44 Knowledge of the molecular biology of cells makes it possible to experimentally move from gene to protein and from protein to gene. A small quantity of a purified protein is used to obtain a partial amino acid sequence. This provides sequence information that enables the corresponding gene to be cloned from a DNA library. Once the gene has been cloned, its protein-coding sequence can be inserted into an expression vector and used to produce large quantities of the protein from genetically engineered cells.

Summary

DNA cloning allows a copy of any specific part of a DNA or RNA sequence to be selected from the millions of other sequences in a cell and produced in unlimited amounts in pure form. DNA sequences can be amplified after cutting chromosomal DNA with a restriction nuclease and inserting the resulting DNA fragments into the chromosome of a self-replicating genetic element. Plasmid vectors are generally used and the resulting “genomic DNA library” is housed in millions of bacterial cells, each carrying a different cloned DNA fragment. Individual cells that are allowed to proliferate produce large amounts of a single cloned DNA fragment from this library. As an alternative, the polymerase chain reaction (PCR) allows DNA cloning to be performed directly with a purified, thermostable DNA polymerase—providing that the DNA sequence of interest is already known.

The procedures used to obtain DNA clones that correspond in sequence to mRNA molecules are the same except that a DNA copy of the mRNA sequence, called cDNA, is first made. Unlike genomic DNA clones, cDNA clones lack intron sequences, making them the clones of choice for analyzing the protein product of a gene.

Nucleic acid hybridization reactions provide a sensitive means of detecting a gene or any other nucleotide sequence of choice. Under stringent hybridization conditions (a combination of solvent and temperature where a perfect double helix is barely stable), two strands can pair to form a “hybrid” helix only if their nucleotide sequences are almost perfectly complementary. The enormous specificity of this hybridization reaction allows any single-stranded sequence of nucleotides to be labeled with a radioisotope or chemical and used as a probe to find a complementary partner strand, even in a cell or cell extract that contains millions of different DNA and RNA sequences. Probes of this type are widely used to detect the nucleic acids corresponding to specific genes, both to facilitate their purification and characterization and to localize them in cells, tissues, and organisms.

The nucleotide sequence of purified DNA fragments can be determined rapidly and simply by using highly automated techniques based on the dideoxy method for sequencing DNA. This technique has made it possible to determine the complete DNA sequences of tens of thousands of genes and to completely sequence the genomes of many organisms. Comparison of the genome sequences of different organisms allows us to trace the evolutionary relationships among genes and organisms, and it has proved valuable for discovering new genes and predicting their function.

Taken together, these techniques have made it possible to identify, isolate, and sequence genes from any organism of interest. Related technologies allow scientists to produce the protein products of these genes in the large quantities needed for detailed analyses of their structure and function, as well as for medical purposes.

ANALYZING PROTEIN STRUCTURE AND FUNCTION

Proteins perform most of the work of living cells. This versatile class of macromolecule is involved in virtually every cellular process: proteins replicate and transcribe DNA, and produce, process, and secrete other proteins. They control cell division, metabolism, and the flow of materials and information into and out of the cell. Understanding how cells work requires understanding how proteins function.