component would be too small to be useful. But if the B lymphocytes that produce the various components of this antiserum are made into hybridomas, it becomes possible to screen individual hybridoma clones from the large mixture to select one that produces the desired type of monoclonal antibody and to propagate the selected hybridoma indefinitely so as to produce that antibody in unlimited quantities. In principle, therefore, a monoclonal antibody can be made against any protein in a biological sample.

Once an antibody has been made, it can be used as a specific probe—both to track down and localize its protein antigen and to purify that protein in order to study its structure and function. Because only a small fraction of the estimated 10,000 –20,000 proteins in a typical mammalian cell have thus far been isolated, many monoclonal antibodies made against impure protein mixtures in fractionated cell extracts identify new proteins. With the use of monoclonal antibodies and the rapid protein identification methods we shall describe shortly, it is no longer difficult to identify and characterize novel proteins and genes. The major problem is instead to determine their function, using a set of powerful tools that we discuss in the last sections of this chapter.

Summary

Tissues can be dissociated into their component cells, from which individual cell types can be purified and used for biochemical analysis or for the establishment of cell cultures. Many animal and plant cells survive and proliferate in a culture dish if they are provided with a suitable medium containing nutrients and specific protein growth factors. Although most animal cells die after a finite number of divisions, immortal cells that arise spontaneously in culture—or are generated by adding genes through genetic manipulation—can be maintained indefinitely as cell lines. Clones can be derived from a single ancestor cell, making it possible to isolate uniform populations of mutant cells with defects in a single protein. Two cells can be fused to produce heterocaryons with two nuclei, enabling interactions between the components of the original two cells to be examined. Heterocaryons eventually form hybrid cells with a single fused nucleus. Because such cells lose chromosomes, they can provide a convenient method for assigning genes to specific chromosomes. One type of hybrid cell, called a hybridoma, is widely employed to produce unlimited quantities of uniform monoclonal antibodies, which are widely used to detect and purify cellular proteins.

FRACTIONATION OF CELLS

Although biochemical analysis requires disruption of the anatomy of the cell, gentle fractionation techniques have been devised to separate the various cell components while preserving their individual functions. Just as a tissue can be separated into its living constituent cell types, so the cell can be separated into its functioning organelles and macromolecules. In this section we consider the methods that allow organelles and proteins to be purified and analyzed biochemically.

Organelles and Macromolecules Can Be Separated by Ultracentrifugation

Cells can be broken up in various ways: they can be subjected to osmotic shock or ultrasonic vibration, forced through a small orifice, or ground up in a blender. These procedures break many of the membranes of the cell (including the plasma membrane and membranes of the endoplasmic reticulum) into fragments that immediately reseal to form small closed vesicles. If carefully applied, however, the disruption procedures leave organelles such as nuclei, mitochondria, the Golgi apparatus, lysosomes, and peroxisomes largely intact. The suspension of cells is thereby reduced to a thick slurry (called a *homogenate* or *extract)* that contains a variety of membrane-enclosed organelles, each with a distinctive

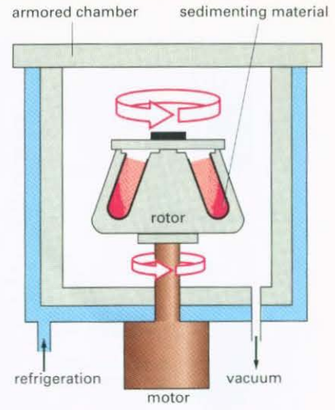


Figure 8–7 The preparative ultracentrifuge. The sample is contained in tubes that are inserted into a ring of cylindrical holes in a metal *rotor*. Rapid rotation of the rotor generates enormous centrifugal forces, which cause particles in the sample to sediment. The vacuum reduces friction, preventing heating of the rotor and allowing the refrigeration system to maintain the sample at 4°C.

size, charge, and density. Provided that the homogenization medium has been carefully chosen (by trial and error for each organelle), the various components—including the vesicles derived from the endoplasmic reticulum, called microsomes—retain most of their original biochemical properties.

The different components of the homogenate must then be separated. Such cell fractionations became possible only after the commercial development in the early 1940s of an instrument known as the *preparative ultracentrifuge*, in which extracts of broken cells are rotated at high speeds (Figure 8–7). This treatment separates cell components by size and density: in general, the largest units experience the largest centrifugal force and move the most rapidly. At relatively low speed, large components such as nuclei sediment to form a pellet at the bottom of the centrifuge tube; at slightly higher speed, a pellet of mitochondria is deposited; and at even higher speeds and with longer periods of centrifugation, first the small closed vesicles and then the ribosomes can be collected (Figure 8–8). All of these fractions are impure, but many of the contaminants can be removed by resuspending the pellet and repeating the centrifugation procedure several times.

Centrifugation is the first step in most fractionations, but it separates only components that differ greatly in size. A finer degree of separation can be achieved by layering the homogenate in a thin band on top of a dilute salt solution that fills a centrifuge tube. When centrifuged, the various components in the mixture move as a series of distinct bands through the salt solution, each at a different rate, in a process called *velocity sedimentation* (Figure 8–9A). For the procedure to work effectively, the bands must be protected from convective mixing, which would normally occur whenever a denser solution (for example, one containing organelles) finds itself on top of a lighter one (the salt solution). This is achieved by filling the centrifuge tube with a shallow gradient of sucrose prepared by a special mixing device. The resulting density gradient—with the dense end at the bottom of the tube—keeps each region of the salt solution denser than any solution above it, and it thereby prevents convective mixing from distorting the separation.

When sedimented through such dilute sucrose gradients, different cell components separate into distinct bands that can be collected individually. The relative rate at which each component sediments depends primarily on its size and shape—being normally described in terms of its sedimentation coefficient, or s value. Present-day ultracentrifuges rotate at speeds of up to 80,000 rpm and produce forces as high as 500,000 times gravity. With these enormous forces, even small macromolecules, such as tRNA molecules and simple enzymes, can be driven to sediment at an appreciable rate and so can be separated from one another by size. Measurements of sedimentation coefficients are routinely used to help in determining the size and subunit composition of the organized assemblies of macromolecules found in cells.

The ultracentrifuge is also used to separate cellular components on the basis of their buoyant density, independently of their size and shape. In this case the

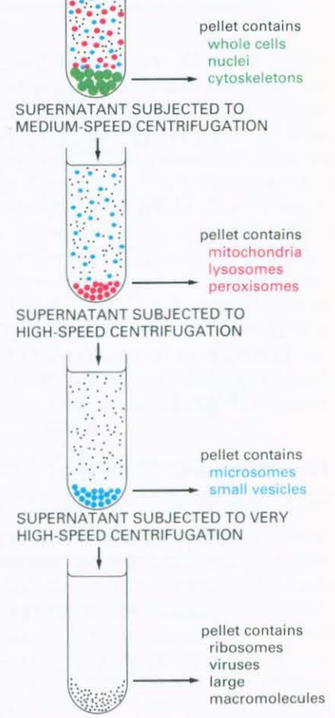


Figure 8–8 Cell fractionation by centrifugation. Repeated centrifugation at progressively higher speeds will fractionate homogenates of cells into their components. In general, the smaller the subcellular component, the greater is the centrifugal force required to sediment it. Typical values for the various centrifugation steps referred to in the figure are:

low speed	1000 times gravity for 10 minutes
medium speed	20,000 times gravity for 20 minutes
high speed	80,000 times gravity for 1 hour
very high speed	150,000 times gravity for 3 hours

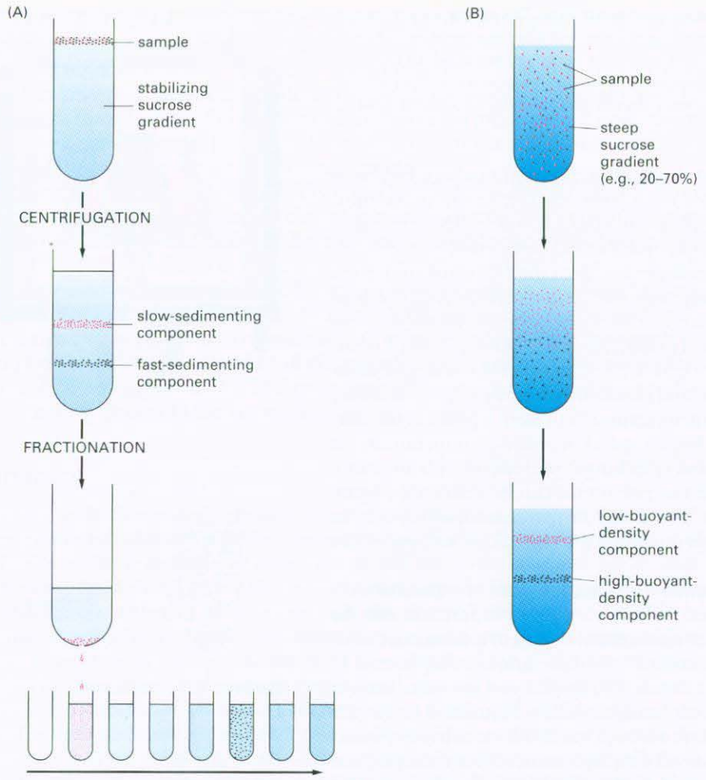


Figure 8–9 Comparison of velocity sedimentation and equilibrium sedimentation. In velocity sedimentation (A) subcellular components sediment at different speeds according to their size and shape when layered over a dilute solution containing sucrose. To stabilize the sedimenting bands against convective mixing caused by small differences in temperature or solute concentration, the tube contains a continuous shallow gradient of sucrose that increases in concentration toward the bottom of the tube (typically from 5% to 20% sucrose). Following centrifugation, the different components can be collected individually, most simply by puncturing the plastic centrifuge tube and collecting drops from the bottom, as illustrated here. In equilibrium sedimentation (B) subcellular components move up or down when centrifuged in a gradient until they reach a position where their density matches their surroundings. Although a sucrose gradient is shown here, denser gradients, which are especially useful for protein and nucleic acid separation, can be formed from cesium chloride. The final bands, at equilibrium, can be collected as in (A).

sample is usually sedimented through a steep density gradient that contains a very high concentration of sucrose or cesium chloride. Each cellular component begins to move down the gradient as in Figure 8–9A, but it eventually reaches a position where the density of the solution is equal to its own density. At this point the component floats and can move no farther. A series of distinct bands is thereby produced in the centrifuge tube, with the bands closest to the bottom of the tube containing the components of highest buoyant density (Figure 8–9B). This method, called *equilibrium sedimentation*, is so sensitive that it is capable of separating macromolecules that have incorporated heavy isotopes, such as ^{13}C or ^{15}N, from the same macromolecules that contain the lighter, common isotopes (^{12}C or ^{14}N). In fact, the cesium-chloride method was developed in 1957 to separate the labeled from the unlabeled DNA produced after exposure of a growing population of bacteria to nucleotide precursors containing ^{15}N; this classic experiment provided direct evidence for the semiconservative replication of DNA (see Figure 5–5).

The Molecular Details of Complex Cellular Processes Can Be Deciphered in Cell-Free Systems

Studies of organelles and other large subcellular components isolated in the ultracentrifuge have contributed enormously to our understanding of the functions of different cellular components. Experiments on mitochondria and chloroplasts purified by centrifugation, for example, demonstrated the central function of these organelles in converting energy into forms that the cell can use. Similarly, resealed vesicles formed from fragments of rough and smooth endoplasmic reticulum (microsomes) have been separated from each other and analyzed as functional models of these compartments of the intact cell.

An extension of this approach makes it possible to study many other biological processes free from all of the complex side reactions that occur in a living cell, by using purified **cell-free systems**. In this case, cell homogenates are fractionated with the aim of purifying each of the individual macromolecules that are needed to catalyze a biological process of interest. For example, the mechanisms of protein synthesis were deciphered in experiments that began with a cell homogenate that could translate RNA molecules to produce proteins. Fractionation of this homogenate, step by step, produced in turn the ribosomes, tRNAs, and various enzymes that together constitute the protein-synthetic machinery. Once individual pure components were available, each could be added or withheld separately to define its exact role in the overall process. A major goal today is the reconstitution of every biological process in a purified cell-free system, so as to be able to define all of its components and their mechanism of action. Some landmarks in the development of this critical approach for understanding the cell are listed in Table 8–4.

Much of what we know about the molecular biology of the cell has been discovered by studying cell-free systems. As a few of many examples, they have been used to decipher the molecular details of DNA replication and DNA transcription, RNA splicing, protein translation, muscle contraction, and particle transport along microtubules. Cell-free systems have even been used to study such complex and highly organized processes as the cell-division cycle, the separation of chromosomes on the mitotic spindle, and the vesicular-transport steps involved in the movement of proteins from the endoplasmic reticulum through the Golgi apparatus to the plasma membrane.

Cell homogenates also provide, in principle, the starting material for the complete separation of all of the individual macromolecular components from the cell. We now consider how this separation is achieved, focusing on proteins.

Proteins Can Be Separated by Chromatography

Proteins are most often fractionated by **column chromatography**, in which a mixture of proteins in solution is passed through a column containing a porous solid matrix. The different proteins are retarded to different extents by their

TABLE 8–4 Some Major Events in the Development of Cell-Free Systems

1897	**Buchner** shows that cell-free extracts of yeast can ferment sugars to form carbon dioxide and ethanol, laying the foundations of enzymology.
1926	**Svedberg** develops the first analytical ultracentrifuge and uses it to estimate the mass of hemoglobin as 68,000 daltons.
1935	**Pickels and Beams** introduce several new features of centrifuge design that lead to its use as a preparative instrument.
1938	**Behrens** employs differential centrifugation to separate nuclei and cytoplasm from liver cells, a technique further developed for the fractionation of cell organelles by **Claude**, **Brachet**, **Hogeboom**, and others in the 1940s and early 1950s.
1939	**Hill** shows that isolated chloroplasts, when illuminated, can perform the reactions of photosynthesis.
1949	**Szent-Györgyi** shows that isolated myofibrils from skeletal muscle cells contract upon the addition of ATP. In 1955 a similar cell-free system was developed for ciliary beating by **Hofmann-Berling**.
1951	**Brakke** uses density-gradient centrifugation in sucrose solutions to purify a plant virus.
1954	**de Duve** isolates lysosomes and, later, peroxisomes by centrifugation.
1954	**Zamecnik** and colleagues develop the first cell-free system to perform protein synthesis. A decade of intense research activity, during which the genetic code is elucidated, follows.
1957	**Meselson**, **Stahl**, **and Vinograd** develop equilibrium density-gradient centrifugation in cesium chloride solutions for separating nucleic acids.
1975	**Dobberstein and Blobel** demonstrate protein translocation across membranes in a cell-free system.
1976	**Neher and Sakmann** develop patch-clamp recording to measure the activity of single ion channels.
1983	**Lohka and Masui** makes concentrated extracts from frog eggs that performs the entire cell cycle *in vitro*.
1984	**Rothman** and colleagues reconstitute Golgi vesicle trafficking *in vitro* with a cell-free system.

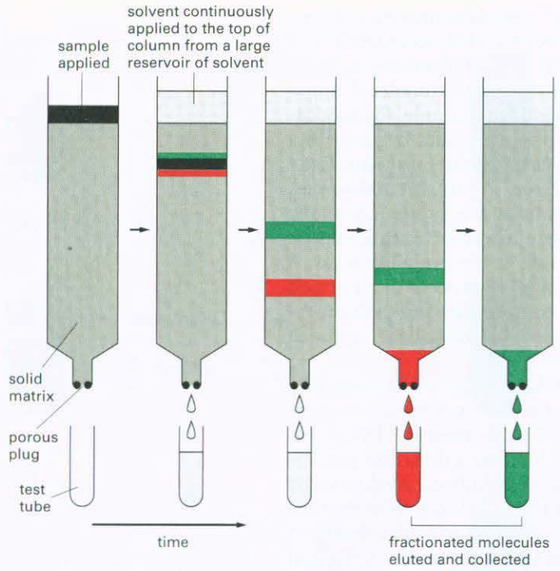


Figure 8–10 The separation of molecules by column chromatography. The sample, a mixture of different molecules, is applied to the top of a cylindrical glass or plastic column filled with a permeable solid matrix, such as cellulose, immersed in solvent. A large amount of solvent is then pumped slowly through the column and collected in separate tubes as it emerges from the bottom. Because various components of the sample travel at different rates through the column, they are fractionated into different tubes.

interaction with the matrix, and they can be collected separately as they flow out of the bottom of the column (Figure 8–10). Depending on the choice of matrix, proteins can be separated according to their charge *(ion-exchange chromatography)*, their hydrophobicity *(hydrophobic chromatography)*, their size *(gel-filtration chromatography)*, or their ability to bind to particular small molecules or to other macromolecules *(affinity chromatography)*.

Many types of matrices are commercially available (Figure 8–11). Ion-exchange columns are packed with small beads that carry either a positive or negative charge, so that proteins are fractionated according to the arrangement of charges on their surface. Hydrophobic columns are packed with beads from which hydrophobic side chains protrude, so that proteins with exposed hydrophobic regions are retarded. Gel-filtration columns, which separate proteins according to their size, are packed with tiny porous beads: molecules that are small enough to enter the pores linger inside successive beads as they pass down the column, while larger molecules remain in the solution flowing between the beads and therefore move more rapidly, emerging from the column first. Besides providing a means of separating molecules, gel-filtration chromatography is a convenient way to determine their size.

The resolution of conventional column chromatography is limited by inhomogeneities in the matrices (such as cellulose), which cause an uneven flow of solvent through the column. Newer chromatography resins (usually silica-based) have been developed in the form of tiny spheres (3 to 10 μm in diameter) that can be packed with a special apparatus to form a uniform column bed. A high degree of resolution is attainable on such **high-performance liquid chromatography (HPLC)** columns. Because they contain such tightly packed particles, HPLC columns have negligible flow rates unless high pressures are applied. For this reason these columns are typically packed in steel cylinders and require an elaborate system of pumps and valves to force the solvent through them at sufficient pressure to produce the desired rapid flow rates of about one column volume per minute. In conventional column chromatography, flow rates must be kept slow (often about one column volume per hour) to give the solutes being fractionated time to equilibrate with the interior of the large matrix particles. In HPLC the solutes equilibrate very rapidly with the interior of the tiny spheres, so solutes with different affinities for the matrix are efficiently separated from one another even at fast flow rates. This allows most fractionations to be carried out in minutes, whereas hours are required to obtain a

poorer separation by conventional chromatography. HPLC has therefore become the method of choice for separating many proteins and small molecules.

Affinity Chromatography Exploits Specific Binding Sites on Proteins

If one starts with a complex mixture of proteins, these types of column chromatography do not produce very highly purified fractions: a single passage through the column generally increases the proportion of a given protein in the mixture no more than twentyfold. Because most individual proteins represent less than 1/1000 of the total cellular protein, it is usually necessary to use several different types of column in succession to attain sufficient purity (Figure 8–12). A more efficient procedure, known as **affinity chromatography**, takes advantage of the biologically important binding interactions that occur on protein surfaces. If a substrate molecule is covalently coupled to an inert matrix such as a polysaccharide bead, for example, the enzyme that operates on that substrate will often be specifically retained by the matrix and can then be eluted (washed out) in nearly pure form. Likewise, short DNA oligonucleotides of a specifically designed sequence can be immobilized in this way and used to purify DNA-binding proteins that normally recognize this sequence of nucleotides in chromosomes (see Figure 7–30). Alternatively, specific antibodies can be coupled to a matrix to purify protein molecules recognized by the antibodies. Because of the great specificity of all such affinity columns, 1000- to 10,000-fold purifications can sometimes be achieved in a single pass.

Any gene can be modified, using the recombinant DNA methods discussed in the next section, to produce its protein with a molecular tag attached to it, making subsequent purification of the protein by affinity chromatography simple and rapid (see Figure 8–48, below). For example, the amino acid histidine

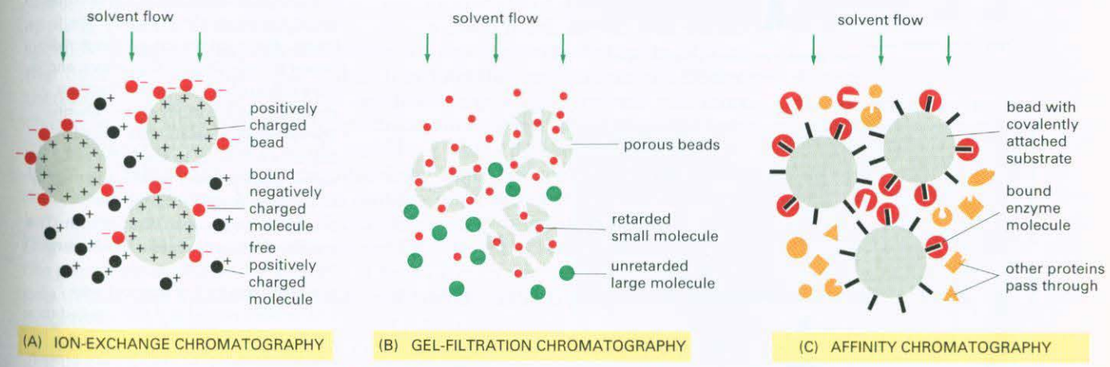


Figure 8–11 Three types of matrices used for chromatography. In ion-exchange chromatography (A) the insoluble matrix carries ionic charges that retard the movement of molecules of opposite charge. Matrices used for separating proteins include diethylaminoethylcellulose (DEAE-cellulose), which is positively charged, and carboxymethylcellulose (CM-cellulose) and phosphocellulose, which are negatively charged. Analogous matrices based on agarose or other polymers are also frequently used. The strength of the association between the dissolved molecules and the ion-exchange matrix depends on both the ionic strength and the pH of the solution that is passing down the column, which may therefore be varied systematically (as in Figure 8–12) to achieve an effective separation. In gel-filtration chromatography (B) the matrix is inert but porous. Molecules that are small enough to penetrate into the matrix are thereby delayed and travel more slowly through the column. Beads of cross-linked polysaccharide (dextran, agarose, or acrylamide) are available commercially in a wide range of pore sizes, making them suitable for the fractionation of molecules of various molecular weights, from less than 500 to more than 5×10^6. Affinity chromatography (C) uses an insoluble matrix that is covalently linked to a specific ligand, such as an antibody molecule or an enzyme substrate, that will bind a specific protein. Enzyme molecules that bind to immobilized substrates on such columns can be eluted with a concentrated solution of the free form of the substrate molecule, while molecules that bind to immobilized antibodies can be eluted by dissociating the antibody–antigen complex with concentrated salt solutions or solutions of high or low pH. High degrees of purification are often achieved in a single pass through an affinity column.

binds to certain metal ions, including nickel and copper. If genetic engineering techniques are used to attach a short string of histidine residues to either end of a protein, the slightly modified protein can be retained selectively on an affinity column containing immobilized nickel ions. Metal affinity chromatography can thereby be used to purify that modified protein from a complex molecular mixture. In other cases, an entire protein is used as the molecular tag. When the small enzyme glutathione S-transferase (GST) is attached to a target protein, the resulting fusion protein can be purified using an affinity column containing glutathione, a substrate molecule that binds specifically and tightly to GST (see Figure 8–50, below).

As a further refinement of this last technique, an amino acid sequence that forms a cleavage site for a highly specific protease can be engineered between the protein of choice and the histidine or GST tag. The cleavage sites for the proteases that are used, such as factor X that functions during blood clotting, are very rarely found by chance in proteins. Thus, the tag can later be specifically removed by cleavage at the cleavage site without destroying the purified protein.

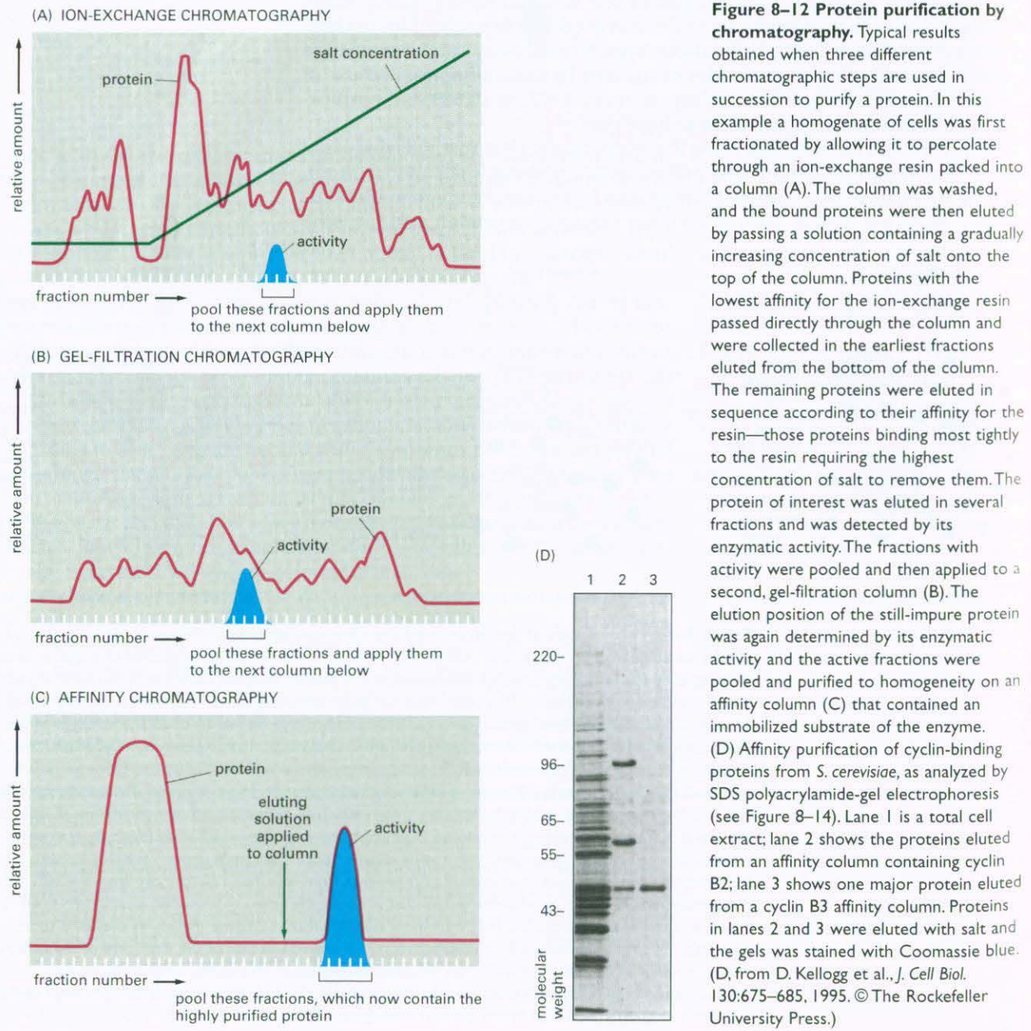


Figure 8–12 Protein purification by chromatography. Typical results obtained when three different chromatographic steps are used in succession to purify a protein. In this example a homogenate of cells was first fractionated by allowing it to percolate through an ion-exchange resin packed into a column (A). The column was washed, and the bound proteins were then eluted by passing a solution containing a gradually increasing concentration of salt onto the top of the column. Proteins with the lowest affinity for the ion-exchange resin passed directly through the column and were collected in the earliest fractions eluted from the bottom of the column. The remaining proteins were eluted in sequence according to their affinity for the resin—those proteins binding most tightly to the resin requiring the highest concentration of salt to remove them. The protein of interest was eluted in several fractions and was detected by its enzymatic activity. The fractions with activity were pooled and then applied to a second, gel-filtration column (B). The elution position of the still-impure protein was again determined by its enzymatic activity and the active fractions were pooled and purified to homogeneity on an affinity column (C) that contained an immobilized substrate of the enzyme. (D) Affinity purification of cyclin-binding proteins from *S. cerevisiae*, as analyzed by SDS polyacrylamide-gel electrophoresis (see Figure 8–14). Lane 1 is a total cell extract; lane 2 shows the proteins eluted from an affinity column containing cyclin B2; lane 3 shows one major protein eluted from a cyclin B3 affinity column. Proteins in lanes 2 and 3 were eluted with salt and the gels was stained with Coomassie blue. (D, from D. Kellogg et al., *J. Cell Biol.* 130:675–685, 1995. © The Rockefeller University Press.)

The Size and Subunit Composition of a Protein Can Be Determined by SDS Polyacrylamide-Gel Electrophoresis

Proteins usually possess a net positive or negative charge, depending on the mixture of charged amino acids they contain. When an electric field is applied to a solution containing a protein molecule, the protein migrates at a rate that depends on its net charge and on its size and shape. This technique, known as electrophoresis, was originally used to separate mixtures of proteins either in free aqueous solution or in solutions held in a solid porous matrix such as starch.

In the mid-1960s a modified version of this method—which is known as **SDS polyacrylamide-gel electrophoresis (SDS-PAGE)**—was developed that has revolutionized routine protein analysis. It uses a highly cross-linked gel of polyacrylamide as the inert matrix through which the proteins migrate. The gel is prepared by polymerization from monomers; the pore size of the gel can be adjusted so that it is small enough to retard the migration of the protein molecules of interest. The proteins themselves are not in a simple aqueous solution but in one that includes a powerful negatively charged detergent, sodium dodecyl sulfate, or SDS (Figure 8–13). Because this detergent binds to hydrophobic regions of the protein molecules, causing them to unfold into extended polypeptide chains, the individual protein molecules are released from their associations with other proteins or lipid molecules and rendered freely soluble in the detergent solution. In addition, a reducing agent such as β-mercaptoethanol (see Figure 8–13) is usually added to break any S–S linkages in the proteins, so that all of the constituent polypeptides in multisubunit molecules can be analyzed separately.

What happens when a mixture of SDS-solubilized proteins is run through a slab of polyacrylamide gel? Each protein molecule binds large numbers of the negatively charged detergent molecules, which mask the protein's intrinsic charge and cause it to migrate toward the positive electrode when a voltage is applied. Proteins of the same size tend to move through the gel with similar speeds because (1) their native structure is completely unfolded by the SDS, so that their shapes are the same, and (2) they bind the same amount of SDS and therefore have the same amount of negative charge. Larger proteins, with more charge, will be subjected to larger electrical forces and also to a larger drag. In free solution the two effects would cancel out, but in the mesh of the polyacrylamide gel, which acts as a molecular sieve, large proteins are retarded much more than small ones. As a result, a complex mixture of proteins is fractionated into a series of discrete protein bands arranged in order of molecular weight (Figure 8–14). The major proteins are readily detected by staining the proteins in the gel with a dye such as Coomassie blue, and even minor proteins are seen in gels treated with a silver or gold stain (with which as little as 10 ng of protein can be detected in a band).

SDS polyacrylamide-gel electrophoresis is a more powerful procedure than any previous method of protein analysis principally because it can be used to separate all types of proteins, including those that are insoluble in water. Membrane proteins, protein components of the cytoskeleton, and proteins that are part of large macromolecular aggregates can all be resolved. Because the method separates polypeptides by size, it also provides information about the molecular weight and the subunit composition of any protein complex. A photograph of a gel that has been used to analyze each of the successive stages in the purification of a protein is shown in Figure 8–15.

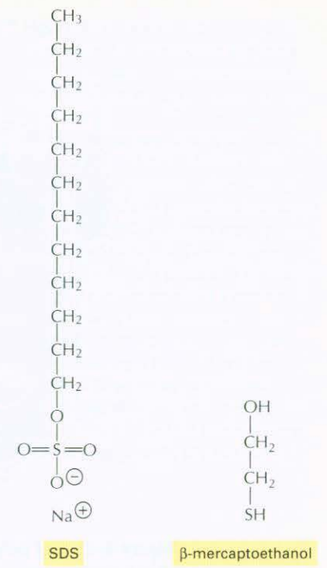


Figure 8–13 The detergent sodium dodecyl sulfate (SDS) and the reducing agent β-mercaptoethanol. These two chemicals are used to solubilize proteins for SDS polyacrylamide-gel electrophoresis. The SDS is shown here in its ionized form.

More Than 1000 Proteins Can Be Resolved on a Single Gel by Two-dimensional Polyacrylamide-Gel Electrophoresis

Because closely spaced protein bands or peaks tend to overlap, one-dimensional separation methods, such as SDS polyacrylamide-gel electrophoresis or chromatography, can resolve only a relatively small number of proteins (generally fewer than 50). In contrast, **two-dimensional gel electrophoresis**, which

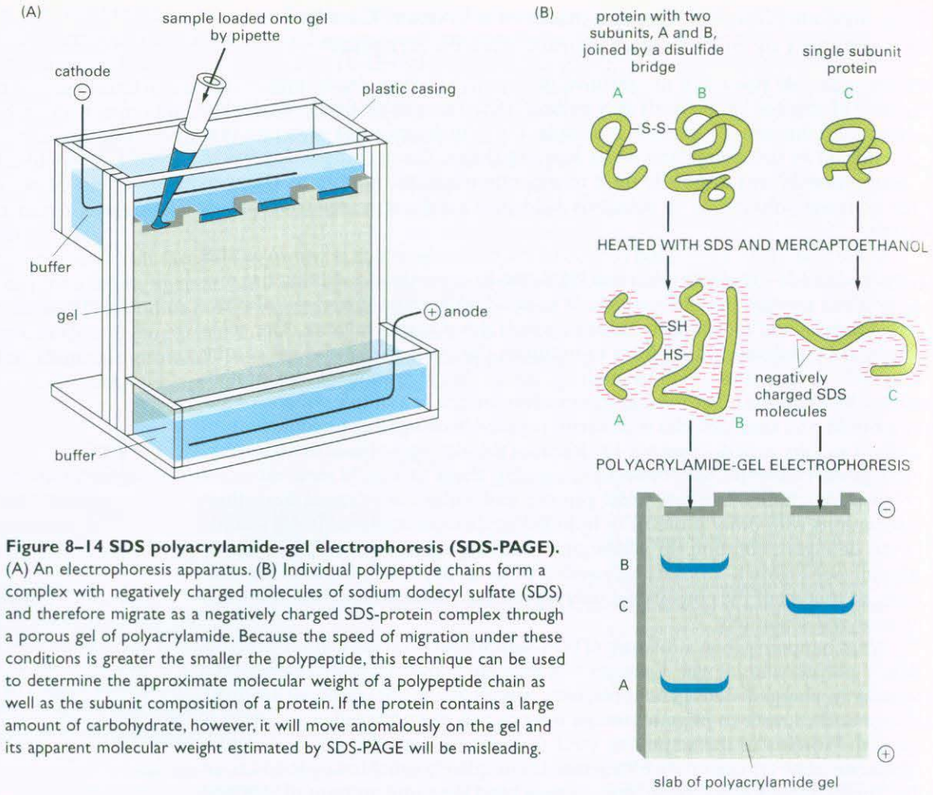


Figure 8–14 SDS polyacrylamide-gel electrophoresis (SDS-PAGE). (A) An electrophoresis apparatus. (B) Individual polypeptide chains form a complex with negatively charged molecules of sodium dodecyl sulfate (SDS) and therefore migrate as a negatively charged SDS-protein complex through a porous gel of polyacrylamide. Because the speed of migration under these conditions is greater the smaller the polypeptide, this technique can be used to determine the approximate molecular weight of a polypeptide chain as well as the subunit composition of a protein. If the protein contains a large amount of carbohydrate, however, it will move anomalously on the gel and its apparent molecular weight estimated by SDS-PAGE will be misleading.

combines two different separation procedures, can resolve up to 2000 proteins—the total number of different proteins in a simple bacterium—in the form of a two-dimensional protein map.

In the first step, the proteins are separated by their intrinsic charges. The sample is dissolved in a small volume of a solution containing a nonionic (uncharged) detergent, together with β-mercaptoethanol and the denaturing reagent urea. This solution solubilizes, denatures, and dissociates all the polypeptide chains but leaves their intrinsic charge unchanged. The polypeptide chains are then separated by a procedure called isoelectric focusing, which takes advantage of the fact that the net charge on a protein molecule varies with the pH of the surrounding solution. Every protein has a characteristic isoelectric point, the pH at which the protein has no net charge and therefore does not migrate in an electric field. In isoelectric focusing, proteins are separated electrophoretically in a narrow tube of polyacrylamide gel in which a gradient of pH is established by a mixture of special buffers. Each protein moves to a position in the gradient that corresponds to its isoelectric point and stays there (Figure 8–16). This is the first dimension of two-dimensional gel electrophoresis.

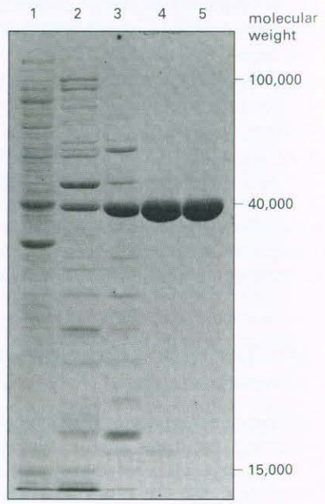


Figure 8–15 Analysis of protein samples by SDS polyacrylamide-gel electrophoresis. The photograph shows a Coomassie-stained gel that has been used to detect the proteins present at successive stages in the purification of an enzyme. The leftmost lane (lane 1) contains the complex mixture of proteins in the starting cell extract, and each succeeding lane analyzes the proteins obtained after a chromatographic fractionation of the protein sample analyzed in the previous lane (see Figure 8–12). The same total amount of protein (10 μg) was loaded onto the gel at the top of each lane. Individual proteins normally appear as sharp, dye-stained bands; a band broadens, however, when it contains too much protein. (From T. Formosa and B.M. Alberts, *J. Biol. Chem.* 261:6107–6118, 1986.)

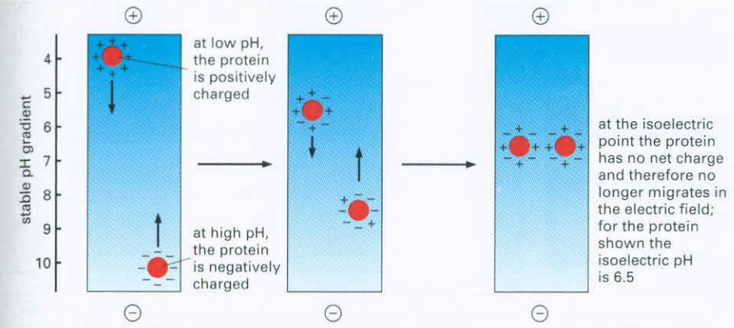


Figure 8–16 Separation of protein molecules by isoelectric focusing. At low pH (high H^+ concentration) the carboxylic acid groups of proteins tend to be uncharged (–COOH) and their nitrogen-containing basic groups fully charged (for example, $-NH_3^+$), giving most proteins a net positive charge. At high pH the carboxylic acid groups are negatively charged ($-COO^-$) and the basic groups tend to be uncharged (for example, $-NH_2$), giving most proteins a net negative charge. At its *isoelectric pH* a protein has no net charge since the positive and negative charges balance. Thus, when a tube containing a fixed pH gradient is subjected to a strong electric field in the appropriate direction, each protein species present migrates until it forms a sharp band at its isoelectric pH, as shown.

In the second step the narrow gel containing the separated proteins is again subjected to electrophoresis but in a direction that is at a right angle to the direction that used in the first step. This time SDS is added, and the proteins are separated according to their size, as in one-dimensional SDS-PAGE: the original narrow gel is soaked in SDS and then placed on one edge of an SDS polyacrylamide-gel slab, through which each polypeptide chain migrates to form a discrete spot. This is the second dimension of two-dimensional polyacrylamide-gel electrophoresis. The only proteins left unresolved are those that have both identical sizes and identical isoelectric points, a relatively rare situation. Even trace amounts of each polypeptide chain can be detected on the gel by various staining procedures—or by autoradiography if the protein sample was initially labeled with a radioisotope (Figure 8–17). The technique has such great resolving power that it can distinguish between two proteins that differ in only a single charged amino acid.

A specific protein can be identified after its fractionation on either one-dimensional or two-dimensional gels by exposing all the proteins present on the gel to a specific antibody that has been coupled to a radioactive isotope, to an easily detectable enzyme, or to a fluorescent dye. For convenience, this is normally done after all the separated proteins present in the gel have been transferred (by "blotting") onto a sheet of nitrocellulose paper, as described later for nucleic acids (see Figure 8–27). This protein-detection method is called **Western blotting** (Figure 8–18).

Some landmarks in the development of chromatography and electrophoresis are listed in Table 8–5.

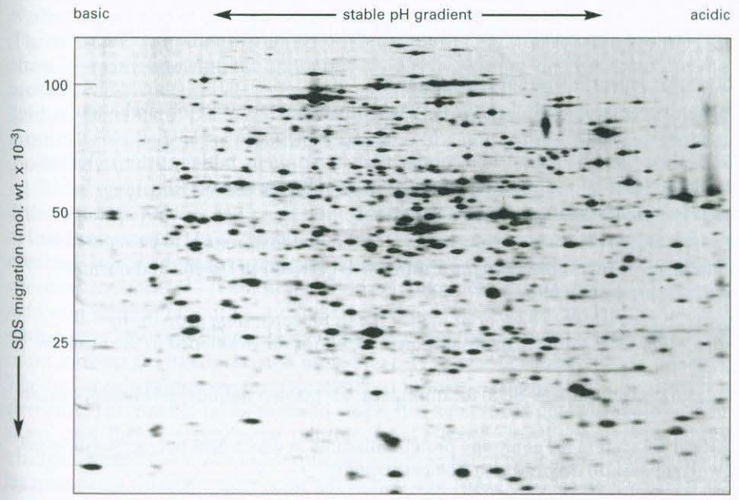
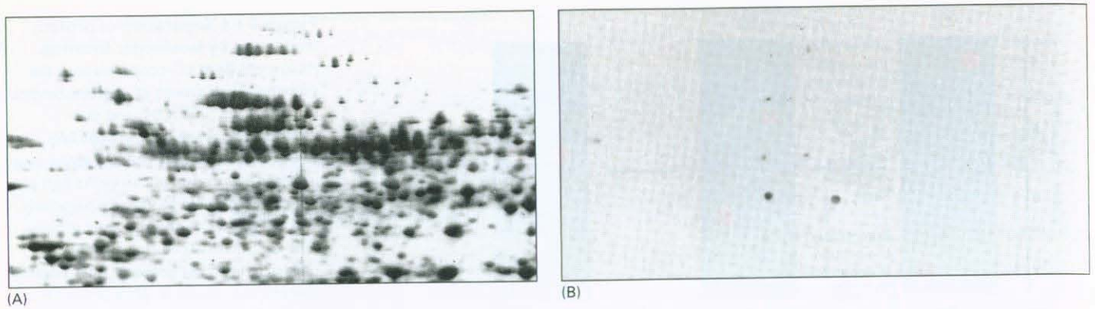


Figure 8–17 Two-dimensional polyacrylamide-gel electrophoresis. All the proteins in an *E. coli* bacterial cell are separated in this gel, in which each spot corresponds to a different polypeptide chain. The proteins were first separated on the basis of their isoelectric points by isoelectric focusing from left to right. They were then further fractionated according to their molecular weights by electrophoresis from top to bottom in the presence of SDS. Note that different proteins are present in very different amounts. The bacteria were fed with a mixture of radioisotope-labeled amino acids so that all of their proteins were radioactive and could be detected by autoradiography (see pp. 578–579). (Courtesy of Patrick O'Farrell.)


(A)

(B)

Figure 8–18 Western blotting. The total proteins from dividing tobacco cells in culture are first separated by two-dimensional polyacrylamide-gel electrophoresis and in (A) their positions are revealed by a sensitive protein stain. In (B) the separated proteins on an identical gel were then transferred to a sheet of nitrocellulose and exposed to an antibody that recognizes only those proteins that are phosphorylated on threonine residues during mitosis. The positions of the dozen or so proteins that are recognized by this antibody are revealed by an enzyme-linked second antibody. This technique is also known as immunoblotting. (From J.A. Traas, A.F. Bevan, J.H. Doonan, J. Cordewener, and P.J. Shaw, *Plant J.* 2:723–732, 1992. © Blackwell Scientific Publications.)

Selective Cleavage of a Protein Generates a Distinctive Set of Peptide Fragments

Although proteins have distinctive molecular weights and isoelectric points, unambiguous identification ultimately depends on determining their amino acid sequences. This can be most easily accomplished by determining the nucleotide sequence of the gene encoding the protein and using the genetic code to deduce the amino acid sequence of the protein, as discussed later in this chapter. It can also be done by directly analyzing the protein, although the complete amino acid sequences of proteins are rarely determined directly today.

There are several more rapid techniques that are used to reveal crucial information about the identity of purified proteins. For example, simply cleaving the protein into smaller fragments can provide information that helps to characterize the molecule. Proteolytic enzymes and chemical reagents are available that cleave proteins between specific amino acid residues (Table 8–6). The enzyme trypsin, for instance, cuts on the carboxyl side of lysine or arginine residues, whereas the chemical cyanogen bromide cuts peptide bonds next to methionine residues. Because these enzymes and chemicals cleave at relatively few sites, they tend to produce a few relatively large peptides when applied to a purified protein. If such a mixture of peptides is separated by chromatographic or electrophoretic procedures, the resulting pattern, or peptide map, is diagnostic of

TABLE 8–5 Landmarks in the Development of Chromatography and Electrophoresis and their Applications to Protein Molecules

1833	**Faraday** describes the fundamental laws concerning the passage of electricity through ionic solutions.
1850	**Runge** separates inorganic chemicals by their differential adsorption to paper, a forerunner of later chromatographic separations.
1906	**Tswett** invents column chromatography, passing petroleum extracts of plant leaves through columns of powdered chalk.
1933	**Tiselius** introduces electrophoresis for separating proteins in solution.
1942	**Martin and Synge** develop partition chromatography, leading to paper chromatography on ion-exchange resins.
1946	**Stein and Moore** determine for the first time the amino acid composition of a protein, initially using column chromatography on starch and later developing chromatography on ion-exchange resins.
1955	**Smithies** uses gels made of starch to separate proteins by electrophoresis.
	Sanger completes the analysis of the amino acid sequence of bovine insulin, the first protein to be sequenced.
1956	**Ingram** produces the first protein fingerprints, showing that the difference between sickle-cell and normal hemoglobin is due to a change in a single amino acid.
1959	**Raymond** introduces polyacrylamide gels, which are superior to starch gels for separating proteins by electrophoresis; improved buffer systems allowing high-resolution separations are developed in the next few years by **Ornstein and Davis**.
1966	**Maizel** introduces the use of sodium dodecyl sulfate (SDS) for improving the polyacrylamide-gel electrophoresis of proteins.
1975	**O'Farrell** devises a two-dimensional gel system for analyzing protein mixtures in which SDS polyacrylamide-gel electrophoresis is combined with separation according to isoelectric point.

TABLE 8–6 Some Reagents Commonly Used to Cleave Peptide Bonds in Proteins

	AMINO ACID 1	AMINO ACID 2
Enzyme		
Trypsin	Lys or Arg	any
Chymotrypsin	Phe, Trp, or Tyr	any
V8 protease	Glu	any
Chemical		
Cyanogen bromide	Met	any
2-Nitro-5-thiocyanobenzoate	any	Cys

The specificity for the amino acids on either side of the cleaved bond is indicated; amino acid 2 is linked to the C-terminus of amino acid 1.

the protein from which the peptides were generated and is sometimes referred to as the protein's "fingerprint" (Figure 8–19).

Protein fingerprinting was developed in 1956 to compare normal hemoglobin with the mutant form of the protein found in patients suffering from sickle-cell anemia. A single peptide difference was found and was eventually traced to a single amino acid change, providing the first demonstration that a mutation can change a single amino acid in a protein. Nowadays it is most often used to map the position of posttranslational modifications, such as phosphorylation sites.

Historically, cleaving a protein into a set of smaller peptides was an essential step in determining its amino acid sequence. This was ultimately accomplished through a series of repeated chemical reactions that removed one amino acid at a time from each peptide's N-terminus. After each cycle, the identity of the excised amino acid was determined by chromatographic methods. Now that the complete genome sequences for many organisms are available, mass spectrometry has become the method of choice for identifying proteins and matching each to its corresponding gene, thereby also determining its amino acid sequence as we discuss next.

Mass Spectrometry Can Be Used to Sequence Peptide Fragments and Identify Proteins

Mass spectrometry allows one to determine the precise mass of intact proteins and of peptides derived from them by enzymatic or chemical cleavage. This information can then be used to search genomic databases, in which the masses of all proteins and of all their predicted peptide fragments have been tabulated (Figure 8–20A). An unambiguous match to a particular open reading frame can often be made knowing the mass of only a few peptides derived from a given protein. Mass spectrometric methods are therefore critically important for the field of *proteomics*, the large-scale effort to identify and characterize all of the proteins encoded in an organism's genome, including their posttranslational modifications.

Mass spectrometry is an enormously sensitive technique that requires very little material. Masses can be obtained with great accuracy, often with an error of less than one part in a million. The most commonly used mass spectrometric method is called *matrix-assisted laser desorption ionization–time-of-flight spectrometry (MALDI-TOF)*. In this method, peptides are mixed with an organic acid and then dried onto a metal or ceramic slide. The sample is then blasted with a laser, causing the peptides to become ejected from the slide in the form of an ionized gas in which each molecule carries one or more positive charges. The ionized peptides are then accelerated in an electric field and fly toward a detector. The time it takes them to reach the detector is determined by their mass and their charge: large peptides move more slowly, and more highly charged molecules move more quickly. The precise mass is readily determined by analysis of those peptides with a single charge. MALDI–TOF can even be used

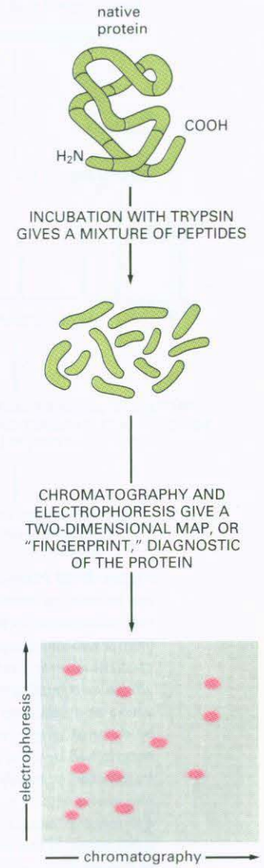


Figure 8–19 Production of a peptide map, or fingerprint, of a protein. Here, the protein was digested with trypsin to generate a mixture of polypeptide fragments, which was then fractionated in two dimensions by electrophoresis and partition chromatography. The latter technique separates peptides on the basis of their differential solubilities in water, which is preferentially bound to the solid matrix, as compared to the solvent in which they are applied. The resulting pattern of spots obtained from such a digest is diagnostic of the protein analyzed. It is also used to detect posttranslational modifications of proteins.

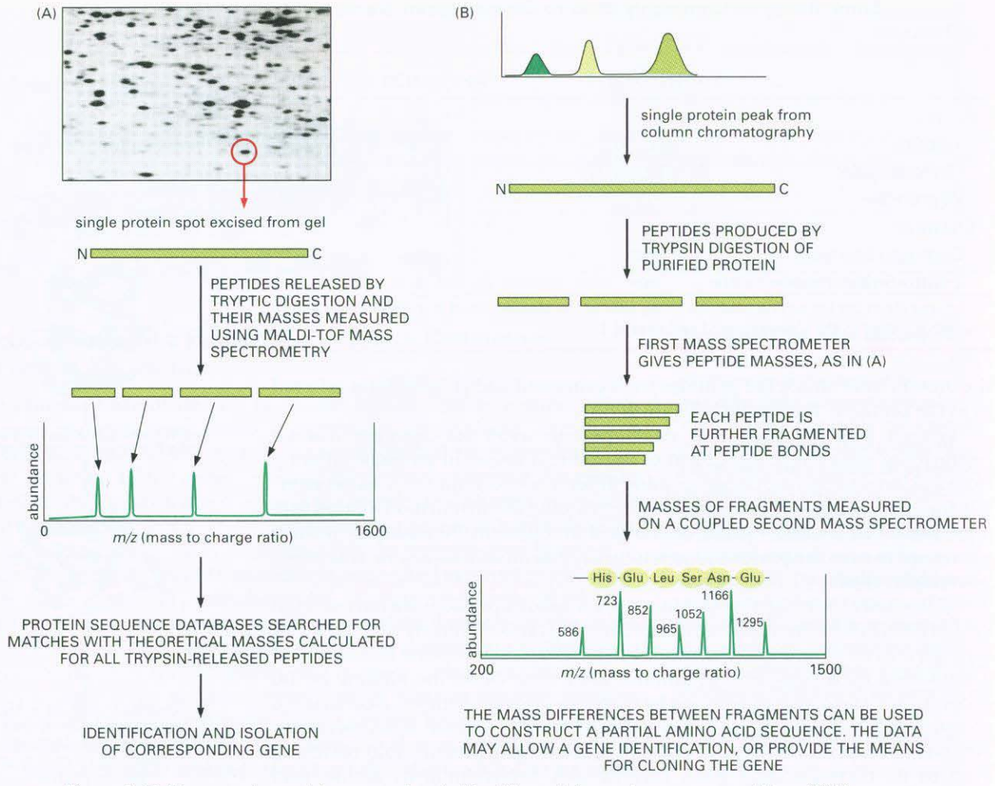


Figure 8–20 Mass-spectrometric approaches to identify proteins and sequence peptides. (A) Mass spectrometry can be used to identify proteins by determining their precise masses, and the masses of peptides derived from them, and using that information to search a genomic database for the corresponding gene. In this example, the protein of interest is excised from a two-dimensional polyacrylamide gel and then digested with trypsin. The peptide fragments are loaded into the mass spectrometer and their masses are measured. Sequence databases are then searched to find the gene that encodes a protein whose calculated tryptic digest profile matches these values. (B) Mass spectrometry can also be used to determine directly the amino acid sequence of peptide fragments. In this example, proteins that form a macromolecular complex have been separated by chromatography, and a single protein selected for digestion with trypsin. The masses of these tryptic fragments are then determined by mass spectrometry as in (A). To determine their exact amino acid sequence, each peptide is further fragmented, primarily by cleaving its peptide bonds. This treatment generates a nested set of peptides, each differing in size by one amino acid. These fragments are fed into a second coupled mass spectrometer and their masses are determined. The difference in masses between two closely related peptides can be used to deduce the "missing" amino acid. By repeated applications of this procedure, a partial amino acid sequence of the original protein can be determined. (Micrograph courtesy of Patrick O'Farrell.)

to measure the mass of intact proteins as large as 200,000 daltons, which corresponds to a polypeptide about 2000 amino acids in length.

Mass spectrometry is also used to determine the sequence of amino acids of individual peptide fragments. This method is particularly useful when the genome for the organism of interest has not yet been fully sequenced; the partial amino acid sequence obtained in this way can then be used to identify and clone the gene. Peptide sequencing is also important if proteins contain modifications, such as attached carbohydrates, phosphates, or methyl groups. In this case, the precise amino acids that are the sites of modifications can be determined.

To obtain such peptide sequence information, two mass spectrometers are required in tandem. The first separates peptides obtained after digestion of the protein of interest and allows one to zoom in on one peptide at a time. This peptide is then further fragmented by collision with high-energy gas atoms. This method of fragmentation preferentially cleaves the peptide bonds, generating a ladder of fragments, each differing by a single amino acid. The second mass

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 hydridization 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 *Haemophilus influenzae*.
1996	**Goffeau** and an international consortium of researchers announce the completion of the first genome sequence of a eucaryote, the yeast *Saccharomyces cerevisiae*.
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 *Caenorhabditis elegans*.
2001	Consortia of researchers announce the completion of the draft human genome sequence.

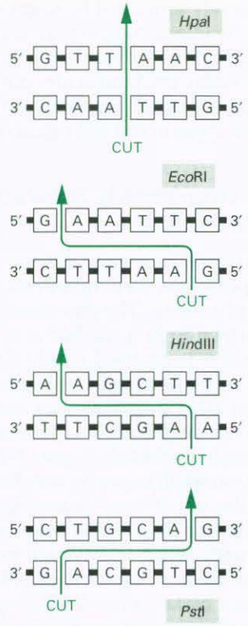


Figure 8–21 The DNA nucleotide sequences recognized by four widely used restriction nucleases. As in the examples shown, such sequences are often six base pairs long and "palindromic" (that is, the nucleotide sequence is the same if the helix is turned by 180 degrees around the center of the short region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

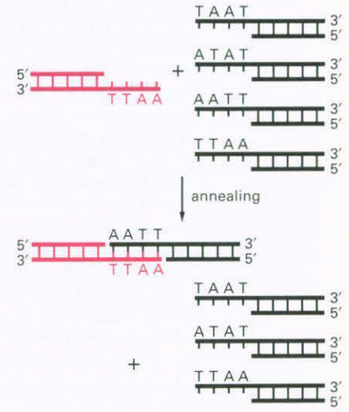


Figure 8–22 Restriction nucleases produce DNA fragments that can be easily joined together. Fragments with the same cohesive ends can readily join by complementary base-pairing between their cohesive ends, as illustrated. The two DNA fragments that join in this example were both produced by the *Eco*RI 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 *Hpa*I (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

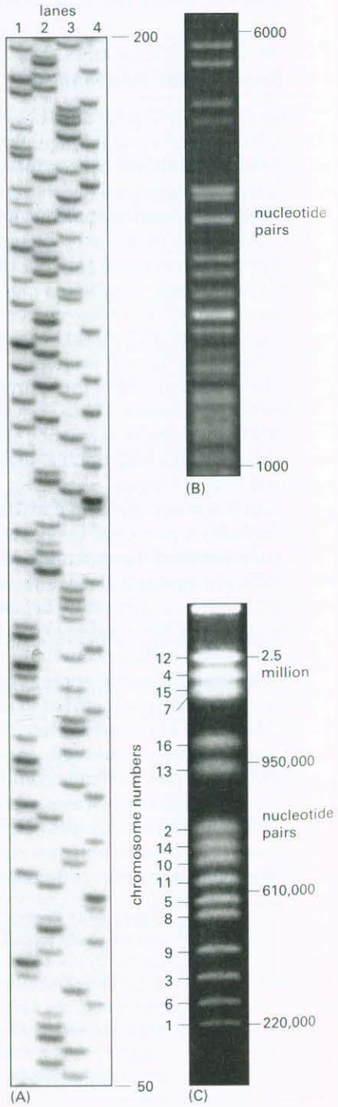


Figure 8–23 Gel electrophoresis techniques for separating DNA molecules by size. In the three examples shown, electrophoresis is from top to bottom, so that the largest—and thus slowest-moving—DNA molecules are near the top of the gel. In (A) a polyacrylamide gel with small pores is used to fractionate 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 (pH ≥ 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

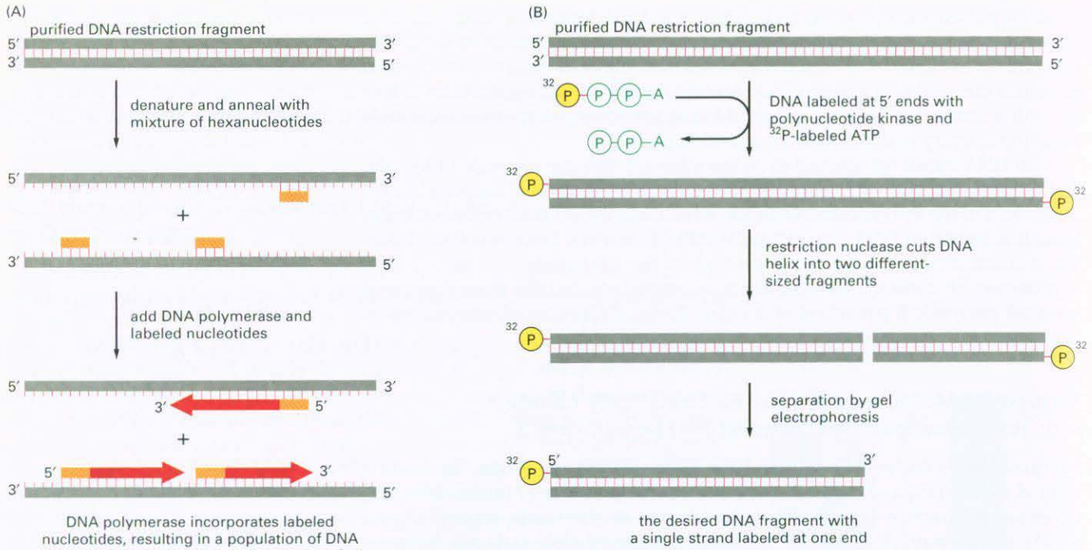


Figure 8–24 Methods for labeling DNA molecules *in vitro*. (A) A purified DNA polymerase enzyme labels all the nucleotides in a DNA molecule and can thereby produce highly radioactive DNA probes. (B) Polynucleotide kinase labels only the 5′ ends of DNA strands; therefore, when labeling is followed by 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.

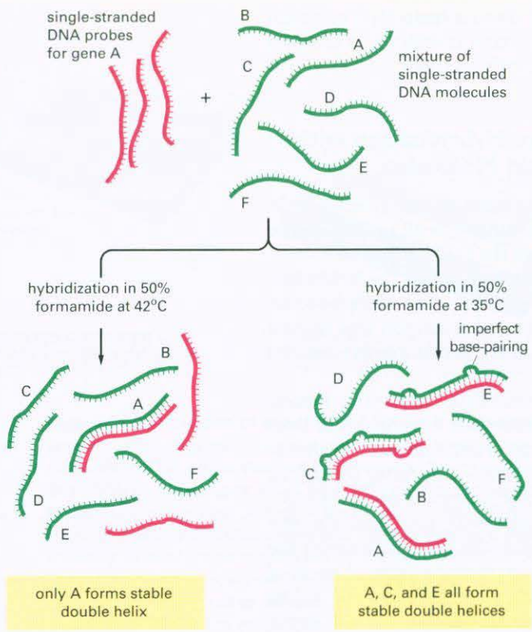


Figure 8–25 Different hybridization conditions allow less than perfect DNA matching. When only an identical match with a DNA probe is desired, the hybridization reaction is kept just a few degrees below the temperature at which a perfect DNA helix denatures in the solvent used (its *melting temperature*), 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

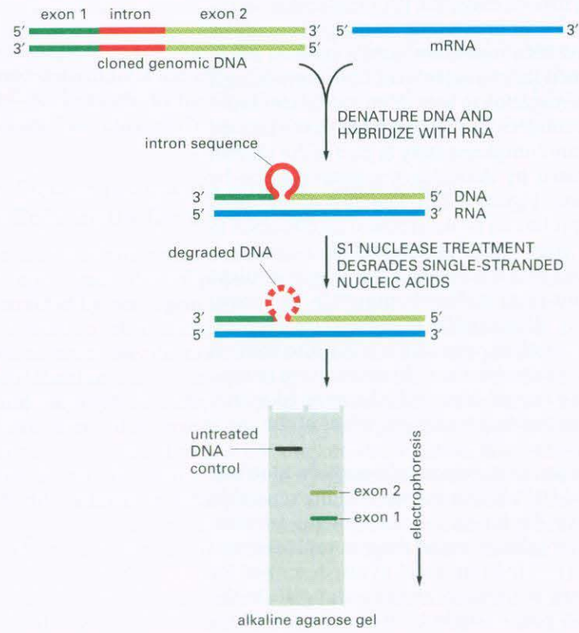


Figure 8–26 The use of nucleic acid hybridization to determine the region of a cloned DNA fragment that is present in an mRNA molecule. The method shown requires a nuclease that cuts the DNA chain only where it is not base-paired to a complementary RNA chain. The positions of the introns in eucaryotic 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

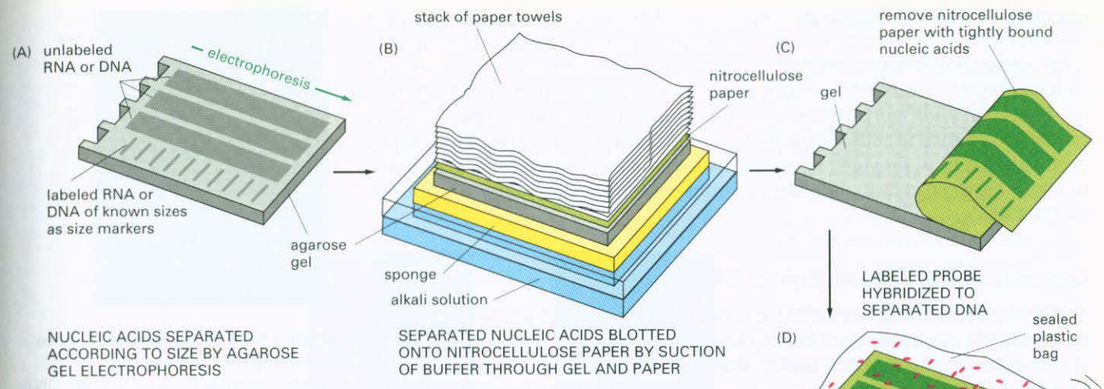


Figure 8–27 Detection of specific RNA or DNA molecules by gel-transfer hybridization. In this example, the DNA probe is detected by its radioactivity. DNA probes detected by chemical or fluorescence methods are also widely used (see Figure 8–24). (A) A mixture of either single-stranded RNA molecules *(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

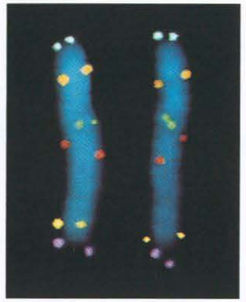


Figure 8–28 *In situ* hybridization to locate specific genes on chromosomes. Here, six different DNA probes have been used to mark the location of their respective nucleotide sequences on human chromosome 5 at metaphase. The probes have been chemically labeled and detected with fluorescent antibodies. 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.)

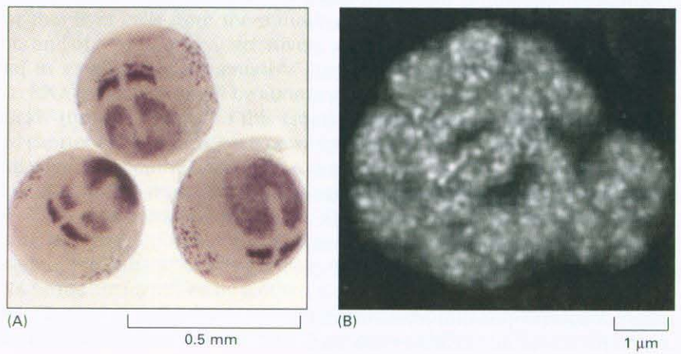


Figure 8–29 *In situ* hybridization for RNA localization. (A) Expression pattern of *deltaC* in the early zebrafish embryo. This gene codes for a ligand in the Notch signaling pathway (discussed in Chapter 15), and the pattern shown here reflects its role in the development of somites—the future segments 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.)

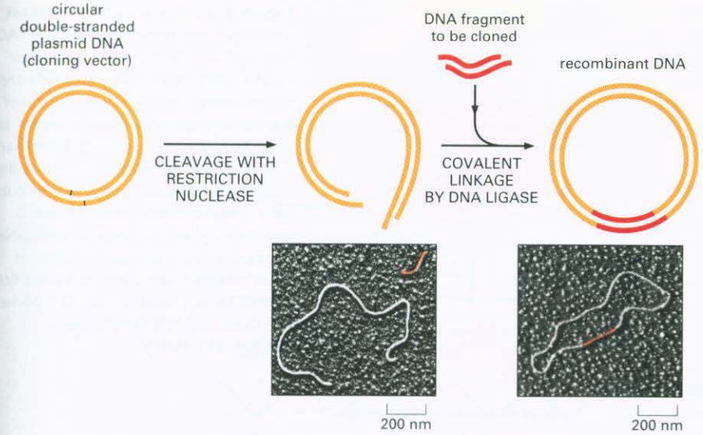


Figure 8–30 The insertion of a DNA fragment into a bacterial plasmid with the enzyme DNA ligase. The plasmid is cut open with a restriction nuclease (in this case one that produces cohesive ends) and is mixed with the DNA fragment to be cloned (which has been prepared with the same restriction nuclease), 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

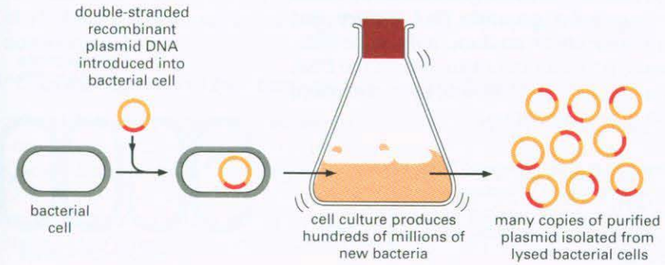


Figure 8–31 Purification and amplification of a specific DNA sequence by DNA cloning in a bacterium. To produce many copies of a particular DNA sequence, the fragment is first inserted into a plasmid vector, as shown in Figure 8–30. The resulting recombinant plasmid DNA is then introduced into a bacterium,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

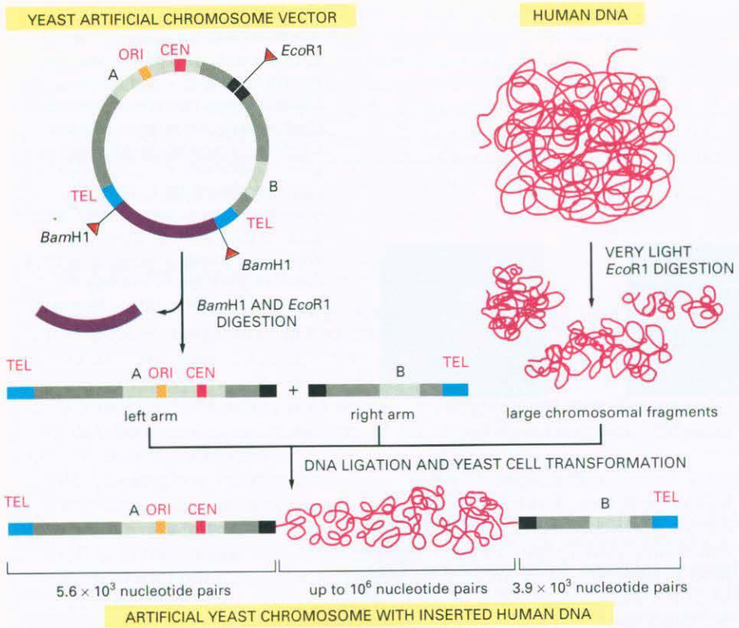


Figure 8–32 The making of a yeast artificial chromosome (YAC). A YAC vector allows the cloning of very large DNA molecules. TEL, CEN, and ORI are the telomere, centromere, and origin of replication sequences, respectively, for the yeast *Saccharomyces cerevisiae*. *Bam*H1 and *Eco*R1 are sites where 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.

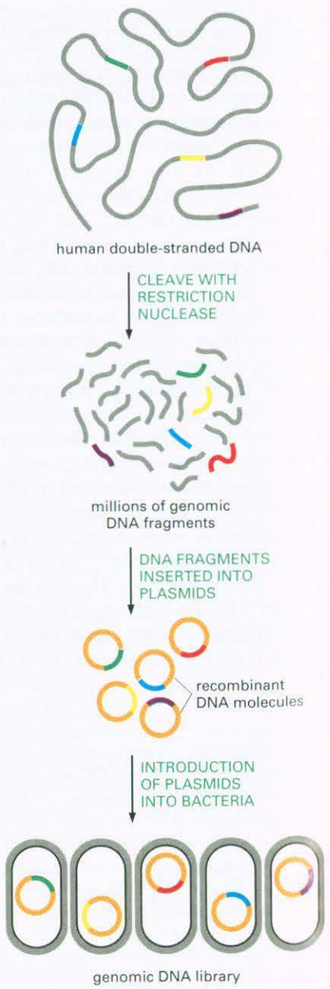


Figure 8–33 Construction of a human genomic DNA library. A genomic library is usually stored as a set of bacteria, each carrying a different fragment of human DNA. For simplicity, cloning of just a few representative fragments *(colored)* is shown. In reality, all of the *gray* DNA fragments would also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

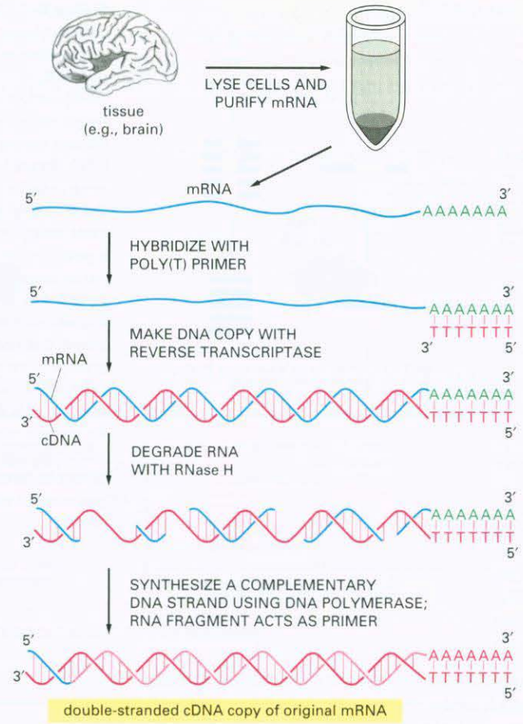


Figure 8–34 The synthesis of cDNA. Total mRNA is extracted from a particular tissue, and DNA copies (cDNA) of the mRNA molecules are produced by the enzyme reverse transcriptase (see p. 289). For simplicity, the copying of just one of these mRNAs into cDNA is illustrated. A short oligonucleotide complementary 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 helix. 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

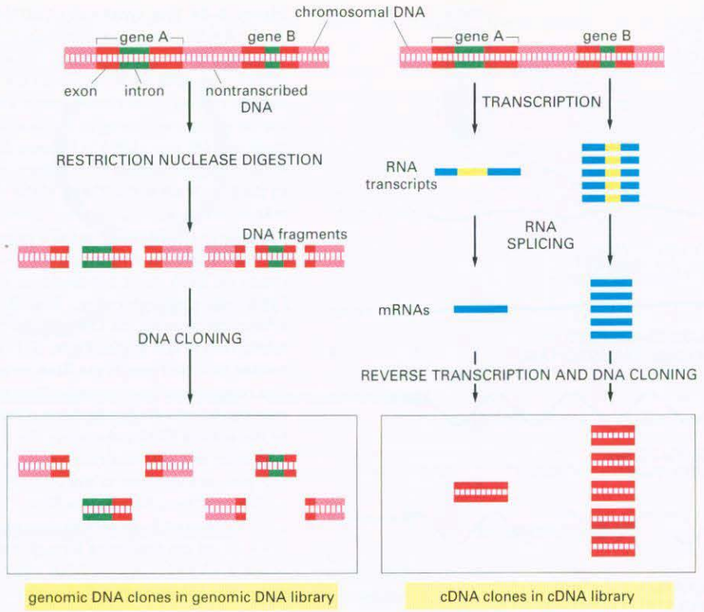


Figure 8–35 The differences between cDNA clones and genomic DNA clones derived from the same region of DNA. In this example gene A is infrequently transcribed, whereas gene B is frequently transcribed, and both genes contain introns *(green)*. In the genomic DNA library, both the introns and the nontranscribed 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 in 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

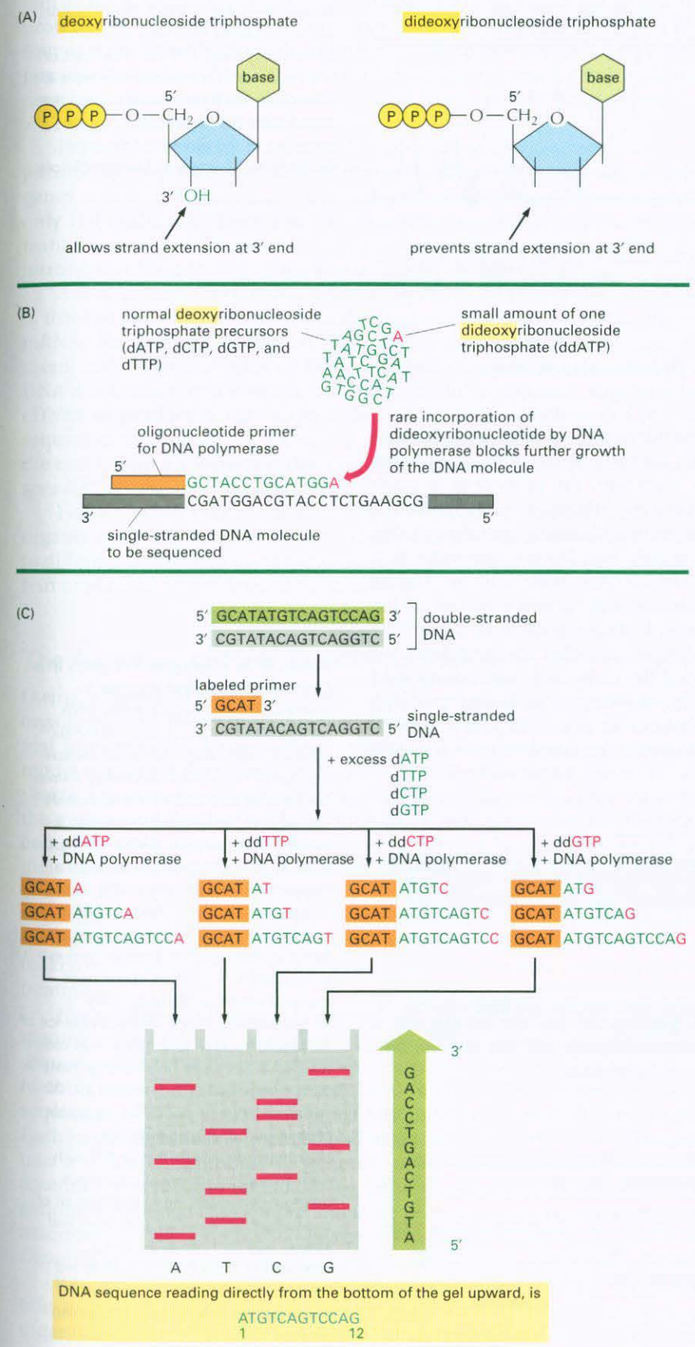


Figure 8–36 The enzymatic—or dideoxy—method of sequencing DNA. (A) This method relies on the use of dideoxyribonucleoside triphosphates, derivatives of the normal deoxyribonucleoside triphosphates that lack the 3′ hydroxyl group. (B) Purified DNA is synthesized *in vitro* in a mixture that contains single-stranded 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.

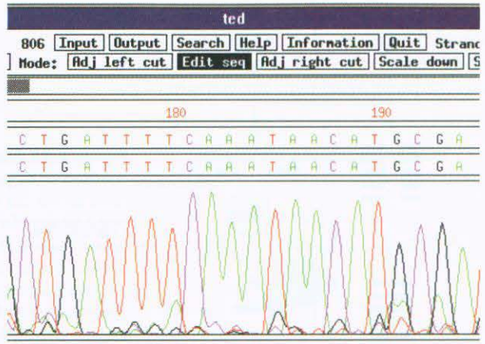


Figure 8–37 Automated DNA sequencing. Shown here is a tiny part of the data from an automated DNA-sequencing run as it appears on the computer screen. Each colored peak represents a nucleotide in the DNA sequence—a clear stretch of nucleotide sequence can be read here between positions 173 and 194 from 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.

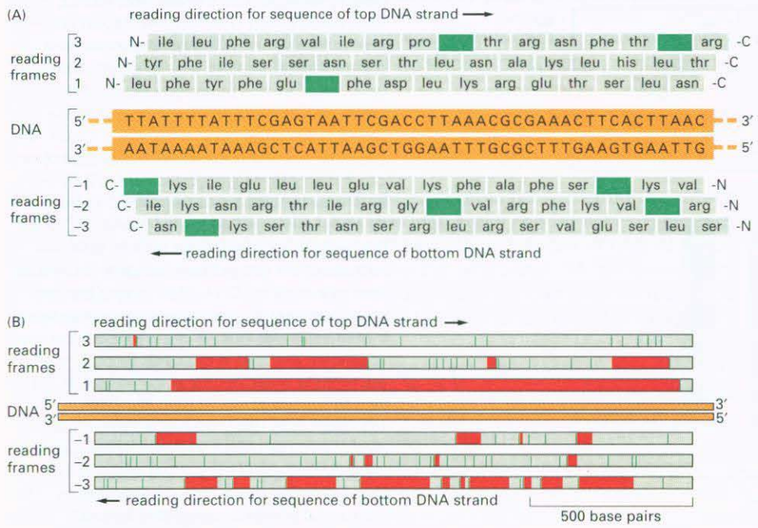


Figure 8–38 Finding the regions in a DNA sequence that encode a protein. (A) Any region of the DNA sequence can, in principle, code for six different amino acid sequences, because any one of three different reading frames can be used to interpret the nucleotide sequence on each strand. Note that a nucleotide 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 archea, 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

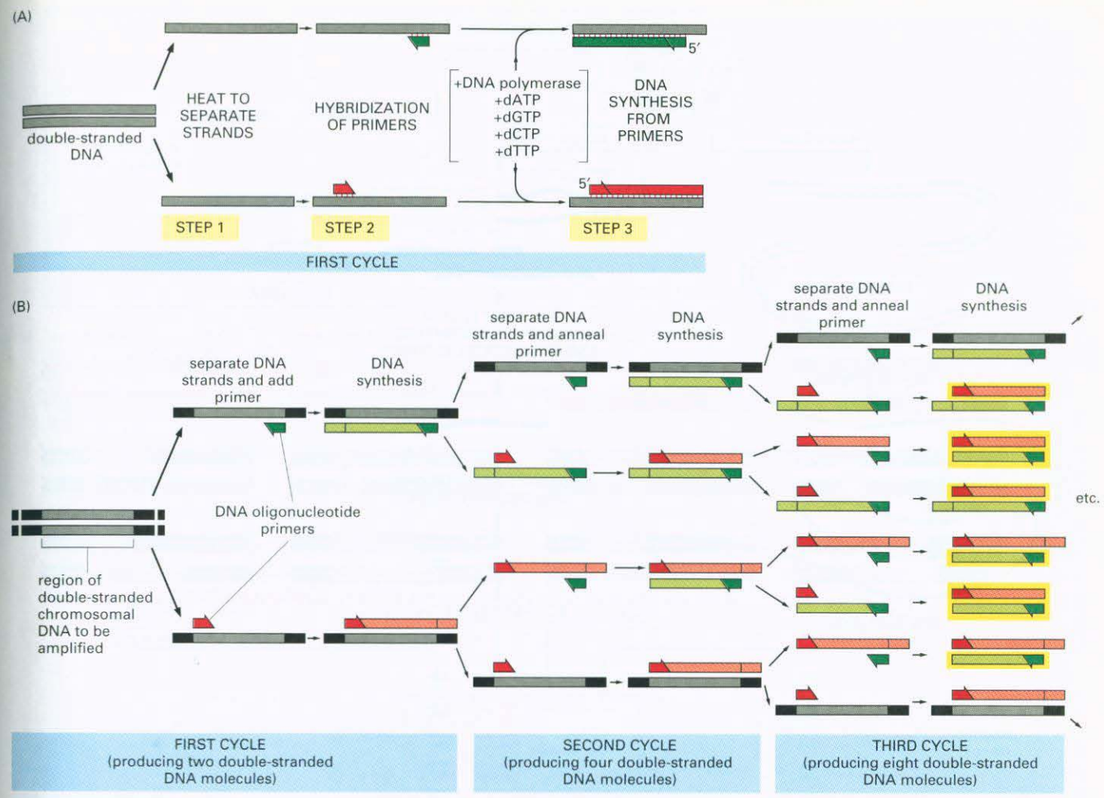


Figure 8–39 Amplification of DNA using the PCR technique. Knowledge of the DNA sequence to be amplified is used to design two synthetic DNA oligonucleotides, each complementary to the sequence on one strand of the DNA double helix at opposite ends of the region to be amplified. These oligonucleotides 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).

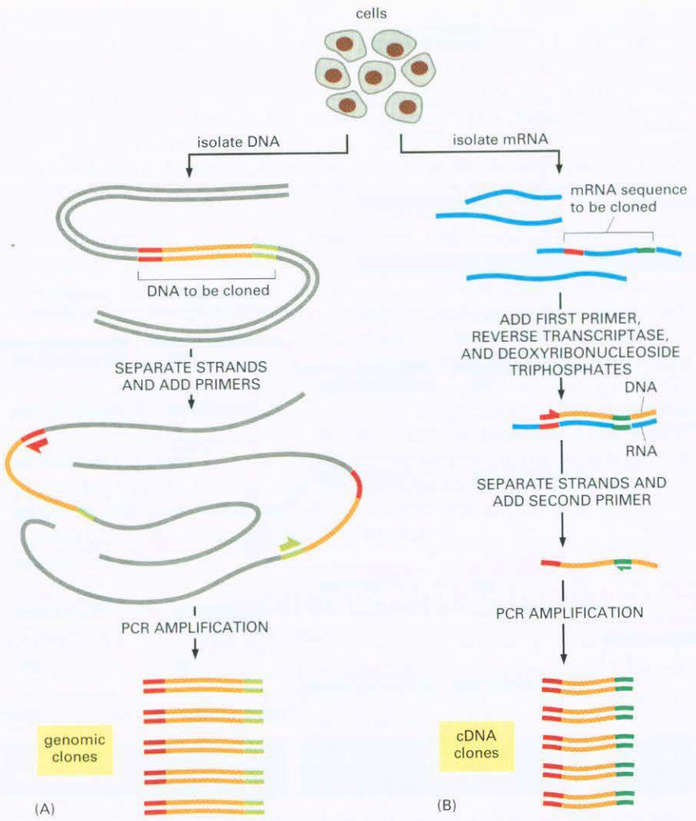


Figure 8–40 Use of PCR to obtain a genomic or cDNA clone. (A) To obtain a genomic clone by using PCR, chromosomal DNA is first purified from cells. PCR primers that flank the stretch of DNA to be cloned are added, and many cycles of the reaction are completed (see Figure 8–39). Since only the DNA between (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

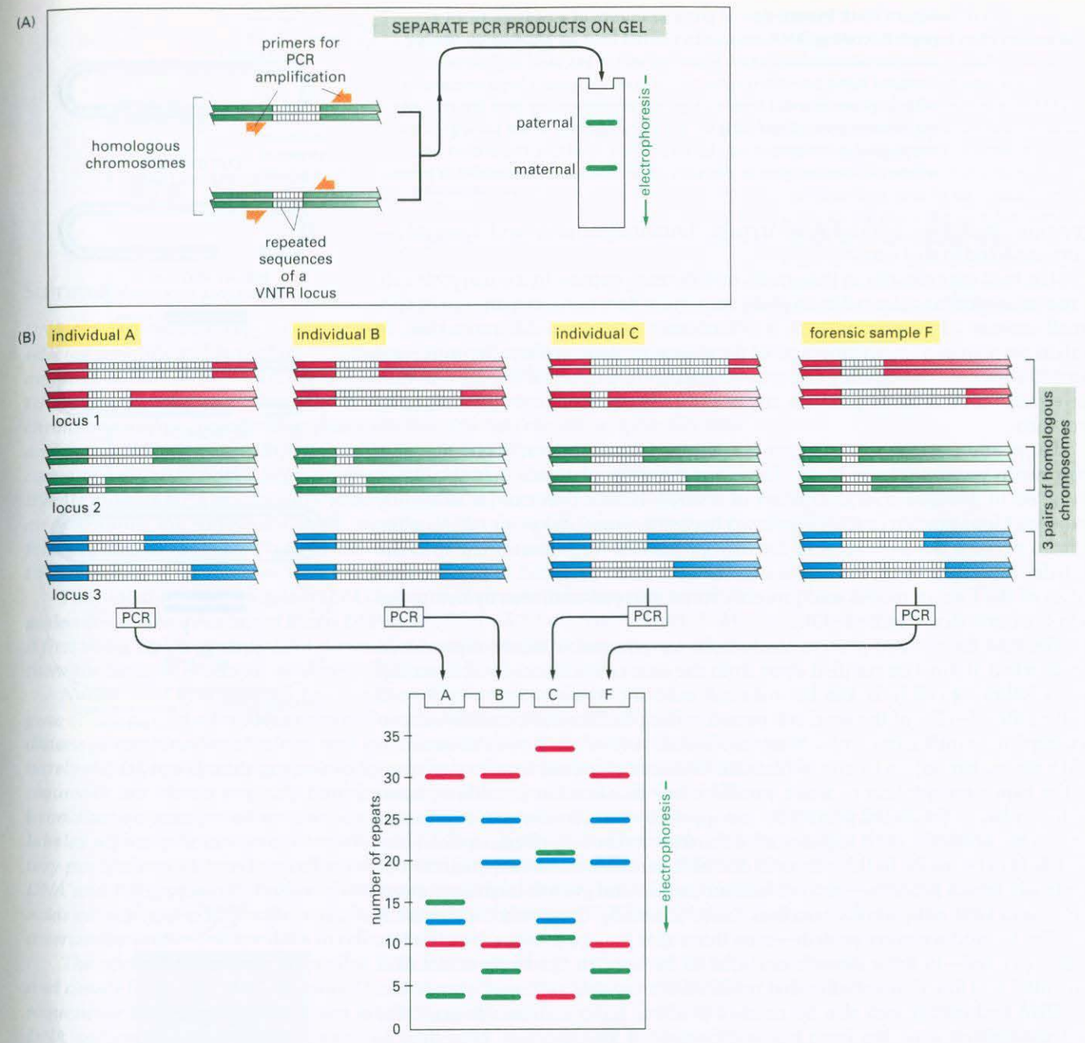
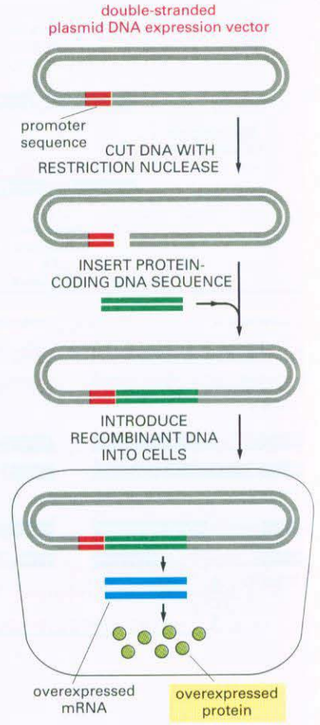


Figure 8–41 How PCR is used in forensic science. (A) The DNA sequences that create the variability used in this analysis contain runs of short, repeated sequences, such as CACACA . . . , which are found in various positions (loci) in the human genome. The number of repeats in each run can be highly variable 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.

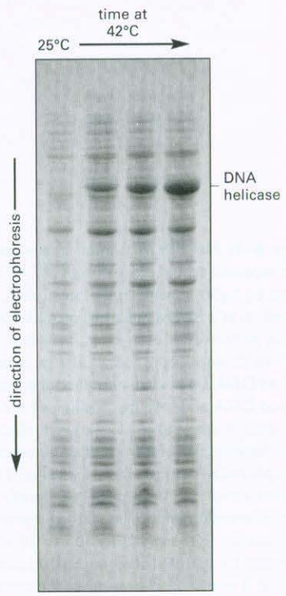
The vast majority of the thousands of different proteins in a eucaryotic cell, however, including many with crucially important functions, are present in very small amounts. For most of them it is extremely difficult, if not impossible, to obtain more than a few micrograms of pure material. One of the 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, ar 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.)



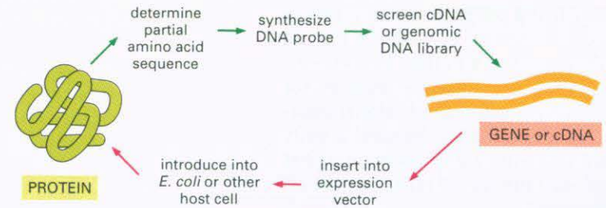


Figure 8–44 Knowledge of the molecular biology of cells makes it possible to experimentally move from gene to protein and from protein to gene. A small quantity of a purified protein is used to obtain a partial amino acid sequence. This provides sequence information that enables the corresponding gene 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.

The question of what a protein does inside a living cell is not a simple one to answer. Imagine isolating an uncharacterized protein and discovering that its structure and amino acid sequence suggest that it acts as a protein kinase. Simply knowing that the protein can add a phosphate group to serine residues, for example, does not reveal how it functions in a living organism. Additional information is required to understand the context in which the biochemical activity is used. Where is this kinase located in the cell and what are its protein targets? In which tissues is it active? Which pathways does it influence? What role does it have in the growth or development of the organism?

In this section, we discuss the methods currently used to characterize protein structure and function. We begin with an examination of the techniques used to determine the three-dimensional structure of purified proteins. We then discuss methods that are used to predict how a protein functions, based on its homology to other known proteins and its location inside the cell. Finally, because most proteins act in concert with other proteins, we present techniques for detecting protein–protein interactions. But these approaches only begin to define how a protein might work inside a cell. In the last section of this chapter, we discuss how genetic approaches are used to dissect and analyze the biological processes in which a given protein functions.

The Diffraction of X-rays by Protein Crystals Can Reveal a Protein's Exact Structure

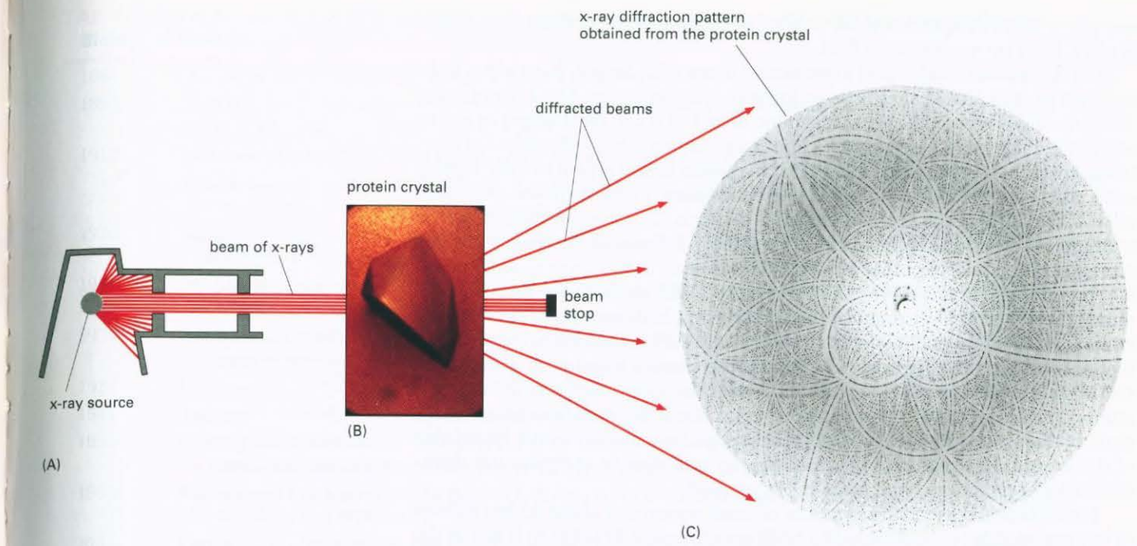
Starting with the amino acid sequence of a protein, one can often predict which secondary structural elements, such as membrane-spanning α helices, will be present in the protein. It is presently not possible, however, to deduce reliably the three-dimensional folded structure of a protein from its amino acid sequence unless its amino acid sequence is very similar to that of a protein whose three-dimensional structure is already known. The main technique that has been used to discover the three-dimensional structure of molecules, including proteins, at atomic resolution is **x-ray crystallography**.

X-rays, like light, are a form of electromagnetic radiation, but they have a much shorter wavelength, typically around 0.1 nm (the diameter of a hydrogen atom). If a narrow parallel beam of x-rays is directed at a sample of a pure protein, most of the x-rays pass straight through it. A small fraction, however, is scattered by the atoms in the sample. If the sample is a well-ordered crystal, the scattered waves reinforce one another at certain points and appear as diffraction spots when the x-rays are recorded by a suitable detector (Figure 8–45).

The position and intensity of each spot in the x-ray diffraction pattern contain information about the locations of the atoms in the crystal that gave rise to it. Deducing the three-dimensional structure of a large molecule from the diffraction pattern of its crystal is a complex task and was not achieved for a protein molecule until 1960. But in recent years x-ray diffraction analysis has become increasingly automated, and now the slowest step is likely to be the generation of suitable protein crystals. This requires large amounts of very pure protein and often involves years of trial and error, searching for the proper crystallization conditions. There are still many proteins, especially membrane proteins, that have so far resisted all attempts to crystallize them.

Analysis of the resulting diffraction pattern produces a complex three-dimensional electron-density map. Interpreting this map—translating its contours into a three-dimensional structure—is a complicated procedure that requires knowledge of the amino acid sequence of the protein. Largely by trial and error, the sequence and the electron-density map are correlated by computer to give the best possible fit. The reliability of the final atomic model depends on the resolution of the original crystallographic data: 0.5 nm resolution might produce a low-resolution map of the polypeptide backbone, whereas a resolution of 0.15 nm allows all of the non-hydrogen atoms in the molecule to be reliably positioned.

A complete atomic model is often too complex to appreciate directly, but simplified versions that show a protein's essential structural features can be



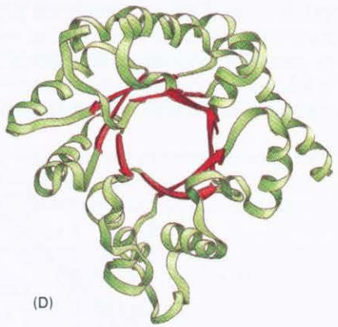

(D)

Figure 8–45 X-ray crystallography. (A) A narrow parallel beam of x-rays is directed at a well-ordered crystal (B). Shown here is a protein crystal of ribulose bisphosphate carboxylase, an enzyme with a central role in CO_2 fixation during photosynthesis. Some of the beam is scattered by the atoms in the crystal. The scattered waves reinforce one another at certain points and appear as a pattern of diffraction spots (C). This diffraction pattern, together with the amino acid sequence of the protein, can be used to produce an atomic model (D). The complete atomic model is hard to interpret, but this simplified version, derived from the x-ray diffraction data, shows the protein's structural features clearly (α helices, *green*; β strands, *red*). Note that the components pictured in A to D are not shown to scale. (B, courtesy of C. Branden; C, courtesy of J. Hajdu and I. Andersson; D, adapted from original provided by B. Furugren.)

readily derived from it (see Panel 3–2, pp. 138–139). The three-dimensional structures of about 10,000 different proteins have now been determined by x-ray crystallography or by NMR spectroscopy (see below)—enough to begin to see families of common structures emerging. These structures or protein folds often seem to be more conserved in evolution than are the amino acid sequences that form them (see Figure 3–15).

X-ray crystallographic techniques can also be applied to the study of macromolecular complexes. In a recent triumph, the method was used to solve the structure of the ribosome, a large and complex cellular machine made of several RNAs and more than 50 proteins (see Figure 6–64). The determination required the use of a synchrotron, a radiation source that generates x-rays with the intensity needed to analyze the crystals of such large macromolecular complexes.

Molecular Structure Can Also Be Determined Using Nuclear Magnetic Resonance (NMR) Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy has been widely used for many years to analyze the structure of small molecules. This technique is now also increasingly applied to the study of small proteins or protein domains. Unlike x-ray crystallography, NMR does not depend on having a crystalline

sample; it simply requires a small volume of concentrated protein solution that is placed in a strong magnetic field.

Certain atomic nuclei, and in particular those of hydrogen, have a magnetic moment or spin: that is, they have an intrinsic magnetization, like a bar magnet. The spin aligns along the strong magnetic field, but it can be changed to a misaligned, excited state in response to applied radiofrequency (RF) pulses of electromagnetic radiation. When the excited hydrogen nuclei return to their aligned state, they emit RF radiation, which can be measured and displayed as a spectrum. The nature of the emitted radiation depends on the environment of each hydrogen nucleus, and if one nucleus is excited, it influences the absorption and emission of radiation by other nuclei that lie close to it. It is consequently possible, by an ingenious elaboration of the basic NMR technique known as two-dimensional NMR, to distinguish the signals from hydrogen nuclei in different amino acid residues and to identify and measure the small shifts in these signals that occur when these hydrogen nuclei lie close enough together to interact: the size of such a shift reveals the distance between the interacting pair of hydrogen atoms. In this way NMR can give information about the distances between the parts of the protein molecule. By combining this information with a knowledge of the amino acid sequence, it is possible in principle to compute the three-dimensional structure of the protein (Figure 8–46).

For technical reasons the structure of small proteins of about 20,000 daltons or less can readily be determined by NMR spectroscopy. Resolution is lost as the size of a macromolecule increases. But recent technical advances have now pushed the limit to about 100,000 daltons, thereby making the majority of proteins accessible for structural analysis by NMR.

The NMR method is especially useful when a protein of interest has resisted attempts at crystallization, a common problem for many membrane proteins. Because NMR studies are performed in solution, this method also offers a convenient means of monitoring changes in protein structure, for example during protein folding or when a substrate binds to the protein. NMR is also used widely to investigate molecules other than proteins and is valuable, for example, as a

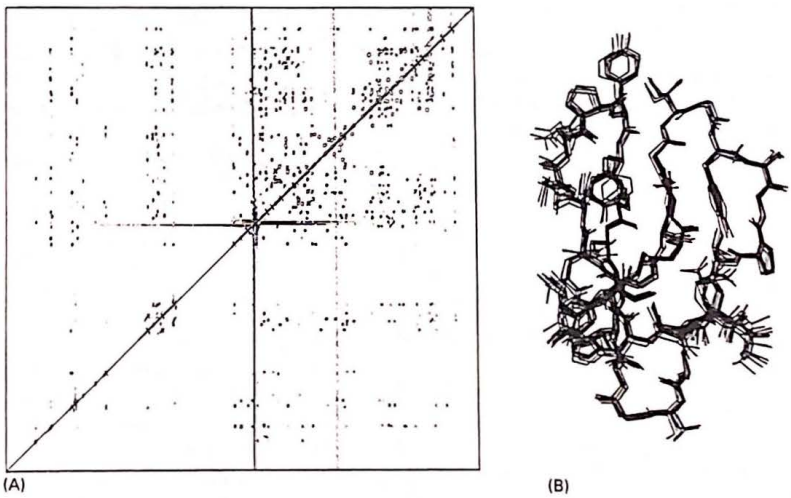


Figure 8–46 NMR spectroscopy. (A) An example of the data from an NMR machine. This two-dimensional NMR spectrum is derived from the C-terminal domain of the enzyme cellulase. The spots represent interactions between hydrogen atoms that are near neighbors in the protein and hence reflects the distance that separates them. Complex computing methods, in conjunction with the known amino acid sequence, enable possible compatible structures to be derived. In (B) 10 structures, which all satisfy the distance constraints equally well, are shown superimposed on one another, giving a good indication of the probable three-dimensional structure. (Courtesy of P. Kraulis.)

TABLE 8–8 Landmarks in the Development of X-ray Crystallography and NMR and Their Application to Biological Molecules

1864	**Hoppe-Seyler** crystallizes, and names, the protein hemoglobin.
1895	**Röntgen** observes that a new form of penetrating radiation, which he names x-rays, is produced when cathode rays (electrons) hit a metal target.
1912	**Von Laue** obtains the first x-ray diffraction patterns by passing x-rays through a crystal of zinc sulfide.
	W.L. Bragg proposes a simple relationship between an x-ray diffraction pattern and the arrangement of atoms in a crystal that produce the pattern.
1926	**Summer** obtains crystals of the enzyme urease from extracts of jack beans and demonstrates that proteins possess catalytic activity.
1931	**Pauling** publishes his first essays on 'The Nature of the Chemical Bond,' detailing the rules of covalent bonding.
1934	**Bernal and Crowfoot** present the first detailed x-ray diffraction patterns of a protein obtained from crystals of the enzyme pepsin.
1935	**Patterson** develops an analytical method for determining interatomic spacings from x-ray data.
1941	**Astbury** obtains the first x-ray diffraction pattern of DNA.
1951	**Pauling and Corey** propose the structure of a helical conformation of a chain of L-amino acids—the α helix—and the structure of the β sheet, both of which were later found in many proteins.
1953	**Watson and Crick** propose the double-helix model of DNA, based on x-ray diffraction patterns obtained by **Franklin and Wilkins**.
1954	**Perutz** and colleagues develop heavy-atom methods to solve the phase problem in protein crystallography.
1960	**Kendrew** describes the first detailed structure of a protein (sperm whale myoglobin) to a resolution of 0.2 nm, and **Perutz** presents a lower-resolution structure of the larger protein hemoglobin.
1966	**Phillips** describes the structure of lysozyme, the first enzyme to have its structure analyzed in detail.
1971	**Jeener** proposes the use of two-dimensional NMR, and **Wuthrich** and colleagues first use the method to solve a protein structure in the early 1980s.
1976	**Kim and Rich** and **Klug** and colleagues describe the detailed three-dimensional structure of tRNA determined by x-ray diffraction.
1977–1978	**Holmes and Klug** determine the structure of tobacco mosaic virus (TMV), and **Harrison** and **Rossman** determine the structure of two small spherical viruses.
1985	**Michel, Deisenhofer** and colleagues determine the first structure of a transmembrane protein (a bacterial reaction center) by x-ray crystallography. **Henderson** and colleagues obtain the structure of bacteriorhodopsin, a transmembrane protein, by high-resolution electron-microscopy methods between 1975 and 1990.

method to determine the three-dimensional structures of RNA molecules and the complex carbohydrate side chains of glycoproteins.

Some landmarks in the development of x-ray crystallography and NMR are listed in Table 8–8.

Sequence Similarity Can Provide Clues About Protein Function

Thanks to the proliferation of protein and nucleic acid sequences that are catalogued in genome databases, the function of a gene—and its encoded protein—can often be predicted by simply comparing its sequence with those of previously characterized genes. Because amino acid sequence determines protein structure and structure dictates biochemical function, proteins that share a similar amino acid sequence usually perform similar biochemical functions, even when they are found in distantly related organisms. At present, determining what a newly discovered protein does therefore usually begins with a search for previously identified proteins that are similar in their amino acid sequences.

Searching a collection of known sequences for homologous genes or proteins is typically done over the World-Wide Web, and it simply involves selecting a database and entering the desired sequence. A sequence alignment program—the most popular are BLAST and FASTA—scans the database for similar sequences by sliding the submitted sequence along the archived sequences until a cluster of residues falls into full or partial alignment (Figure 8–17). The results of even a complex search—which can be performed on either a

```
Score =  399 bits (1025), Expect = e-111
Identities = 198/290 (68%), Positives = 241/290 (82%), Gaps = 1/290

Query: 57  MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREISLLKELNH 116
           ME ++KVEKIGEGTYGVVYKA +K T E +ALKKIRL+ E EGVPSTAIREISLLKE+NH
Sbjct: 1   MEQYEKVEKIGEGTYGVVYKALDKATNETIALKKIRLEQEDEGVPSTAIREISLLKEMNH 60

Query: 117 PNIVKLLDVIHTENKLYLVFEFLHQDLKKFMDASALTGIPLPLIKSYLFQLLQGLAFCHS 176
            NIV+L  DV+H+E ++YLVFE+L  DLKKFMD+          LIKSYL+Q+L  G+A+CHS
Sbjct: 61  GNIVRLHDVVHSEKRLYLVFEYLDLDLKKFMDSCPEFAKNPTLIKSYLYQILHGVAYCHS 120

Query: 177 HRVLHRDLKPQNLLINTE-GAIKLADFGLARAFGVPVRTYTHEVVTLWYRAPEILLGCKY 235
           HRVLHRDLKPQNLLI+     A+KLADFGLARAFG+PVRT+THEVVTLWYRAPEILLG +
Sbjct: 121 HRVLHRDLKPQNLLIDRRTNALKLADFGLARAFGIPVRTFTHEVVTLWYRAPEILLGARQ 180

Query: 236 YSTAVDIWSLGCIFAEMVTRRALFPGDSEIDQLFRIFRTLGTPDEVVWPGVTSMPDYKPS 295
           YST VD+WS+GCIFAEMV ++ LFPGDSEID+LF+IFR LGTP+E  WPGV+ +PD+K +
Sbjct: 181 YSTPVDVWSVGCIFAEMVNQKPLFPGDSEIDELFKIFRILGTPNEQSWPGVSCLPDFKTA 240

Query: 296 FPKWARQDFSKVVPPLDEDGRSLLSQMLHYDPNKRISAKAALAHPFFQDV 345
           FP+W  QD + VVP LD  G  LLS+ML Y+P+KRI+A+ AL H +F+D+
Sbjct: 241 FPRWQAQDLATVVPNLDPAGLDLLSKMLRYEPSKRITARQALEHEYFKDL 290
```

Figure 8–47 Results of a BLAST search. Sequence databases can be searched to find similar amino acid or nucleic acid sequences. Here a search for proteins similar to the human cell-cycle regulatory protein cdc2 *(Query)* locates maize cdc2 *(Subject)*, which is 68% identical (and 82% similar) to human cdc2 in its amino acid sequence. The alignment begins at residue 57 of the Query protein, suggesting that the human protein has an N-terminal region that is absent from the maize protein. The *green* blocks indicate differences in sequence, and the *yellow* bar summarizes the similarities: when the two amino acid sequences are identical, the residue is shown; conservative amino acid substitutions are indicated by a plus sign (+). Only one small gap has been introduced—indicated by the *red arrow* at position 194 in the Query sequence—to align the two sequences maximally. The alignment score *(Score)* takes into account penalties for substitution and gaps; the higher the alignment score, the better the match. The significance of the alignment is reflected in the *Expectation* (E) value, which represents the number of alignments with scores equal to or better than the given score that are expected to occur by chance. The lower the E value, the more significant the match; the extremely low value here indicates certain significance. E values much higher than 0.1 are unlikely to reflect true relatedness.

nucleotide or an amino acid sequence—are returned within minutes. Such comparisons can be used to predict the functions of individual proteins, families of proteins, or even the entire protein complement of a newly sequenced organism.

In the end, however, the predictions that emerge from sequence analysis are often only a tool to direct further experimental investigations.

Fusion Proteins Can Be Used to Analyze Protein Function and to Track Proteins in Living Cells

The location of a protein within the cell often suggests something about its function. Proteins that travel from the cytoplasm to the nucleus when a cell is exposed to a growth factor, for example, may have a role in regulating gene expression in response to that factor. A protein often contains short amino acid sequences that determine its location in a cell. Most nuclear proteins, for example, contain one or more specific short sequences of amino acids that serve as signals for their import into the nucleus after their synthesis in the cytosol (discussed in Chapter 12). These special regions of the protein can be identified by fusing them to an easily detectable protein that lacks such regions and then following the behavior of this surrogate protein in a cell. Such fusion proteins can be readily produced by the recombinant DNA techniques discussed previously.

Another common strategy used both to follow proteins in cells and to purify them rapidly is *epitope tagging*. In this case, a fusion protein is produced that contains the entire protein being analyzed plus a short peptide of 8 to 12 amino acids (an "epitope") that can be recognized by a commercially available antibody. The fusion protein can therefore be specifically detected, even in the presence of a large excess of the normal protein, using the anti-epitope antibody and a labeled secondary antibody that can be monitored by light or electron microscopy (Figure 8–48).

Today large numbers of proteins are being tracked in living cells by using a fluorescent marker called **green fluorescent protein** (**GFP**). Tagging proteins with GFP is as simple as attaching the gene for GFP to one end of the gene that encodes a protein of interest. In most cases, the resulting GFP fusion protein behaves in the same way as the original protein, and its movement can be monitored by following its fluorescence inside the cell by fluorescence microscopy.

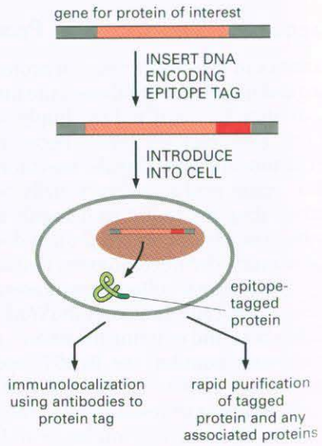


Figure 8–48 Epitope tagging allows the localization or purification of proteins. Using standard genetic engineering techniques, a short epitope tag can be added to a protein of interest. The resulting protein contains the protein being analyzed plus a short peptide that can be recognized by commercially available antibodies. The labeled antibody can be used to follow the cellular localization of the protein or to purify it by immunoprecipitation or affinity chromatography.

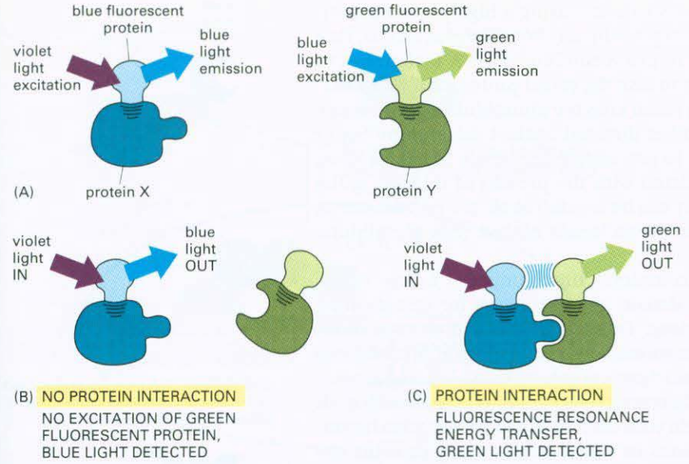


Figure 8–49 Fluorescence resonance energy transfer (FRET). To determine whether (and when) two proteins interact inside the cell, the proteins are first produced as fusion proteins attached to different variants of GFP. (A) In this example, protein X is coupled to a blue fluorescent protein, which is excited by violet light (370–440 nm) and emits blue light (440–480 nm); protein Y is coupled to a green fluorescent protein, which is excited by blue light and emits green light (510 nm). (B) If protein X and Y do not interact, illuminating the sample with violet light yields fluorescence from the blue fluorescent protein only. (C) When protein X and protein Y interact, FRET can now occur. Illuminating the sample with violet light excites the blue fluorescent protein, whose emission in turn excites the green fluorescent protein, resulting in an emission of green light. The fluorochromes must be quite close together—within about 1–10 nm of one another—for FRET to occur. Because not every molecule of protein X and protein Y is bound at all times, some blue light may still be detected. But as the two proteins begin to interact, emission from the donor GFP falls as the emission from the acceptor GFP rises.

The GFP fusion protein strategy has become a standard way to determine the distribution and dynamics of any protein of interest in living cells. We discuss its use further in Chapter 9.

GFP, and its derivatives of different color, can also be used to monitor protein–protein interactions. In this application, two proteins of interest are each labeled with a different fluorochrome, such that the emission spectrum of one fluorochrome overlaps the absorption spectrum of the second fluorochrome. If the two proteins—and their attached fluorochromes—come very close to each other (within about 1–10 nm), the energy of the absorbed light will be transferred from one fluorochrome to the other. The energy transfer, called **fluorescence resonance energy transfer** (**FRET**), is determined by illuminating the first fluorochrome and measuring emission from the second (Figure 8–49). By using two different spectral variants of GFP as the fluorochromes in such studies, one can monitor the interaction of any two protein molecules inside a living cell.

Affinity Chromatography and Immunoprecipitation Allow Identification of Associated Proteins

Because most proteins in the cell function as part of a complex with other proteins, an important way to begin to characterize their biological roles is to identify their binding partners. If an uncharacterized protein binds to a protein whose role in the cell is understood, its function is likely to be related. For example, if a protein is found to be part of the proteasome complex, it is likely to be involved somehow in degrading damaged or misfolded proteins.

Protein affinity chromatography is one method that can be used to isolate and identify proteins that interact physically. To capture interacting proteins, a target protein is attached to polymer beads that are packed into a column. Cellular proteins are washed through the column and those proteins that interact with the target adhere to the affinity matrix (see Figure 8–11C). These proteins can then be eluted and their identity determined by mass spectrometry or another suitable method.

Perhaps the simplest method for identifying proteins that bind to one another tightly is **co-immunoprecipitation**. In this case, an antibody is used to recognize a specific target protein; affinity reagents that bind to the antibody and are coupled to a solid matrix are then used to drag the complex out of solution to the bottom of a test tube. If this protein is associated tightly enough with another protein when it is captured by the antibody, the partner precipitates as well. This method is useful for identifying proteins that are part of a complex inside cells, including those that interact only transiently—for example when cells are stimulated by signal molecules (discussed in Chapter 15).

Co-immunoprecipitation techniques require having a highly specific antibody against a known cellular protein target, which is not always available. One way to overcome this requirement is to use recombinant DNA techniques to add an epitope tag (see Figure 8–48) or to fuse the target protein to a well-characterized marker protein, such as the small enzyme glutathione S-transferase (GST). Commercially available antibodies directed against the epitope tag or the marker protein can then be used to precipitate the whole fusion protein, including any cellular proteins associated with the protein of interest. If the protein is fused to GST, antibodies may not be needed at all: the hybrid and its binding partners can be readily selected on beads coated with glutathione (Figure 8–50).

In addition to capturing protein complexes on columns or in test tubes, researchers are also developing high-density protein arrays for investigating protein function and protein interactions. These arrays, which contain thousands of different proteins or antibodies spotted onto glass slides or immobilized in tiny wells, allow one to examine the biochemical activities and binding profiles of a large number of proteins at once. To examine protein interactions with such an array, one incubates a labeled protein with each of the target proteins immobilized on the slide and then determines to which of the many proteins the labeled molecule binds.

Protein–Protein Interactions Can Be Identified by Use of the Two-Hybrid System

Methods such as co-immunoprecipitation and affinity chromatography allow the physical isolation of interacting proteins. A successful isolation yields a protein whose identity must then be ascertained by mass spectrometry, and whose gene must be retrieved and cloned before further studies characterizing its activity—or the nature of the protein–protein interaction—can be performed.

Other techniques allow the simultaneous isolation of interacting proteins along with the genes that encode them. The first method we discuss, called the **two-hybrid system**, uses a reporter gene to detect the physical interaction of a pair of proteins inside a yeast cell nucleus. This system has been designed so that when a target protein binds to another protein in the cell, their interaction brings together two halves of a transcriptional activator, which is then able to switch on the expression of the reporter gene.

The technique takes advantage of the modular nature of gene activator proteins (see Figure 7–42). These proteins both bind to DNA and activate transcription—activities that are often performed by two separate protein domains. Using recombinant DNA techniques, the DNA sequence that codes for a target protein is fused with DNA that encodes the DNA-binding domain of a gene activator protein. When this construct is introduced into yeast, the cells produce the target protein attached to this DNA-binding domain (Figure 8–51). This protein binds to the regulatory region of a reporter gene, where it serves as "bait" to fish for proteins that interact with the target protein inside a yeast cell. To prepare a set of potential binding partners, DNA encoding the activation domain of a gene activator protein is ligated to a large mixture of DNA fragments from a cDNA library. Members of this collection of genes—the "prey"—are introduced individually into yeast cells containing the bait. If the yeast cell has received a DNA clone that expresses a prey partner for the bait protein, the two halves of a transcriptional activator are united, switching on the reporter gene (see Figure 8–51). Cells that express this reporter are selected and grown, and the gene (or gene fragment) encoding the prey protein is retrieved and identified through nucleotide sequencing.

Although it sounds complex, the two-hybrid system is relatively simple to use in the laboratory. Although the protein–protein interactions occur in the yeast cell nucleus, proteins from every part of the cell and from any organism can be studied in this way. Of the thousands of protein–protein interactions that have been catalogued in yeast, half have been discovered with such two-hybrid screens.

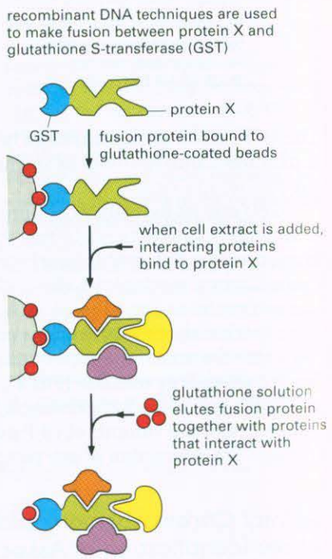


Figure 8–50 Purification of protein complexes using a GST-tagged fusion protein. GST fusion proteins, generated by standard recombinant DNA techniques, can be captured on an affinity column containing beads coated with glutathione. To look for proteins that bind to protein X, cell extracts can be passed through this column. The hybrid protein and its binding partners can then be eluted with glutathione. The identities of these interacting proteins can be determined by mass spectrometry (see Figure 8–20). In an alternative approach, a cell extract can be made from a cell producing the GST fusion protein and passed directly through the glutathione affinity column. The GST fusion protein, along with proteins that have associated with it in the cell, are thereby retained. Affinity columns can also be made to contain antibodies against GST or another convenient small protein or epitope tag (see Figure 8–48).

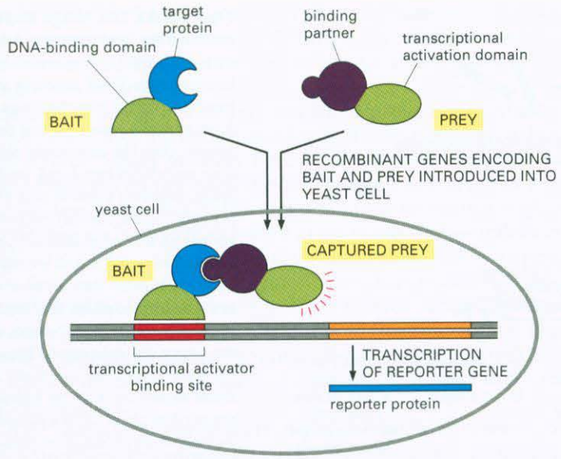


Figure 8–51 The yeast two-hybrid system for detecting protein–protein interactions. The target protein is fused to a DNA-binding domain that localizes it to the regulatory region of a reporter gene as "bait." When this target protein binds to another specially designed protein in the cell nucleus ("prey"), their interaction brings together two halves of a transcriptional activator, which then switches on the expression of the reporter gene. The reporter gene is often one that will permit growth on a selective medium. Bait and prey fusion proteins are generated by standard recombinant DNA techniques. In most cases, a single bait protein is used to fish for interacting partners among a large collection of prey proteins produced by ligating DNA encoding the activation domain of a transcriptional activator to a large mixture of DNA fragments from a cDNA library.

The two-hybrid system can be scaled up to map the interactions that occur among all of the proteins produced by an organism. In this case, a set of bait fusions is produced for every cellular protein, and each of these constructs is introduced into a separate yeast cell. These cells are then mated to yeast containing the prey library. Those rare cells that are positive for a protein–protein interaction are then characterized. In this way a protein linkage map has been generated for most of the 6,000 proteins in yeast (see Figure 3–78), and similar projects are underway to catalog the protein interactions in *C. elegans* and *Drosophila*.

A related technique, called a *reverse two-hybrid system*, can be used to identify mutations—or chemical compounds—that are able to disrupt specific protein–protein interactions. In this case the reporter gene can be replaced by a gene that kills cells in which the bait and prey proteins interact. Only those cells in which the proteins no longer bind—because an engineered mutation or a test compound prevents them from doing so—can survive. Like knocking out a gene (which we discuss shortly), eliminating a particular molecular interaction can reveal something about the role of the participating proteins in the cell. In addition, compounds that selectively interrupt protein interactions can be medically useful: a drug that prevents a virus from binding to its receptor protein on human cells could help people to avoid infections, for example.

Phage Display Methods Also Detect Protein Interactions

Another powerful method for detecting protein–protein interactions involves introducing genes into a virus that infects the *E. coli* bacterium (a bacteriophage, or "phage"). In this case the DNA encoding the protein of interest (or a smaller peptide fragment of this protein) is fused with a gene encoding one of the proteins that forms the viral coat. When this virus infects *E. coli*, it replicates, producing phage particles that display the hybrid protein on the outside of their coats (Figure 8–52A). This bacteriophage can then be used to fish for binding partners in a large pool of potential target proteins.

However, the most powerful use of this **phage display** method allows one to screen large collections of proteins or peptides for binding to selected targets. This approach requires first generating a library of fusion proteins, much like the prey library in the two-hybrid system. This collection of phage is then screened for binding to a purified protein of interest. For example, the phage library can be passed through an affinity column containing an immobilized target protein. Viruses that display a protein or peptide that binds tightly to the target are captured on the column and can be eluted with excess target protein. Those phage containing a DNA fragment that encodes an interacting protein or peptide are

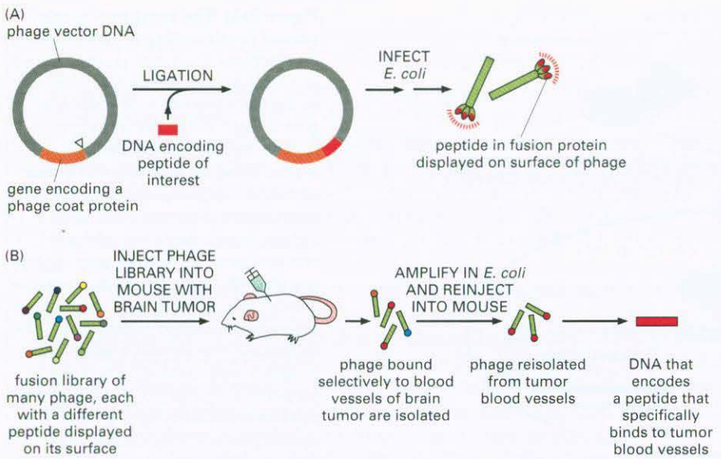


Figure 8–52 The phage display method for investigating protein interactions. (A) Preparation of the bacteriophage. DNA encoding the desired peptide is ligated into the phage vector, fused with the gene encoding the viral protein coat. The engineered phage are then introduced into *E. coli*, which produce phage displaying a hybrid coat protein that contains the peptide. (B) Phage libraries containing billions of different peptides can also be generated. In this example, the library is injected into a mouse with a brain tumor and phage that bind selectively to the blood vessels that supply the tumor are isolated and amplified. A peptide that binds specifically to tumor blood vessels can then be isolated from the purified phage. Such a peptide could be used to target drugs or toxins to the tumor.

collected and allowed to replicate in *E. coli*. The DNA from each phage can then be recovered and its nucleotide sequence determined to identify the protein or peptide partner that bound to the target protein. A similar technique has been used to isolate peptides that bind specifically to the inside of the blood vessels associated with human tumors. These peptides are presently being tested as agents for delivering therapeutic anti-cancer compounds directly to such tumors (Figure 8–52B).

Phage display has also been used to generate monoclonal antibodies that recognize a specific target molecule or cell. In this case, a library of phage expressing the appropriate parts of antibody molecules is screened for those phage that bind to a target antigen.

Protein Interactions Can Be Monitored in Real Time Using Surface Plasmon Resonance

Once two proteins—or a protein and a small molecule—are known to associate, it becomes important to characterize their interaction in more detail. Proteins can bind to one another permanently, or engage in transient encounters in which proteins remain associated only temporarily. These dynamic interactions are often regulated through reversible modifications (such as phosphorylation), through ligand binding, or through the presence or absence of other proteins that compete for the same binding site.

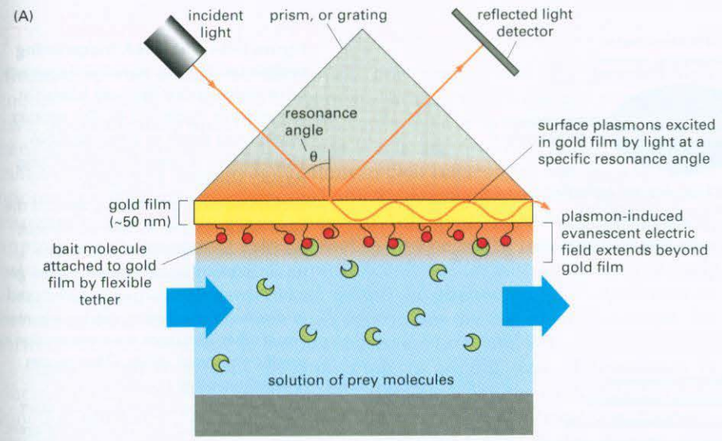
To begin to understand these intricacies, one must determine how tightly two proteins associate, how slowly or rapidly molecular complexes assemble and break down over time, and how outside influences can affect these parameters. As we have seen in this chapter, there are many different techniques available to study protein–protein interactions, each with its individual advantages and disadvantages. One particularly useful method for monitoring the dynamics of protein association is called **surface plasmon resonance (SPR)**. The SPR method has been used to characterize a wide variety of molecular interactions, including antibody–antigen binding, ligand–receptor coupling, and the binding of proteins to DNA, carbohydrates, small molecules, and other proteins.

SPR detects binding interactions by monitoring the reflection of a beam of light off the interface between an aqueous solution of potential binding molecules and a biosensor surface carrying immobilized bait protein. The bait protein is attached to a very thin layer of metal that coats one side of a glass prism (Figure 8–53). A light beam is passed through the prism; at a certain angle, called the resonance angle, some of the energy from the light interacts with the cloud of electrons in the metal film, generating a plasmon—an oscillation of the electrons at right angles to the plane of the film, bouncing up and down between its upper and lower surfaces like a weight on a spring. The plasmon, in turn,

generates an electrical field that extends a short distance—about the wavelength of the light—above and below the metal surface. Any change in the composition of the environment within the range of the electrical field causes a measurable change in the resonance angle.

To measure binding, a solution containing proteins (or other molecules) that might interact with the immobilized bait protein is allowed to flow past the biosensor surface. When proteins bind to the bait, the composition of the molecular complexes on the metal surface change, causing a change in the resonance angle (see Figure 8–53). The changes in the resonance angle are monitored in real time and reflect the kinetics of the association—or dissociation—of molecules with the bait protein. The association rate (k_{on}) is measured as the molecules interact, and the dissociation rate (k_{off}) is determined as buffer washes the bound molecules from the sensor surface. A binding constant (K) is calculated by dividing k_{off} by k_{on}. In addition to determining the kinetics, SPR can be used to determine the number of molecules that are bound in each complex: the magnitude of the SPR signal change is proportional to the mass of the immobilized complex.

The SPR method is particularly useful because it requires only small amounts of proteins, the proteins do not have to be labeled in any way, and protein–protein interactions can be monitored in real time.



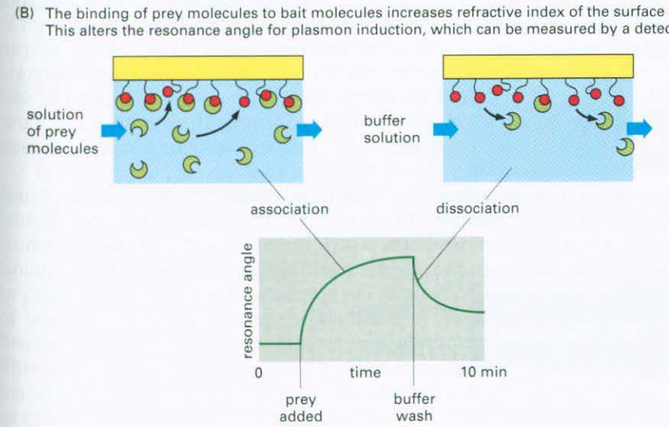


Figure 8–53 Surface plasmon resonance. (A) SPR can detect binding interactions by monitoring the reflection of a beam of light off the interface between an aqueous solution of potential binding molecules *(green)* and a biosensor surface coated with an immobilized bait protein *(red)*. (B) A solution of prey proteins is allowed to flow past the immobilized bait protein. Binding of prey molecules to bait proteins produces a measurable change in the resonance angle. These changes, monitored in real time, reflect the association and dissociation of the molecular complexes.

DNA Footprinting Reveals the Sites Where Proteins Bind on a DNA Molecule

So far we have concentrated on examining protein–protein interactions. But some proteins act by binding to DNA. Most of these proteins have a central role in determining which genes are active in a particular cell by binding to regulatory DNA sequences, which are usually located outside the coding regions of a gene.

In analyzing how such a protein functions, it is important to identify the specific nucleotide sequences to which it binds. A method used for this purpose is called **DNA footprinting**. First, a pure DNA fragment that is labeled at one end with ^{32}P is isolated (see Figure 8–24B); this molecule is then cleaved with a nuclease or a chemical that makes random single-stranded cuts in the DNA. After the DNA molecule is denatured to separate its two strands, the resultant fragments from the labeled strand are separated on a gel and detected by autoradiography. The pattern of bands from DNA cut in the presence of a DNA-binding protein is compared with that from DNA cut in its absence. When the protein is present, it covers the nucleotides at its binding site and protects their phosphodiester bonds from cleavage. As a result, the labeled fragments that terminate in the binding site are missing, leaving a gap in the gel pattern called a "footprint" (Figure 8–54). Similar methods can be used to determine the binding sites of proteins on RNA.

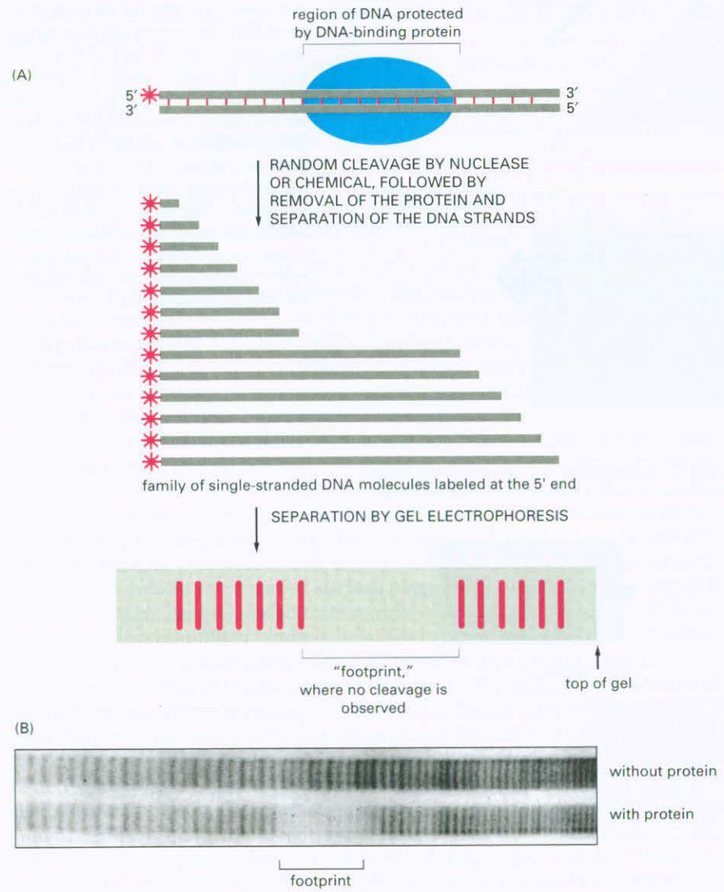


Figure 8–54 The DNA footprinting technique. (A) This technique requires a DNA molecule that has been labeled at one end (see Figure 8–24B). The protein shown binds tightly to a specific DNA sequence that is seven nucleotides long, thereby protecting these seven nucleotides from the cleaving agent. If the same reaction were performed without the DNA-binding protein, a complete ladder of bands would be seen on the gel (not shown). (B) An actual footprint used to determine the binding site for a human protein that stimulates the transcription of specific eucaryotic genes. These results locate the binding site about 60 nucleotides upstream from the start site for RNA synthesis. The cleaving agent was a small, iron-containing organic molecule that normally cuts at every phosphodiester bond with nearly equal frequency. (B, courtesy of Michele Sawadogo and Robert Roeder.)

Summary

Many powerful techniques are used to study the structure and function of a protein. To determine the three-dimensional structure of a protein at atomic resolution, large proteins have to be crystallized and studied by x-ray diffraction. The structure of small proteins in solution can be determined by nuclear magnetic resonance analysis. Because proteins with similar structures often have similar functions, the biochemical activity of a protein can sometimes be predicted by searching for known proteins that are similar in their amino acid sequences.

Further clues to the function of a protein can be derived from examining its subcellular distribution. Fusion of the protein with a molecular tag, such as the green fluorescent protein (GFP), allows one to track its movement inside the cell. Proteins that enter the nucleus and bind to DNA can be further characterized by footprint analysis, a technique used to determine which regulatory sequences the protein binds to as it controls gene transcription.

All proteins function by binding to other proteins or molecules, and many methods exist for studying protein–protein interactions and identifying potential protein partners. Either protein affinity chromatography or co-immunoprecipitation by antibodies directed against a target protein will allow physical isolation of interacting proteins. Other techniques, such as the two-hybrid system or phage display, permit the simultaneous isolation of interacting proteins and the genes that encode them. The identity of the proteins recovered from any of these approaches is then ascertained by determining the sequence of the protein or its corresponding gene.

STUDYING GENE EXPRESSION AND FUNCTION

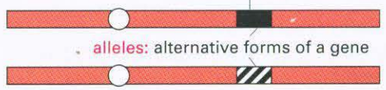
Ultimately, one wishes to determine how genes—and the proteins they encode—function in the intact organism. Although it may sound counterintuitive, one of the most direct ways to find out what a gene does is to see what happens to the organism when that gene is missing. Studying mutant organisms that have acquired changes or deletions in their nucleotide sequences is a time-honored practice in biology. Because mutations can interrupt cellular processes, mutants often hold the key to understanding gene function. In the classical approach to the important field of **genetics**, one begins by isolating mutants that have an interesting or unusual appearance: fruit flies with white eyes or curly wings, for example. Working backward from the **phenotype**—the appearance or behavior of the individual—one then determines the organism's **genotype**, the form of the gene responsible for that characteristic (Panel 8–1).

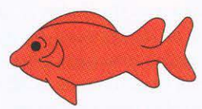
Today, with numerous genome projects adding tens of thousands of nucleotide sequences to the public databases each day, the exploration of gene function often begins with a DNA sequence. Here the challenge is to translate sequence into function. One approach, discussed earlier in the chapter, is to search databases for well-characterized proteins that have similar amino acid sequences to the protein encoded by a new gene, and from there employ some of the methods described in the previous section to explore the gene's function further. But to tackle directly the problem of how a gene functions in a cell or organism, the most effective approach involves studying mutants that either lack the gene or express an altered version of it. Determining which cellular processes have been disrupted or compromised in such mutants will then frequently provide a window to a gene's biological role.

In this section, we describe several different approaches to determining a gene's function, whether one starts from a DNA sequence or from an organism with an interesting phenotype. We begin with the classical genetic approach to studying genes and gene function. These studies start with a *genetic screen* for isolating mutants of interest, and then proceed toward identification of the gene or genes responsible for the observed phenotype. We then review the collection of techniques that fall under the umbrella of *reverse genetics*, in which one begins with a gene or gene sequence and attempts to determine its function. This approach often involves some intelligent guesswork—searching for homologous sequences and determining when and where a gene is expressed—as well as generating mutant organisms and characterizing their phenotype.

GENES AND PHENOTYPES

Gene: a functional unit of inheritance, usually corresponding to the segment of DNA coding for a single protein.
Genome: an organism's set of genes.



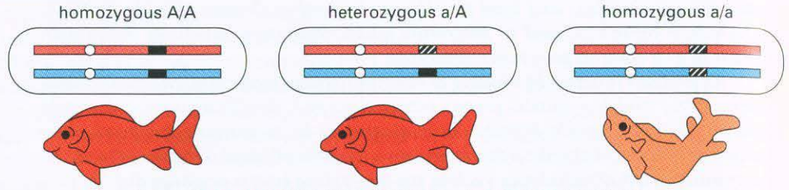


Wild-type: the normal, naturally occurring type

Mutant: differing from the wild-type because of a genetic change (a mutation)

GENOTYPE: the specific set of alleles forming the genome of an individual

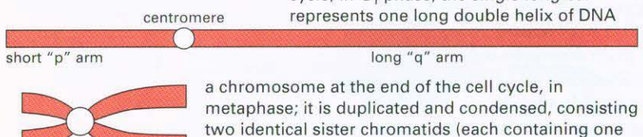
PHENOTYPE: the visible character of the individual



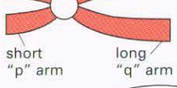
allele A is dominant (relative to a); allele a is recessive (relative to A)

In the example above, the phenotype of the heterozygote is the same as that of one of the homozygotes; in cases where it is different from both, the two alleles are said to be co-dominant.

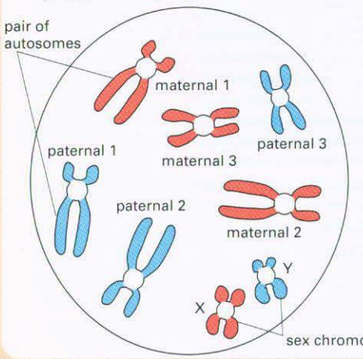
CHROMOSOMES



a chromosome at the beginning of the cell cycle, in G_1 phase; the single long bar represents one long double helix of DNA

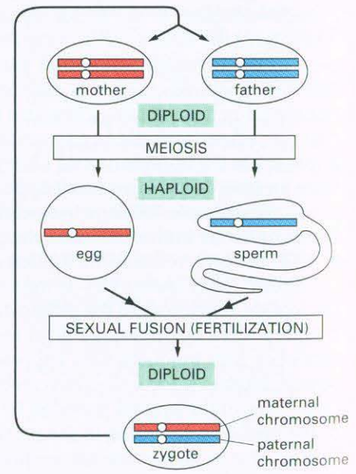


a chromosome at the end of the cell cycle, in metaphase; it is duplicated and condensed, consisting of two identical sister chromatids (each containing one DNA double helix) joined at the centromere.



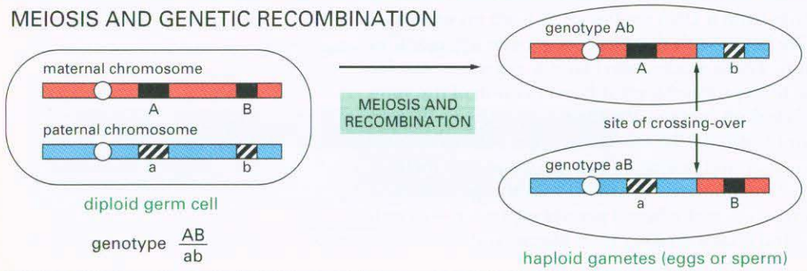
A normal diploid chromosome set, as seen in a metaphase spread, prepared by bursting open a cell at metaphase and staining the scattered chromosomes. In the example shown schematically here, there are three pairs of autosomes (chromosomes inherited symmetrically from both parents, regardless of sex) and two sex chromosomes—an X from the mother and a Y from the father. The numbers and types of sex chromosomes and their role in sex determination are variable from one class of organisms to another, as is the number of pairs of autosomes.

THE HAPLOID–DIPLOID CYCLE OF SEXUAL REPRODUCTION



For simplicity, the cycle is shown for only one chromosome/chromosome pair.

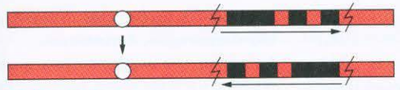
MEIOSIS AND GENETIC RECOMBINATION



The greater the distance between two loci on a single chromosome, the greater is the chance that they will be separated by crossing-over occurring at a site between them. If two genes are thus reassorted in x% of gametes, they are said to be separated on a chromosome by a genetic map distance of x map units (or x centimorgans).

TYPES OF MUTATIONS

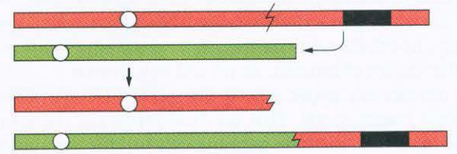
POINT MUTATION: maps to a single site in the genome, corresponding to a single nucleotide pair or a very small part of a single gene



INVERSION: inverts a segment of a chromosome

lethal mutation: causes the developing organism to die prematurely.
conditional mutation: produces its phenotypic effect only under certain conditions, called the *restrictive* conditions. Under other conditions—the *permissive* conditions—the effect is not seen. For a *temperature-sensitive* mutation, the restrictive condition typically is high temperature, while the permissive condition is low temperature.
loss-of-function mutation: either reduces or abolishes the activity of the gene. These are the commonest class of mutations. Loss-of-function mutations are usually *recessive*—the organism can usually function normally as long as it retains at least one normal copy of the affected gene.
null mutation: a loss-of-function mutation that completely abolishes the activity of the gene.

DELETION: deletes a segment of a chromosome



TRANSLOCATION: breaks off a segment from one chromosome and attaches it to another

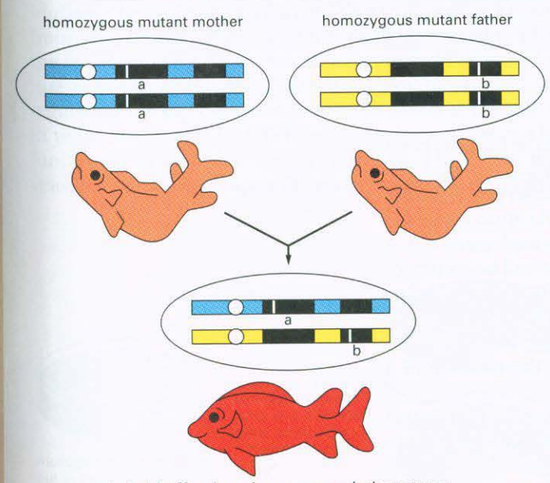
gain-of-function mutation: increases the activity of the gene or makes it active in inappropriate circumstances; these mutations are usually *dominant*.
dominant negative mutation: dominant-acting mutation that blocks gene activity, causing a loss-of-function phenotype even in the presence of a normal copy of the gene. This phenomenon occurs when the mutant gene product interferes with the function of the normal gene product.
suppressor mutation: suppresses the phenotypic effect of another mutation, so that the double mutant seems normal. An *intragenic* suppressor mutation lies within the gene affected by the first mutation; an *extragenic* suppressor mutation lies in a second gene—often one whose product interacts directly with the product of the first.

TWO GENES OR ONE?

Given two mutations that produce the same phenotype, how can we tell whether they are mutations in the same gene? If the mutations are recessive (as they most often are), the answer can be found by a complementation test.

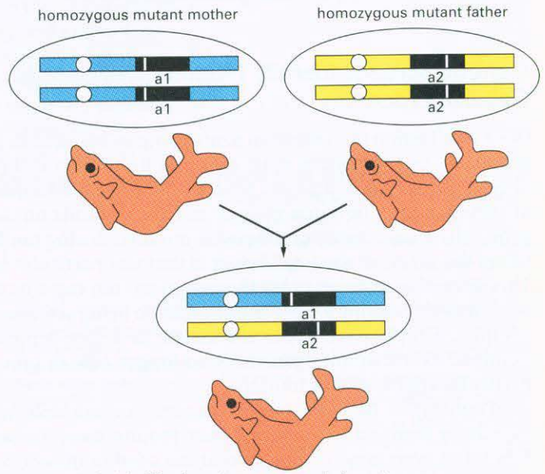
In the simplest type of complementation test, an individual who is homozygous for one mutation is mated with an individual who is homozygous for the other. The phenotype of the offspring gives the answer to the question.

COMPLEMENTATION: MUTATIONS IN TWO DIFFERENT GENES



hybrid offspring shows normal phenotype: one normal copy of each gene is present

NONCOMPLEMENTATION: TWO INDEPENDENT MUTATIONS IN THE SAME GENE



hybrid offspring shows mutant phenotype: no normal copies of the mutated gene are present

The Classical Approach Begins with Random Mutagenesis

Before the advent of gene cloning technology, most genes were identified by the processes disrupted when the gene was mutated. This classical genetic approach—identifying the genes responsible for mutant phenotypes—is most easily performed in organisms that reproduce rapidly and are amenable to genetic manipulation, such as bacteria, yeasts, nematode worms, and fruit flies. Although spontaneous mutants can sometimes be found by examining extremely large populations—thousands or tens of thousands of individual organisms—the process of isolating mutants can be made much more efficient by generating mutations with agents that damage DNA. By treating organisms with mutagens, very large numbers of mutants can be created quickly and then screened for a particular defect of interest, as we will see shortly.

An alternative approach to chemical or radiation mutagenesis is called *insertional mutagenesis*. This method relies on the fact that exogenous DNA inserted randomly into the genome can produce mutations if the inserted fragment interrupts a gene or its regulatory sequences. The inserted DNA, whose sequence is known, then serves as a molecular tag that aids in the subsequent identification and cloning of the disrupted gene (Figure 8–55). In *Drosophila*, the use of the transposable P element to inactivate genes has revolutionized the study of gene function in the fruit fly. Transposable elements (see Table 5–3, p. 287) have also been used to generate mutants in bacteria, yeast, and in the flowering plant *Arabidopsis*. Retroviruses, which copy themselves into the host genome (see Figure 5–73), have been used to disrupt genes in zebrafish and in mice.

Such studies are well suited for dissecting biological processes in worms and flies, but how can we study gene function in humans? Unlike the organisms we have been discussing, humans do not reproduce rapidly, and they are not intentionally treated with mutagens. Moreover, any human with a serious defect in an essential process, such as DNA replication, would die long before birth.

There are two answers to the question of how we study human genes. First, because genes and gene functions have been so highly conserved throughout evolution, the study of less complex model organisms reveals critical information about similar genes and processes in humans. The corresponding human genes can then be studied further in cultured human cells. Second, many mutations that are not lethal—tissue-specific defects in lysosomes or in cell-surface receptors, for example—have arisen spontaneously in the human population. Analyses of the phenotypes of the affected individuals, together with studies of their cultured cells, have provided many unique insights into important human cell functions. Although such mutations are rare, they are very efficiently discovered because of a unique human property: the mutant individuals call attention to themselves by seeking special medical care.

Figure 8–55 Insertional mutant of the snapdragon, *Antirrhinum*. A mutation in a single gene coding for a regulatory protein causes leafy shoots to develop in place of flowers. The mutation allows cells to adopt a character that would be appropriate to a different part of the normal plant. The mutant plant is on the left, the normal plant on the right. (Courtesy of Enrico Coen and Rosemary Carpenter.)

Genetic Screens Identify Mutants Deficient in Cellular Processes

Once a collection of mutants in a model organism such as yeast or flies has been produced, one generally must examine thousands of individuals to find the altered phenotype of interest. Such a search is called a **genetic screen**. Because obtaining a mutation in a gene of interest depends on the likelihood that the gene will be inactivated or otherwise mutated during random mutagenesis, the larger the genome, the less likely it is that any particular gene will be mutated. Therefore, the more complex the organism, the more mutants must be examined to avoid missing genes. The phenotype being screened for can be simple or complex. Simple phenotypes are easiest to detect: a metabolic deficiency, for example, in which an organism is no longer able to grow in the absence of a particular amino acid or nutrient.

Phenotypes that are more complex, for example mutations that cause defects in learning or memory, may require more elaborate screens (Figure 8–56). But even genetic screens that are used to dissect complex physiological systems should be as simple as possible in design, and, if possible, should permit

Figure 8–56 Screens can detect mutations that affect an animal's behavior. (A) Wild-type *C. elegans* engage in social feeding. The worms swim around until they encounter their neighbors and commence feeding. (B) Mutant animals feed by themselves. (Courtesy of Cornelia Bargmann, *Cell* 94:cover, 1998. © Elsevier.)

the examination of large numbers of mutants simultaneously. As an example, one particularly elegant screen was designed to search for genes involved in visual processing in the zebrafish. The basis of this screen, which monitors the fishes' response to motion, is a change in behavior. Wild-type fish tend to swim in the direction of a perceived motion, while mutants with defects in their visual systems swim in random directions—a behavior that is easily detected. One mutant discovered in this screen is called *lakritz*, which is missing 80% of the retinal ganglion cells that help to relay visual signals from the eye to the brain. As the cellular organization of the zebrafish retina mirrors that of all vertebrates, the study of such mutants should also provide insights into visual processing in humans.

Because defects in genes that are required for fundamental cell processes—RNA synthesis and processing or cell cycle control, for example—are usually lethal, the functions of these genes are often studied in temperature-sensitive mutants. In these mutants the protein product of the mutant gene functions normally at a medium temperature, but can be inactivated by a small increase or decrease in temperature. Thus the abnormality can be switched on and off experimentally simply by changing the temperature. A cell containing a temperature-sensitive mutation in a gene essential for survival at a non-permissive temperature can nevertheless grow at the normal or permissive temperature (Figure 8–57). The temperature-sensitive gene in such a mutant usually contains a point mutation that causes a subtle change in its protein product.

Many temperature-sensitive mutants were isolated in the genes that encode the bacterial proteins required for DNA replication by screening populations of mutagen-treated bacteria for cells that stop making DNA when they are warmed from 30°C to 42°C. These mutants were later used to identify and characterize the corresponding DNA replication proteins (discussed in Chapter 5). Temperature-sensitive mutants also led to the identification of many proteins involved in regulating the cell cycle and in moving proteins through the secretory pathway in yeast (see Panel 13–1). Related screening approaches have demonstrated the function of enzymes involved in the principal metabolic pathways of bacteria and yeast (discussed in Chapter 2), as well as discovering many of the gene products

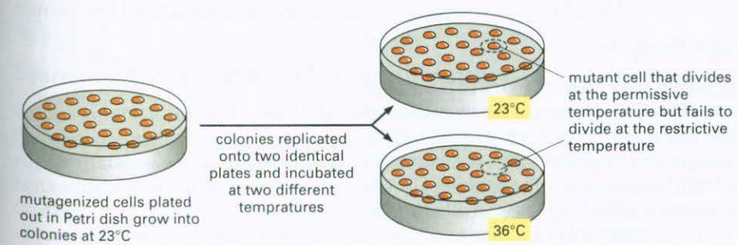


Figure 8–57 Screening for temperature-sensitive bacterial or yeast mutants. Mutagenized cells are plated out at the permissive temperature. The resulting colonies are transferred to two identical Petri dishes by replica plating; one of these plates is incubated at the permissive temperature, the other at the non-permissive temperature. Cells containing a temperature-sensitive mutation in a gene essential for proliferation can divide at the normal, permissive temperature but fail to divide at the elevated, non-permissive temperature.

responsible for the orderly development of the *Drosophila* embryo (discussed in Chapter 21).

A Complementation Test Reveals Whether Two Mutations Are in the Same or in Different Genes

A large-scale genetic screen can turn up many different mutants that show the same phenotype. These defects might lie in different genes that function in the same process, or they might represent different mutations in the same gene. How can we tell, then, whether two mutations that produce the same phenotype occur in the same gene or in different genes? If the mutations are recessive—if, for example, they represent a loss of function of a particular gene—a complementation test can be used to ascertain whether the mutations fall in the same or in different genes. In the simplest type of complementation test, an individual that is homozygous for one mutation—that is, it possesses two identical alleles of the mutant gene in question—is mated with an individual that is homozygous for the other mutation. If the two mutations are in the same gene, the offspring show the mutant phenotype, because they still will have no normal copies of the gene in question (see Panel 8–1, pp. 526–527). If, in contrast, the mutations fall in different genes, the resulting offspring show a normal phenotype. They retain one normal copy (and one mutant copy) of each gene. The mutations thereby complement one another and restore a normal phenotype. Complementation testing of mutants identified during genetic screens has revealed, for example, that 5 genes are required for yeast to digest the sugar galactose; that 20 genes are needed for *E. coli* to build a functional flagellum; that 48 genes are involved in assembling bacteriophage T4 viral particles; and that hundreds of genes are involved in the development of an adult nematode worm from a fertilized egg.

Once a set of genes involved in a particular biological process has been identified, the next step is to determine in which order the genes function. Determining when a gene acts can facilitate the reconstruction of entire genetic or biochemical pathways, and such studies have been central to our understanding of metabolism, signal transduction, and many other developmental and physiological processes. In essence, untangling the order in which genes function requires careful characterization of the phenotype caused by mutations in each different gene. Imagine, for example, that mutations in a handful of genes all cause an arrest in cell division during early embryo development. Close examination of each mutant may reveal that some act extremely early, preventing the fertilized egg from dividing into two cells. Other mutations may allow early cell divisions but prevent the embryo from reaching the blastula stage.

To test predictions made about the order in which genes function, organisms can be made that are mutant in two different genes. If these mutations affect two different steps in the same process, such *double mutants* should have a phenotype identical to that of the mutation that acts earliest in the pathway. As an example, the pathway of protein secretion in yeast has been deciphered in this manner. Different mutations in this pathway cause proteins to accumulate aberrantly in the endoplasmic reticulum (ER) or in the Golgi apparatus. When a cell is engineered to harbor both a mutation that blocks protein processing in the ER *and* a mutation that blocks processing in the Golgi compartment, proteins accumulate in the ER. This indicates that proteins must pass through the ER before being sent to the Golgi before secretion (Figure 8–58).

Genes Can Be Located by Linkage Analysis

With mutants in hand, the next step is to identify the gene or genes that seem to be responsible for the altered phenotype. If insertional mutagenesis was used for the original mutagenesis, locating the disrupted gene is fairly simple. DNA fragments containing the insertion (a transposon or a retrovirus, for example) are collected and amplified, and the nucleotide sequence of the flanking DNA is determined. This sequence is then used to search a DNA database to identify the gene that was interrupted by insertion of the transposable element.

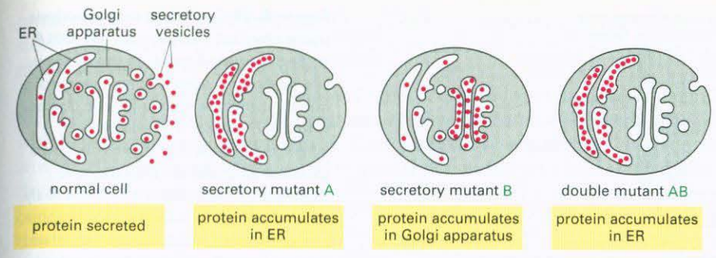


Figure 8–58 Using genetics to determine the order of function of genes. In normal cells, proteins are loaded into vesicles, which fuse with the plasma membrane and secrete their contents into the extracellular medium. In secretory mutant A, proteins accumulate in the ER. In a different secretory mutant B, proteins accumulate in the Golgi. In the double mutant AB, proteins accumulate in the ER; this indicates that the gene defective in mutant A acts before the gene defective in mutant B in the secretory pathway.

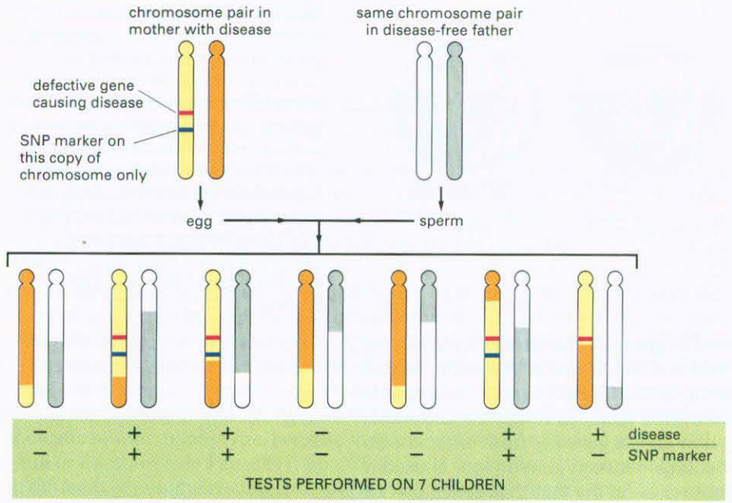
If a DNA-damaging chemical was used to generate the mutants, identifying the inactivated gene is often more laborious and can be accomplished by several different approaches. In one, the first step is to determine where on the genome the gene is located. To map a newly discovered gene, its rough chromosomal location is first determined by assessing how far the gene lies from other known genes in the genome. Estimating the distance between genetic loci is usually done by linkage analysis, a technique that relies on the fact that genes that lie near one another on a chromosome tend to be inherited together. The closer the genes are, the greater the likelihood they will be passed to offspring as a pair. Even closely linked genes, however, can be separated by recombination during meiosis. The larger the distance between two genetic loci, the greater the chance that they will be separated by a crossover (see Panel 8–1, pp. 526–527). By calculating the recombination frequency between two genes, the approximate distance between them can be determined.

Because genes are not always located close enough to one another to allow a precise pinpointing of their position, linkage analyses often rely on physical markers along the genome for estimating the location of an unknown gene. These markers are generally nucleotide fragments, with a known sequence and genome location, that can exist in at least two allelic forms. Single-nucleotide polymorphisms (SNPs), for example, are short sequences that differ by one or more nucleotides among individuals in a population. SNPs can be detected by hybridization techniques. Many such physical markers, distributed all along the length of chromosomes, have been collected for a variety of organisms, including more than 10^6 for humans. If the distribution of these markers is sufficiently dense, one can, through a linkage analysis that tests for the tight coinheritance of one or more SNPs with the mutant phenotype, narrow the potential location of a gene to a chromosomal region that may contain only a few gene sequences. These are then considered candidate genes, and their structure and function can be tested directly to determine which gene is responsible for the original mutant phenotype.

Linkage analysis can be used in the same way to identify the genes responsible for heritable human disorders. Such studies require that DNA samples be collected from a large number of families affected by the disease. These samples are examined for the presence of physical markers such as SNPs that seem to be closely linked to the disease gene—these sequences would always be inherited by individuals who have the disease, and not by their unaffected relatives. The disease gene is then located as described above (Figure 8–59). The genes for cystic fibrosis and Huntington's disease, for example, were discovered in this manner.

Searching for Homology Can Help Predict a Gene's Function

Once a gene has been identified, its function can often be predicted by identifying homologous genes whose functions are already known. As we discussed earlier, databases containing nucleotide sequences from a variety of organisms—including the complete genome sequences of many dozens of microbes, *C. elegans*, *A. thaliana*, *D. melanogaster*, and human—can be searched for sequences that are similar to those of the uncharacterized target gene.



CONCLUSION: gene causing disease is coinherited with SNP marker from diseased mother in 75% of the diseased progeny. If this same correlation is observed in other families that have been examined, the gene causing disease is mapped to this chromosome close to the SNP. Note that a SNP that is either far away from the gene on the same chromosome, or located on a different chromosome than the gene of interest, will be coinherited only 50% of the time.

Figure 8–59 Genetic linkage analysis using physical markers on the DNA to find a human gene. In this example, one studies the coinheritance of a specific human phenotype (here a genetic disease) with a SNP marker. If individuals who inherit the disease nearly always inherit a particular SNP marker, then the gene causing the disease and the SNP are likely to be close together on the chromosome, as shown here. To prove that an observed linkage is statistically significant, hundreds of individuals may need to be examined. Note that the linkage will not be absolute unless the SNP marker is located in the gene itself. Thus, occasionally the SNP will be separated from the disease gene by meiotic crossing-over during the formation of the egg or sperm: this has happened in the case of the chromosome pair on the far right. When working with a sequenced genome, this procedure would be repeated with SNPs located on either side of the initial SNP, until a 100% coinheritance is found.

When analyzing a newly sequenced genome, such a search serves as a first-pass attempt to assign functions to as many genes as possible, a process called *annotation*. Further genetic and biochemical studies are then performed to confirm whether the gene encodes a product with the predicted function, as we discuss shortly. Homology analysis does not always reveal information about function: in the case of the yeast genome, 30% of the previously uncharacterized genes could be assigned a putative function by homology analysis; 10% had homologues whose function was also unknown; and another 30% had no homologues in any existing databases. (The remaining 30% of the genes had been identified before sequencing the yeast genome.)

In some cases, a homology search turns up a gene in organism A which produces a protein that, in a different organism, is fused to a second protein that is produced by an independent gene in organism A. In yeast, for example, two separate genes encode two proteins that are involved in the synthesis of tryptophan; in *E. coli*, however, these two genes are fused into one (Figure 8–60). Knowledge that these two proteins in yeast correspond to two domains in a single bacterial protein means that they are likely to be functionally associated, and probably work together in a protein complex. More generally, this approach is used to establish functional links between genes that, for most organisms, are widely separated in the genome.

Reporter Genes Reveal When and Where a Gene Is Expressed

Clues to gene function can often be obtained by examining when and where a gene is expressed in the cell or in the whole organism. Determining the pattern and timing of gene expression can be accomplished by replacing the coding

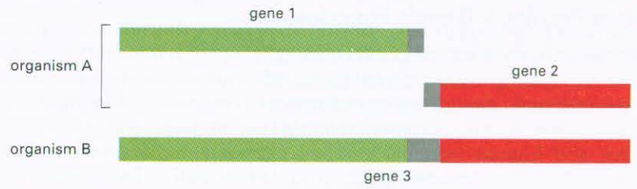


Figure 8–60 Domain fusions reveal relationships between functionally linked genes. In this example, the functional interaction of genes 1 and 2 in organism A is inferred by the fusion of homologous domains into a single gene (gene 3) in organism B.

portion of the gene under study with a reporter gene. In most cases, the expression of the reporter gene is then monitored by tracking the fluorescence or enzymatic activity of its protein product (pp. 518–519).

As discussed in detail in Chapter 7, gene expression is controlled by regulatory DNA sequences, located upstream or downstream of the coding region, which are not generally transcribed. These regulatory sequences, which control which cells will express a gene and under what conditions, can also be made to drive the expression of a reporter gene. One simply replaces the target gene's coding sequence with that of the reporter gene, and introduces these recombinant DNA molecules into cells. The level, timing, and cell specificity of reporter protein production reflect the action of the regulatory sequences that belong to the original gene (Figure 8–61).

Several other techniques, discussed previously, can also be used to determine the expression pattern of a gene. Hybridization techniques such as Northern analysis (see Figure 8–27) and *in situ* hybridization for RNA detection (see Figure 8–29) can reveal when genes are transcribed and in which tissue, and how much mRNA they produce.

Microarrays Monitor the Expression of Thousands of Genes at Once

So far we have discussed techniques that can be used to monitor the expression of only a single gene at a time. Many of these methods are fairly labor-intensive: generating reporter gene constructs or GFP fusions requires manipulating DNA and transfecting cells with the resulting recombinant molecules. Even Northern analyses are limited in scope by the number of samples that can be run on an agarose gel. Developed in the 1990s, **DNA microarrays** have revolutionized the way in which gene expression is now analyzed by allowing the RNA products of thousands of genes to be monitored at once. By examining the expression of so many genes simultaneously, we can now begin to identify and study the gene expression patterns that underlie cellular physiology: we can see which genes are switched on (or off) as cells grow, divide, or respond to hormones or to toxins.

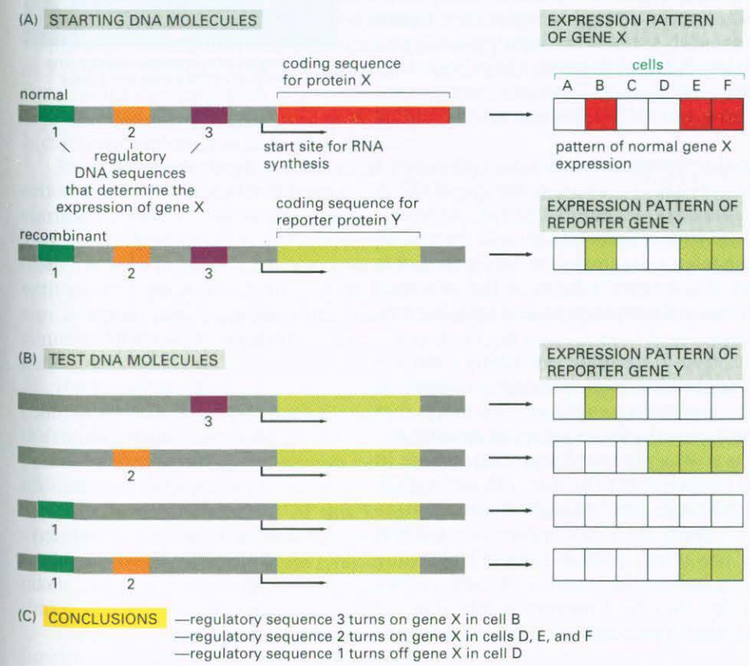
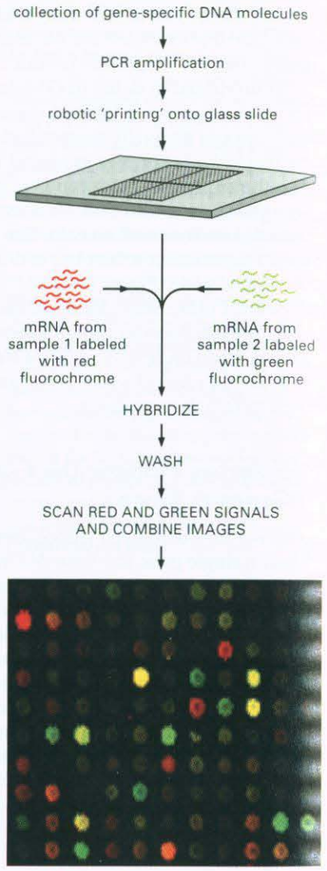


Figure 8–61 Using a reporter protein to determine the pattern of a gene's expression. (A) In this example the coding sequence for protein X is replaced by the coding sequence for protein Y. (B) Various fragments of DNA containing candidate regulatory sequences are added in combinations. The recombinant DNA molecules are then tested for expression after their transfection into a variety of different types of mammalian cells, and the results are summarized in (C). For experiments in eucaryotic cells, two commonly used reporter proteins are the enzymes β-galactosidase *(β-gal)* and green fluorescent protein or GFP (see Figure 9–44). Because these are bacterial enzymes, their presence can be monitored by simple and sensitive assays of enzyme activity, without any interference from host cell enzymes. Figure 7–39 shows an example in which the β-gal receptor gene is used to monitor the activity of the eve gene regulatory sequence in a *Drosophila* embryo.

Figure 8–62 Using DNA microarrays to monitor the expression of thousands of genes simultaneously. To prepare the microarray, DNA fragments—each corresponding to a gene—are spotted onto a slide by a robot. Prepared arrays are also available commercially. In this example, mRNA is collected from two different cell samples for a direct comparison of their relative levels of gene expression. These samples are converted to cDNA and labeled, one with a red fluorochrome, the other with a green fluorochrome. The labeled samples are mixed and then allowed to hybridize to the microarray. After incubation, the array is washed and the fluorescence scanned. In the portion of a microarray shown, which represents the expression of 110 yeast genes, *red* spots indicate that the gene in sample 1 is expressed at a higher level than the corresponding gene in sample 2; *green* spots indicate that expression of the gene is higher in sample 2 than in sample 1. *Yellow* spots reveal genes that are expressed at equal levels in both cell samples. Dark spots indicate little or no expression in either sample of the gene whose fragment is located at that position in the array. For details see Figure 1–45. (Microarray courtesy of J.L. DeRisi et al., *Science* 278:680–686, 1997. © AAAS.)



small region of microarray representing expression of 110 genes from yeast

DNA microarrays are little more than glass microscope slides studded with a large number of DNA fragments, each containing a nucleotide sequence that serves as a probe for a specific gene. The most dense arrays may contain tens of thousands of these fragments in an area smaller than a postage stamp, allowing thousands of hybridization reactions to be performed in parallel (Figure 8–62). Some microarrays are generated from large DNA fragments that have been generated by PCR and then spotted onto the slides by a robot. Others contain short oligonucleotides that are synthesized on the surface of the glass wafer with techniques similar to those that are used to etch circuits onto computer chips. In either case, the exact sequence—and position—of every probe on the chip is known. Thus any nucleotide fragment that hybridizes to a probe on the array can be identified as the product of a specific gene simply by detecting the position to which it is bound.

To use a DNA microarray to monitor gene expression, mRNA from the cells being studied is first extracted and converted to cDNA (see Figure 8–34). The cDNA is then labeled with a fluorescent probe. The microarray is incubated with this labeled cDNA sample and hybridization is allowed to occur (see Figure 8–62). The array is then washed to remove cDNA that is not tightly bound, and the positions in the microarray to which labeled DNA fragments have bound are identified by an automated scanning-laser microscope. The array positions are then matched to the particular gene whose sample of DNA was spotted in this location.

Typically the fluorescent DNA from the experimental samples (labeled, for example, with a red fluorescent dye) are mixed with a reference sample of cDNA fragments labeled with a differently colored fluorescent dye (green, for example). Thus, if the amount of RNA expressed from a particular gene in the cells of interest is increased relative to that of the reference sample, the resulting spot is red. Conversely, if the gene's expression is decreased relative to the reference sample, the spot is green. Using such an internal reference, gene expression profiles can be tabulated with great precision.

So far, DNA microarrays have been used to examine everything from the change in gene expression that make strawberries ripen to the gene expression "signatures" of different types of human cancer cells (see Figure 7–3). Arrays that contain probes representing all 6000 yeast genes have been used to monitor the changes that occur in gene expression as yeast shift from fermenting glucose to growing on ethanol; as they respond to a sudden shift to heat or cold; and as they proceed through different stages of the cell cycle. The first study showed that, as yeast use up the last glucose in their medium, their gene expression pattern changes markedly: nearly 900 genes are more actively transcribed, while another 1200 decrease in activity. About half of these genes have no known function, although this study suggests that they are somehow involved in the metabolic reprogramming that occurs when yeast cells shift from fermentation to respiration.

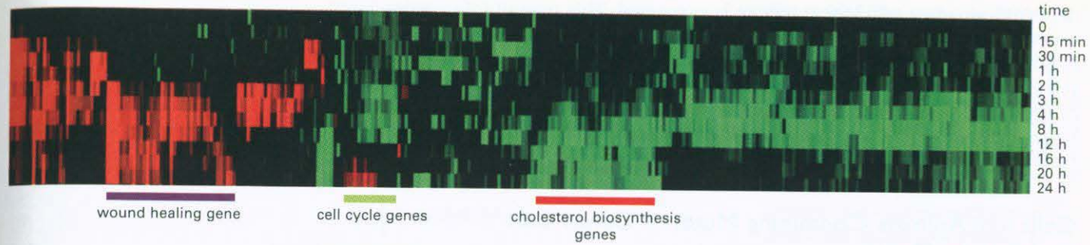


Figure 8–63 Using cluster analysis to identify sets of genes that are coordinately regulated. Genes that belong to the same cluster may be involved in common cellular pathways or processes. To perform a cluster analysis, microarray data are obtained from cell samples exposed to a variety of different conditions, and genes that show coordinate changes in their expression pattern are grouped together. In this experiment, human fibroblasts were deprived of serum for 48 hours; serum was then added back to the cultures at time 0 and the cells were harvested for microarray analysis at different time points. Of the 8600 genes analyzed on the DNA microarray, just over 300 showed threefold or greater variation in their expression patterns in response to serum reintroduction. Here, *red* indicates an increase in expression; *green* is a decrease in expression. On the basis of the results of many microarray experiments, the 8600 genes have been grouped in clusters based on similar patterns of expression. The results of this analysis show that genes involved in wound healing are turned on in response to serum, while genes involved in regulating cell cycle progression and cholesterol biosynthesis are shut down. (From M.B. Eisen et al., *Proc. Natl. Acad. Sci. USA* 95:14863–14868, 1998. © National Academy of Sciences.)

Comprehensive studies of gene expression also provide an additional layer of information that is useful for predicting gene function. Earlier we discussed how identifying a protein's interaction partners can yield clues about that protein's function. A similar principle holds true for genes: information about a gene's function can be deduced by identifying genes that share its expression pattern. Using a technique called *cluster analysis*, one can identify sets of genes that are coordinately regulated. Genes that are turned on or turned off together under a variety of different circumstances may work in concert in the cell: they may encode proteins that are part of the same multiprotein machine, or proteins that are involved in a complex coordinated activity, such as DNA replication or RNA splicing. Characterizing an unknown gene's function by grouping it with known genes that share its transcriptional behavior is sometimes called "guilt by association." Cluster analyses have been used to analyze the gene expression profiles that underlie many interesting biological processes, including wound healing in humans (Figure 8–63).

Targeted Mutations Can Reveal Gene Function

Although in rapidly reproducing organisms it is often not difficult to obtain mutants that are deficient in a particular process, such as DNA replication or eye development, it can take a long time to trace the defect to a particular altered protein. Recently, recombinant DNA technology and the explosion in genome sequencing have made possible a different type of genetic approach. Instead of beginning with a randomly generated mutant and using it to identify a gene and its protein, one can start with a particular gene and proceed to make mutations in it, creating mutant cells or organisms so as to analyze the gene's function. Because the new approach reverses the traditional direction of genetic discovery—proceeding from genes and proteins to mutants, rather than vice versa—it is commonly referred to as **reverse genetics**.

Reverse genetics begins with a cloned gene, a protein with interesting properties that has been isolated from a cell, or simply a genome sequence. If the starting point is a protein, the gene encoding it is first identified and, if necessary, its nucleotide sequence is determined. The gene sequence can then be altered *in vitro* to create a mutant version. This engineered mutant gene, together with an appropriate regulatory region, is transferred into a cell. Inside the cell, it can integrate into a chromosome, becoming a permanent part of the cell's genome. All of the descendants of the modified cell will now contain the mutant gene.

If the original cell used for the gene transfer is a fertilized egg, whole multicellular organisms can be obtained that contain the mutant gene, provided that the mutation does not cause lethality. In some of these animals, the altered gene will be incorporated into the germ cells—a germline mutation—allowing the mutant gene to be passed on to their progeny.

Genetic transformations of this kind are now routinely performed with organisms as complex as fruit flies and mammals. Technically, even humans could now be transformed in this way, although such procedures are not undertaken, even for therapeutic purposes, for fear of the unpredictable aberrations that might occur in such individuals.

Earlier in this chapter we discussed other approaches to discover a gene's function, including searching for homologous genes in other organisms and

determining when and where a gene is expressed. This type of information is especially useful in suggesting what sort of phenotypes to look for in the mutant organisms. A gene that is expressed only in adult liver, for example, may have a role in degrading toxins, but is not likely to affect the development of the eye. All of these approaches can be used either to study single genes or to attempt a large-scale analysis of the function of every gene in an organism—a burgeoning field known as *functional genomics*.

Cells and Animals Containing Mutated Genes Can Be Made to Order

We have seen that searching for homologous genes and analyzing gene expression patterns can provide clues about gene function, but they do not reveal what exactly a gene does inside a cell. Genetics provides a powerful solution to this problem, because mutants that lack a particular gene may quickly reveal the function of the protein that it encodes. Genetic engineering techniques allow one to specifically produce such gene knockouts, as we will see. However, one can also generate mutants that express a gene at abnormally high levels (overexpression), in the wrong tissue or at the wrong time (misexpression), or in a slightly altered form that exerts a dominant phenotype. To facilitate such studies of gene function, the coding sequence of a gene and its regulatory regions can be engineered to change the functional properties of the protein product, the amount of protein made, or the particular cell type in which the protein is produced.

Altered genes are introduced into cells in a variety of ways, some of which are described in detail in Chapter 9. DNA can be microinjected into mammalian cells with a glass micropipette or introduced by a virus that has been engineered to carry foreign genes. In plant cells, genes are frequently introduced by a technique called particle bombardment: DNA samples are painted onto tiny gold beads and then literally shot through the cell wall with a specially modified gun. Electroporation is the method of choice for introducing DNA into bacteria and some other cells. In this technique, a brief electric shock renders the cell membrane temporarily permeable, allowing foreign DNA to enter the cytoplasm.

We will now examine how the study of such mutant cells and organisms allows the dissection of biological pathways.

The Normal Gene in a Cell Can Be Directly Replaced by an Engineered Mutant Gene in Bacteria and Some Lower Eucaryotes

Unlike higher eucaryotes (which are multicellular and diploid), bacteria, yeasts, and the cellular slime mold *Dictyostelium* generally exist as haploid single cells. In these organisms an artificially introduced DNA molecule carrying a mutant gene can, with a relatively high frequency, replace the single copy of the normal gene by homologous recombination (see p. 276), so that it is easy to produce cells in which the mutant gene has replaced the normal gene (Figure 8–64A). In this way cells can be made to order that produce an altered form of any specific protein or RNA molecule instead of the normal form of the molecule. If the mutant gene is completely inactive and the gene product normally performs an essential function, the cell dies; but in this case a less severely mutated version of the gene can be used to replace the normal gene, so that the mutant cell survives but is abnormal in the process for which the gene is required. Often the mutant of choice is one that produces a temperature-sensitive gene product, which functions normally at one temperature but is inactivated when cells are shifted to a higher or lower temperature.

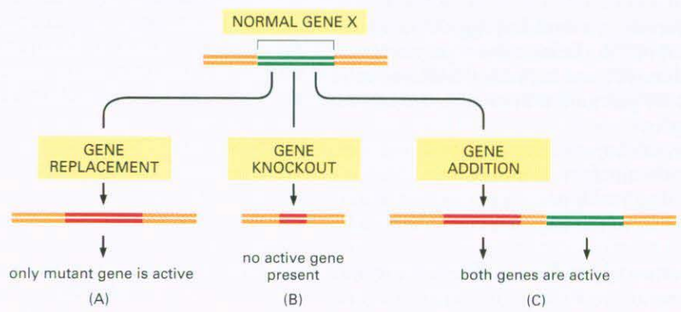


Figure 8–64 Gene replacement, gene knockout, and gene addition. A normal gene can be altered in several ways in a genetically engineered organism. (A) The normal gene *(green)* can be completely replaced by a mutant copy of the gene *(red)*, a process called gene replacement. This provides information on the activity of the mutant gene without interference from the normal gene, and thus the effects of small and subtle mutations can be determined. (B) The normal gene can be inactivated completely, for example, by making a large deletion in it; the gene is said to have suffered a knockout. (C) A mutant gene can simply be added to the genome. In some organisms this is the easiest type of genetic engineering to perform. This approach can provide useful information when the introduced mutant gene overrides the function of the normal gene.

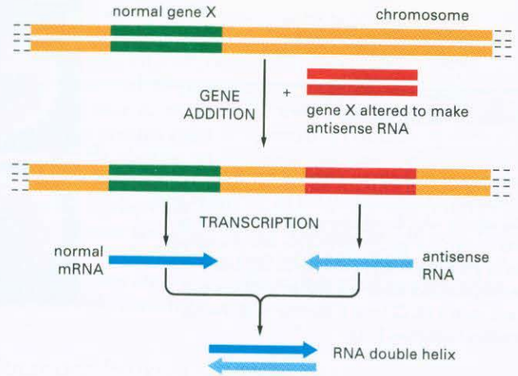


Figure 8–65 The antisense RNA strategy for generating dominant negative mutations. Mutant genes that have been engineered to produce antisense RNA, which is complementary in sequence to the RNA made by the normal gene X, can cause double-stranded RNA to form inside cells. If a large excess of the antisense RNA is produced, it can hybridize with—and thereby inactivate—most of the normal RNA produced by gene X. Although in the future it may become possible to inactivate any gene in this way, at present the technique seems to work for some genes but not others.

The ability to perform direct gene replacements in lower eucaryotes, combined with the power of standard genetic analyses in these haploid organisms, explains in large part why studies in these types of cells have been so important for working out the details of those processes that are shared by all eucaryotes. As we shall see, gene replacements are possible, but more difficult to perform in higher eucaryotes, for reasons that are not entirely understood.

Engineered Genes Can Be Used to Create Specific Dominant Negative Mutations in Diploid Organisms

Higher eucaryotes, such as mammals, fruit flies, or worms, are diploid and therefore have two copies of each chromosome. Moreover, transfection with an altered gene generally leads to gene addition rather than gene replacement: the altered gene inserts at a random location in the genome, so that the cell (or the organism) ends up with the mutated gene in addition to its normal gene copies.

Because gene addition is much more easily accomplished than gene replacement in higher eucaryotic cells, it is useful to create specific dominant negative mutations in which a mutant gene eliminates the activity of its normal counterparts in the cell. One ingenious approach exploits the specificity of hybridization reactions between two complementary nucleic acid chains. Normally, only one of the two DNA strands in a given portion of double helix is transcribed into RNA, and it is always the same strand for a given gene (see Figure 6–14). If a cloned gene is engineered so that the opposite DNA strand is transcribed instead, it will produce antisense RNA molecules that have a sequence complementary to the normal RNA transcripts. Such *antisense RNA*, when synthesized in large enough amounts, can often hybridize with the "sense" RNA made by the normal genes and thereby inhibit the synthesis of the corresponding protein (Figure 8–65). A related method involves synthesizing short antisense nucleic acid molecules chemically or enzymatically and then injecting (or otherwise delivering) them into cells, again blocking (although only temporarily) production of the corresponding protein. To avoid degradation of the

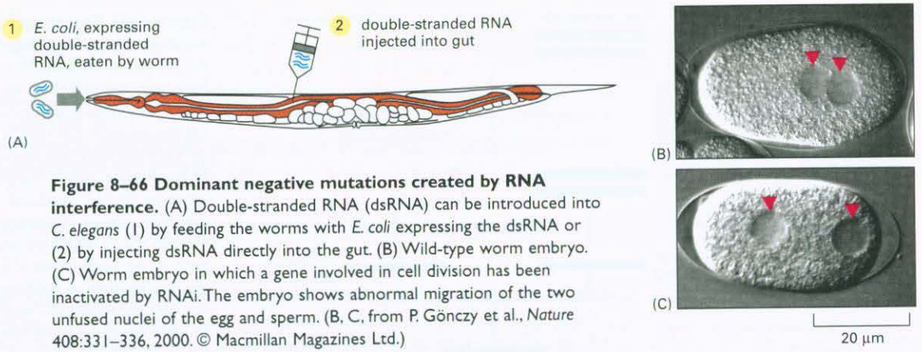


Figure 8–66 Dominant negative mutations created by RNA interference. (A) Double-stranded RNA (dsRNA) can be introduced into *C. elegans* (1) by feeding the worms with *E. coli* expressing the dsRNA or (2) by injecting dsRNA directly into the gut. (B) Wild-type worm embryo. (C) Worm embryo in which a gene involved in cell division has been inactivated by RNAi. The embryo shows abnormal migration of the two unfused nuclei of the egg and sperm. (B, C, from P. Gönczy et al., *Nature* 408:331–336, 2000. © Macmillan Magazines Ltd.)

injected nucleic acid, a stable synthetic RNA analog, called morpholino-RNA, is often used instead of ordinary RNA.

As investigators continued to explore the antisense RNA strategy, they made an interesting discovery. An antisense RNA strand can block gene expression, but a preparation of double-stranded RNA (dsRNA), containing both the sense and antisense strands of a target gene, inhibit the activity of target genes even more effectively (see Figure 7–107). This phenomenon, dubbed *RNA interference (RNAi)*, has now been exploited for examining gene function in several organisms.

The RNAi technique has been widely used to study gene function in the nematode *C. elegans*. When working with worms, introducing the dsRNA is quite simple: RNA can be injected directly into the intestine of the animal, or the worm can be fed with *E. coli* expressing the target gene dsRNA (Figure 8–66A). The RNA is distributed throughout the body of the worm and is found to inhibit expression of the target gene in different tissue types. Further, as explained in Figure 7–107, the interference is frequently inherited by the progeny of the injected animal. Because the entire genome of *C. elegans* has been sequenced, RNAi is being used to help in assigning functions to the entire complement of worm genes. In one study, researchers were able to inhibit 96% of the approximately 2300 predicted genes on *C. elegans* chromosome III. In this way, they identified 133 genes involved in cell division in *C. elegans* embryos (Figure 8–66C). Of these, only 11 had been previously ascribed a function by direct experimentation.

For unknown reasons, RNA interference does not efficiently inactivate all genes. And interference can sometimes suppress the activity of a target gene in one tissue and not another. An alternative way to produce a dominant negative mutation takes advantage of the fact that most proteins function as part of a larger protein complex. Such complexes can often be inactivated by the inclusion of just one nonfunctional component. Therefore, by designing a gene that produces large quantities of a mutant protein that is inactive but still able to assemble into the complex, it is often possible to produce a cell in which all the complexes are inactivated despite the presence of the normal protein (Figure 8–67).

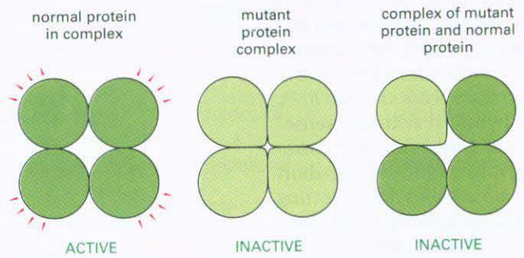


Figure 8–67 A dominant negative effect of a protein. Here a gene is engineered to produce a mutant protein that prevents the normal copies of the same protein from performing their function. In this simple example, the normal protein must form a multisubunit complex to be active, and the mutant protein blocks function by forming a mixed complex that is inactive. In this way a single copy of a mutant gene located anywhere in the genome can inactivate the normal products produced by other gene copies.

If a protein is required for the survival of the cell (or the organism), a dominant negative mutant dies, making it impossible to test the function of the protein. To avoid this problem, one can couple the mutant gene to control sequences that have been engineered to produce the gene product only on command—for example, in response to an increase in temperature or to the presence of a specific signaling molecule. Cells or organisms containing such a dominant mutant gene under the control of an *inducible promoter* can be deprived of a specific protein at a particular time, and the effect can then be followed. Inducible promoters also allow genes to be switched on or off in specific tissues, allowing one to examine the effect of the mutant gene in selected parts of the organism. In the future, techniques for producing dominant negative mutations to inactivate specific genes are likely to be widely used to determine the functions of proteins in higher organisms.

Gain-of-Function Mutations Provide Clues to the Role Genes Play in a Cell or Organism

In the same way that cells can be engineered to express a dominant negative version of a protein, resulting in a loss-of-function phenotype, they can also be engineered to display a novel phenotype through a *gain-of-function* mutation. Such mutations may confer a novel activity on a particular protein, or they may cause a protein with normal activity to be expressed at an inappropriate time or in the wrong tissue in an animal. Regardless of the mechanism, gain-of-function mutations can produce a new phenotype in a cell, tissue, or organism.

Often, gain-of-function mutants are generated by expressing a gene at a much higher level than normal in cells. Such overexpression can be achieved by coupling a gene to a powerful promoter sequence and placing it on a multicopy plasmid—or integrating it in multiple copies in the genome. In either case, the gene is present in many copies and each copy directs the transcription of unusually large numbers of mRNA molecules. Although the effect that such overexpression has on the phenotype of an organism must be interpreted with caution, this approach has provided invaluable insights into the activity of many genes. In an alternate type of gain-of-function mutation, the mutant protein is made in normal amounts, but is much more active than its normal counterpart. Such proteins are frequently found in tumors, and they have been exploited to study signal transduction pathways in cells (discussed in Chapter 15).

Genes can also be expressed at the wrong time or in the wrong place in an organism—often with striking results (Figure 8–68). Such misexpression is most often accomplished by re-engineering the genes themselves, thereby supplying them with the regulatory sequences needed to alter their expression.

Genes Can Be Redesigned to Produce Proteins of Any Desired Sequence

In studying the action of a gene and the protein it encodes, one does not always wish to make drastic changes—flooding cells with huge quantities of hyperactive protein or eliminating a gene product entirely. It is sometimes useful to make slight changes in a protein's structure so that one can begin to dissect which portions of a protein are important for its function. The activity of an enzyme, for example, can be studied by changing a single amino acid in its active site. Special techniques are required to alter genes, and their protein products, in such subtle ways. The first step is often the chemical synthesis of a short DNA molecule containing the desired altered portion of the gene's nucleotide sequence. This synthetic DNA oligonucleotide is hybridized with single-stranded plasmid DNA that contains the DNA sequence to be altered, using conditions that allow imperfectly matched DNA strands to pair (Figure 8–69). The synthetic oligonucleotide will now serve as a primer for DNA synthesis by DNA polymerase, thereby generating a DNA double helix that incorporates the altered sequence into one of its two strands. After transfection, plasmids that carry the fully modified gene sequence are obtained. The appropriate DNA is

Figure 8–68 Ectopic misexpression of Wnt, a signaling protein that affects development of the body axis in the early *Xenopus* embryo. In this experiment, mRNA coding for Wnt was injected into the ventral vegetal blastomere, inducing a second body axis (discussed in Chapter 21). (From S. Sokol et al., *Cell* 67:741–752, 1992. © Elsevier.)

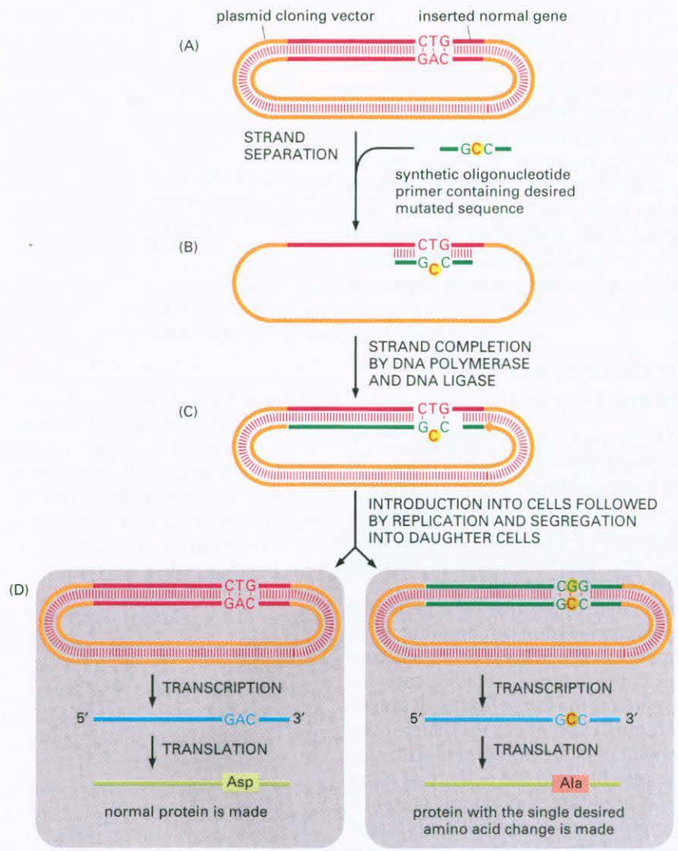


Figure 8–69 The use of a synthetic oligonucleotide to modify the protein-coding region of a gene by site-directed mutagenesis. (A) A recombinant plasmid containing a gene insert is separated into its two DNA strands. A synthetic oligonucleotide primer corresponding to part of the gene sequence but containing a single altered nucleotide at a predetermined point is added to the single-stranded DNA under conditions that permit less than perfect DNA hybridization. (B) The primer hybridizes to the DNA, forming a single mismatched nucleotide pair. (C) The recombinant plasmid is made double-stranded by *in vitro* DNA synthesis starting from the primer and sealed by DNA ligase. (D) The double-stranded DNA is introduced into a cell, where it is replicated. Replication using one strand of the template produces a normal DNA molecule, but replication using the other (the strand that contains the primer) produces a DNA molecule carrying the desired mutation. Only half of the progeny cells will end up with a plasmid that contains the desired mutant gene. However, a progeny cell that contains the mutated gene can be identified, separated from other cells, and cultured to produce a pure population of cells, all of which carry the mutated gene. Only one of the many changes that can be engineered in this way is shown here. With an oligonucleotide of the appropriate sequence, more than one amino acid substitution can be made at a time, or one or more amino acids can be inserted or deleted. Although not shown in this figure, it is also possible to create a site-directed mutation by using the appropriate oligonucleotides and PCR (instead of plasmid replication) to amplify the mutated gene.

then inserted into an expression vector so that the redesigned protein can be produced in the appropriate type of cells for detailed studies of its function. By changing selected amino acids in a protein in this way—a technique called **site-directed mutagenesis**—one can determine exactly which parts of the polypeptide chain are important for such processes as protein folding, interactions with other proteins, and enzymatic catalysis.

Engineered Genes Can Be Easily Inserted into the Germ Line of Many Animals

When engineering an organism that is to express an altered gene, ideally one would like to be able to replace the normal gene with the altered one so that the function of the mutant protein can be analyzed in the absence of the normal protein. As discussed above, this can be readily accomplished in some haploid, single-celled organisms. We shall see in the following section that much more complicated procedures have been developed that allow gene replacements of this type in mice. Foreign DNA can, however, be rather easily integrated into random positions of many animal genomes. In mammals, for example, linear DNA fragments introduced into cells are rapidly ligated end-to-end by intracellular enzymes to form long tandem arrays, which usually become integrated into a chromosome at an apparently random site. Fertilized mammalian eggs behave like other mammalian cells in this respect. A mouse egg injected with 200 copies of a linear DNA molecule often develops into a mouse containing, in many of its cells, a tandem array of copies of the injected gene integrated at a single random

site in one of its chromosomes. If the modified chromosome is present in the germ line cells (eggs or sperm), the mouse will pass these foreign genes on to its progeny.

Animals that have been permanently reengineered by either gene insertion, gene deletion, or gene replacement are called **transgenic organisms**, and any foreign or modified genes that are added are called *transgenes*. When the normal gene remains present, only dominant effects of the alteration will show up in phenotypic analyses. Nevertheless, transgenic animals with inserted genes have provided important insights into how mammalian genes are regulated and how certain altered genes (called oncogenes) cause cancer.

It is also possible to produce transgenic fruit flies, in which single copies of a gene are inserted at random into the *Drosophila* genome. In this case the DNA fragment is first inserted between the two terminal sequences of a *Drosophila* transposon called the P element. The terminal sequences enable the P element to integrate into *Drosophila* chromosomes when the P element transposase enzyme is also present (see p. 288). To make transgenic fruit flies, therefore, the appropriately modified DNA fragment is injected into a very young fruit fly embryo along with a separate plasmid containing the gene encoding the transposase. When this is done, the injected gene often enters the germ line in a single copy as the result of a transposition event.

Gene Targeting Makes It Possible to Produce Transgenic Mice That Are Missing Specific Genes

If a DNA molecule carrying a mutated mouse gene is transferred into a mouse cell, it usually inserts into the chromosomes at random, but about once in a thousand times, it replaces one of the two copies of the normal gene by homologous recombination. By exploiting these rare "gene targeting" events, any specific gene can be altered or inactivated in a mouse cell by a direct gene replacement. In the special case in which the gene of interest is inactivated, the resulting animal is called a "knockout" mouse.

The technique works as follows: in the first step, a DNA fragment containing a desired mutant gene (or a DNA fragment designed to interrupt a target gene) is inserted into a vector and then introduced into a special line of embryo-derived mouse stem cells, called **embryonic stem cells** or ES cells, that grow in cell culture and are capable of producing cells of many different tissue types. After a period of cell proliferation, the rare colonies of cells in which a homologous recombination event is likely to have caused a gene replacement to occur are isolated. The correct colonies among these are identified by PCR or by Southern blotting: they contain recombinant DNA sequences in which the inserted fragment has replaced all or part of one copy of the normal gene. In the second step, individual cells from the identified colony are taken up into a fine micropipette and injected into an early mouse embryo. The transfected embryo-derived stem cells collaborate with the cells of the host embryo to produce a normal-looking mouse; large parts of this chimeric animal, including—in favorable cases—cells of the germ line, often derive from the artificially altered stem cells (Figure 8–70).

The mice with the transgene in their germ line are bred to produce both a male and a female animal, each heterozygous for the gene replacement (that is, they have one normal and one mutant copy of the gene). When these two mice are in turn mated, one-fourth of their progeny will be homozygous for the altered gene. Studies of these homozygotes allow the function of the altered gene—or the effects of eliminating a gene activity—to be examined in the absence of the corresponding normal gene.

The ability to prepare transgenic mice lacking a known normal gene has been a major advance, and the technique is now being used to dissect the functions of a large number of mammalian genes (Figure 8–71). Related techniques can be used to produce conditional mutants, in which a selected gene becomes disrupted in a specific tissue at a certain time in development. The strategy takes advantage of a site-specific recombination system to excise—and thus disable—

the target gene in a particular place or at a particular time. The most common of these recombination systems called **Cre/lox**, is widely used to engineer gene replacements in mice and in plants (see Figure 5–82). In this case the target gene in ES cells is replaced by a fully functional version of the gene that is flanked by a pair of the short DNA sequences, called lox sites, that are recognized by the Cre recombinase protein. The transgenic mice that result are phenotypically normal. They are then mated with transgenic mice that express the Cre recombinase gene under the control of an inducible promoter. In the specific cells or tissues in which Cre is switched on, it catalyzes recombination between the lox sequences—excising a target gene and eliminating its activity. Similar recombination systems are used to generate conditional mutants in *Drosophila* (see Figure 21–48).

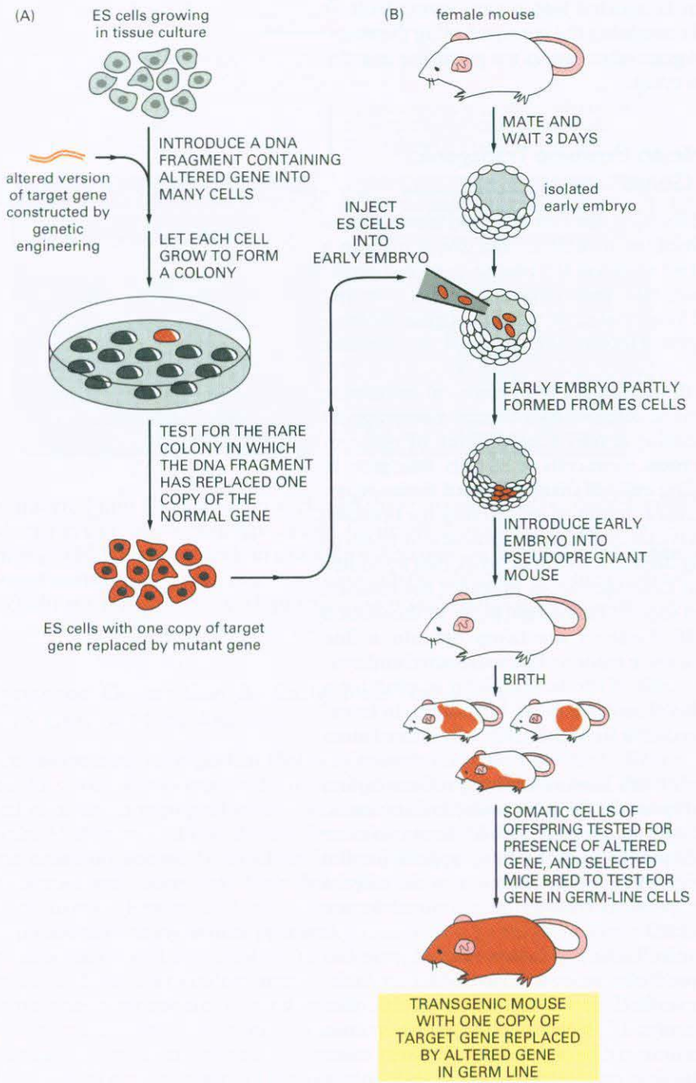


Figure 8–70 Summary of the procedures used for making gene replacements in mice. In the first step (A), an altered version of the gene is introduced into cultured ES (embryonic stem) cells. Only a few rare ES cells will have their corresponding normal genes replaced by the altered gene through a homologous recombination event. Although the procedure is often laborious, these rare cells can be identified and cultured to produce many descendants, each of which carries an altered gene in place of one of its two normal corresponding genes. In the next step of the procedure (B), these altered ES cells are injected into a very early mouse embryo; the cells are incorporated into the growing embryo, and a mouse produced by such an embryo will contain some somatic cells (indicated by *orange*) that carry the altered gene. Some of these mice will also contain germ-line cells that contain the altered gene. When bred with a normal mouse, some of the progeny of these mice will contain the altered gene in all of their cells. If two such mice are in turn bred (not shown), some of the progeny will contain two altered genes (one on each chromosome) in all of their cells.

If the original gene alteration completely inactivates the function of the gene, these mice are known as knockout mice. When such mice are missing genes that function during development, they often die with specific defects long before they reach adulthood. These defects are carefully analyzed to help decipher the normal function of the missing gene.

Transgenic Plants Are Important for Both Cell Biology and Agriculture

When a plant is damaged, it can often repair itself by a process in which mature differentiated cells "dedifferentiate," proliferate, and then redifferentiate into other cell types. In some circumstances the dedifferentiated cells can even form an apical meristem, which can then give rise to an entire new plant, including gametes. This remarkable plasticity of plant cells can be exploited to generate transgenic plants from cells growing in culture.

When a piece of plant tissue is cultured in a sterile medium containing nutrients and appropriate growth regulators, many of the cells are stimulated to proliferate indefinitely in a disorganized manner, producing a mass of relatively undifferentiated cells called a callus. If the nutrients and growth regulators are carefully manipulated, one can induce the formation of a shoot and then root apical meristems within the callus, and, in many species, a whole new plant can be regenerated.

Callus cultures can also be mechanically dissociated into single cells, which will grow and divide as a suspension culture. In several plants—including tobacco, petunia, carrot, potato, and *Arabidopsis*—a single cell from such a suspension culture can be grown into a small clump (a clone) from which a whole plant can be regenerated. Such a cell, which has the ability to give rise to all parts of the organism, is considered **totipotent**. Just as mutant mice can be derived by genetic manipulation of embryonic stem cells in culture, so transgenic plants can be created from single totipotent plant cells transfected with DNA in culture (Figure 8–72).

The ability to produce transgenic plants has greatly accelerated progress in many areas of plant cell biology. It has had an important role, for example, in isolating receptors for growth regulators and in analyzing the mechanisms of morphogenesis and of gene expression in plants. It has also opened up many new possibilities in agriculture that could benefit both the farmer and the consumer. It has made it possible, for example, to modify the lipid, starch, and protein storage reserved in seeds, to impart pest and virus resistance to plants, and to create modified plants that tolerate extreme habitats such as salt marshes or water-stressed soil.

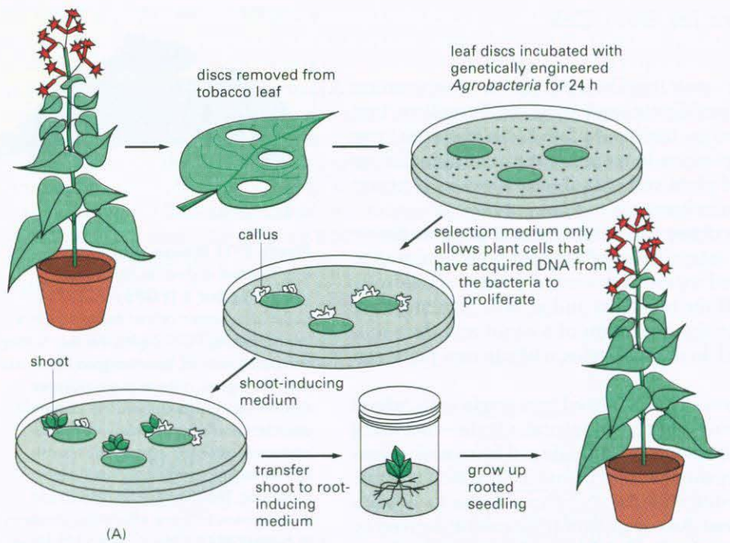
Many of the major advances in understanding animal development have come from studies on the fruit fly *Drosophila* and the nematode worm *Caenorhabditis elegans*, which are amenable to extensive genetic analysis as well as to experimental manipulation. Progress in plant developmental biology has, in the past, been relatively slow by comparison. Many of the plants that have proved most amenable to genetic analysis—such as maize and tomato—have long life cycles and very large genomes, making both classical and molecular genetic analysis time-consuming. Increasing attention is consequently being paid to a fast-growing small weed, the common wall cress *(Arabidopsis thaliana)*, which has several major advantages as a "model plant" (see Figures 1–46 and 21–107). The relatively small *Arabidopsis* genome was the first plant genome to be completely sequenced.

Figure 8–71 Mouse with an engineered defect in fibroblast growth factor 5 (FGF5). FGF5 is a negative regulator of hair formation. In a mouse lacking FGF5 *(right)*, the hair is long compared with its heterozygous littermate *(left)*. Transgenic mice with phenotypes that mimic aspects of a variety of human disorders, including Alzheimer's disease, atherosclerosis, diabetes, cystic fibrosis, and some type of cancers, have been generated. Their study may lead to the development of more effective treatments. (Courtesy of Gail Martin, from J.M. Hebert et al., *Cell* 78:1017–1025, 1994. © Elsevier.)

Large Collections of Tagged Knockouts Provide a Tool for Examining the Function of Every Gene in an Organism

Extensive collaborative efforts are underway to generate comprehensive libraries of mutations in several model organisms, including *S. cerevisiae, C. elegans, Drosophila, Arabidopsis*, and the mouse. The ultimate aim in each case is to produce a collection of mutant strains in which every gene in the organism has either been systematically deleted, or altered such that it can be conditionally disrupted. Collections of this type will provide an invaluable tool for investigating gene function on a genomic scale. In some cases, each of the individual mutants within the collection will sport a distinct molecular tag—a unique DNA sequence designed to make identification of the altered gene rapid and routine.

In *S. cerevisiae*, the task of generating a set of 6000 mutants, each missing



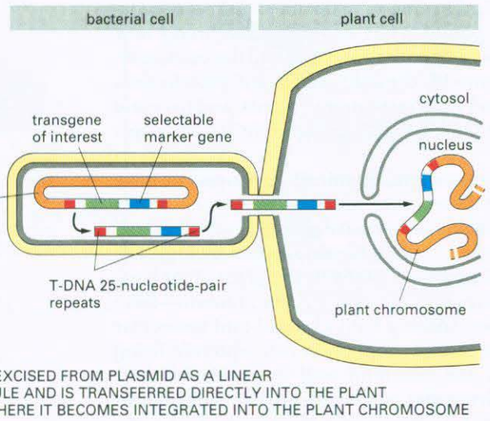
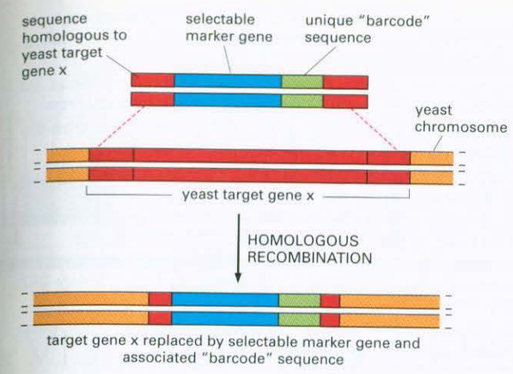


Figure 8–72 A procedure used to make a transgenic plant. (A) Outline of the process. A disc is cut out of a leaf and incubated in culture with *Agrobacteria* that carry a recombinant plasmid with both a selectable marker and a desired transgene. The wounded cells at the edge of the disc release substances that attract the *Agrobacteria* and cause them to inject DNA into these cells. Only those plant cells that take up the appropriate DNA and express the selectable marker gene survive to proliferate and form a callus. The manipulation of growth factors supplied to the callus induces it to form shoots that subsequently root and grow into adult plants carrying the transgene. (B) The preparation of the recombinant plasmid and its transfer to plant cells. An *Agrobacterium* plasmid that normally carries the T-DNA sequence is modified by substituting a selectable marker (such as the kanamycin-resistance gene) and a desired transgene between the 25-nucleotide-pair T-DNA repeats. When the *Agrobacterium* recognizes a plant cell, it efficiently passes a DNA strand that carries these sequences into the plant cell, using the special machinery that normally transfers the plasmid's T-DNA sequence.

only one gene, is made simpler by yeast's propensity for homologous recombination. For each gene, a "deletion cassette" is prepared. The cassette consists of a special DNA molecule that contains 50 nucleotides identical in sequence to each end of the targeted gene, surrounding a selectable marker. In addition, a special "barcode" sequence tag is embedded in this DNA molecule to facilitate the later rapid identification of each resulting mutant strain (Figure 8–73). A large mixture of such gene knockout mutants can then be grown under various selective test conditions—such as nutritional deprivation, temperature shift, or the presence of various drugs—and the cells that survive can be rapidly identified by their unique sequence tags. By assessing how well each mutant in the mixture fares, one can begin to assess which genes are essential, useful, or irrelevant for growth under various conditions.

The challenge in deriving information from the study of such yeast mutants lies in deducing a gene's activity or biological role based on a mutant phenotype.



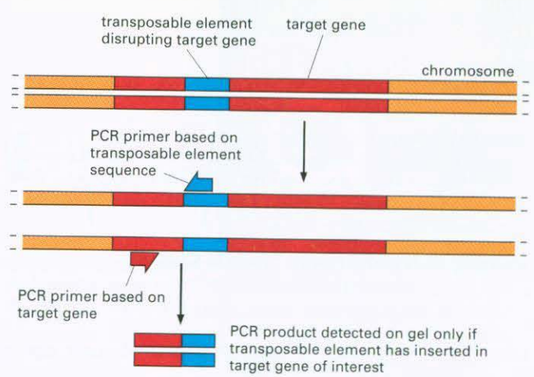


Figure 8–73 Making collections of mutant organisms. (A) A deletion cassette for use in yeast contains sequences homologous to each end of a target gene x *(red)*, a selectable marker *(blue)*, and a unique "barcode" sequence, approximately 20 nucleotide pairs in length *(green)*. This DNA is introduced into yeast, where it readily replaces the target gene by homologous recombination. (B) A similar approach can be taken to prepare tagged knockout mutants in *Arabidopsis* and *Drosophila*. In this case, mutants are generated by the accidental insertion of a transposable element into a target gene. The total DNA from the resulting organism can be collected and quickly screened for disruption of a gene of interest by using PCR primers that bind to the transposable element and to the target gene. A PCR product is detected on the gel only if the transposable element has inserted into the target gene.

Some defects—an inability to live without histidine, for example—point directly to the function of the wild-type gene. Other connections may not be so obvious. What might a sudden sensitivity to cold indicate about the role that a particular gene plays in the yeast cell? Such problems are even greater in organisms that are more complex than yeast. The loss of function of a single gene in the mouse, for example, can affect many different tissue types at different stages of development—whereas the loss of other genes is found to have no obvious effect. Adequately characterizing mutant phenotypes in mice often requires a thorough examination, along with extensive knowledge of mouse anatomy, histology, pathology, physiology, and complex behavior.

The insights generated by examination of mutant libraries, however, will be great. For example, studies of an extensive collection of mutants in *Mycoplasma genitalium*—the organism with the smallest known genome—have identified the minimum complement of genes essential for cellular life. Analysis of the mutant pool suggests that 265–350 of the 480 protein-coding genes in *M. genitalium* are required for growth under laboratory conditions. Approximately 100 of these essential genes are of unknown function, which suggests that a surprising number of the basic molecular mechanisms that underlie cellular life have yet to be discovered.

Summary

Genetics and genetic engineering provide powerful tools for the study of gene function in both cells and organisms. In the classical genetic approach, random mutagenesis is coupled with screening to identify mutants that are deficient in a particular biological process. These mutants are then used to locate and study the genes responsible for that process.

Gene function can also be ascertained by reverse genetic techniques. DNA engineering methods can be used to mutate any gene and to re-insert it into a cell's chromosomes so that it becomes a permanent part of the genome. If the cell used for this gene transfer is a fertilized egg (for an animal) or a totipotent plant cell in culture, transgenic organisms can be produced that express the mutant gene and pass it on to their progeny. Especially important for cell biology is the ability to alter cells and organisms in highly specific ways—allowing one to discern the effect on the cell or the organism of a designed change in a single protein or RNA molecule.

Many of these methods are being expanded to investigate gene function on a genome-wide scale. Technologies such as DNA microarrays can be used to monitor the expression of thousands of genes simultaneously, providing detailed, comprehensive snapshots of the dynamic patterns of gene expression that underlie complex cellular processes. And the generation of mutant libraries in which every gene in an organism has been systematically deleted or disrupted will provide an invaluable tool for exploring the role of each gene in the elaborate molecular collaboration that gives rise to life.

References

General

Ausubel FM, Brent R, Kingston RE et al. (eds) (1999) Short Protocols in Molecular Biology, 4th edn. New York: Wiley.

Brown TA (1999) Genomes. New York: Wiley-Liss.

Spector DL, Goldman RD, Leinwand LA (eds) (1998) Cells: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Watson JD, Gilman M, Witkowski J & Zoller M (1992) Recombinant DNA, 2nd edn. New York: WH Freeman.

Isolating Cells and Growing Them in Culture

Cohen S, Chang A, Boyer H & Helling R (1973) Construction of biologically functional bacterial plasmids *in vitro*. *Proc. Natl. Acad. Sci. USA* 70, 3240–3244.

Emmert-Buck, MR, Bonner RF, Smith PD et al. (1996) Laser capture microdissection. *Science* 274, 998–1001.

Freshney RI (2000) Culture of Animal Cells: A Manual of Basic Technique, 4th edn. New York: Wiley.

Ham RG (1965) Clonal growth of mammalian cells in a chemically defined, synthetic medium. *Proc. Natl. Acad. Sci. USA* 53, 288–293.

Harlow E & Lane D (1999) Using Antibodies: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Herzenberg LA, Sweet RG & Herzenberg LA (1976) Fluorescence-activated cell sorting. *Sci. Am.* 234(3), 108–116.

Jackson D, Symons R & Berg P (1972) Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 69, 2904–2909.

Levi-Montalcini R (1987) The nerve growth factor thirty-five years later. *Science* 237, 1154–1162.

Milstein C (1980) Monoclonal antibodies. *Sci. Am.* 243(4), 66–74.

Fractionation of Cells

de Duve C & Beaufay H (1981) A short history of tissue fractionation. *J. Cell Biol.* 91, 293s–299s.

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227, 680–685.

Nirenberg MW & Matthaei JH (1961) The dependence of cell-free protein synthesis in *E. coli* on naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* 47, 1588–1602.

O'Farrell PH (1975) High-resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 250, 4007–4021.

Palade G (1975) Intracellular aspects of the process of protein synthesis. *Science* 189, 347–358.

Pandey A & Mann M (2000) Proteomics to study genes and genomes. *Nature* 405, 837–846.

Scopes RK & Cantor CR (1993) Protein Purification: Principles and Practice, 3rd edn. New York: Springer-Verlag.

Yates JR (1998) Mass spectrometry and the age of the proteome. *J. Mass. Spectrom.* 33, 1–19.

Isolating, Cloning, and Sequencing DNA

Adams MD, Celniker SE, Holt RA et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.

Alwine JC, Kemp DJ & Stark GR (1977) Method for detection of specific RNAs in agarose gels by transfer to diabenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. USA* 74, 5350–5354.

Blattner FR, Plunkett G, Bloch CA et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474.

Hunkapiller T, Kaiser RJ, Koop BK & Hood L Large-scale and automated DNA sequence determination. *Science* 254, 59–67.

International Human Genome Sequencing Consortium (2000) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Maniatis T et al. (1978) The isolation of structural genes from libraries of eucaryotic DNA. *Cell* 15, 687–701.

Saiki RK, Gelfand DH, Stoffel S et al. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239, 487–491.

Sambrook J, Russell D (2001) Molecular Cloning: A Laboratory Manual, 3rd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sanger F, Nicklen S & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98, 503–517.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018.

Venter JC, Adams MA, Myers EW et al. (2000) The sequence of the human genome. *Science* 291, 1304–1351.

Analyzing Protein Structure and Function

Bamdad C (1997) Surface plasmon resonance for measurements of biological interest, in Current Protocols in Molecular Biology (FM Ausubel, R Brent, RE Kingston et al. eds), pp 20.4.1–20.4.12. New York: Wiley.

Branden C & Tooze J (1999) Introduction to Protein Structure, 2nd edn. New York: Garland Publishing.

Fields S & Song O (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246.

Kendrew JC (1961) The three-dimensional structure of a protein molecule. *Sci. Am.* 205(6), 96–111.

MacBeath G & Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763.

Miyawaki A, Tsien RY (2000) Monitoring protein conformations and interactions by fluorescence resonance energy transfer between mutants of green fluorescent protein. *Methods Enzymol.* 327, 472–500.

Sali A & Kuriyan J (1999) Challenges at the frontiers of structural biology. *Trends Genet.* 15, M20–M24.

Wuthrich K (1989) Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* 243, 45–50.

Studying Gene Expression and Function

Botstein D, White RL, Skolnick M & Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Brent R (2000) Genomic Biology. *Cell* 100, 169–182.

Capecchi MR (2001) Generating Mice with Targeted Mutations. *Nat. Med.* 7, 1086-1090.

Coelho PS, Kumar A & Snyder M (2000) Genome-wide mutant collections: toolboxes for functional genomics. *Curr. Opin. Microbiol.* 3, 309–315.

DeRisi JL, Iyer VR & Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686.

Eisenberg D, Marcotte EM, Xenarios I & Yeates TO (2000) Protein function in the post-genomic era. *Nature* 415, 823–826.

Enright AJ, Illiopoulos I, Kyrpides NC & Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90.

Hartwell LH, Hood L, Goldberg ML et al. (2000) Genetics: From Genes to Genomes. Boston: McGraw-Hill.

Lockhart DJ & Winzeler EA (2000) Genomics, gene expression and DNA arrays. *Nature* 405, 827–836, 2000.

Nusslein-Volhard C & Weischaus E (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287, 795–801.

Palmiter RD & Brinster RL (1986) Germ line transformation of mice. *Annu. Rev. Genet.* 20, 465–499.

Rubin GM & Sprading AC (1982) Genetic transformation of *Drosophila* with transposable element vectors. *Science* 218, 348–353.

Tabara H, Grishok A & Mello CC (1998) RNAi in *C. elegans*: soaking in the genome sequence. *Science* 282, 430–431.

Weigel D & Glazebrook J (2001) *Arabidopsis*: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Garland
Vice President: Denise Schanck
Managing Editor: Sarah Gibbs
Senior Editorial Assistant: Kirsten Jenner
Managing Production Editor: Emma Hunt
Proofreader and Layout: Emma Hunt
Production Assistant: Angela Bennett
Text Editors: Marjorie Singer Anderson and Betsy Dilernia
Copy Editor: Bruce Goatly
Word Processors: Fran Dependahl, Misty Landers and Carol Winter
Designer: Blink Studio, London
Illustrator: Nigel Orme
Indexer: Janine Ross and Sherry Granum
Manufacturing: Nigel Eyre and Marion Morrow

Cell Biology Interactive
Artistic and Scientific Direction: Peter Walter
Narrated by: Julie Theriot
Production, Design, and Development: Mike Morales

Bruce Alberts received his Ph.D. from Harvard University and is President of the National Academy of Sciences and Professor of Biochemistry and Biophysics at the University of California, San Francisco. **Alexander Johnson** received his Ph.D. from Harvard University and is a Professor of Microbiology and Immunology at the University of California, San Francisco. **Julian Lewis** received his D.Phil. from the University of Oxford and is a Principal Scientist at the Imperial Cancer Research Fund, London. **Martin Raff** received his M.D. from McGill University and is at the Medical Research Council Laboratory for Molecular Cell Biology and Cell Biology Unit and in the Biology Department at University College London. **Keith Roberts** received his Ph.D. from the University of Cambridge and is Associate Research Director at the John Innes Centre, Norwich. **Peter Walter** received his Ph.D. from The Rockefeller University in New York and is Professor and Chairman of the Department of Biochemistry and Biophysics at the University of California, San Francisco, and an Investigator of the Howard Hughes Medical Institute.

Library of Congress Cataloging-in-Publicaton Data
Molecular biology of the cell / Bruce Alberts ... [et al.].-- 4th ed.
p. cm
Includes bibliographical references and index.
ISBN 0-8153-3218-1 (hardbound) -- ISBN 0-8153-4072-9 (pbk.)
1. Cytology. 2. Molecular biology. I. Alberts, Bruce.
[DNLM: 1. Cells. 2. Molecular Biology.]
QH581.2 .M64 2002
571.6--dc21

2001054471 CIP

Published by Garland Science, a member of the Taylor & Francis Group, 29 West 35th Street, New York, NY 10001-2299

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Front cover Human Genome: Reprinted by permission from *Nature*, International Human Genome Sequencing Consortium, 409:860–921, 2001 © Macmillan Magazines Ltd. Adapted from an image by Francis Collins, NHGRI; Jim Kent, UCSC; Ewan Birney, EBI; and Darryl Leja, NHGRI; showing a portion of Chromosome 1 from the initial sequencing of the human genome.

Back cover In 1967, the British artist Peter Blake created a design classic. Nearly 35 years later Nigel Orme (illustrator), Richard Denyer (photographer), and the authors have together produced an affectionate tribute to Mr Blake's image. With its gallery of icons and influences, its assembly created almost as much complexity, intrigue and mystery as the original. *Drosophila, Arabidopsis,* Dolly and the assembled company tempt you to dip inside where, as in the original, "a splendid time is guaranteed for all."
(Gunter Blobel, courtesy of The Rockefeller University; Marie Curie, Keystone Press Agency Inc; Darwin bust, by permission of the President and Council of the Royal Society; Rosalind Franklin, courtesy of Cold Spring Harbor Laboratory Archives; Dorothy Hodgkin, © The Nobel Foundation, 1964; James Joyce, etching by Peter Blake; Robert Johnson, photo booth self-portrait early 1930s, © 1986 Delta Haze Corporation all rights reserved, used by permission; Albert L. Lehninger, (unidentified photographer) courtesy of The Alan Mason Chesney Medical Archives of The Johns Hopkins Medical Institutions; Linus Pauling, from Ava Helen and Linus Pauling Papers, Special Collections, Oregon State University; Nicholas Poussin, courtesy of ArtToday.com; Barbara McClintock, © David Micklos, 1983; Andrei Sakharov, courtesy of Elena Bonner; Frederick Sanger, © The Nobel Foundation, 1958.)

MOLECULAR BIOLOGY OF THE CELL

fourth edition

Bruce Alberts

Alexander Johnson

Julian Lewis

Martin Raff

Keith Roberts

Peter Walter

Garland Science
Taylor & Francis Group