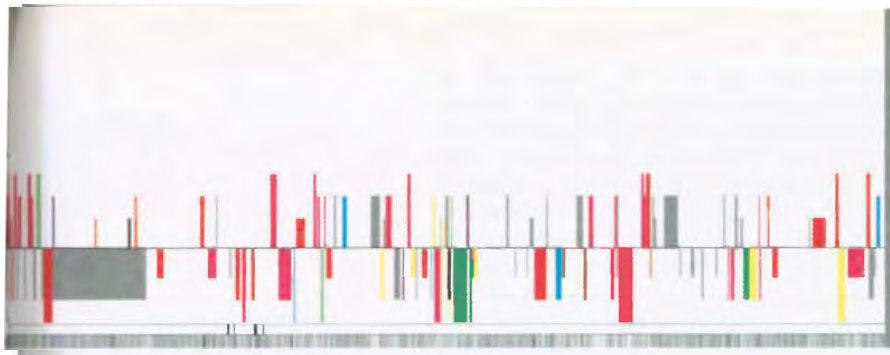


4

DNA AND CHROMOSOMES

THE STRUCTURE AND FUNCTION OF DNA

CHROMOSOMAL DNA AND ITS PACKAGING IN THE CHROMATIN FIBER

THE GLOBAL STRUCTURE OF CHROMOSOMES

Life depends on the ability of cells to store, retrieve, and translate the genetic instructions required to make and maintain a living organism. This *hereditary* information is passed on from a cell to its daughter cells at cell division, and from one generation of an organism to the next through the organism's reproductive cells. These instructions are stored within every living cell as its **genes**, the information-containing elements that determine the characteristics of a species as a whole and of the individuals within it.

As soon as genetics emerged as a science at the beginning of the twentieth century, scientists became intrigued by the chemical structure of genes. The information in genes is copied and transmitted from cell to daughter cell millions of times during the life of a multicellular organism, and it survives the process essentially unchanged. What form of molecule could be capable of such accurate and almost unlimited replication and also be able to direct the development of an organism and the daily life of a cell? What kind of instructions does the genetic information contain? How are these instructions physically organized so that the enormous amount of information required for the development and maintenance of even the simplest organism can be contained within the tiny space of a cell?

The answers to some of these questions began to emerge in the 1940s, when researchers discovered, from studies in simple fungi, that genetic information consists primarily of instructions for making proteins. Proteins are the macromolecules that perform most cellular functions: they serve as building blocks for cellular structures and form the enzymes that catalyze all of the cell's chemical reactions (Chapter 3), they regulate gene expression (Chapter 7), and they enable cells to move (Chapter 16) and to communicate with each other (Chapter 15). The properties and functions of a cell are determined almost entirely by the proteins it is able to make. With hindsight, it is hard to imagine what other type of instructions the genetic information could have contained.

Illumina Exhibit 1031, Page 1 of 53

The other crucial advance made in the 1940s was the identification of **deoxyribonucleic acid (DNA)** as the likely carrier of genetic information. But the mechanism whereby the hereditary information is copied for transmission from cell to cell, and how proteins are specified by the instructions in the DNA, remained completely mysterious. Suddenly, in 1953, the mystery was solved when the structure of DNA was determined by James Watson and Francis Crick. As mentioned in Chapter 1, the structure of DNA immediately solved the problem of how the information in this molecule might be copied, or *replicated*. It also provided the first clues as to how a molecule of DNA might encode the instructions for making proteins. Today, the fact that DNA is the genetic material is so fundamental to biological thought that it is difficult to realize what an enormous intellectual gap this discovery filled.

Well before biologists understood the structure of DNA, they had recognized that genes are carried on *chromosomes*, which were discovered in the nineteenth century as threadlike structures in the nucleus of a eucaryotic cell that become visible as the cell begins to divide (Figure 4–1). Later, as biochemical analysis became possible, chromosomes were found to consist of both DNA and protein. We now know that the DNA carries the hereditary information of the cell (Figure 4–2). In contrast, the protein components of chromosomes function largely to package and control the enormously long DNA molecules so that they fit inside cells and can easily be accessed by them.

In this chapter we begin by describing the structure of DNA. We see how, despite its chemical simplicity, the structure and chemical properties of DNA make it ideally suited as the raw material of genes. The genes of every cell on Earth are made of DNA, and insights into the relationship between DNA and genes have come from experiments in a wide variety of organisms. We then consider how genes and other important segments of DNA are arranged on the long molecules of DNA that are present in chromosomes. Finally, we discuss how eucaryotic cells fold these long DNA molecules into compact chromosomes. This packing has to be done in an orderly fashion so that the chromosomes can be replicated and apportioned correctly between the two daughter cells at each cell division. It must also allow access of chromosomal DNA to enzymes that repair it when it is damaged and to the specialized proteins that direct the expression of its many genes.

This is the first of four chapters that deal with basic genetic mechanisms—the ways in which the cell maintains, replicates, expresses, and occasionally improves the genetic information carried in its DNA. In the following chapter (Chapter 5) we discuss the mechanisms by which the cell accurately replicates and repairs DNA; we also describe how DNA sequences can be rearranged through the process of genetic recombination. Gene expression—the process through which the information encoded in DNA is interpreted by the cell to guide the synthesis of proteins—is the main topic of Chapter 6. In Chapter 7, we describe how gene expression is controlled by the cell to ensure that each of the many thousands of proteins encrypted in its DNA is manufactured only at the proper time and place in the life of the cell. Following these four chapters on basic genetic mechanisms, we present an account of the experimental techniques used to study these and other processes that are fundamental to all cells (Chapter 8).

THE STRUCTURE AND FUNCTION OF DNA

Biologists in the 1940s had difficulty in accepting DNA as the genetic material because of the apparent simplicity of its chemistry. DNA was known to be a long polymer composed of only four types of subunits, which resemble one another chemically. Early in the 1950s, DNA was first examined by x-ray diffraction analysis, a technique for determining the three-dimensional atomic structure of a molecule (discussed in Chapter 8). The early x-ray diffraction results indicated that DNA was composed of two strands of the polymer wound into a helix. The observation that DNA was double-stranded was of crucial significance and

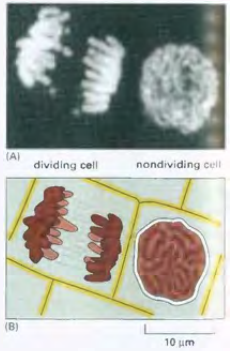


Figure 4–1 Chromosomes in cells. (A) Two adjacent plant cells photographed through a light microscope. The DNA has been stained with a fluorescent dye (DAPI) that binds to it. The DNA is present in chromosomes, which become visible as distinct structures in the light microscope only when they become compact structures in preparation for cell division, as shown on the left. The cell on the right, which is not dividing, contains identical chromosomes, but they cannot be clearly distinguished in the light microscope at this phase in the cell's life cycle, because they are in a more extended conformation. (B) Schematic diagram of the outlines of the two cells along with their chromosomes. (A, courtesy of Peter Shaw.)

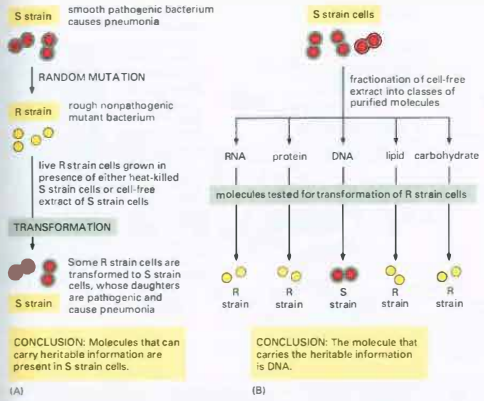


Figure 4–2 Experimental demonstration that DNA is the genetic material. These experiments, carried out in the 1940s, showed that adding purified DNA to a bacterium changed its properties and that this change was faithfully passed on to subsequent generations. Two closely related strains of the bacterium *Streptococcus pneumoniae* differ from each other in both their appearance under the microscope and their pathogenicity. One strain appears smooth (S) and causes death when injected into mice, and the other appears rough (R) and is nonlethal. (A) This experiment shows that a substance present in the S strain can change (or transform) the R strain into the S strain and that this change is inherited by subsequent generations of bacteria. (B) This experiment, in which the R strain has been incubated with various classes of biological molecules obtained from the S strain, identifies the substance as DNA.

provided one of the major clues that led to the Watson–Crick structure of DNA. Only when this model was proposed did DNA's potential for replication and information encoding become apparent. In this section we examine the structure of the DNA molecule and explain in general terms how it is able to store hereditary information.

A DNA Molecule Consists of Two Complementary Chains of Nucleotides

A DNA molecule consists of two long polynucleotide chains composed of four types of nucleotide subunits. Each of these chains is known as a *DNA chain*, or a *DNA strand*. *Hydrogen bonds* between the base portions of the nucleotides hold the two chains together (Figure 4–3). As we saw in Chapter 2 (Panel 2–6, pp. 120–121), nucleotides are composed of a five-carbon sugar to which are attached one or more phosphate groups and a nitrogen-containing base. In the case of the nucleotides in DNA, the sugar is deoxyribose attached to a single phosphate group (hence the name deoxyribonucleic acid), and the base may be either *adenine (A)*, *cytosine (C)*, *guanine (G)*, or *thymine (T)*. The nucleotides are covalently linked together in a chain through the sugars and phosphates, which thus form a "backbone" of alternating sugar–phosphate–sugar–phosphate (see Figure 4–3). Because only the base differs in each of the four types of subunits, each polynucleotide chain in DNA is analogous to a necklace (the backbone) strung with four types of beads (the four bases A, C, G, and T). These same symbols (A, C, G, and T) are also commonly used to denote the four different nucleotides—that is, the bases with their attached sugar and phosphate groups.

The way in which the nucleotide subunits are lined together gives a DNA strand a chemical polarity. If we think of each sugar as a block with a protruding knob (the 5′ phosphate) on one side and a hole (the 3′ hydroxyl) on the other (see Figure 4–3), each completed chain, formed by interlocking knobs with holes, will have all of its subunits lined up in the same orientation. Moreover, the two ends of the chain will be easily distinguishable, as one has a hole (the 3′ hydroxyl) and the other a knob (the 5′ phosphate) at its terminus. This polarity in a DNA chain is indicated by referring to one end as the *3′ end* and the other as the *5′ end*.

The three-dimensional structure of DNA—**the double helix**—arises from the chemical and structural features of its two polynucleotide chains. Because

Illumina Exhibit 1031, Page 2 of 53

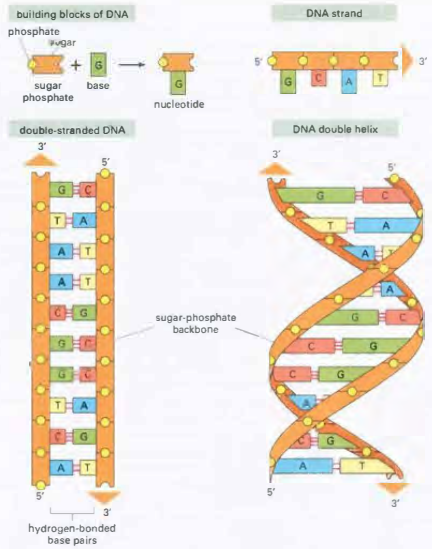


Figure 4–3 DNA and its building blocks. DNA is made of four types of nucleotides, which are linked covalently into a polynucleotide chain (a DNA strand) with a sugar-phosphate backbone from which the bases (A, C, G, and T) extend. A DNA molecule is composed of two DNA strands held together by hydrogen bonds between the paired bases. The *arrowheads* at the ends of the DNA strands indicate the polarities of the two strands, which run antiparallel to each other in the DNA molecule. In the diagram at the bottom left of the figure, the DNA molecule is shown straightened out; in reality, it is twisted into a double helix, as shown on the right. For details, see Figure 4–5.

these two chains are held together by hydrogen bonding between the bases on the different strands, all the bases are on the inside of the double helix, and the sugar-phosphate backbones are on the outside (see Figure 4–3). In each case, a bulkier two-ring base (a purine; see Panel 2–6, pp. 120–121) is paired with a single-ring base (a pyrimidine); A always pairs with T, and G with C (Figure 4–4). This *complementary base-pairing* enables the **base pairs** to be packed in the

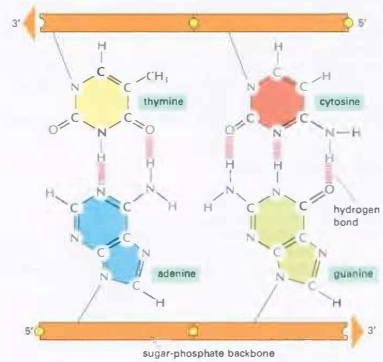


Figure 4–4 Complementary base pairs in the DNA double helix. The shapes and chemical structure of the bases allow hydrogen bonds to form efficiently only between A and T and between G and C, where atoms that are able to form hydrogen bonds (see Panel 2–3, pp. 114–115) can be brought close together without distorting the double helix. As indicated, two hydrogen bonds form between A and T, while three form between G and C. The bases can pair in this way only if the two polynucleotide chains that contain them are antiparallel to each other.

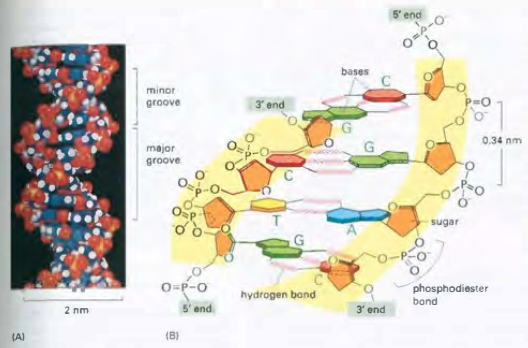


Figure 4–5 The DNA double helix. (A) A space-filling model of 1.5 turns of the DNA double helix. Each turn of DNA is made up of 10.4 nucleotide pairs and the center-to-center distance between adjacent nucleotide pairs is 3.4 nm. The coiling of the two strands around each other creates two grooves in the double helix. As indicated in the figure, the wider groove is called the major groove, and the smaller the minor groove. (B) A short section of the double helix viewed from its side, showing four base pairs. The nucleotides are linked together covalently by phosphodiester bonds through the 3′-hydroxyl (–OH) group of one sugar and the 5′-phosphate (P) of the next. Thus, each polynucleotide strand has a chemical polarity; that is, its two ends are chemically different. The 3′ end carries an unlinked –OH group attached to the 3′ position on the sugar ring; the 5′ end carries a free phosphate group attached to the 5′ position on the sugar ring.

energetically most favorable arrangement in the interior of the double helix. In this arrangement, each base pair is of similar width, thus holding the sugar-phosphate backbones an equal distance apart along the DNA molecule. To maximize the efficiency of base-pair packing, the two sugar-phosphate backbones wind around each other to form a double helix, with one complete turn every ten base pairs (Figure 4–5).

The members of each base pair can fit together within the double helix only if the two strands of the helix are **antiparallel**—that is, only if the polarity of one strand is oriented opposite to that of the other strand (see Figures 4–3 and 4–4). A consequence of these base-pairing requirements is that each strand of a DNA molecule contains a sequence of nucleotides that is exactly **complementary** to the nucleotide sequence of its partner strand.

The Structure of DNA Provides a Mechanism for Heredity

Genes carry biological information that must be copied accurately for transmission to the next generation each time a cell divides to form two daughter cells. Two central biological questions arise from these requirements: how can the information for specifying an organism be carried in chemical form, and how is it accurately copied? The discovery of the structure of the DNA double helix was a landmark in twentieth-century biology because it immediately suggested answers to both questions, thereby resolving at the molecular level the problem of heredity. We discuss briefly the answers to these questions in this section, and we shall examine them in more detail in subsequent chapters.

DNA encodes information through the order, or sequence, of the nucleotides along each strand. Each base—A, C, T, or G—can be considered as a letter in a four-letter alphabet that spells out biological messages in the chemical structure of the DNA. As we saw in Chapter 1, organisms differ from one another because their respective DNA molecules have different nucleotide sequences and, consequently, carry different biological messages. But how is the nucleotide alphabet used to make messages, and what do they spell out?

As discussed above, it was known well before the structure of DNA was determined that genes contain the instructions for producing proteins. The DNA messages must therefore somehow encode proteins (Figure 4–6). This relationship immediately makes the problem easier to understand, because of the chemical character of proteins. As discussed in Chapter 3, the properties of a protein, which are responsible for its biological function, are determined by its three-dimensional structure, and its structure is determined in turn by the linear

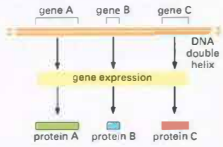


Figure 4–6 The relationship between genetic information carried in DNA and proteins.

Illumina Exhibit 1031, Page 3 of 53

Figure 4–7 The nucleotide sequence of the human β-globin gene. This gene carries the information for the amino acid sequence of one of the two types of subunits of the hemoglobin molecule, which carries oxygen in the blood. A different gene, the α-globin gene, carries the information for the other type of hemoglobin subunit (a hemoglobin molecule has four subunits, two of each type). Only one of the two strands of the DNA double helix containing the β-globin gene is shown; the other strand has the exact complementary sequence. By convention, a nucleotide sequence is written from its 5′ end to its 3′ end, and it should be read from left to right in successive lines down the page as though it were normal English text. The DNA sequences highlighted in *yellow* show the three regions of the gene that specify the amino sequence for the β-globin protein. We see in Chapter 6 how the cell connects these three sequences together to synthesize a full-length β-globin protein.

sequence of the amino acids of which it is composed. The linear sequence of nucleotides in a gene must therefore somehow spell out the linear sequence of amino acids in a protein. The exact correspondence between the four-letter nucleotide alphabet of DNA and the twenty-letter amino acid alphabet of proteins—the genetic code—is not obvious from the DNA structure, and it took over a decade after the discovery of the double helix before it was worked out. In Chapter 6 we describe this code in detail in the course of elaborating the process, known as *gene expression*, through which a cell translates the nucleotide sequence of a gene into the amino acid sequence of a protein.

The complete set of information in an organism's DNA is called its **genome**, and it carries the information for all the proteins the organism will ever synthesize. (The term genome is also used to describe the DNA that carries this information.) The amount of information contained in genomes is staggering: for example, a typical human cell contains 2 meters of DNA. Written out in the four-letter nucleotide alphabet, the nucleotide sequence of a very small human gene occupies a quarter of a page of text (Figure 4–7), while the complete sequence of nucleotides in the human genome would fill more than a thousand books the size of this one. In addition to other critical information, it carries the instructions for about 30,000 distinct proteins.

At each cell division, the cell must copy its genome to pass it to both daughter cells. The discovery of the structure of DNA also revealed the principle that makes this copying possible: because each strand of DNA contains a sequence of nucleotides that is exactly complementary to the nucleotide sequence of its partner strand, each strand can act as a **template**, or mold, for the synthesis of a new complementary strand. In other words, if we designate the two DNA strands as S and S′, strand S can serve as a template for making a new strand S′, while strand S′ can serve as a template for making a new strand S (Figure 4–8).

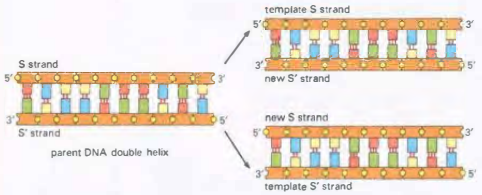


Figure 4–8 DNA as a template for its own duplication. As the nucleotide A successfully pairs only with T, and G with C, each strand of DNA can specify the sequence of nucleotides in its complementary strand. In this way, double-helical DNA can be copied precisely.

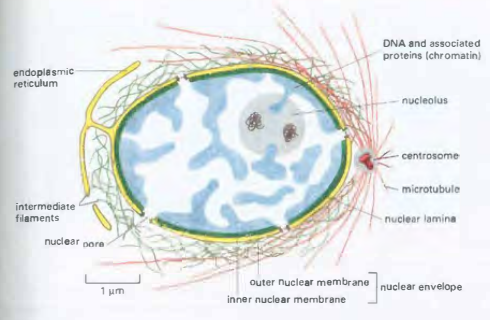


Figure 4–9 A cross-sectional view of a typical cell nucleus. The nuclear envelope consists of two membranes, the outer one being continuous with the endoplasmic reticulum membrane (see also Figure 12–9). The space inside the endoplasmic reticulum (the ER lumen) is colored *yellow*; it is continuous with the space between the two nuclear membranes. The lipid bilayers of the inner and outer nuclear membranes are connected at each nuclear pore. Two networks of intermediate filaments (*green*) provide mechanical support for the nuclear envelope; the intermediate filaments inside the nucleus form a special supporting structure called the nuclear lamina.

Thus, the genetic information in DNA can be accurately copied by the beautifully simple process in which strand S separates from strand S′, and each separated strand then serves as a template for the production of a new complementary partner strand that is identical to its former partner.

The ability of each strand of a DNA molecule to act as a template for producing a complementary strand enables a cell to copy, or *replicate*, its genes before passing them on to its descendants. In the next chapter we describe the elegant machinery the cell uses to perform this enormous task.

In Eucaryotes, DNA Is Enclosed in a Cell Nucleus

Nearly all the DNA in a eucaryotic cell is sequestered in a nucleus, which occupies about 10% of the total cell volume. This compartment is delimited by a *nuclear envelope* formed by two concentric lipid bilayer membranes that are punctured at intervals by large nuclear pores, which transport molecules between the nucleus and the cytosol. The nuclear envelope is directly connected to the extensive membranes of the endoplasmic reticulum. It is mechanically supported by two networks of intermediate filaments: one, called the *nuclear lamina*, forms a thin sheetlike meshwork inside the nucleus, just beneath the inner nuclear membrane; the other surrounds the outer nuclear membrane and is less regularly organized (Figure 4–9).

The nuclear envelope allows the many proteins that act on DNA to be concentrated where they are needed in the cell, and, as we see in subsequent chapters, it also keeps nuclear and cytosolic enzymes separate, a feature that is crucial for the proper functioning of eucaryotic cells. Compartmentalization, of which the nucleus is an example, is an important principle of biology; it serves to establish an environment in which biochemical reactions are facilitated by the high concentration of both substrates and the enzymes that act on them.

Summary

Genetic information is carried in the linear sequence of nucleotides in DNA. Each molecule of DNA is a double helix formed from two complementary strands of nucleotides held together by hydrogen bonds between G-C and A-T base pairs. Duplication of the genetic information occurs by the use of one DNA strand as a template for formation of a complementary strand. The genetic information stored in an organism's DNA contains the instructions for all the proteins the organism will ever synthesize. In eucaryotes, DNA is contained in the cell nucleus.

Illumina Exhibit 1031, Page 4 of 53

CHROMOSOMAL DNA AND ITS PACKAGING IN THE CHROMATIN FIBER

The most important function of DNA is to carry genes, the information that specifies all the proteins that make up an organism—including information about when, in what types of cells, and in what quantity each protein is to be made. The genomes of eucaryotes are divided up into chromosomes, and in this section we see how genes are typically arranged on each chromosome. In addition, we describe the specialized DNA sequences that allow a chromosome to be accurately duplicated and passed on from one generation to the next.

We also confront the serious challenge of DNA packaging. Each human cell contains approximately 2 meters of DNA if stretched end-to-end; yet the nucleus of a human cell, which contains the DNA, is only about 6 μm in diameter. This is geometrically equivalent to packing 40 km (24 miles) of extremely fine thread into a tennis ball! The complex task of packaging DNA is accomplished by specialized proteins that bind to and fold the DNA, generating a series of coils and loops that provide increasingly higher levels of organization, preventing the DNA from becoming an unmanageable tangle. Amazingly, although the DNA is very tightly folded, it is compacted in a way that allows it to easily become available to the many enzymes in the cell that replicate it, repair it, and use its genes to produce proteins.

Eucaryotic DNA Is Packaged into a Set of Chromosomes

In eucaryotes, the DNA in the nucleus is divided between a set of different **chromosomes**. For example, the human genome—approximately 3.2×10^9 nucleotides—is distributed over 24 different chromosomes. Each chromosome consists of a single, enormously long linear DNA molecule associated with proteins that fold and pack the fine DNA thread into a more compact structure. The complex of DNA and protein is called *chromatin* (from the Greek *chroma*, "colored," because of its staining properties). In addition to the proteins involved in packaging the DNA, chromosomes are also associated with many proteins required for the processes of gene expression, DNA replication, and DNA repair.

Bacteria carry their genes on a single DNA molecule, which is usually circular (see Figure 1–30). This DNA is associated with proteins that package and condense the DNA, but they are different from the proteins that perform these functions in eucaryotes. Although often called the bacterial "chromosome," it does not have the same structure as eucaryotic chromosomes, and less is known about how the bacterial DNA is packaged. Even less is known about how DNA is compacted in archaea. Therefore, our discussion of chromosome structure will focus almost entirely on eucaryotic chromosomes.

With the exception of the germ cells, and a few highly specialized cell types that cannot multiply and lack DNA altogether (for example, red blood cells), each human cell contains two copies of each chromosome, one inherited from the mother and one from the father. The maternal and paternal chromosomes of a pair are called **homologous chromosomes (homologs)**. The only nonhomologous chromosome pairs are the sex chromosomes in males, where a *Y chromosome* is inherited from the father and an *X chromosome* from the mother. Thus, each human cell contains a total of 46 chromosomes—22 pairs common to both males and females, plus two so-called sex chromosomes (X and Y in males, two Xs in females). *DNA hybridization* (described in detail in Chapter 8) can be used to distinguish these human chromosomes by "painting" each one a different color (Figure 4–10). Chromosome painting is typically done at the stage in the cell cycle when chromosomes are especially compacted and easy to visualize (mitosis, see below).

Another more traditional way to distinguish one chromosome from another is to stain them with dyes that produce a striking and reliable pattern of bands along each mitotic chromosome (Figure 4–11). The structural bases for these banding patterns are not well understood, and we return to this issue at the end of the chapter. Nevertheless, the pattern of bands on each type of chromosome is unique, allowing each chromosome to be identified and numbered.



Figure 4–10 Human chromosomes. These chromosomes, from a male, were isolated from a cell undergoing nuclear division (mitosis) and are therefore highly compacted. Each chromosome has been "painted" a different color to permit its unambiguous identification under the light microscope. Chromosome painting is performed by exposing the chromosomes to a collection of human DNA molecules that have been coupled to a combination of fluorescent dyes. For example, DNA molecules derived from chromosome 1 are labeled with one specific dye combination, those from chromosome 2 with another, and so on. Because the labeled DNA can form base pairs, or hybridize, only to the chromosome from which it was derived (discussed in Chapter 8), each chromosome is differently labeled. For such experiments, the chromosomes are subjected to treatments that separate the double-helical DNA into individual strands, designed to permit base-pairing with the single-stranded labeled DNA while keeping the chromosome structure relatively intact. (A) The chromosomes visualized as they originally spilled from the lysed cell. (B) The same chromosomes artificially lined up in their numerical order. This arrangement of the full chromosome set is called a karyotype. (From E. Schrock et al., *Science* 273:494–497, 1996. © AAAS.)

The display of the 46 human chromosomes at mitosis is called the human **karyotype**. If parts of chromosomes are lost, or switched between chromosomes, these changes can be detected by changes in the banding patterns or by changes in the pattern of chromosome painting (Figure 4–12). Cytogeneticists use these alterations to detect chromosome abnormalities that are associated with inherited defects or with certain types of cancer that arise through the rearrangement of chromosomes in somatic cells.



Figure 4–11 The banding patterns of human chromosomes. Chromosomes 1–22 are numbered in approximate order of size. A typical human somatic (non-germ line) cell contains two of each of these chromosomes, plus two sex chromosomes—two X chromosomes in a female, one X and one Y chromosome in a male. The chromosomes used to make these maps were stained at an early stage in mitosis, when the chromosomes are incompletely compacted. The *horizontal green line* represents the position of the centromere (see Figure 4–22), which appears as a constriction on mitotic chromosomes; the knobs on chromosomes 13, 14, 15, 21, and 22 indicate the positions of genes that code for the large ribosomal RNAs (discussed in Chapter 6). These patterns are obtained by staining chromosomes with Giemsa stain, and they can be observed under the light microscope. (Adapted from U. Franke, *Cytogenet. Cell Genet.* 31:24–32, 1981.)

Illumina Exhibit 1031, Page 5 of 53

Chromosomes Contain Long Strings of Genes

The most important function of chromosomes is to carry genes—the functional units of heredity. A gene is usually defined as a segment of DNA that contains the instructions for making a particular protein (or a set of closely related proteins). Although this definition holds for the majority of genes, several percent of genes produce an RNA molecule, instead of a protein, as their final product. Like proteins, these RNA molecules perform a diverse set of structural and catalytic functions in the cell, and we discuss them in detail in subsequent chapters.

As might be expected, a correlation exists between the complexity of an organism and the number of genes in its genome (see Table 1–1). For example, total gene numbers range from less than 500 for simple bacteria to about 30,000 for humans. Bacteria and some single-celled eucaryotes have especially compact genomes; the complete nucleotide sequence of their genomes reveals that the DNA molecules that make up their chromosomes are little more than strings of closely packed genes (Figure 4–13; see also Figure 1–30). However, chromosomes from many eucaryotes (including humans) contain, in addition to genes, a large excess of interspersed DNA that does not seem to carry critical information. Sometimes called junk DNA to signify that its usefulness to the cell has not been demonstrated, the particular nucleotide sequence of this DNA may not be important; but the DNA itself, by acting as spacer material, may be crucial for the long-term evolution of the species and for the proper expression of genes. These issues are taken up in detail in Chapter 7.

In general, the more complex the organism, the larger its genome, but because of differences in the amount of excess DNA, the relationship is not systematic (see Figure 1–38). For example, the human genome is 200 times larger than that of the yeast *S. cerevisiae*, but 30 times smaller than that of some plants and amphibians and 200 times smaller than a species of amoeba. Moreover, because of differences in the amount of excess DNA, the genomes of similar organisms (bony fish, for example) can vary several hundredfold in their DNA content, even though they contain roughly the same number of genes. Whatever the excess DNA may do, it seems clear that it is not a great handicap for a higher eucaryotic cell to carry a large amount of it.

The apportionment of the genome over chromosomes also differs from one eucaryotic species to the next. For example, compared with 46 for humans,



Figure 4–12 An aberrant human chromosome. (A) Two pairs of chromosomes, stained with Giemsa (see Figure 4–11), from a patient with ataxia, a disease characterized by progressive deterioration of motor skills. The patient has a normal pair of chromosome 4s (*left-hand pair*), but one normal chromosome 12 and one aberrant chromosome 12, as seen by its greater length (*right-hand pair*). The additional material contained on the aberrant chromosome 12 was deduced, from its pattern of bands, as a piece of chromosome 4 that had become attached to chromosome 12 through an abnormal recombination event, called a chromosomal translocation. (B) The same two chromosome pairs, "painted" *blue* for chromosome 4 DNA and *purple* for chromosome 12 DNA. The two techniques give rise to the same conclusion regarding the nature of the aberrant chromosome 12, but chromosome painting provides better resolution, and the clear identification of even short pieces of chromosomes that have become translocated. However, Giemsa staining is easier to perform. (From E. Schrock et al., *Science* 273:494–497, 1996. © AAAS.)

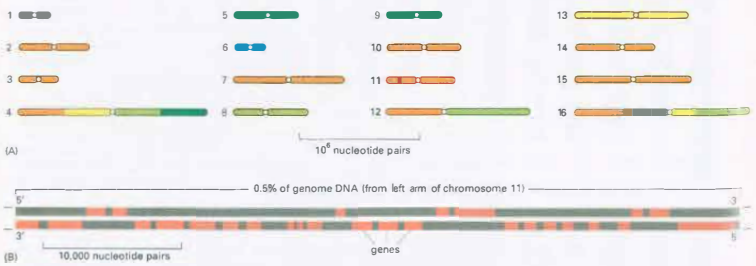


Figure 4–13 The genome of *S. cerevisiae* (budding yeast). (A) The genome is distributed over 16 chromosomes, and its complete nucleotide sequence was determined by a cooperative effort involving scientists working in many different locations, as indicated (*gray*, Canada; *orange*, European Union; *yellow*, United Kingdom; *blue*, Japan; *light green*, St Louis, Missouri; *dark green*, Stanford, California). The constriction present on each chromosome represents the position of its centromere (see Figure 4–22). (B) A small region of chromosome 11, highlighted in *red* in part A, is magnified to show the high density of genes characteristic of this species. As indicated by *orange*, some genes are transcribed from the lower strand (see Figure 1–5), while others are transcribed from the upper strand. There are about 6000 genes in the complete genome, which is 12,147,813 nucleotide pairs long.

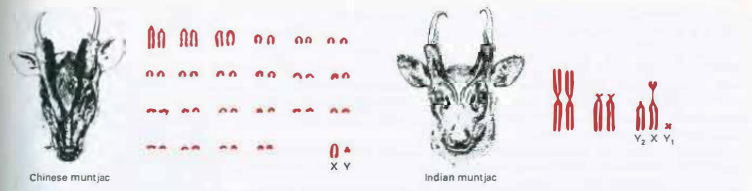


Figure 4–14 Two closely related species of deer with very different chromosome numbers. In the evolution of the Indian muntjac, initially separate chromosomes fused, without having a major effect on the animal. These two species have roughly the same number of genes. (Adapted from M.W. Strickberger, Evolution. 3rd edition, 2000, Sudbury, MA: Jones & Bartlett Publishers.)

somatic cells from a species of small deer contain only 6 chromosomes, while those from a species of carp contain over 100. Even closely related species with similar genome sizes can have very different numbers and sizes of chromosomes (Figure 4–14). Thus, there is no simple relationship between chromosome number, species complexity, and total genome size. Rather, the genomes and chromosomes of modern-day species have each been shaped by a unique history of seemingly random genetic events, acted on by selection pressures.

The Nucleotide Sequence of the Human Genome Shows How Genes Are Arranged in Humans

When the DNA sequence of human chromosome 22, one of the smallest human chromosomes (see Figure 4–11), was completed in 1999, it became possible for the first time to see exactly how genes are arranged along an entire vertebrate chromosome (Figure 4–15 and Table 4–1). With the publication of the "first draft" of the entire human genome in 2001, the genetic landscape of all human chromosomes suddenly came into sharp focus. The sheer quantity of information

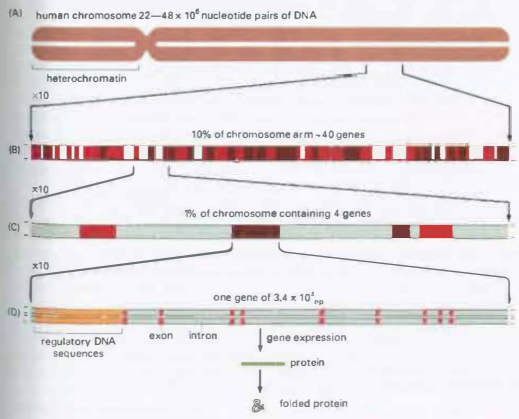


Figure 4–15 The organization of genes on a human chromosome. (A) Chromosome 22, one of the smallest human chromosomes, contains 48 x 10^6 nucleotide pairs and makes up approximately 1.5% of the entire human genome. Most of the left arm of chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin), which is discussed later in this chapter. (B) A tenfold expansion of a portion of chromosome 22, with about 40 genes indicated. Those in *dark brown* are known genes and those in *light brown* are predicted genes. (C) An expanded portion of (B) shows the entire length of several genes. (D) The intron–exon arrangement of a typical gene is shown after a further tenfold expansion. Each exon (*red*) codes for a portion of the protein, while the DNA sequence of the introns (*gray*) is relatively unimportant. The entire human genome (3.2 x 10^9 nucleotide pairs) is distributed over 22 autosomes and 2 sex chromosomes (see Figures 4–10 and 4–11). The term *human genome sequence* refers to the complete nucleotide sequence of DNA in these 24 chromosomes. Being diploid, a human somatic cell therefore contains roughly twice this amount of DNA. Humans differ from one another by an average of one nucleotide in every thousand, and a wide variety of humans contributed DNA for the genome sequencing project. The published human genome sequence is therefore a composite of many individual sequences. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001.)

Illumina Exhibit 1031, Page 6 of 53

provided by the Human Genome Project is unprecedented in biology (Figure 4–16 and Table 4–1); the human genome is 25 times larger than any other genome sequenced so far, and is 8 times as large as the sum of all previously sequenced genomes. At its peak, the Human Genome Project generated raw nucleotide sequences at a rate of 1000 nucleotides per second around the clock. It will be many decades before this information is fully analyzed, but it will continue to stimulate many new experiments and has already affected the content of all the chapters in this book.

Although there are many aspects to analyzing the human genome, here we simply make a few generalizations regarding the arrangement of genes in human chromosomes. The first striking feature of the human genome is how little of it (only a few percent) codes for proteins or structural and catalytic RNAs (Figure 4–17). Much of the remaining chromosomal DNA is made up of short, mobile pieces of DNA that have gradually inserted themselves in the chromosome over evolutionary time. We discuss these *transposable elements* in detail in later chapters.

A second notable feature of the human genome is the large average gene size of 27,000 nucleotide pairs. As discussed above, a typical gene carries in its linear sequence of nucleotides the information for the linear sequence of the amino acids of a protein. Only about 1300 nucleotide pairs are required to encode a protein of average size (about 430 amino acids in humans). Most of the remaining DNA in a gene consists of long stretches of noncoding DNA that interrupt the relatively short segments of DNA that code for protein. The coding sequences are called *exons*; the intervening (noncoding) sequences are called *introns* (see Figure 4–15 and Table 4–1).

The majority of human genes thus consist of a long string of alternating exons and introns, with most of the gene consisting of introns. In contrast, the majority of genes from organisms with compact genomes lack introns. This accounts for the much smaller size of their genes (about one-twentieth that of human genes), as well as for the much higher fraction of coding DNA in their chromosomes. In addition to introns and exons, each gene is associated with *regulatory DNA sequences*, which are responsible for ensuring that the gene is



Figure 4–16 Scale of the human genome. If each nucleotide pair is drawn as 1 mm as in (A), then the human genome would extend 3200 km (approximately 2000 miles), far enough to stretch across the center of Africa, the site of our human origins (*red line* in B). At this scale, there would be, on average, a protein-coding gene every 300 m. An average gene would extend for 30 m, but the coding sequences in this gene would add up to only just over a meter.

TABLE 4–1 Vital Statistics of Human Chromosome 22 and the Entire Human Genome

	CHROMOSOME 22	HUMAN GENOME
DNA length	48×10^6 nucleotide pairs*	3.2×10^9
Number of genes	approximately 700	approximately 30,000
Smallest protein-coding gene	1000 nucleotide pairs	not analyzed
Largest gene	583,000 nucleotide pairs	2.4×10^6 nucleotide pairs
Mean gene size	19,000 nucleotide pairs	27,000 nucleotide pairs
Smallest number of exons per gene	1	1
Largest number of exons per gene	54	178
Mean number of exons per gene	5.4	8.8
Smallest exon size	8 nucleotide pairs	not analyzed
Largest exon size	7600 nucleotide pairs	17,106 nucleotide pairs
Mean exon size	266 nucleotide pairs	145 nucleotide pairs
Number of pseudogenes**	more than 134	not analyzed
Percentage of DNA sequence in exons (protein coding sequences)	3%	1.5%
Percentage of DNA in high-copy repetitive elements	42%	approximately 50%
Percentage of total human genome	1.5%	100%

* The nucleotide sequence of 33.8×10^6 nucleotides is known; the rest of the chromosome consists primarily of very short repeated sequences that do not code for proteins or RNA.

** A pseudogene is a nucleotide sequence of DNA closely resembling that of a functional gene, but containing numerous deletion mutations that prevent its proper expression. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.

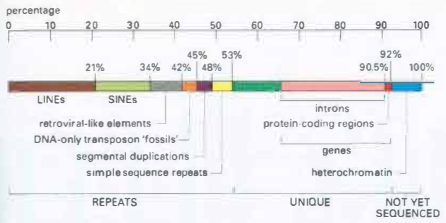


Figure 4–17 Representation of the nucleotide sequence content of the human genome. LINES, SINES, retroviral-like elements, and DNA-only transposons are all mobile genetic elements that have multiplied in our genome by replicating themselves and inserting the new copies in different positions. Mobile genetic elements are discussed in Chapter 5. Simple sequence repeats are short nucleotide sequences (less than 14 nucleotide pairs) that are repeated again and again for long stretches. Segmental duplications are large blocks of the genome (1000–200,000 nucleotide pairs) that are present at two or more locations in the genome. Over half of the unique sequence consists of genes and the remainder is probably regulatory DNA. Most of the DNA present in heterochromatin, a specialized type of chromatin (discussed later in this chapter) that contains relatively few genes, has not yet been sequenced. (Adapted from Unveiling the Human Genome, Supplement to the Wellcome Trust Newsletter. London: Wellcome Trust, February 2001.)

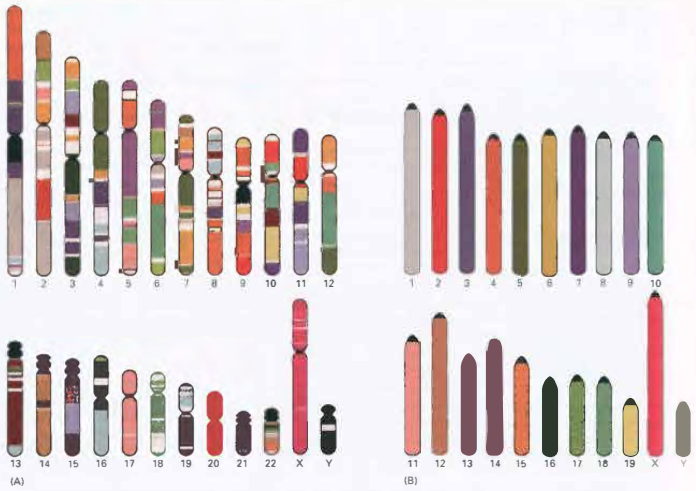
expressed at the proper level and time, and the proper type of cell. In humans, the regulatory sequences for a typical gene are spread out over tens of thousands of nucleotide pairs. As would be expected, these regulatory sequences are more compressed in organisms with compact genomes. We discuss in Chapter 7 how regulatory DNA sequences work.

Finally, the nucleotide sequence of the human genome has revealed that the critical information seems to be in an alarming state of disarray. As one commentator described our genome, "In some ways it may resemble your garage/bedroom/refrigerator/life: highly individualistic, but unkempt; little evidence of organization; much accumulated clutter (referred to by the uninitiated as 'junk'); virtually nothing ever discarded; and the few patently valuable items indiscriminately, apparently carelessly, scattered throughout."

Comparisons Between the DNAs of Related Organisms Distinguish Conserved and Nonconserved Regions of DNA Sequence

A major obstacle in interpreting the nucleotide sequences of human chromosomes is the fact that much of the sequence is probably unimportant. Moreover, the coding regions of the genome (the exons) are typically found in short segments (average size about 145 nucleotide pairs) floating in a sea of DNA whose exact nucleotide sequence is of little consequence. This arrangement makes it very difficult to identify all the exons in a stretch of DNA sequence; even harder is the determination of where a gene begins and ends and how many exons it spans. Accurate gene identification requires approaches that extract information from the inherently low signal-to-noise ratio of the human genome, and we describe some of them in Chapter 8. Here we discuss the most general approach, one that has the potential to identify not only coding sequences but also additional DNA sequences that are important. It is based on the observation that sequences that have a function are conserved during evolution, whereas those without a function are free to mutate randomly. The strategy is therefore to compare the human sequence with that of the corresponding regions of a related genome, such as that of the mouse. Humans and mice are thought to have diverged from a common mammalian ancestor about 100×10^6 years ago, which is long enough for the majority of nucleotides in their genomes to have been changed by random mutational events. Consequently, the only regions that will have remained closely similar in the two genomes are those in which mutations would have impaired function and put the animals carrying them at a disadvantage, resulting in their elimination from the population by natural selection. Such closely similar regions are known as *conserved regions*. In general, conserved regions represent functionally important exons and regulatory sequences. In contrast, *nonconserved regions* represent DNA whose sequence is generally not critical for function. By revealing in this way the results of a very long natural "experiment," comparative DNA sequencing studies highlight the most interesting regions in genomes.

Illumina Exhibit 1031, Page 7 of 53



Comparative studies of this kind have revealed not only that mice and humans share most of the same genes, but also that large blocks of the mouse and human genomes contain these genes in the same order, a feature called *conserved synteny* (Figure 4–18). Conserved synteny can also be revealed by chromosome painting, and this technique has been used to reconstruct the evolutionary history of our own chromosomes by comparing them with those from other mammals (Figure 4–19).

Figure 4–18 Conserved synteny between the human and mouse genomes. Regions from different mouse chromosomes (indicated by the colors of each mouse in B) show conserved synteny (gene order) with the indicated regions of the human genome (A). For example the genes present in the upper portion of human chromosome 1 (*orange*) are present in the same order in a portion of mouse chromosome 4. Regions of human chromosomes that are composed primarily of short, repeated sequences are shown in *black*. Mouse centromeres (indicated in *black* in B) are located at the ends of chromosomes; no known genes lie beyond the centromere on any mouse chromosome. For the most part, human centromeres, indicated by constrictions, occupy more internal positions on chromosomes (see Figure 4–11). (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001.)

Chromosomes Exist in Different States Throughout the Life of a Cell

We have seen how genes are arranged in chromosomes, but to form a functional chromosome, a DNA molecule must be able to do more than simply carry genes: it must be able to replicate, and the replicated copies must be separated and reliably partitioned into daughter cells at each cell division. This process occurs through an ordered series of stages, collectively known as the **cell cycle**. The cell cycle is briefly summarized in Figure 4–20, and discussed in detail in Chapter 17. Only two of the stages of the cycle concern us in this chapter. During *interphase* chromosomes are replicated, and during *mitosis* they become highly condensed and then are separated and distributed to the two daughter nuclei. The highly condensed chromosomes in a dividing cell are known as *mitotic chromosomes*. This is the form in which chromosomes are most easily visualized; in fact, all the images of chromosomes shown so far in the chapter are of chromosomes in mitosis. This condensed state is important in allowing the duplicated chromosomes to be separated by the mitotic spindle during cell division, as discussed in Chapter 18.

During the portions of the cell cycle when the cell is not dividing, the chromosomes are extended and much of their chromatin exists as long, thin tangled

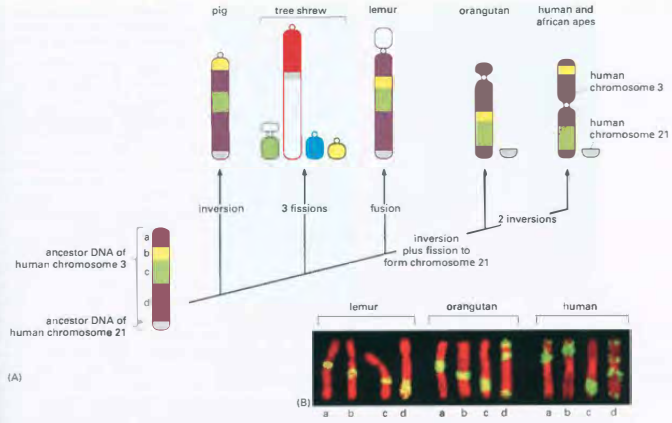


Figure 4–19 A proposed evolutionary history of human chromosome 3 and its relatives in other mammals. (A) At the *lower left* is the order of chromosome 3 segments hypothesized to be present on a chromosome of a mammalian ancestor. Along the *top* are the patterns of chromosome sequences found in the chromosomes of modern mammals. The minimum changes necessary to account for the appearance of the modern chromosomes from the hypothetical ancestor are marked along each branch. In mammals, these types of changes in chromosome organization are thought to occur once every 5–10 x 10^6 years. The *small circles* depicted in the modern chromosomes represent the positions of centromeres. (B) Some of the chromosome painting experiments that led to the diagram in (A). Each image shows the chromosome most closely related to human chromosome 3, painted *green* by hybridization with different segments of DNA, lettered a, b, c, and d along the *bottom* of the figure. These letters correspond to the colored segments of the diagram in (A). (From S. Muller et al., *Proc. Natl. Acad. Sci. USA* 97:206–211, 2000. © National Academy of Sciences.)

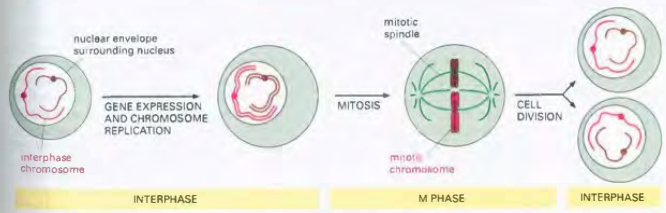


Figure 4–20 A simplified view of the eucaryotic cell cycle. During interphase, the cell is actively expressing its genes and is therefore synthesizing proteins. Also, during interphase and before cell division, the DNA is replicated and the chromosomes are duplicated. Once DNA replication is complete, the cell can enter *M phase*, when mitosis occurs and the nucleus is divided into two daughter nuclei. During this stage, the chromosomes condense, the nuclear envelope breaks down, and the mitotic spindle forms from microtubules and other proteins. The condensed mitotic chromosomes are captured by the mitotic spindle, and one complete set of chromosomes is then pulled to each end of the cell. A nuclear envelope re-forms around each chromosome set, and in the final step of M phase, the cell divides to produce two daughter cells. Most of the time in the cell cycle is spent in interphase; M phase is brief in comparison, occupying only about an hour in many mammalian cells.

Illumina Exhibit 1031, Page 8 of 53

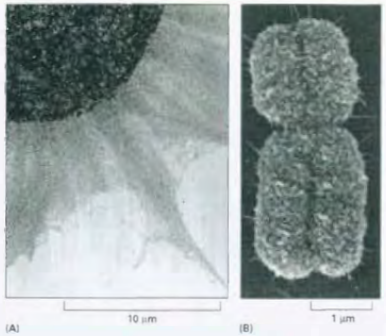


Figure 4–21 A comparison of extended interphase chromatin with the chromatin in a mitotic chromosome. (A) An electron micrograph showing an enormous tangle of chromatin spilling out of a lysed interphase nucleus. (B) A scanning electron micrograph of a mitotic chromosome: a condensed duplicated chromosome in which the two new chromosomes are still linked together (see Figure 4–22). The constricted region indicates the position of the centromere. Note the difference in scales. (A, courtesy of Victoria Foe; B, courtesy of Terry D. Allen.)

threads in the nucleus so that individual chromosomes cannot be easily distinguished (Figure 4–21). We refer to chromosomes in this extended state as *interphase chromosomes*.

Each DNA Molecule That Forms a Linear Chromosome Must Contain a Centromere, Two Telomeres, and Replication Origins

A chromosome operates as a distinct structural unit: for a copy to be passed on to each daughter cell at division, each chromosome must be able to replicate, and the newly replicated copies must subsequently be separated and partitioned correctly into the two daughter cells. These basic functions are controlled by three types of specialized nucleotide sequence in the DNA, each of which binds specific proteins that guide the machinery that replicates and segregates chromosomes (Figure 4–22).

Experiments in yeasts, whose chromosomes are relatively small and easy to manipulate, have identified the minimal DNA sequence elements responsible for each of these functions. One type of nucleotide sequence acts as a DNA **replication origin**, the location at which duplication of the DNA begins. Eucaryotic chromosomes contain many origins of replication to ensure that the entire chromosome can be replicated rapidly, as discussed in detail in Chapter 5.

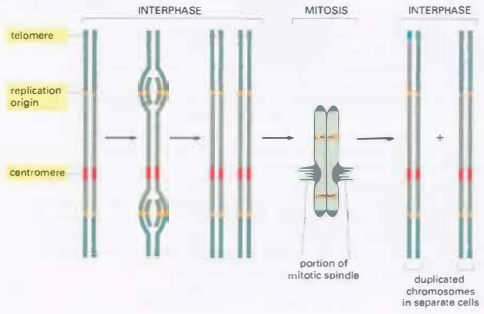


Figure 4–22 The three DNA sequences required to produce a eucaryotic chromosome that can be replicated and then segregated at mitosis. Each chromosome has multiple origins of replication, one centromere, and two telomeres. Shown here is the sequence of events a typical chromosome follows during the cell cycle. The DNA replicates in interphase beginning at the origins of replication and proceeding bidirectionally from the origins across the chromosome. In M phase, the centromere attaches the duplicated chromosomes to the mitotic spindle so that one copy is distributed to each daughter cell during mitosis. The centromere also helps to hold the duplicated chromosomes together until they are ready to be moved apart. The telomeres form special caps at each chromosome end.

After replication, the two daughter chromosomes remain attached to one another and, as the cell cycle proceeds, are condensed further to produce mitotic chromosomes. The presence of a second specialized DNA sequence, called a **centromere**, allows one copy of each duplicated and condensed chromosome to be pulled into each daughter cell when a cell divides. A protein complex called a *kinetochore* forms at the centromere and attaches the duplicated chromosomes to the mitotic spindle, allowing them to be pulled apart (discussed in Chapter 18).

The third specialized DNA sequence forms **telomeres**, the ends of a chromosome. Telomeres contain repeated nucleotide sequences that enable the ends of chromosomes to be efficiently replicated. Telomeres also perform another function: the repeated telomere DNA sequences, together with the regions adjoining them, form structures that protect the end of the chromosome from being recognized by the cell as a broken DNA molecule in need of repair. We discuss this type of repair and the other features of telomeres in Chapter 5.

In yeast cells, the three types of sequences required to propagate a chromosome are relatively short (typically less than 1000 base pairs each) and therefore use only a tiny fraction of the information-carrying capacity of a chromosome. Although telomere sequences are fairly simple and short in all eucaryotes, the DNA sequences that specify centromeres and replication origins in more complex organisms are much longer than their yeast counterparts. For example, experiments suggest that human centromeres may contain up to 100,000 nucleotide pairs. It has been proposed that human centromeres may not even require a stretch of DNA with a defined nucleotide sequence; instead, they may simply create a large, regularly repeating protein–nucleic acid structure. We return to this issue at the end of the chapter when we discuss in more general terms the proteins that, along with DNA, make up chromosomes.

DNA Molecules Are Highly Condensed in Chromosomes

All eucaryotic organisms have elaborate ways of packaging DNA into chromosomes. Recall from earlier in this chapter that human chromosome 22 contains about 48 million nucleotide pairs. Stretched out end to end, its DNA would extend about 1.5 cm. Yet, when it exists as a mitotic chromosome, chromosome 22 measures only about 2 μm in length (see Figures 4–10 and 4–11), giving an end-to-end compaction ratio of nearly 10,000-fold. This remarkable feat of compression is performed by proteins that successively coil and fold the DNA into higher and higher levels of organization. Although less condensed than mitotic chromosomes, the DNA of interphase chromosomes is still tightly packed, with an overall compaction ratio of approximately 1000-fold. In the next sections we discuss the specialized proteins that make the compression possible.

In reading these sections it is important to keep in mind that chromosome structure is dynamic. Not only do chromosomes globally condense in accord with the cell cycle, but different regions of the interphase chromosomes condense and decondense as the cells gain access to specific DNA sequences for gene expression, DNA repair, and replication. The packaging of chromosomes must therefore be accomplished in a way that allows rapid localized, on-demand access to the DNA.

Nucleosomes Are the Basic Unit of Eucaryotic Chromosome Structure

The proteins that bind to the DNA to form eucaryotic chromosomes are traditionally divided into two general classes: the **histones** and the *nonhistone chromosomal proteins*. The complex of both classes of protein with the nuclear DNA of eucaryotic cells is known as **chromatin**. Histones are present in such enormous quantities in the cell (about 60 million molecules of each type per human cell) that their total mass in chromatin is about equal to that of the DNA.

Histones are responsible for the first and most basic level of chromosome organization, the **nucleosome**, which was discovered in 1974. When interphase

Illumina Exhibit 1031, Page 9 of 53

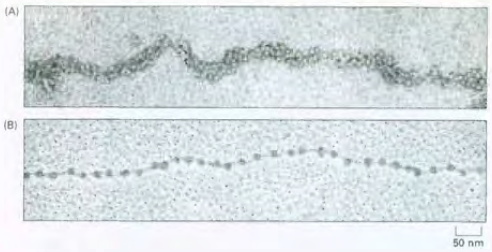


Figure 4–23 Nucleosomes as seen in the electron microscope. (A) Chromatin isolated directly from an interphase nucleus appears in the electron microscope as a thread 30 nm thick. (B) This electron micrograph shows a length of chromatin that has been experimentally unpacked, or decondensed, after isolation to show the nucleosomes. (A, courtesy of Barbara Hamkalo; B, courtesy of Victoria Foe.)

nuclei are broken open very gently and their contents examined under the electron microscope, most of the chromatin is in the form of a fiber with a diameter of about 30 nm (Figure 4–23A). If this chromatin is subjected to treatments that cause it to unfold partially, it can be seen under the electron microscope as a series of "beads on a string" (Figure 4–23B). The string is DNA, and each bead is a "nucleosome core particle" that consists of DNA wound around a protein core formed from histones. The beads on a string represent the first level of chromosomal DNA packing.

The structural organization of nucleosomes was determined after first isolating them from unfolded chromatin by digestion with particular enzymes (called nucleases) that break down DNA by cutting between the nucleosomes. After digestion for a short period, the exposed DNA between the nucleosome core particles, the *linker DNA*, is degraded. Each individual nucleosome core particle consists of a complex of eight histone proteins—two molecules each of histones H2A, H2B, H3, and H4—and double-stranded DNA that is 146 nucleotide pairs long. The *histone octamer* forms a protein core around which the double-stranded DNA is wound (Figure 4–24).

Each nucleosome core particle is separated from the next by a region of linker DNA, which can vary in length from a few nucleotide pairs up to about 80. (The term *nucleosome* technically refers to a nucleosome core particle plus one of its adjacent DNA linkers, but it is often used synonymously with nucleosome core particle.) On average, therefore, nucleosomes repeat at intervals of about 200 nucleotide pairs. For example, a diploid human cell with 6.4×10^9 nucleotide pairs contains approximately 30 million nucleosomes. The formation of nucleosomes converts a DNA molecule into a chromatin thread about one-third of its initial length, and this provides the first level of DNA packing.

The Structure of the Nucleosome Core Particle Reveals How DNA Is Packaged

The high-resolution structure of a nucleosome core particle, solved in 1997, revealed a disc-shaped histone core around which the DNA was tightly wrapped 1.65 turns in a left-handed coil (Figure 4–25). All four of the histones that make up the core of the nucleosome are relatively small proteins (102–135 amino

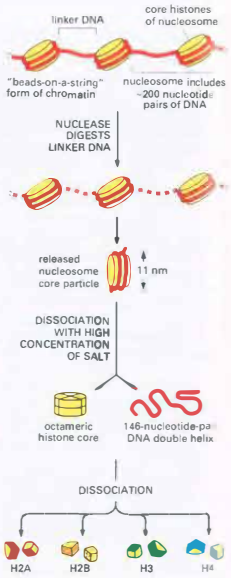


Figure 4–24 Structural organization of the nucleosome. A nucleosome contains a protein core made of eight histone molecules. As indicated, the nucleosome core particle is released from chromatin by digestion of the linker DNA with a nuclease, an enzyme that breaks down DNA. (The nuclease can degrade the exposed linker DNA but cannot attack the DNA wound tightly around the nucleosome core.) After dissociation of the isolated nucleosome into its protein core and DNA, the length of the DNA that was wound around the core can be determined. This length of 146 nucleotide pairs is sufficient to wrap 1.65 times around the histone core.

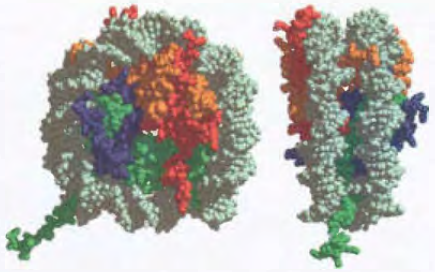


Figure 4–25 The structure of a nucleosome core particle, as determined by x-ray diffraction analyses of crystals. Each histone is colored according to the scheme of Figure 4–26, with the DNA double helix in *light gray*. (Reprinted by permission from K. Luger et al., *Nature* 389:251–260, 1997. © Macmillan Magazines Ltd.)

acids), and they share a structural motif, known as the *histone fold*, formed from three α helices connected by two loops (Figure 4–26). In assembling a nucleosome, the histone folds first bind to each other to form H3–H4 and H2A–H2B dimers, and the H3–H4 dimers combine to form tetramers. An H3–H4 tetramer then further combines with two H2A–H2B dimers to form the compact octamer core, around which the DNA is wound (Figure 4–27).

The interface between DNA and histone is extensive: 142 hydrogen bonds are formed between DNA and the histone core in each nucleosome. Nearly half of these bonds form between the amino acid backbone of the histones and the phosphodiester backbone of the DNA. Numerous hydrophobic interactions and salt linkages also hold DNA and protein together in the nucleosome. For example, all the core histones are rich in lysine and arginine (two amino acids with basic side chains), and their positive charges can effectively neutralize the negatively charged DNA backbone. These numerous interactions explain in part why DNA of virtually any sequence can be bound on a histone octamer core. The path of the DNA around the histone core is not smooth; rather, several kinks are seen in the DNA, as expected from the nonuniform surface of the core.

In addition to its histone fold, each of the core histones has a long N-terminal amino acid "tail", which extends out from the DNA-histone core (see Figure 4–27). These histone tails are subject to several different types of covalent modifications, which control many aspects of chromatin structure. We discuss these issues later in the chapter.

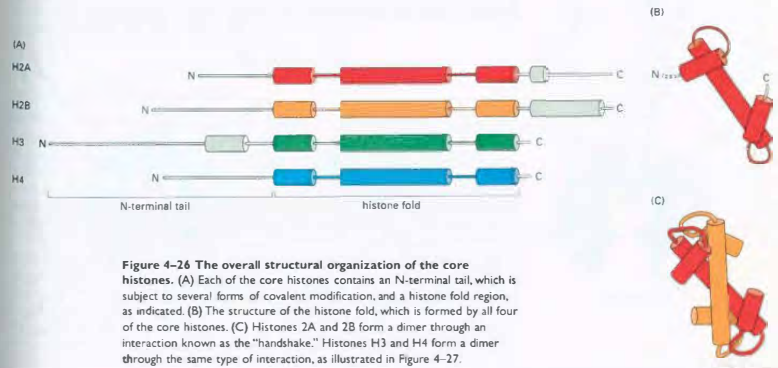


Figure 4–26 The overall structural organization of the core histones. (A) Each of the core histones contains an N-terminal tail, which is subject to several forms of covalent modification, and a histone fold region, as indicated. (B) The structure of the histone fold, which is formed by all four of the core histones. (C) Histones 2A and 2B form a dimer through an interaction known as the "handshake." Histones H3 and H4 form a dimer through the same type of interaction, as illustrated in Figure 4–27.

Illumina Exhibit 1031, Page 10 of 53

As might be expected from their fundamental role in DNA packaging, the histones are among the most highly conserved eucaryotic proteins. For example, the amino acid sequence of histone H4 from a pea and a cow differ at only at 2 of the 102 positions. This strong evolutionary conservation suggests that the functions of histones involve nearly all of their amino acids, so that a change in any position is deleterious to the cell. This suggestion has been tested directly in yeast cells, in which it is possible to mutate a given histone gene *in vitro* and introduce it into the yeast genome in place of the normal gene. As might be expected, most changes in histone sequences are lethal; the few that are not lethal cause changes in the normal pattern of gene expression, as well as other abnormalities.

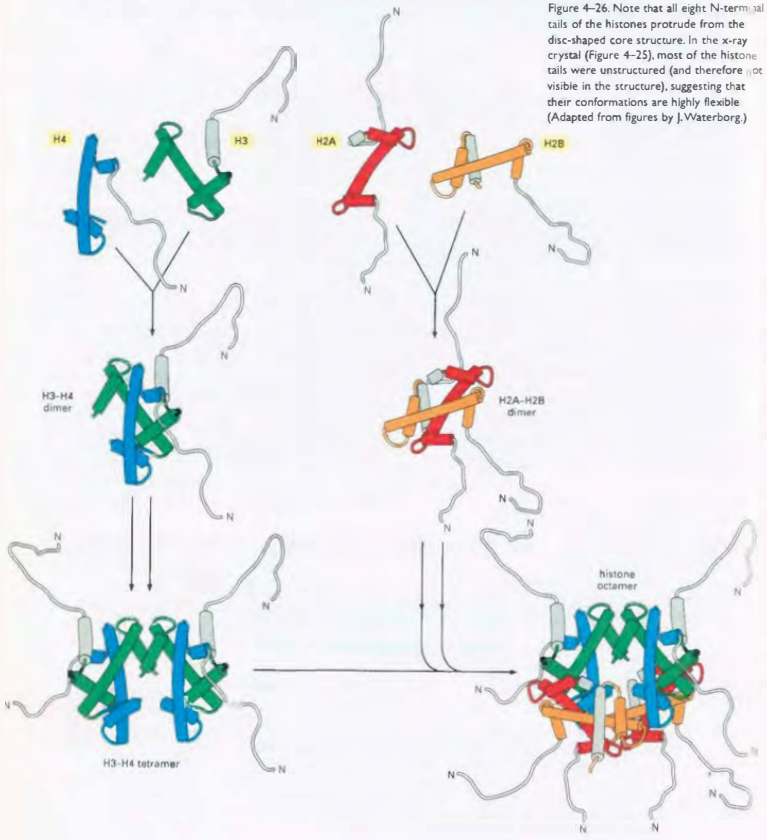


Figure 4–27 The assembly of a histone octamer. The histone H3–H4 dimer and the H2A–H2B dimer are formed from the handshake interaction. An H3–H4 tetramer forms the scaffold of the octamer onto which two H2A–H2B dimers are added, to complete the assembly. The histones are colored as in Figure 4–26. Note that all eight N-terminal tails of the histones protrude from the disc-shaped core structure. In the x-ray crystal (Figure 4–25), most of the histone tails were unstructured (and therefore not visible in the structure), suggesting that their conformations are highly flexible. (Adapted from figures by J. Waterborg.)

Despite the high conservation of the core histones, many eucaryotic organisms also produce specialized variant core histones that differ in amino acid sequence from the main ones. For example, the sea urchin has five histone H2A variants, each of which is expressed at a different time during development. It is thought that nucleosomes that have incorporated these variant histones differ in stability from regular nucleosomes, and they may be particularly well suited for the high rates of DNA transcription and DNA replication that occur during these early stages of development.

The Positioning of Nucleosomes on DNA Is Determined by Both DNA Flexibility and Other DNA-bound Proteins

Although nearly every DNA sequence can, in principle, be folded into a nucleosome, the spacing of nucleosomes in the cell can be irregular. Two main influences determine where nucleosomes form in the DNA. One is the difficulty of bending the DNA double helix into two tight turns around the outside of the histone octamer, a process that requires substantial compression of the minor groove of the DNA helix. Because A-T-rich sequences in the minor groove are easier to compress than G-C-rich sequences, each histone octamer tends to position itself on the DNA so as to maximize A-T-rich minor grooves on the inside of the DNA coil (Figure 4–28). Thus, a segment of DNA that contains short A-T-rich sequences spaced by an integral number of DNA turns is easier to bend around the nucleosome than a segment of DNA lacking this feature. In addition, because the DNA in a nucleosome is kinked in several places, the ability of a given nucleotide sequence to accommodate this deformation can also influence the position of DNA on the nucleosome.

These features of DNA probably explain some striking, but unusual, cases of very precise positioning of nucleosomes along a stretch of DNA. For most of the DNA sequences found in chromosomes, however, there is no strongly preferred nucleosome-binding site; a nucleosome can occupy any one of a number of positions relative to the DNA sequence.

The second, and probably most important, influence on nucleosome positioning is the presence of other tightly bound proteins on the DNA. Some bound proteins favor the formation of a nucleosome adjacent to them. Others create obstacles that force the nucleosomes to assemble at positions between them. Finally, some proteins can bind tightly to DNA even when their DNA-binding site is part of a nucleosome. The exact positions of nucleosomes along a stretch of DNA therefore depend on factors that include the DNA sequence and the presence and nature of other proteins bound to the DNA. Moreover, as we see below, the arrangement of nucleosomes on DNA is highly dynamic, changing rapidly according to the needs of the cell.

Nucleosomes Are Usually Packed Together into a Compact Chromatin Fiber

Although long strings of nucleosomes form on most chromosomal DNA, chromatin in a living cell probably rarely adopts the extended "beads on a string" form. Instead, the nucleosomes are packed on top of one another, generating regular arrays in which the DNA is even more highly condensed. Thus, when nuclei are very gently lysed onto an electron microscope grid, most of the chromatin is seen to be in the form of a fiber with a diameter of about 30 nm, which is considerably wider than chromatin in the "beads on a string" form (see Figure 4–23).

Several models have been proposed to explain how nucleosomes are packed in the 30-nm chromatin fiber; the one most consistent with the available data is a series of structural variations known collectively as the Zigzag model (Figure 4–29). In reality, the 30-nm structure found in chromosomes is probably a fluid mosaic of the different zigzag variations. We saw earlier that the linker DNA that connects adjacent nucleosomes can vary in length; these differences in linker length probably introduce further local perturbations into the zigzag structure. Finally, the presence of other DNA-binding proteins and DNA sequence that

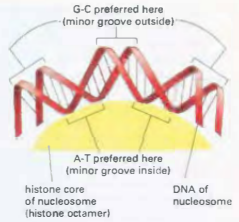
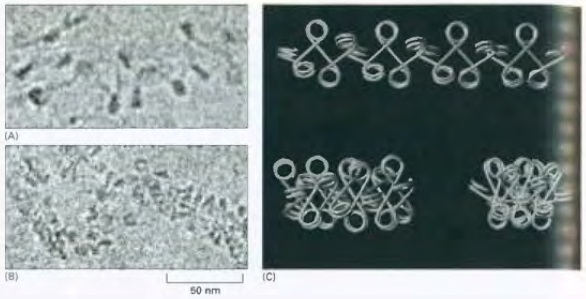


Figure 4–28 The bending of DNA in a nucleosome. The DNA helix makes 1.65 tight turns around the histone octamer. This diagram is drawn approximately to scale, illustrating how the minor groove is compressed on the inside of the turn. Owing to certain structural features of the DNA molecule, A-T base pairs are preferentially accommodated in such a narrow minor groove.

Illumina Exhibit 1031, Page 11 of 53



are difficult to fold into nucleosomes punctuate the 30-nm fiber with irregular features (Figure 4–30).

Several mechanisms probably act together to form the 30-nm fiber from a linear string of nucleosomes. First, an additional histone, called histone H1, is involved in this process. H1 is larger than the core histones and is considerably less well conserved. In fact, the cells of most eucaryotic organisms make several histone H1 proteins of related but quite distinct amino acid sequences. A single histone H1 molecule binds to each nucleosome, contacting both DNA and protein, and changing the path of the DNA as it exits from the nucleosome. Although it is not understood in detail how H1 pulls nucleosomes together into the 30-nm fiber, a change in the exit path in DNA seems crucial for compacting nucleosomal DNA so that it interlocks to form the 30-nm fiber (Figure 4–31).

A second mechanism for forming the 30-nm fiber probably involves the tails of the core histones, which, as we saw above, extend from the nucleosome. It is thought that these tails may help attach one nucleosome to another—thereby allowing a string of them, with the aid of histone H1, to condense into the 30-nm fiber (Figure 4–32).

Figure 4–29 Variations on the Zigzag model for the 30-nm chromatin fiber. (A and B) Electron microscopic evidence for the *top* and *bottom-left* model structures depicted in (C). (C) Zigzag variations. An interconversion between these three variations is proposed to occur by an accordion-like expansion and contraction of the fiber length. Differences in the length of the linker between adjacent nucleosome beads can be accommodated by snaking or coiling of the linker DNA, or by small local changes in the width of the fiber. Formation of the 30-nm fiber requires both histone H1 and the core histone tails; for simplicity, neither is shown here, but see Figures 4–30 and 4–32. (From J. Bednar et al., *Proc. Natl. Acad. Sci. USA* 95:14173–14178, 1998. © National Academy of Sciences.)

ATP-driven Chromatin Remodeling Machines Change Nucleosome Structure

For many years biologists thought that, once formed in a particular position on DNA, a nucleosome remained fixed in place because of the tight association

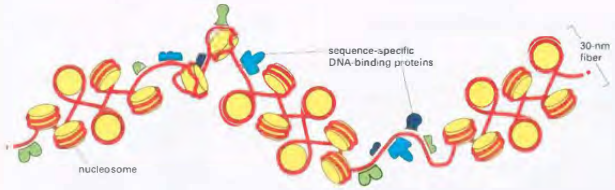


Figure 4–30 Irregularities in the 30-nm fiber. This schematic view of the 30-nm fiber illustrates its interruption by sequence-specific DNA-binding proteins. How these proteins bind tightly to DNA is explained in Chapter 7. The interruptions in the 30-nm fiber may be due to regions of DNA that lack nucleosomes altogether or, more probably, to regions that contain altered or remodeled nucleosomes. Regions of chromatin that are nucleosome free or contain remodeled nucleosome can often be detected experimentally by the unusually high susceptibility of their DNA to digestion by nucleases—as compared with the DNA in nucleosomes.

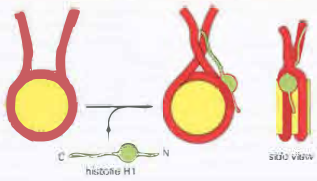


Figure 4–31 A speculative model for how histone H1 could change the path of DNA as it exits from the nucleosome. Histone H1 (*green*) consists of a globular core and two extended tails. Part of the effect of H1 on the compaction of nucleosome organization may result from charge neutralization, like the core histones, H1 is positively charged (especially its C-terminal tail), and this helps to compact the negatively charged DNA. Unlike the core histones, H1 does not seem to be essential for cell viability; in one ciliated protozoan the nucleus expands nearly twofold in the absence of H1, but the cells otherwise appear normal.

between the core histones and DNA. But it has recently been discovered that eucaryotic cells contain *chromatin remodeling complexes*, protein machines that use the energy of ATP hydrolysis to change the structure of nucleosomes temporarily so that DNA becomes less tightly bound to the histone core. The remodeled state may result from movement of the H2A–H2B dimers in the nucleosome core; the H3–H4 tetramer is particularly stable and would be difficult to rearrange (see Figure 4–27).

The remodeling of nucleosome structure has two important consequences. First, it permits ready access to nucleosomal DNA by other proteins in the cell, particularly those involved in gene expression, DNA replication, and repair. Even after the remodeling complex has dissociated, the nucleosome can remain in a "remodeled state" that contains DNA and the full complement of histones—but one in which the DNA–histone contacts have been loosened; only gradually does this remodeled state revert to that of a standard nucleosome. Second, remodeling complexes can catalyze changes in the positions of nucleosomes along DNA (Figure 4–33); some can even transfer a histone core from one DNA molecule to another.

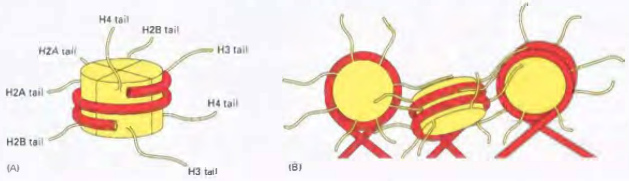
Cells have several different chromatin remodeling complexes that differ subtly in their properties. Most are large protein complexes that can contain more than ten subunits. It is likely that they are used whenever a eucaryotic cell needs direct access to nucleosome DNA for gene expression, DNA replication, or DNA repair. Different remodeling complexes may have features specialized for each of these roles. It is thought that the primary role of some remodeling complexes is to allow access to nucleosomal DNA, whereas that of others is to re-form nucleosomes when access to DNA is no longer required (Figure 4–34).

Chromatin remodeling complexes are carefully controlled by the cell. We shall see in Chapter 7 that, when genes are turned on and off, these complexes can be brought to specific regions of DNA where they act locally to influence chromatin structure. During mitosis, at least some of the chromatin-remodeling complexes are inactivated by phosphorylation. This may help the tightly packaged mitotic chromosomes maintain their structure.

Covalent Modification of the Histone Tails Can Profoundly Affect Chromatin

The N-terminal tails of each of the four core histones are highly conserved in their sequence, and perform crucial functions in regulating chromatin structure.

Figure 4–32 A speculative model for histone tails in the formation of the 30-nm fiber. (A) The approximate exit points of the eight histone tails, four from each histone subunit, that extend from each nucleosome. In the high-resolution structure of the nucleosome (see Figure 4–25), the tails are largely unstructured, suggesting that they are highly flexible. (B) A speculative model showing how the histone tails may help to pack nucleosomes together into the 30-nm fiber. This model is based on (1) experimental evidence that histone tails aid in the formation of the 30-nm fiber, (2) the x-ray crystal structure of the nucleosome, which showed that the tails of one nucleosome contact the histone core of an adjacent nucleosome in the crystal lattice, and (3) evidence that the histone tails interact with DNA.



Illumina Exhibit 1031, Page 12 of 53

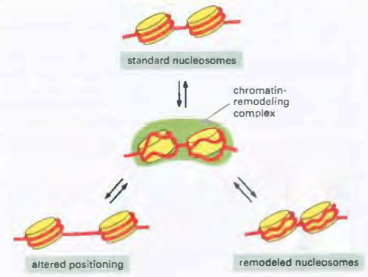


Figure 4–33 Model for the mechanism of some chromatin remodeling complexes. In the absence of remodeling complexes, the interconversion between the three nucleosomal states shown is very slow because of a high activation energy barrier. Using ATP hydrolysis, chromatin-remodeling complexes *(green)* create an activated intermediate (shown in the center of the figure) in which the histone–DNA contacts have been partly disrupted. This activated state can then decay to any one of the three nucleosomal configurations shown. In this way, the remodeling complexes greatly increase the rate of interconversion between different nucleosomal states. The remodeled state, in which the histone–DNA contacts have been loosened, has a higher free energy level than that of standard nucleosomes and will slowly revert to the standard nucleosome conformation, even in the absence of a remodeling complex. Cells have many different chromatin remodeling complexes, and they differ in their detailed biochemical properties; for example, not all can change the position of a nucleosome, but all use the energy of ATP hydrolysis to alter nucleosome structure. (Adapted from R.E. Kingston and G.J. Narlikar, *Genes Dev.* 13:2339–2352, 1999.)

Each tail is subject to several types of covalent modifications, including acetylation of lysines, methylation of lysines, and phosphorylation of serines (Figure 4–35A). Histones are synthesized in the cytosol and then assembled into nucleosomes. Some of the modifications of histone tails occur just after their synthesis, but before their assembly. The modifications that concern us, however, take place once the nucleosome has been assembled. These nucleosome modifications are added and removed by enzymes that reside in the nucleus; for example, acetyl groups are added to the histone tails by histone acetyl transferases (HATs) and taken off by histone deacetylases (HDACs).

The various modifications of the histone tails have several important consequences. Although modifications of the tails have little direct effect on the stability of an individual nucleosome, they seem to affect the stability of the

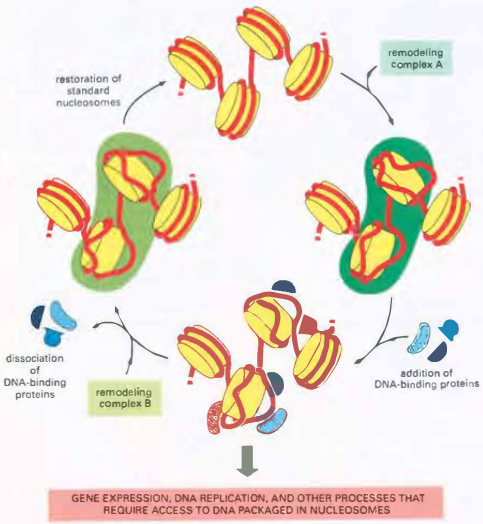


Figure 4–34 A cyclic mechanism for nucleosome disruption and re-formation. According to this model, different chromatin remodeling complexes disrupt and re-form nucleosomes, although, in principle, the same complex might catalyze both reactions. The DNA-binding proteins could function in gene expression, DNA replication, or DNA repair, and in some cases their binding could lead to the dissociation of the histone core to form nucleosome-free regions of DNA like those illustrated in Figure 4–30. (Adapted from A. Travers, *Cell* 96:311–314, 1999.)

30-nm chromatin fiber and of the higher-order structures discussed below. For example, histone acetylation tends to destabilize chromatin structure, perhaps in part because adding an acetyl group removes the positive charge from the lysine, thereby making it more difficult for histones to neutralize the charges on DNA as chromatin is compacted. However, the most profound effect of modified histone tails is their ability to attract specific proteins to a stretch of chromatin that has been appropriately modified. Depending on the precise tail modifications, these additional proteins can either cause further compaction of the chromatin or can facilitate access to the DNA. If combinations of modifications are taken into account, the number of possible distinct markings for each histone tail is very large. Thus, it has been proposed that, through covalent modification of the histone tails, a given stretch of chromatin can convey a particular meaning to

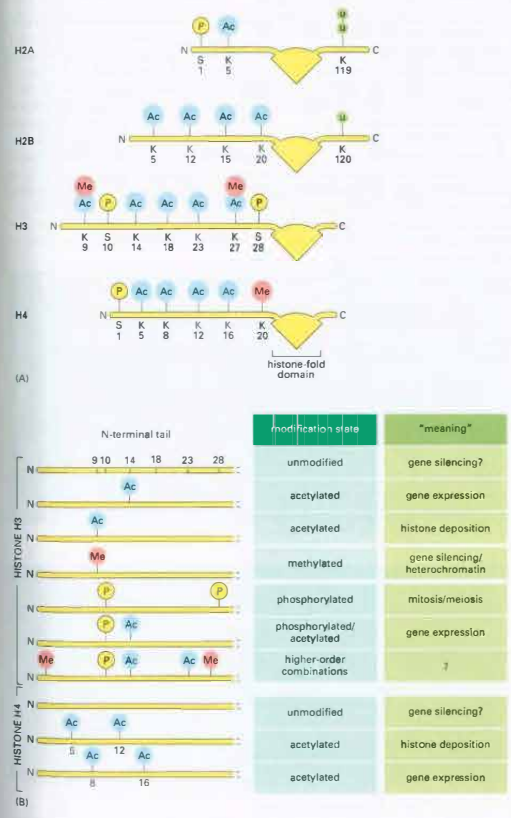


Figure 4–35 Covalent modification of core histone tails. (A) Known modifications of the four histone core proteins are indicated: Me = methyl group, Ac = acetyl group, P = phosphate, u = ubiquitin. Note that some positions (e.g., lysine 9 of H3) can be modified in more than one way. Most of these modifications add a relatively small molecule onto the histone tails; the exception is ubiquitin, a 76 amino acid protein also used in other cellular processes (see Figure 6–87). The function of ubiquitin in chromatin is not well understood; histone H2B can be modified by a single ubiquitin molecule; H2A can be modified by the addition of several ubiquitins. **(B)** A histone code hypothesis. Histone tails can be marked by different combinations of modifications. According to this hypothesis, each marking conveys a specific meaning to the stretch of chromatin on which it occurs. Only a few of the meanings of the modifications are known. In Chapter 7, we discuss the way a doubly-acetylated H4 tail is "read" by a protein required for gene expression. In another well-studied case, an H3 tail methylated at lysine 9 is recognized by a set of proteins that create an especially compact form of chromatin, which silences gene expression.

The acetylation of lysine 14 of histone H3 and lysines 8 and 16 of histone H4—usually associated with gene expression—is performed by the type A histone acetylases (HATs) in the nucleus. In contrast, the acetylation of lysines 5 and 12 of histone H4 and a lysine of histone H3 takes place in the cytosol, after the histones have been synthesized but before they have been incorporated into nucleosomes; these modifications are catalyzed by type B HATs. These modified histones are deposited onto DNA after DNA replication (see Figure 5-41), and their acetyl groups are taken off shortly afterwards by histone deacetylases (HDACs). Thus, the acetylation at these positions signals newly replicated chromatin.

Modification of a particular position in a histone tail can take on different meanings depending on other features of the local chromatin structure. For example, the phosphorylation of position 10 of histone H3 is associated not only with the condensation of chromosomes that takes place in mitosis and meiosis but also with the expression of certain genes. Some histone tail modifications are interdependent. For example methylation of H3 position 9 blocks the phosphorylation of H3 position 10, and vice versa.

Illumina Exhibit 1031, Page 13 of 53

the cell (Figure 4–35B). For example, one type of marking could signal that the stretch of chromatin has been newly replicated, and another could signal that gene expression should not take place. According to this idea, each different marking would attract those proteins that would then execute the appropriate functions. Because the histone tails are extended, and are therefore probably accessible even when chromatin is condensed, they provide an especially apt format for such messages.

As with chromatin remodeling complexes, the enzymes that modify (and remove modifications from) histone tails are usually multisubunit proteins, and they are tightly regulated. They are brought to a particular region of chromatin by other cues, particularly by sequence-specific DNA-binding proteins. We can thus imagine how cycles of histone tail modification and demodification can allow chromatin structure to be dynamic—locally compacting and decompacting it, and, in addition, attracting other proteins specific for each modification state. It is likely that histone-modifying enzymes and chromatin remodeling complexes work in concert to condense and recondense stretches of chromatin; for example, evidence suggests that a particular modification of the histone tail attracts a particular type of remodeling complex. Moreover, some chromatin remodeling complexes contain histone modification enzymes as subunits, directly connecting the two processes.

Summary

A gene is a nucleotide sequence in a DNA molecule that acts as a functional unit for the production of a protein, a structural RNA, or a catalytic RNA molecule. In eucaryotes, protein-coding genes are usually composed of a string of alternating introns and exons. A chromosome is formed from a single, enormously long DNA molecule that contains a linear array of many genes. The human genome contains 3.2 × 10^9 DNA nucleotide pairs, divided between 22 different autosomes and 2 sex chromosomes. Only a small percentage of this DNA codes for proteins or structural and catalytic RNAs. A chromosomal DNA molecule also contains three other types of functionally important nucleotide sequences: replication origins and telomeres allow the DNA molecule to be completely replicated, while a centromere attaches the daughter DNA molecules to the mitotic spindle, ensuring their accurate segregation to daughter cells during the M phase of the cell cycle.

The DNA in eucaryotes is tightly bound to an equal mass of histones, which form a repeating array of DNA–protein particles called nucleosomes. The nucleosome is composed of an octameric core of histone proteins around which the DNA double helix is wrapped. Despite irregularities in the positioning of nucleosomes along DNA, nucleosomes are usually packed together (with the aid of histone H1 molecules) into quasi-regular arrays to form a 30-nm fiber. Despite the high degree of compaction in chromatin, its structure must be highly dynamic to allow the cell access to the DNA. Two general strategies for reversibly changing local chromatin structures are important for this purpose: ATP-driven chromatin remodeling complexes, and an enzymatically catalyzed covalent modification of the N-terminal tails of the four core histones.

THE GLOBAL STRUCTURE OF CHROMOSOMES

Having discussed the DNA and protein molecules from which the 30-nm chromatin fiber is made, we now turn to the organization of the chromosome on a more global scale. As a 30-nm fiber, the typical human chromosome would still be 0.1 cm in length and able to span the nucleus more than 100 times. Clearly, there must be a still higher level of folding, even in interphase chromosomes. This higher-order packaging is one of the most fascinating—but also one of the most poorly understood—aspects of chromosome structure. Although its molecular basis is still largely a mystery, it almost certainly involves the folding of the 30-nm fiber into a series of loops and coils, as we see below. Our discussion of this higher-order packing continues an important theme in chromosome

architecture: interphase chromatin structure is fluid, exposing at any given moment the DNA sequences directly needed by the cell.

We first describe several rare cases in which the overall structure and organization of interphase chromosomes can be easily visualized, and we explain that certain features of these exceptional cases may be representative of the structures of all interphase chromosomes. Next we describe the different forms of chromatin that make up a typical interphase chromosome. Finally we discuss the additional compaction that interphase chromosomes undergo during the process of mitosis.

Lampbrush Chromosomes Contain Loops of Decondensed Chromatin

Most chromosomes in interphase cells are too fine and too tangled to be visualized clearly. In a few exceptional cases, however, *interphase chromosomes* can be seen to have a precisely defined higher-order structure, and it is thought that certain characteristics of these higher-order structures are representative of all interphase chromosomes. The meiotically paired chromosomes in growing amphibian oocytes (immature eggs), for example, are highly active in gene expression, and they form unusually stiff and extended chromatin loops. These so-called **lampbrush chromosomes** (the largest chromosomes known) are clearly visible even in the light microscope, where they are seen to be organized into a series of large chromatin loops emanating from a linear chromosomal axis (Figure 4–36).

The organization of a lampbrush chromosome is illustrated in Figure 4–37. A given loop always contains the same DNA sequence, and it remains extended in the same manner as the oocyte grows. Other experiments demonstrate that

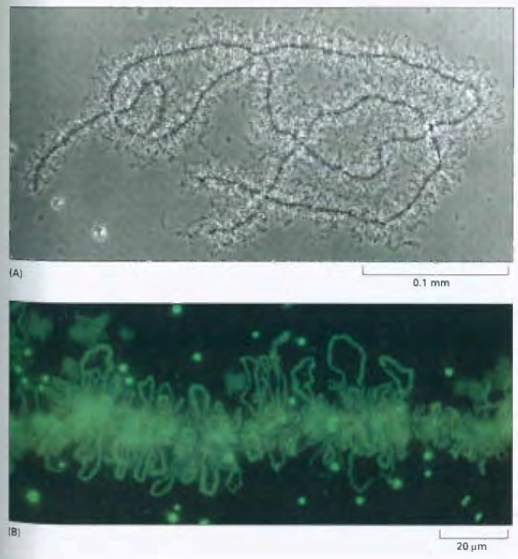


Figure 4–36 Lampbrush chromosomes. **(A)** A light micrograph of lampbrush chromosomes in an amphibian oocyte. Early in oocyte differentiation, each chromosome replicates to begin meiosis, and the homologous replicated chromosomes pair to form this highly extended structure containing a total of four replicated DNA molecules, or chromatids. The lampbrush chromosome stage persists for months or years, while the oocyte builds up a supply of materials required for its ultimate development into a new individual. **(B)** Fluorescence light micrograph showing a portion of an amphibian lampbrush chromosome. The regions of the chromosome that are being actively expressed are stained green by using antibodies against proteins that process RNA during one of the steps of gene expression (discussed in Chapter 6). The round granules are thought to correspond to large complexes of the RNA-splicing machinery that will also be discussed in Chapter 6. (A, courtesy of Joseph G. Gall; B, courtesy of Joseph G. Gall and Christine Murphy.)

Illumina Exhibit 1031, Page 14 of 53

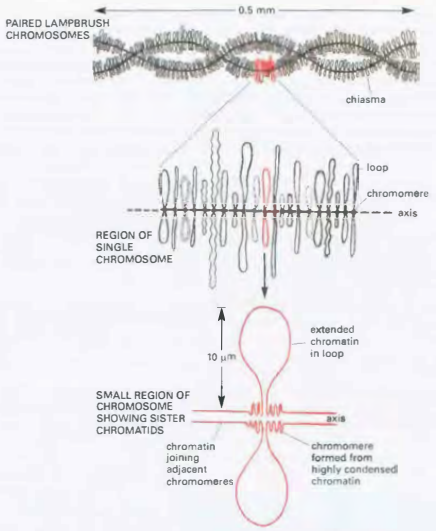


Figure 4–37 A model for the structure of a lampbrush chromosome. The set of lampbrush chromosomes in many amphibians contains a total of about 10,000 chromatin loops, although most of the DNA in each chromosome remains highly condensed in the chromomeres. Each loop corresponds to a particular DNA sequence. Four copies of each loop are present in each cell, since each of the two chromosomes shown at the top consists of two closely apposed, newly replicated chromosomes. This four-stranded structure is characteristic of this stage of development of the oocyte, the diplotene stage of meiosis; see Figure 20–12.

most of the genes present in the DNA loops are being actively expressed (see Figure 4–36B). Most of the DNA, however, is not in loops but remains highly condensed in the *chromomeres* on the axis, which are generally not expressed. Lampbrush chromosomes illustrate a recurrent theme of this chapter—when the DNA in a region of chromatin is in use (in this case, for gene expression), that part of the chromatin has an extended structure; otherwise, the chromatin is condensed. In lampbrush chromosomes, the structural units of this regulation are large, precisely defined loops.

Relatively few species undergo the specialization that produces lampbrush chromosomes. However, when injected into amphibian oocytes, the DNA from organisms that normally do not produce lampbrush chromosomes (e.g., DNA from a fish) is packaged into lampbrush chromosomes. On the basis of this type of experiment, it has been proposed that the interphase chromosomes of all eucaryotes are arranged in loops that are normally too small and fragile to be easily observed. It may be possible in the future to coax the DNA from a mammal such as a mouse to form lampbrush chromosomes by introducing it into amphibian oocytes. This could allow a detailed correlation of loop structure, gene arrangement, and DNA sequence, and we could begin to learn how the packaging into loops reflects the sequence content of our DNA.

Drosophila Polytene Chromosomes Are Arranged in Alternating Bands and Interbands

Certain insect cells also have specialized interphase chromosomes that are readily visible, although this type of specialization differs from that of lampbrush chromosomes. For example, many of the cells of certain fly larvae grow to an enormous size through multiple cycles of DNA synthesis without cell division. The resulting giant cells contain as much as several thousand times the normal DNA complement. Cells with more than the normal DNA complement are said to be *polyploid* when they contain increased numbers of standard chromosomes. In several types of secretory cells of fly larvae, however, all the homologous chromosome copies are held side by side, like drinking straws in a box, creating a single **polytene chromosome**. The fact that, in some large insect cells, polytene chromosomes can disperse to form a conventional polyploid cell demonstrates that these two chromosomal states are closely related, and that the basic structure of a polytene chromosome must be similar to that of a normal chromosome.

Polytene chromosomes are often easy to see in the light microscope because they are so large and because the precisely aligned side-by-side adherence of individual chromatin strands greatly elongates the chromosome axis and prevents tangling. Polyteny has been most studied in the salivary gland cells of *Drosophila* larvae, in which the DNA in each of the four *Drosophila* chromosomes has been replicated through 10 cycles without separation of the daughter chromosomes, so that 1024 (2^{10}) identical strands of chromatin are lined up side by side (Figure 4–38).

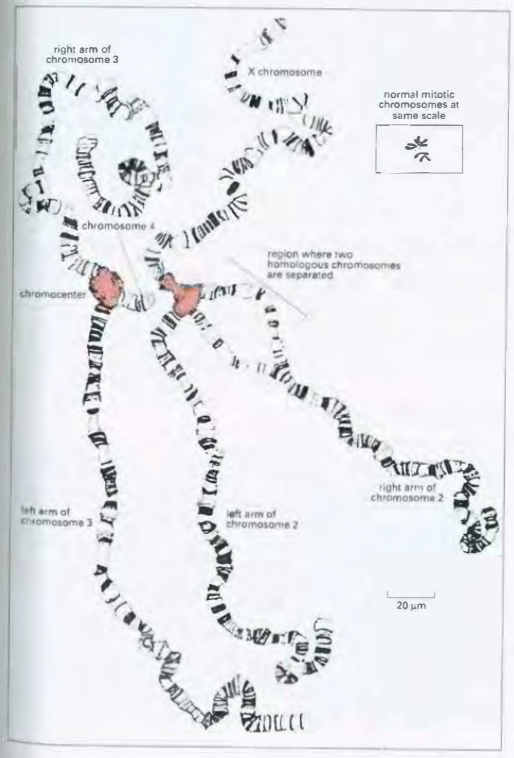


Figure 4–38 The entire set of polytene chromosomes in one *Drosophila* salivary cell. These chromosomes have been spread out for viewing by squashing them against a microscope slide. *Drosophila* has four chromosomes, and there are four different chromosome pairs present. But each chromosome is tightly paired with its homolog (so that each pair appears as a single structure), which is not true in most nuclei (except in meiosis). The four polytene chromosomes are normally linked together by regions near their centromeres that aggregate to create a single large chromocenter *(pink region)*. In this preparation, however, the chromocenter has been split into two halves by the squashing procedure used. (Modified from T.S. Painter, *J. Hered.* 25:465–476, 1934.)

Illumina Exhibit 1031, Page 15 of 53

When polytene chromosomes are viewed in the light microscope, distinct alternating dark *bands* and light *interbands* are visible (Figure 4–39). Each band and interband represents a set of 1024 identical DNA sequences arranged in register. About 95% of the DNA in polytene chromosomes is in bands, and 5% is in interbands. The chromatin in each band appears dark, either because it is much more condensed than the chromatin in the interbands, or because it contains a higher proportion of proteins, or both (Figure 4–40). Depending on their size, individual bands are estimated to contain 3000–300,000 nucleotide pairs in a chromatin strand. The bands of *Drosophila* polytene chromosomes can be recognized by their different thicknesses and spacings, and each one has been given a number to generate a chromosome "map." There are approximately 5000 bands and 5000 interbands in the complete set of *Drosophila* polytene chromosomes.

Both Bands and Interbands in Polytene Chromosomes Contain Genes

The reproducible pattern of bands and interbands seen in *Drosophila* polytene chromosomes means that these interphase chromosomes are highly organized. Since the 1930s, scientists have debated the nature of this organization, and we still do not have a clear answer. Because the number of bands in *Drosophila* chromosomes was once thought to be roughly equal to the number of genes in the genome, it was initially thought that each band might correspond to a single gene; however, we now know this simple idea is incorrect. There are nearly three times more genes in *Drosophila* than chromosome bands, and genes are found in both band and interband regions. Moreover, some bands contain multiple genes, and some bands seem to lack genes altogether.

It seems likely that the band–interband pattern reflects different levels of gene expression and chromatin structure along the chromosome, with genes in the less compact interbands being expressed more highly than those in the more compact bands. In any case, the remarkable appearance of fly polytene chromosomes is thought to reflect the heterogeneous nature of the chromatin compaction found along all interphase chromosomes. In the next section we see how the appearance of a band can change dramatically when the gene or genes within it become highly expressed.

Individual Polytene Chromosome Bands Can Unfold and Refold as a Unit

A major factor controlling gene expression in the polytene chromosomes of *Drosophila* is the insect steroid hormone *ecdysone*, the levels of which rise and fall periodically during larval development. When ecdysone concentrations rise,

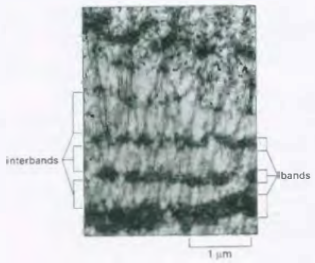


Figure 4–40 An electron micrograph of a small section of a *Drosophila* polytene chromosome seen in thin section. Bands of very different thickness can be readily distinguished, separated by interbands, which contain less condensed chromatin. (Courtesy of Veikko Sorsa.)



Figure 4–39 A light micrograph of a portion of a polytene chromosome from *Drosophila* salivary glands. The distinct pattern produced by bands and interbands is readily seen. The bands are regions of increased chromatin concentration that occur in interphase chromosomes. Although they are detectable only in polytene chromosomes, it is thought that they reflect a structure common to the chromosomes of most eucaryotes. (Courtesy of Joseph G. Gall.)

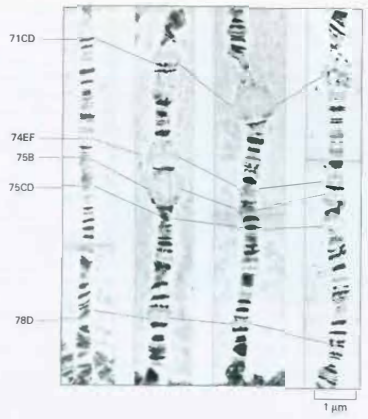


Figure 4–41 Chromosome puffs. This series of time-lapse photographs shows how puffs arise and recede in the polytene chromosomes of *Drosophila* during larval development. A region of the *left* arm of chromosome 3 is shown. It exhibits five very large puffs in salivary gland cells, each active for only a short developmental period. The series of changes shown occur over a period of 22 hours, appearing in a reproducible pattern as the organism develops. The designations of the indicated bands are given at the left of the photographs. (Courtesy of Michael Ashburner.)

they induce the expression of genes coding for proteins that the larva requires for each molt and for pupation. As the organism progresses from one developmental stage to another, distinctive *chromosome puffs* arise and old puffs recede as new genes become expressed and old ones are turned off (Figure 4–41). From inspection of each puff when it is relatively small and the banding pattern is still discernible, it seems that most puffs arise from the decondensation of a single chromosome band.

The individual chromatin fibers that make up a puff can be visualized with an electron microscope. For technical reasons, this is easier in the polytene chromosomes from a different insect, *Chironomus tentans*, a midge. Electron micrographs of certain puffs, called Balbiani rings, of *Chironomus* salivary gland polytene chromosomes show the chromatin arranged in loops (Figures 4–42 and 4–43), much like those observed in the amphibian lampbrush chromosomes discussed earlier. Additional experiments suggest that each loop contains a

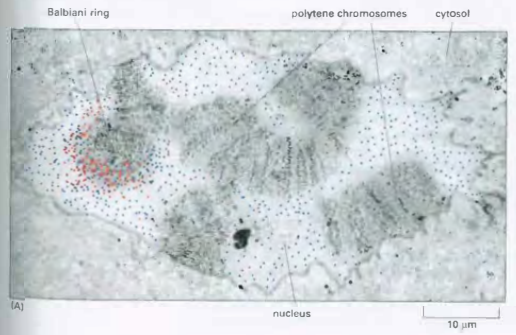
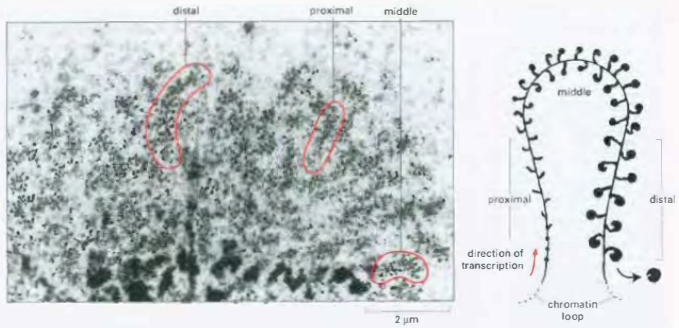


Figure 4–42 RNA synthesis in chromosome puffs. **(A)** Polytene chromosomes from the salivary glands of the insect *C. tentans*. As outlined in Chapter 1 and described in detail in Chapter 6, the first step in gene expression is the synthesis of an RNA molecule using the DNA as a template. In this electron micrograph, newly synthesized RNA from a Balbiani ring gene is indicated in *red*. Cells were exposed to a brief pulse of BrUTP (an analog of UTP), which was incorporated into RNA. Cells were then fixed and the newly synthesized RNAs were identified by using antibodies against BrU. Balbiani ring RNAs could be distinguished from other RNAs by their characteristic shape (see Figure 4–43). The *blue dots* in the figure represent positions of Balbiani ring RNAs that were synthesized before the addition of BrUTP. The experiment shows that Balbiani ring RNAs are synthesized in puffs and then diffuse through the nucleoplasm. **(B)** An autoradiogram of a single puff in a polytene chromosome. The portion of the chromosome indicated is undergoing RNA synthesis and has therefore become labeled with ^{3}H-uridine. (A, courtesy of B. Daneholt, from O.P. Singh et al., *Exp. Cell Res.* 251:135–146, 1999. © Academic Press; B, courtesy of José Bonner.)

Illumina Exhibit 1031, Page 16 of 53



single gene. When not expressed, the loop of DNA assumes a thickened structure, possibly a folded 30-nm fiber, but when gene expression is occurring, the loop becomes more extended. Both types of loops contain the four core histones and histone H1.

It seems likely that the default loop structure is a folded 30-nm fiber and that the histone modifying enzymes, chromatin remodeling complexes, and other proteins required for gene expression all help to convert it to a more extended form whenever a gene is expressed. In electron micrographs, the chromatin located on either side of the decondensed loop appears considerably more compact, which is consistent with the idea that a loop constitutes an independent functional domain of chromatin structure.

Although controversial, it has been proposed that all of the DNA in polytene chromosomes is arranged in loops that condense and decondense according to when the genes within them are expressed. It may be that all interphase chromosomes from all eucaryotes are also packaged into an orderly series of looped domains, each containing a small number of genes whose expression is regulated in a coordinated way (Figure 4–44). We shall return to this issue in Chapter 7 when we discuss the ways in which gene expression is regulated by the cell.

Figure 4–43 Polytene chromosomes from *C. tentans*. The electron micrograph shows a thin section of the chromatin in a Balbiani ring, a chromosome puff very active in gene expression. The Balbiani ring gene codes for secretory proteins the larvae uses to spin a protective tube. The chromatin is arranged in loops, but because the sample has been sectioned, only portions of the loops are visible. As they are synthesized on the chromatin, the RNA molecules are bound up by protein molecules, making them visible as knobs on stalks in the electron microscope. From the size of the RNA–protein complex, the extent of RNA synthesis (transcription) can be inferred; a whole chromatin loop (shown on the right) can then be reconstructed from a set of electron micrograph sections such as that shown here. (Courtesy of B. Daneholt, from U. Skoglund et al., *Cell* 34:847–855, 1983. © Elsevier.)

Heterochromatin Is Highly Organized and Usually Resistant to Gene Expression

Having described some features of interphase chromosomes inferred from a few rare cases, we now turn to characteristics of interphase chromosomes that can be observed in a wide variety of organisms. Light-microscope studies in the 1930s distinguished between two types of chromatin in the interphase nuclei of many higher eucaryotic cells: a highly condensed form, called **heterochromatin**, and all the rest, which is less condensed, called **euchromatin**. Euchromatin is composed of the types of chromosomal structures—30-nm fibers and looped domains—that we have discussed so far. Heterochromatin, in contrast, includes additional proteins and probably represents more compact levels of organization that are just beginning to be understood. In a typical mammalian cell, approximately 10% of the genome is packaged into heterochromatin. Although present in many locations along chromosomes, it is concentrated in specific regions, including the centromeres and telomeres.

Most DNA that is folded into heterochromatin does not contain genes. However, genes that do become packaged into heterochromatin are usually resistant to being expressed, because heterochromatin is unusually compact. This does not mean that heterochromatin is useless or deleterious to the cell; as we see below, regions of heterochromatin are responsible for the proper functioning of telomeres and centromeres (which lack genes), and its formation may even help protect the genome from being overtaken by "parasitic" mobile elements of DNA. Moreover, a few genes require location in heterochromatin regions if they are to be expressed. In fact, the term *heterochromatin* (which was first defined cytologically) is likely to encompass several distinct types of chromatin structures whose common feature is an especially high degree of organization. Thus, heterochromatin should not be thought of as encapsulating "dead" DNA, but rather as creating different types of compact chromatin with distinct features and roles.

Heterochromatin's resistance to gene expression makes it amenable to study even in organisms in which it cannot be directly observed. When a gene that is normally expressed in euchromatin is experimentally relocated into a region of heterochromatin, it ceases to be expressed, and the gene is said to be *silenced*. These differences in gene expression are examples of **position effects**, in which the activity of a gene depends on its position along a chromosome. First recognized in *Drosophila*, position effects have now been observed in many organisms and they are thought to reflect an influence of the different states of chromatin structure along chromosomes on gene expression. Thus, chromosomes can be considered as mosaics of distinct forms of chromatin, each of which has a special effect on the ability of the DNA it contains to be addressed by the cell.

Many position effects exhibit an additional feature called *position effect variegation*, which is responsible for the mottled appearance of the fly eye and the sectoring of the yeast colony in the examples shown in Figure 4–45. These patterns can result from patches of cells in which a silenced gene has become reactivated; once reactivated, the gene is inherited stably in this form in daughter cells. Alternatively, a gene can start out in euchromatin early in development, and then be selected more or less randomly for packaging into heterochromatin, causing its inactivation in a cell and all of its daughters.

The study of position effect variegation has revealed two important characteristics of heterochromatin. First, heterochromatin is dynamic; it can "spread" into a region and later "retract" from it at low but observable frequencies. Second, the state of chromatin—whether heterochromatin or euchromatin—tends to be inherited from a cell to its progeny. These two features are responsible for position effect variegation, as explained in Figure 4–46. In the next section, we discuss several models to account for the self-sustaining nature of heterochromatin, once it has been formed.

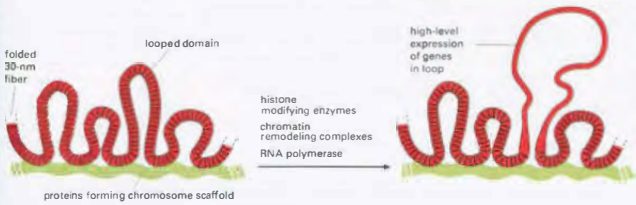


Figure 4–44 A model for the structure of an interphase chromosome. A section of an interphase chromosome is shown folded into a series of looped domains, each containing 20,000–100,000 nucleotide pairs of double-helical DNA condensed into a 30-nm fiber. Individual loops can decondense, perhaps in part through an accordionlike expansion of the 30-nm fiber (see Figure 4–29), when the cell requires direct access to the DNA packaged in these loops. This decondensation is brought about by enzymes that directly modify chromatin structure—as well as by proteins, such as RNA polymerase (discussed in Chapter 6), that act directly on the underlying DNA. It is not understood how the folded 30-nm fiber is anchored to the chromosome axis, but evidence suggests that the base of chromosomal loops is rich in DNA topoisomerases, which are enzymes that allow DNA to swivel when anchored (see pp. 251–253).

Illumina Exhibit 1031, Page 17 of 53

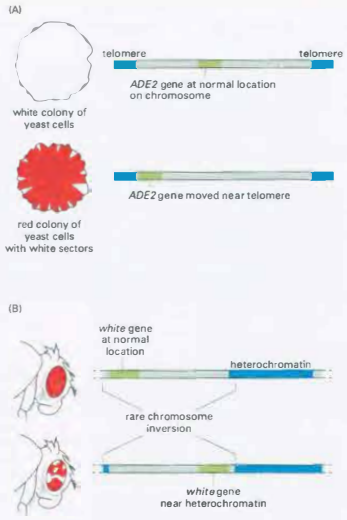


Figure 4–45 Position effects on gene expression in two different eucaryotic organisms. (A) The yeast *ADE2* gene at its normal chromosomal location is expressed in all cells. When moved near the end of a yeast chromosome, which is inferred to be folded into a form of heterochromatin, the gene is no longer expressed in most cells of the population. The *ADE2* gene codes for one of the enzymes of adenine biosynthesis, and the absence of the *ADE2* gene product leads to the accumulation of a red pigment. Therefore, a colony of cells that expresses *ADE2* is *white*, and one composed of cells where the *ADE2* gene is not expressed is *red*. The *white* sectors that fan out from the middle of the *red* colony grown on an agar surface represent the descendants of cells in which the *ADE2* gene has spontaneously become active. These white sectors are thought to result from a heritable change to a less tightly packed state of chromatin near the *ADE2* gene in these cells. Although yeast chromosomes are too small to be seen under the light microscope, the chromatin structure at the ends of yeast chromosomes is thought to have many of the same structural features as the heterochromatin in the chromosomes of larger organisms. **(B)** Position effects can also be observed for the *white* gene in the fruit fly *Drosophila*. The *white* gene controls eye pigment production and is named after the mutation that first identified it. Wild type flies with a normal *white* gene (*white*$^+$) have normal pigment production, which gives them red eyes, but if the *white* gene is mutated and inactivated, the mutant flies (*white*$^-$) make no pigment and have *white* eyes. In flies in which a normal *white*$^+$ gene has been moved near a region of heterochromatin, the eyes are mottled, with both *red* and *white* patches. The *white* patches represent cells in which the *white*$^+$ gene has been silenced by the effects of the heterochromatin. In contrast, the *red* patches represent cells that express the *white*$^+$ gene because the heterochromatin did not spread across this gene at the time, early in development, when the heterochromatin first formed. As in the yeast, the presence of large patches of *red* and *white* cells indicates that the state of transcriptional activity of the gene is inherited, once determined by its chromatin packaging in the early embryo.

The Ends of Chromosomes Have a Special Form of Heterochromatin

Unlike the nucleosome and the 30-nm fiber, heterochromatin is not well understood structurally. It almost certainly involves an additional level of folding of 30-nm fiber and requires many proteins in addition to the histones. Although its chromosomes are too small to be seen under the light microscope, the molecular nature of heterochromatin is probably best understood in the simple yeast *S. cerevisiae*. Many experiments with yeast cells have shown that the chromatin extending inward roughly 5000 nucleotide pairs from each chromosome end is resistant to gene expression, and probably has a structure that corresponds to at least one type of heterochromatin in the chromosomes of more complex organisms. Extensive genetic analysis has led to the identification of many of the yeast proteins required for this type of gene silencing.

Mutations in any one of a set of yeast Silent information regulator (Sir) proteins prevent the silencing of genes located near telomeres, thereby allowing these genes to be expressed. Analysis of these proteins has led to the discovery of a telomere-bound Sir protein complex that recognizes underacetylated N-terminal tails of selected histones (Figure 4–47A). One of the proteins in this complex is a highly conserved histone deacetylase known as Sir2, which has homologs in diverse organisms, including humans, and presumably has a major role in creating a pattern of histone underacetylation unique to heterochromatin. As discussed earlier in this chapter, deacetylation of the histone tails is thought to allow nucleosomes to pack together into tighter arrays and may also render them less susceptible to some chromatin remodeling complexes. In addition, heterochromatin-specific patterns of histone tail modification are likely to attract additional proteins involved in forming and maintaining heterochromatin (see Figure 4–35).

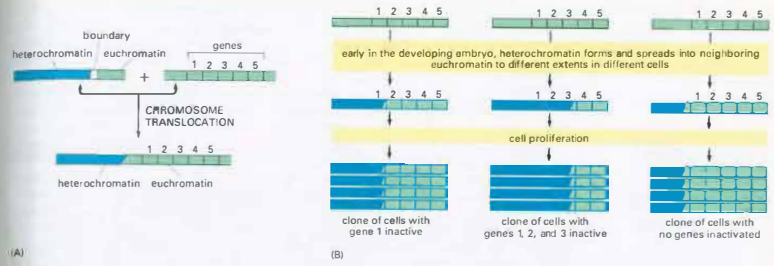


Figure 4–46 The cause of position effect variegation in *Drosophila*. (A) Heterochromatin *(blue)* is normally prevented from spreading into adjacent regions of euchromatin *(green)* by special boundary DNA sequences, which we discuss in Chapter 7. In flies that inherit certain chromosomal rearrangements, however, this barrier is no longer present. **(B)** During the early development of such flies, heterochromatin can spread into neighboring chromosomal DNA, proceeding for different distances in different cells. This spreading soon stops, but the established pattern of heterochromatin is inherited, so that large clones of progeny cells are produced that have the same neighboring genes condensed into heterochromatin and thereby inactivated (hence the "variegated" appearance of some of these flies; see Figure 4–45B). Although "spreading" is used to describe the formation of new heterochromatin near previously existing heterochromatin, the term may not be wholly accurate. There is evidence that during expansion, heterochromatin can "skip over" some regions of chromatin, sparing the genes that lie within them from repressive effects. One possibility is that heterochromatin can expand across the base of some DNA loops, thus bypassing the chromatin contained in the loop.

But how is the Sir2 protein delivered to the ends of chromosomes in the first place? Another series of experiments has suggested the model shown in Figure 4–47B. A DNA-binding protein that recognizes specific DNA sequences in yeast telomeres also binds to one of the Sir proteins, causing the entire Sir protein complex to assemble on the telomeric DNA. The Sir complex then spreads along the chromosome from this site, modifying the N-terminal tails of adjacent histones

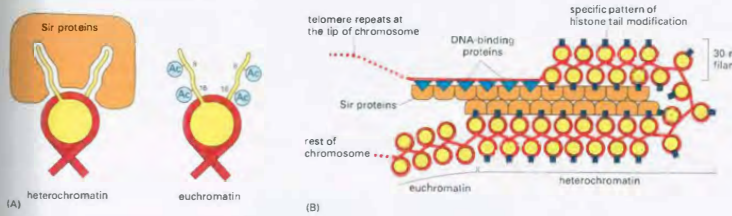


Figure 4–47 Speculative model for the heterochromatin at the ends of yeast chromosomes. (A) Heterochromatin is generally underacetylated, and underacetylated tails of histone H4 are proposed to interact with a complex of Sir proteins, thus stabilizing the association of these proteins with nucleosomes. Although shown as fully unacetylated, the exact pattern of histone H4 tail modification required to bind to the Sir complex is not known with certainty. In some organisms, the methylation of lysine 9 of histone H3 is also a critical signal for heterochromatin formation. In euchromatin, histone tails are typically highly acetylated. Those of H4 are shown as partially acetylated but, in reality, the acetylation state varies across euchromatin. (B) Specialized DNA-binding proteins *(blue triangles)* recognize DNA sequences near the ends of chromosomes and attract the Sir proteins, one of which (Sir2) is an NAD$^+$-dependent histone deacetylase. This then leads to the cooperative spreading of the Sir protein complex down the chromosome. As this complex spreads, the deacetylation catalyzed by Sir2 helps create new binding sites on nucleosomes for more Sir protein complexes. A "fold back" structure of the type shown may also form.

Illumina Exhibit 1031, Page 18 of 53

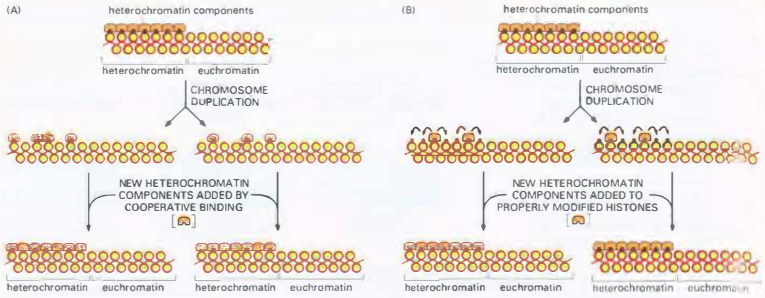


Figure 4–48 Two speculative models for how the tight packaging of DNA in heterochromatin can be inherited during chromosome replication. In both cases, half of the specialized heterochromatin components have been distributed to each daughter chromosome after DNA duplication. (A) In this model, new heterochromatin components bind cooperatively to the inherited components, thereby beginning the process of new heterochromatin formation. The process is completed with the assembly of additional proteins and the eventual covalent modification of the histones (not shown). (B) In this model the inherited heterochromatin components change the pattern of histone modification on the newly formed daughter nucleosomes nearby, creating new binding sites for free heterochromatin components, which assemble and complete the structure. Both models can account for the spreading effects of heterochromatin, and indeed, both processes may occur simultaneously in cells.

to create the nucleosome-binding sites that the complex prefers. This "spreading effect" is thought to be driven by the cooperative binding of adjacent Sir protein complexes, as well as by the folding back of the chromosome on itself to promote Sir binding in nearby regions (see Figure 4–47B). In addition, the formation of heterochromatin probably requires the action of chromatin remodeling complexes to readjust the positions of nucleosomes as they are packed together.

Unlike most deacetylases, Sir2 requires NAD^+ as a cofactor (see Figure 2–60). The NAD^+ levels in the cell fluctuate with the nutritional health of the cell, increasing as cells become nutritionally deprived. This feature might cause the telomeric heterochromatin to spread in response to starvation (perhaps to silence the expression of genes that are not absolutely required for survival) and then to retract when conditions improve.

The properties of the yeast heterochromatin just described may resemble features of heterochromatin in more complex organisms. Certainly, the spreading of yeast heterochromatin from telomeres is similar in principle to the movement of heterochromatin that causes position effect variegation in animals (see Figure 4–46). Moreover, these properties can be used to explain the heritability of heterochromatin, as outlined in Figure 4–48. Whatever the precise mechanism of heterochromatin formation, it has become clear that covalent modifications of the nucleosome core histones have a critical role in this process. Of special importance in many organisms are the *histone methyl transferases*, enzymes that methylate specific lysines on histones including lysine 9 of histone H3 (see Figure 4–35). This modification is "read" by heterochromatin components (including HP1 in *Drosophila*) that specifically bind this modified form of histone H3 to induce the assembly of heterochromatin. It is likely that a spectrum of different histone modifications is used by the cell to distinguish heterochromatin from euchromatin (see Figure 4–35).

Having the ends of chromosomes packaged into heterochromatin provides several advantages to the cell: it helps to protect the ends of chromosomes from being recognized as broken chromosomes by the cellular repair machinery, it may help to regulate telomere length, and it may assist in the accurate pairing and segregation of chromosomes during mitosis. In Chapter 5 we see that telomeres have additional structural features that distinguish them from other parts of chromosomes.

Centromeres Are Also Packaged into Heterochromatin

Heterochromatin is also observed around centromeres, the DNA sequences that direct the movement of each chromosome into daughter cells every time a cell divides (see Figure 4–22). In many complex organisms, including humans, each

centromere seems to be embedded in a very large stretch of heterochromatin that persists throughout interphase, even though the centromere-directed movement of DNA occurs only during mitosis. The structure and biochemical properties of this so-called *centric heterochromatin* are not well understood, but, like other forms of heterochromatin, it silences the expression of genes that are experimentally placed into it. It contains, in addition to histones (which are typically underacetylated and methylated in heterochromatin), several additional structural proteins that compact the nucleosomes into particularly dense arrangements.

As with telomeres, our best understanding of the chromatin structure of a centromere comes from studies of the much simpler centromeres of the yeast *S. cerevisiae*. Earlier in this chapter we saw that a simple DNA sequence of approximately 125 nucleotide pairs was sufficient to serve as a centromere in this organism. Despite its small size, more than a dozen different proteins assemble on this DNA sequence; the proteins include a histone H3 variant that, along with the other core histones, is believed to form a centromere-specific nucleosome (Figure 4–49A). We do not yet understand what properties this variant type of nucleosome provides to the cell, but similar specialized nucleosomes seem to be present in all eucaryotic centromeres (Figure 4-49B). The additional proteins at the yeast centromere attach it to the spindle microtubules and provide signals that ensure that this attachment is complete before the later stages of mitosis are allowed to proceed (discussed in Chapters 17 and 18).

The centromeres in more complex organisms are considerably larger than those in budding yeasts. For example, fly and human centromeres extend over hundreds of thousands of nucleotide pairs and do not seem to contain a centromere-specific DNA sequence. Rather, most consist largely of short, repeated DNA sequences, known as *alpha satellite DNA* in humans (Figure 4–50). But the same repeat sequences are also found at other (noncentromeric) positions on chromosomes, and how they specify a centromere is poorly understood. Somehow the formation of the inner plate of a kinetochore is "seeded," followed by the cooperative assembly of the entire group of special proteins that form the kinetochore (Figure 4–50B). It seems that centromeres in complex organisms are defined more by an assembly of proteins than by a specific DNA sequence.

There are some striking similarities between the formation and maintenance of centromeres and the formation and maintenance of other regions of heterochromatin. The entire centromere forms as an all-or-none entity, suggesting a highly cooperative addition of proteins after a seeding event. Moreover, once formed, the structure seems to be directly inherited on the DNA as part of each round of chromosome replication. Thus, for example, some regions of our chromosomes contain nonfunctional alpha satellite DNA sequences that seem to be identical to those at the centromere; these sequences are presumed to have arisen from a chromosome-joining event that initially created one chromosome with two centromeres (an unstable, dicentric

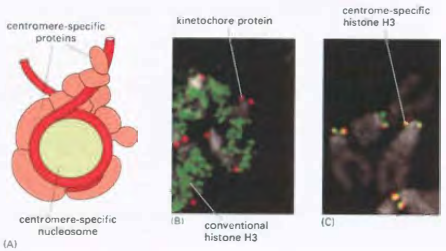


Figure 4–49 The specialized nucleosome formed on centromeres. (A) A model for the proteins that assemble on a yeast centromere. The specialized nucleosome contains an H3 variant (called CENP-A in most organisms), along with core histones H2A, H2B, and H4. The folding of DNA into this nucleosome facilitates the assembly of the other centromere-binding proteins, which form the kinetochore that attaches the centromere to the mitotic spindle (B) The localization of conventional histone H3 on *Drosophila* mitotic chromosomes. The conventional H3 has been fused to a fluorescent protein and appears *green*. A component of the kinetochore has been stained *red* with antibodies against a specific kinetochore protein. (C) The same experiment, but with the centromere-specific histone H3 (instead of the conventional H3) labeled *green*. When the *red* and *green* stains are coincident, the staining appears *yellow*. (A, adapted from P.B. Meluh et al., *Cell* 94:607–613, 1998; B and C, from S. Henikoff et al., *Proc. Natl. Acad. Sci. USA* 97:716–721, 2000. © National Academy of Sciences.)

Illumina Exhibit 1031, Page 19 of 53

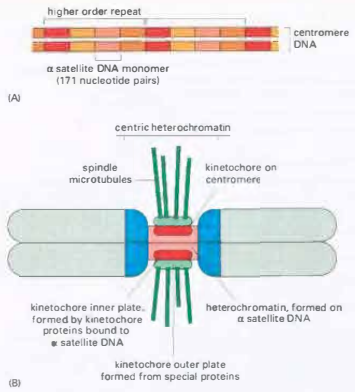


Figure 4–50 The structure of a human centromere. (A) The organization of the alpha satellite DNA sequences, which are repeated many thousands of times at a centromere. (B) An entire chromosome. The alpha satellite DNA sequences *(red)* are AT-rich and consist of a large number of repeats that vary slightly from one another in their DNA sequence. *Blue* represents the position of flanking centric heterochromatin, which contains DNA sequences composed of different types of repeats. As indicated, the kinetochore consists of an inner and an outer plate, formed by a set of kinetochore proteins. The spindle microtubules attach to the kinetochore in M phase of the cell cycle (see Figure 4–22). (B, adapted from T.D. Murphy and G.H. Karpen, *Cell* 93:317–320, 1998.)

chromosome; Figure 4–51A). Moreover, in some unusual cases, new human centromeres (called neocentromeres) have been observed to form spontaneously on fragmented chromosomes. Some of these new positions were originally euchromatic and lack alpha satellite DNA altogether (Figure 4–51B).

To explain these observations it has been proposed that *de novo* centromere formation requires an initial marking (or seeding) event, perhaps the formation of a specialized DNA–protein structure, which, in humans, happens more readily on arrays of alpha satellite DNA than on other DNA sequences. This mark would be duplicated when the chromosome divides, and the same centromere would then function in the next cell division. Very rarely, the mark would be lost after chromosome replication, in which case it would be very difficult to establish again (Figure 4–51C). Although the self-renewing nature of centromeres is not understood in detail, the type of models described for heterochromatin inheritance in Figure 4–48 could also be critical here.

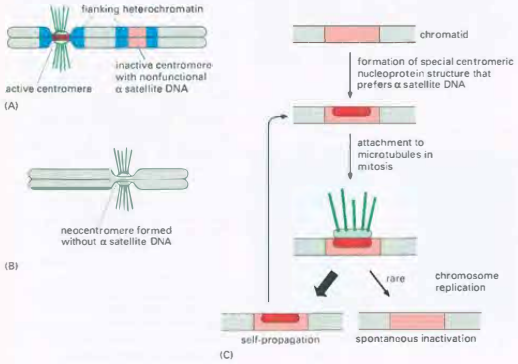


Figure 4–51 The plasticity of human centromere formation. (A) Owing to an ancient chromosome breakage and rejoining event, some human chromosomes contain two blocks of alpha satellite DNA *(red)*, each of which presumably functioned as a centromere in its original chromosome. Usually, these dicentric chromosomes are not stably propagated because they are attached improperly to the spindle and are broken apart during mitosis. In those chromosomes that survive, one of the centromeres has spontaneously inactivated, even though it contains all the necessary DNA sequences. This allows the chromosome to be stably propagated. (B) In a small fraction (1/2000) of human births, extra chromosomes are observed in cells of the offspring. Some of these extra chromosomes, which have formed from a breakage event, lack alpha satellite DNA altogether, yet new centromeres have arisen from what was originally euchromatic DNA. (C) A model to explain the plasticity and inheritance of centromeres.

The plasticity of centromeres may provide an important evolutionary advantage. We have seen that chromosomes evolve in part by breakage and rejoining events (see Figure 4–19). Many of these events produce chromosomes with two centromeres, or chromosome fragments with no centromeres at all. Although rare, the inactivation of centromeres and their ability to be activated *de novo* may occasionally allow newly formed chromosomes to be maintained stably and thereby facilitate the process of chromosome evolution.

Heterochromatin May Provide a Defense Mechanism Against Mobile DNA Elements

DNA packaged in heterochromatin often consists of large tandem arrays of short, repeated sequences that do not code for protein, as we saw above for the heterochromatin of mammalian centromeres. In contrast, euchromatic DNA is rich in genes and other single-copy DNA sequences. Although this correlation is not absolute (some arrays of repeated sequences exist in euchromatin and some genes are present in heterochromatin), this trend suggests that some types of repeated DNA may be a signal for heterochromatin formation. This idea is supported by experiments in which several hundred tandem copies of genes have been artificially introduced into the germ lines of flies and mice. In both organisms these gene arrays are often silenced, and in some cases, they can be observed under the microscope to have formed regions of heterochromatin. In contrast, when single copies of the same genes are introduced into the same position in the chromosome, they are actively expressed.

This feature, called *repeat-induced gene silencing*, may be a mechanism that cells have for protecting their genomes from being overtaken by mobile genetic elements. These elements, which are discussed in Chapter 5, can multiply and insert themselves throughout the genome. According to this idea, once a cluster of such mobile elements has formed, the DNA that contains them would be packaged into heterochromatin to prevent their further proliferation. The same mechanism could be responsible for forming the large regions of heterochromatin that contain large numbers of tandem repeats of a simple sequence, as occurs around centromeres.

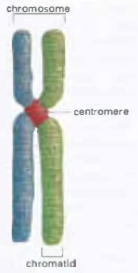


Figure 4–52 A typical mitotic chromosome at metaphase. Each sister chromatid contains one of two identical daughter DNA molecules generated earlier in the cell cycle by DNA replication.

Mitotic Chromosomes Are Formed from Chromatin in Its Most Condensed State

Having discussed the dynamic structure of interphase chromosomes, we now turn to the final level of DNA packaging, that observed for mitotic chromosomes. With the exception of a few specialized cases, such as the lampbrush and polytene chromosomes discussed above, most interphase chromosomes are too extended and entangled for their structures to be clearly seen. In contrast, the chromosomes from nearly all eucaryotic cells are readily visible during mitosis, when they coil up to form highly condensed structures. It is remarkable that this further condensation, which reduces the length of a typical interphase chromosome only about tenfold, produces such a dramatic change in the appearance of chromosomes.

Figure 4–52 depicts a typical **mitotic chromosome** at the metaphase stage of mitosis. The two daughter DNA molecules produced by DNA replication during interphase of the cell-division cycle are separately folded to produce two sister chromosomes, or *sister chromatids*, held together at their centromeres (see also Figure 4–21). These chromosomes are normally covered with a variety of molecules, including large amounts of RNA–protein complexes. Once this covering has been stripped away, each chromatid can be seen in electron micrographs to be organized into loops of chromatin emanating from a central scaffolding (Figures 4–53 and 4–54). Several types of experiment demonstrate that the order of visible features along a mitotic chromosome at least roughly reflects the order of the genes along the DNA molecule. Mitotic chromosome condensation can thus be thought of as the final level in the hierarchy of chromosome packaging (Figure 4–55).



Figure 4–53 A scanning electron micrograph of a region near one end of a typical mitotic chromosome. Each knoblike projection is believed to represent the tip of a separate looped domain. Note that the two identical paired chromatids drawn in Figure 4–52 can be clearly distinguished. (From M.P. Marsden and U.K. Laemmli, *Cell* 17:849–858, 1979. © Elsevier.)

Illumina Exhibit 1031, Page 20 of 53

The compaction of chromosomes during mitosis is a highly organized and dynamic process that serves at least two important purposes. First, when condensation is complete (in metaphase), sister chromatids have been disentangled from each other and lie side by side. Thus, the sister chromatids can easily separate when the mitotic apparatus begins pulling them apart. Second, the compaction of chromosomes protects the relatively fragile DNA molecules from being broken as they are pulled to separate daughter cells.

The condensation of interphase chromosomes into mitotic chromosomes occurs in M phase, and it is intimately connected with the progression of the cell cycle, as discussed in detail in Chapters 17 and 18. It requires a class of proteins called *condensins* which using the energy of ATP hydrolysis, drive the coiling of each interphase chromosome that produces a mitotic chromosome. Condensins are large protein complexes that contain SMC proteins; long, dimeric protein molecules hinged in the center, with globular domains at each end that bind DNA and hydrolyze ATP (Figure 4–56). When added to purified DNA, condensins use the energy of ATP hydrolysis to make large right-handed loops in the DNA. Although it is not yet known how they act on chromatin, the coiling model shown in Figure 4–56C is based on the fact that condensins are a major structural component of mitotic chromosomes, with one molecule of condensin being present for every 10,000 nucleotides of mitotic DNA.

Each Mitotic Chromosome Contains a Characteristic Pattern of Very Large Domains

As mentioned earlier, the display of the 46 human chromosomes at mitosis is called the human karyotype. When stained with dyes such as Giemsa, mitotic



Figure 4–54 An electron micrograph of a mitotic chromosome. This chromosome (from an insect) was treated to reveal loops of chromatin fibers that emanate from a central scaffold of the chromatid. Such micrographs support the idea that the chromatin in all chromosomes is folded into a series of looped domains (see Figure 4–55). (Courtesy of Uli Laemmli.)

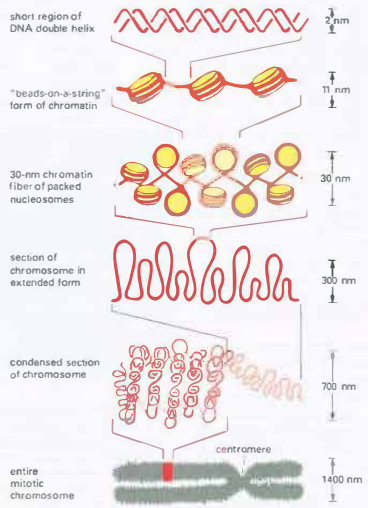


Figure 4–55 Chromatin packing. This model shows some of the many levels of chromatin packing postulated to give rise to the highly condensed mitotic chromosome.

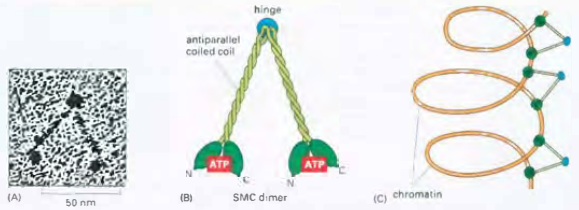


Figure 4–56 The SMC proteins in condensins. (A) Electron micrographs of a purified SMC dimer. (B) The structure of an SMC dimer. The long central region of this protein is an antiparallel coiled coil (see Figure 3–11) with a flexible hinge in its middle, as demonstrated by the electron micrograph in (A). (C) A model for the way in which the SMC proteins in condensins might compact chromatin. In reality, SMC proteins are components of a much larger condensin complex. It has been proposed that, in the cell, condensins coil long strings of looped chromatin domains (see Figure 4–55). In this way the condensins would form a structural framework that maintains the DNA in a highly organized state during M phase of the cell cycle. (A, courtesy of H.P. Erickson; B and C, adapted from T. Hirano, *Genes Dev.* 13:11–19, 1999.)

chromosomes show a striking and reproducible banding pattern along each chromosome, as shown in Figure 4–11. These bands are unrelated to those described earlier for the insect polytene chromosomes, which correspond to relatively small regions of interphase chromatin. In a human mitotic chromosome, all the chromatin is condensed and the bands represent a selective binding of the dyes.

By examining human chromosomes very early in mitosis, when they are less condensed than at metaphase, it has been possible to estimate that the total haploid genome contains about 2000 distinguishable bands. These coalesce progressively as condensation proceeds during mitosis, producing fewer and thicker bands. As we saw earlier, cytogeneticists routinely use the pattern of these chromosome bands to discover in patients genetic alterations such as chromosome inversions, translocations, and other types of chromosomal rearrangements (see Figure 4–12).

Mitotic chromosome bands are detected in chromosomes from species as diverse as humans and flies. Moreover, the pattern of bands in a chromosome has remained unchanged over long periods of evolutionary time. Each human chromosome, for example, has a clearly recognizable counterpart with a nearly identical banding pattern in the chromosomes of the chimpanzee, gorilla, and orangutan—although there are also clear differences, such as chromosome fusion, that give the human 46 chromosomes instead of the ape's 48 (Figure 4–57). This conservation suggests that chromosomes are organized into large domains that may be important for chromosomal function.

Even the thinnest of the bands in Figure 4–11 probably contains more than a million nucleotide pairs, which is nearly the size of a bacterial genome. These bands seem to reflect a rough division of chromosomes into regions of different GC content. The nucleotide sequence of the human genome has revealed large non-random blocks of sequence (some greater than 10^7 nucleotide pairs) that are significantly higher or lower in GC content than the genome-wide average of 41%. The blocks correlate roughly with the staining pattern of metaphase chromosomes. For example, bands that are darkly stained by Giemsa (the so-called G-bands) are correlated with DNA that is low in GC content, whereas lightly stained bands (the R-bands) correspond to DNA of higher than average GC content.

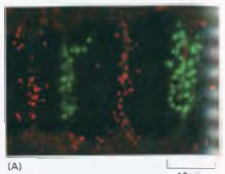
In general, GC-rich regions of the genome have a higher density of genes, especially of "house-keeping" genes, the genes that are expressed in virtually all cell types. On the basis of these observations, it has been proposed that the

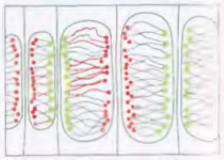


Figure 4–57 Comparison of the Giemsa pattern of the largest human chromosome (chromosome 1) with that of chimpanzee, gorilla, and orangutan. Comparisons among the staining patterns of all the chromosomes indicate that human chromosomes are more closely related to those of chimpanzee than to those of gorilla and that they are more distantly related to those of orangutan. (Adapted from M.W. Strickberger, Evolution, 3rd edn. Sudbury, MA: Jones & Bartlett Publishers, 2000.)

Illumina Exhibit 1031, Page 21 of 53

Figure 4–58 The polarized orientation of chromosomes found in certain interphase nuclei. (A) Fluorescent light micrograph of interphase nuclei from the rapidly growing root tip of a plant. Centromeres are stained *green* and telomeres *red* by *in situ* hybridization of centromere- and telomere-specific DNA sequences coupled to the different fluorescent dyes. (B) Interpretation of (A) showing chromosomes in the Rabl orientation with all the centromeres facing one side of a nucleus and all the telomeres pointing toward the opposite side. (A, from R. Abranches et al., *J. Cell Biol.* 143:5–12, 1998. © The Rockefeller University Press.)





banding pattern may be related to gene expression. Perhaps the differentiation of chromosomes into G- and R-bands reflects subtle differences, determined by GC content, in the way in which chromatin loops are packaged in these areas. If this idea is correct, the rough division of chromosomes can be seen as a form of compartmentalization, in which the particular cellular components involved in gene expression are more concentrated in the R-bands where their activities are required. In any case, it should be obvious from this discussion that we are only beginning to glimpse the principles of large-scale chromosome organization.

Individual Chromosomes Occupy Discrete Territories in an Interphase Nucleus

We saw earlier in this chapter that chromosomes from eucaryotes are contained in the cell nucleus. However, the nucleus is not simply a bag of chromosomes; rather, the chromosomes—as well as the other components inside the nucleus which we shall encounter in subsequent chapters—are highly organized. The way in which chromosomes are organized in the nucleus during interphase, when they are active and difficult to see, has intrigued biologists since the nineteenth century. Although our understanding today is far from complete, we do know some interesting features of these chromosome arrangements.

A certain degree of chromosomal order results from the configuration that the chromosomes always have at the end of mitosis. Just before a cell divides, the condensed chromosomes are pulled to each spindle pole by microtubules attached to the centromeres; thus, as the chromosomes move, the centromeres lead the way and the distal arms (terminating in the telomeres) lag behind. The chromosomes in some nuclei tend to retain this so-called *Rabl orientation* throughout interphase, with their centromeres facing one pole of the nucleus and their telomeres pointing toward the opposite pole (Figures 4–58 and 4–59).

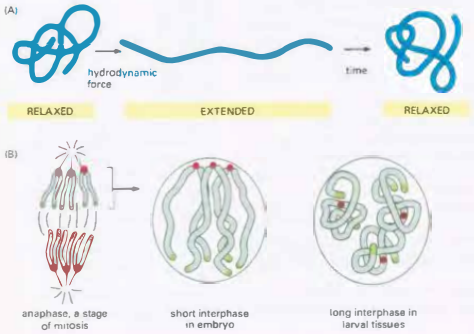


Figure 4–59 A polymer analogy for interphase chromosome organization. (A) The behavior of a polymer in solution. Entropy drives a long polymer into a compact conformation in the absence of an externally applied force. If the polymer is subjected to shear or hydrodynamic force, it becomes extended. But once the force is removed, the polymer chain returns to a more favorable, compact conformation. (B) The behavior of interphase chromosomes may reflect the same simple principles. In *Drosophila* embryos, for example, mitotic divisions occur at intervals of about 10 minutes; during the short intervening interphases, the chromosomes have little time to relax from the Rabl orientation induced by their movement during mitosis. However, in later stages of development, when interphase is much longer, the chromosomes have time to fold up. This folding may be strongly affected by specific associations between different regions of the same chromosome. (Adapted from A.F. Dernburg et al., *Cell* 85:745–759, 1996.)

Figure 4–60 Selective "painting" of two interphase chromosomes in a human peripheral lymphocyte. The fluorescent light micrograph shows that the two copies of human chromosome 18 *(red)* and chromosome 19 *(turquoise)* occupy discrete territories of the nucleus. (From J.A. Croft et al., *J. Cell Biol.* 145:1119–1131, 1999. © The Rockefeller University Press.)



The chromosomes in most interphase cells are not found in the Rabl orientation; instead, the centromeres seem to be dispersed in the nucleus. Most cells have a longer interphase than the specialized cells illustrated above, and this presumably gives their chromosomes time to assume a different conformation (see Figure 4–59). Nevertheless, each interphase chromosome does tend to occupy a discrete and relatively small territory in the nucleus: that is, the different chromosomes are not extensively intertwined (Figure 4–60).

One device for organizing chromosomes in the nucleus may be the attachment of certain portions of each chromosome to the nuclear envelope (Figure 4–61). For example, in many cells, telomeres seem bound in this way. But the exact position of a chromosome in a nucleus is not fixed. In the same tissue, for example, two apparently identical cells can have different chromosomes as nearest neighbors.

Some cell biologists believe that there is an intranuclear framework, analogous to the cytoskeleton, on which chromosomes and other components of the nucleus are organized. The *nuclear matrix*, or *scaffold*, has been defined as the insoluble material left in the nucleus after a series of biochemical extraction steps. Some of the proteins that constitute it can be shown to bind specific DNA sequences called *SARs* or *MARs* (scaffold-associated or matrix-associated regions). These DNA sequences have been postulated to form the base of chromosomal loops (see Figure 4–44), or to attach chromosomes to the nuclear envelope and other structures in the nucleus. By means of such chromosomal attachment sites, the matrix might help to organize chromosomes, localize genes, and regulate gene expression and DNA replication. It still remains uncertain, however, whether the matrix isolated by cell biologists represents a structure that is present in intact cells.

Summary

Chromosomes are generally decondensed during interphase, so that their structure is difficult to visualize directly. Notable exceptions are the specialized lampbrush chromosomes of vertebrate oocytes and the polytene chromosomes in the giant secretory cells of insects. Studies of these two types of interphase chromosomes suggest that each long DNA molecule in a chromosome is divided into a large number of discrete domains organized as loops of chromatin, each loop probably consisting of a folded 30-nm chromatin fiber. When genes contained in a loop are expressed, the loop decondenses and allows the cell's machinery easy access to the DNA.

Euchromatin makes up most of interphase chromosomes and probably corresponds to looped domains of 30-nm fibers. However, euchromatin is interrupted by stretches of heterochromatin, in which 30-nm fibers are subjected to additional levels of packing that usually render it resistant to gene expression. Heterochromatin is commonly found around centromeres and near telomeres, but it is also present at other positions on chromosomes. Although considerably less condensed than mitotic chromosomes, interphase chromosomes occupy discrete territories in the cell nucleus; that is, they are not extensively intertwined.

All chromosomes adopt a highly condensed conformation during mitosis. When they are specially stained, these mitotic chromosomes have a banding structure that allows each individual chromosome to be recognized unambiguously. These bands contain millions of DNA nucleotide pairs, and they reflect a poorly-understood coarse heterogeneity of chromosome structure.

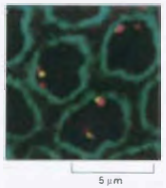


Figure 4–61 Specific regions of interphase chromosomes in close proximity to the nuclear envelope. This high-resolution microscopic view of nuclei from a *Drosophila* embryo shows the localization of two different regions of chromosome 2 *(yellow* and *magenta)* close to the nuclear envelope (stained green with antilamina antibodies). Other regions of the same chromosome are more distant from the envelope. (From W.F. Marshall et al., *Mol. Biol. Cell* 7:825–842, 1996.)

Illumina Exhibit 1031, Page 22 of 53

References

General

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

Lewin B (2000) Genes VII. Oxford: Oxford University Press.

Lodish H, Berk A, Zipursky SL et al. (2000) Molecular Cell Biology. 4th edn. New York: WH Freeman.

Wolfe A (1999) Chromatin: Structure and Function. 3rd edn. New York: Academic Press.

The Structure and Function of DNA

Avery OT, MacLeod CM & McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79, 137.

Meselson M & Stahl FW (1958) The replication of DNA in *E. coli*. *Proc. Natl. Acad. Sci. USA* 44, 671–682.

Watson JD & Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acids. *Nature* 171, 737–738.

Chromosomal DNA and Its Packaging in the Chromatin Fiber

Aalfs JD & Kingston RE (2000) What does 'chromatin remodeling' mean? *Trends Biochem. Sci.* 25, 548–555.

Cairns BR (1998) Chromatin remodeling machines: similar motors, ulterior motives. *Trends Biochem. Sci.* 23, 20–25.

Carter NP (1994) Cytogenetic analysis by chromosome painting. *Cytometry* 18, 2–10.

Cheung P, Allis CD & Sassone-Corsi P (2000) Signaling to chromatin through histone modifications. *Cell* 103, 263–271.

Clark MS (1999) Comparative genomics: the key to understanding the Human Genome Project. *Bioessays* 21, 121–130.

DePamphilis ML (1999) Replication origins in metazoan chromosomes: fact or fiction? *Bioessays* 21, 5–16.

Dunham I, Shimizu N, Roe BA et al. (1999) The DNA sequence of human chromosome 22. *Nature* 402, 489–495.

Felsenfeld G (1985) DNA. *Sci. Am.* 253(4), 58–67.

Grunstein M (1992) Histones as regulators of genes. *Sci. Am.* 267(4), 68–74B.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Jenuwein T & Allis CD (2001) Translating the histone code. *Science* 293:1074–1080.

Kingston RE & Narlikar GJ (1999) ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. *Genes Dev.* 13, 2339–2352.

Kornberg RD & Lorch Y (1999) Chromatin-modifying and -remodeling complexes. *Curr. Opin. Genet. Dev.* 9, 148–151.

Kornberg RD & Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98, 285–294.

Luger K & Richmond TJ (1998) The histone tails of the nucleosome. *Curr. Opin. Genet. Dev.* 8, 140–146.

Luger K, Mader AW, Richmond RK et al. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389, 251–260.

McEachern MJ, Krauskopf A & Blackburn EH (2000) Telomeres and their control. *Annu. Rev. Genet.* 34, 331–358.

Ng HH & Bird A (2000) Histone deacetylases: silencers for hire. *Trends Biochem. Sci.* 25, 121–126.

O'Brien S, Menotti-Raymond M, Murphy W et al. (1999) The promise of comparative genomics in mammals. *Science* 286, 458–480.

Pidoux AL & Allshire RC (2000) Centromeres: getting a grip of chromosomes. *Curr. Opin. Cell Biol.* 12, 308–319.

Rhodes D (1997) Chromatin structure. The nucleosome core all wrapped up. *Nature* 389, 231–233.

Rice JC & Allis CD (2001) Histone methylation versus histone acetylation: new insights into epigenetic regulation. *Curr. Opin. Cell Biol.* 13, 263–273.

Ried T, Schrock E, Ning Y & Wienberg J (1998) Chromosome painting: a useful art. *Hum. Mol. Genet.* 7, 1619–1626.

Rubin GM (2001) Comparing species. *Nature* 409, 820–821.

Stewart A (1990) The functional organization of chromosomes and the nucleus—a special issue. *Trends Genet.* 6, 377–379.

Strahl BD & Allis CD (2000) The language of covalent histone modifications. *Nature* 403, 41–45.

Travers AA (1987) DNA bending and nucleosome positioning. *Trends Biochem. Sci.* 12, 108–112.

Wu J & Grunstein M (2000) 25 years after the nucleosome model: chromatin modifications. *Trends Biochem. Sci.* 25, 619–623.

The Global Structure of Chromosomes

Agard DA & Sedat JW (1983) Three-dimensional architecture of a polytene nucleus. *Nature* 302, 676–681.

Ashburner M, Chihara C, Meltzer P & Richards G (1974) Temporal control of puffing activity in polytene chromosomes. *Cold Spring Harbor Symp. Quant. Biol.* 38, 655–662.

Bickmore WA & Sumner AT (1989) Mammalian chromosome banding—an expression genome organization. *Trends Genet.* 5, 144–148.

Birchler JA, Bhadra MP & Bhadra U (2000) Making noise about silence: repression of repeated genes in animals. *Curr. Opin. Genet. Dev.* 10, 211–216.

Callan HG (1982) Lampbrush chromosomes. *Proc. Roy. Soc. Lond. Ser. B. (Biol.)* 214, 417–448.

Croft JA, Bridger JM, Boyle S et al. (1999) Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.* 145, 1119–1131.

Griffith JD, Comeau L, Rosenfield S et al. (1999) Mammalian telomeres end in a large duplex loop. *Cell* 97, 503–514.

Grunstein M (1997) Molecular model for telomeric heterochromatin in yeast. *Curr. Opin. Cell Biol.* 9, 383–387.

Hart CM & Laemmli UK (1998) Facilitation of chromatin dynamics by SARs. *Curr. Opin. Genet. Dev.* 8, 519–525.

Henikoff S (1990) Position-effect variegation after 60 years. *Trends Genet.* 6, 422–426.

Henikoff S (1998) Conspiracy of silence among repeated transgenes. *Bioessays* 20, 532–535.

Hirano T (1999) SMC-mediated chromosome mechanics: a conserved scheme from bacteria to vertebrates. *Genes Dev.* 13, 11–19.

Hirano T (2000) Chromosome cohesion, condensation, and separation. *Annu. Rev. Biochem.* 69, 115–144.

Lamond AI & Earnshaw WC (1998) Structure and function in the nucleus. *Science* 280, 547–553.

Lyko F & Paro R (1999) Chromosomal elements conferring epigenetic inheritance. *Bioessays* 21, 824–832.

Marsden M & Laemmli UK (1979) Metaphase chromosome structure: evidence for a radial loop model. *Cell* 17, 849–858.

Pluta AF, Mackay AM, Ainsztein AM et al. (1995) The centromere: hub of chromosomal activities. *Science* 270, 1591–1594.

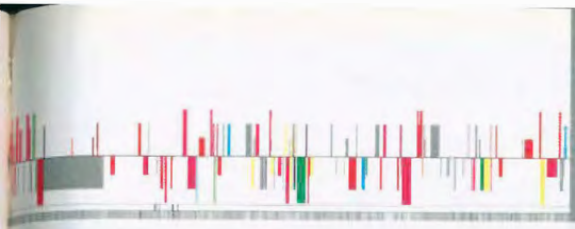
Saitoh N, Goldberg I & Earnshaw WC (1995) The SMC proteins and the coming of age of the chromosome scaffold hypothesis. *Bioessays* 17, 759–766.

Spector DL (1993) Macromolecular domains within the cell nucleus. *Annu. Rev. Cell Biol.* 9, 265–315.

Thummel CS (1990) Puffs and gene regulation—molecular insights into the Drosophila ecdysone regulatory hierarchy. *Bioessays* 12, 561–568.

Weiler KS & Wakimoto BT (1995) Heterochromatin and gene expression in Drosophila. *Annu. Rev. Genet.* 29, 577–605.

Zhimulev IF (1998) Morphology and structure of polytene chromosomes. *Adv. Genet.* 37, 1–566.



5

DNA REPLICATION, REPAIR, AND RECOMBINATION

THE MAINTENANCE OF DNA SEQUENCES

DNA REPLICATION MECHANISMS

THE INITIATION AND COMPLETION OF DNA REPLICATION IN CHROMOSOMES

DNA REPAIR

GENERAL RECOMBINATION

SITE-SPECIFIC RECOMBINATION

The ability of cells to maintain a high degree of order in a chaotic universe depends upon the accurate duplication of vast quantities of genetic information carried in chemical form as DNA. This process, called *DNA replication*, must occur before a cell can produce two genetically identical daughter cells. Maintaining order also requires the continued surveillance and repair of this genetic information because DNA inside cells is repeatedly damaged by chemicals and radiation from the environment, as well as by thermal accidents and reactive molecules. In this chapter we describe the protein machines that replicate and repair the cell's DNA. These machines catalyze some of the most rapid and accurate processes that take place within cells, and their mechanisms clearly demonstrate the elegance and efficiency of cellular chemistry.

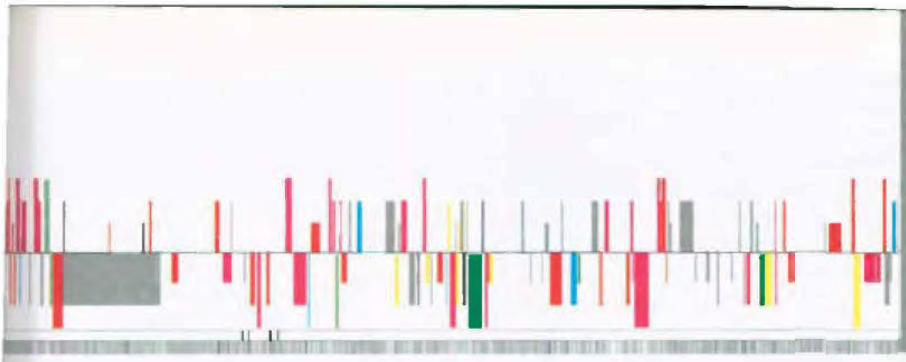
While the short-term survival of a cell can depend on preventing changes in its DNA, the long-term survival of a species requires that DNA sequences be changeable over many generations. Despite the great efforts that cells make to protect their DNA, occasional changes in DNA sequences do occur. Over time, these changes provide the genetic variation upon which selection pressures act during the evolution of organisms.

We begin this chapter with a brief discussion of the changes that occur in DNA as it is passed down from generation to generation. Next, we discuss the cellular mechanisms—DNA replication and DNA repair—that are responsible for keeping these changes to a minimum. Finally, we consider some of the most intriguing ways in which DNA sequences are altered by cells, with a focus on DNA recombination and the movement of special DNA sequences in our chromosomes called transposable elements.

THE MAINTENANCE OF DNA SEQUENCES

Although the long-term survival of a species is enhanced by occasional genetic changes, the survival of the individual demands genetic stability. Only rarely do the cell's DNA-maintenance processes fail, resulting in permanent change in the DNA. Such a change is called a **mutation**, and it can destroy an organism if it

Illumina Exhibit 1031, Page 23 of 53



8

MANIPULATING PROTEINS, DNA, AND RNA

ISOLATING CELLS AND GROWING THEM IN CULTURE

FRACTIONATION OF CELLS

ISOLATING, CLONING, AND SEQUENCING DNA

ANALYZING PROTEIN STRUCTURE AND FUNCTION

STUDYING GENE EXPRESSION AND FUNCTION

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Progress in science is often driven by advances in technology. Biology, for example, entered a new era when Anton van Leeuwenhoek, a Dutch dry-goods dealer, ground the first microscope lens. Peering into his marvelous new looking glass, van Leeuwenhoek discovered a previously unseen cellular world, where tiny creatures tumble and twirl in a small droplet of water (Figure 8–1).

The 21st century promises to be a particularly exciting time for biology. New methods for analyzing proteins, DNA, and RNA are fueling an information explosion and allowing scientists to study cells and their macromolecules in previously unimagined ways. We now have access to the sequences of many billions of nucleotides, providing the complete molecular blueprints for dozens of organisms—from microbes and mustard weeds to worms, flies, and humans. And powerful new techniques are helping us to decipher that information, allowing us not only to compile huge, detailed catalogs of genes and proteins, but to begin to unravel how these components work together to form functional cells and organisms. The goal is nothing short of obtaining a complete understanding of what takes place inside a cell as it responds to its environment and interacts with its neighbors. We want to know which genes are switched on, which mRNA transcripts are present, and which proteins are active—where they are located, with whom they partner, and to which pathways or networks they belong. We also want to understand how the cell successfully manages this staggering number of variables and how it chooses among an almost unlimited number of possibilities in performing its diverse biological tasks. Possession of such information will permit us to begin to build a framework for delineating, and eventually predicting, how genes and proteins operate to lay the foundations for life.

In this chapter we briefly review some of the principal methods used to study cells and their components, particularly proteins, DNA, and RNA. We consider how cells of different types can be separated from tissues and grown outside the body and how cells can be disrupted and their organelles and constituent macromolecules isolated in pure form. We then review the breakthroughs in recombinant DNA technology that continue to revolutionize our understanding

Illumina Exhibit 1031, Page 24 of 53

of cellular function. Finally we present the latest techniques used to determine the structures and functions of proteins and genes, as well as to dissect their complex interactions.

This chapter serves as a bridge from the basics of cell and molecular biology to the detailed discussion of how these macromolecules are organized and function together to coordinate the growth, development, and physiology of cells and organisms. The techniques and methods described here have made possible the discoveries that are presented throughout this book, and they are currently being used by tens of thousands of scientists each day.

ISOLATING CELLS AND GROWING THEM IN CULTURE

Although the organelles and large molecules in a cell can be visualized with microscopes, understanding how these components function requires a detailed biochemical analysis. Most biochemical procedures require obtaining large numbers of cells and then physically disrupting them to isolate their components. If the sample is a piece of tissue, composed of different types of cells, heterogeneous cell populations will be mixed together. To obtain as much information as possible about an individual cell type, biologists have developed ways of dissociating cells from tissues and separating the various types. These manipulations result in a relatively homogeneous population of cells that can then be analyzed—either directly or after their number has been greatly increased by allowing the cells to proliferate as a pure culture.

Cells Can Be Isolated from a Tissue Suspension and Separated into Different Types

The first step in isolating cells of a uniform type from a tissue that contains a mixture of cell types is to disrupt the extracellular matrix that holds the cells together. The best yields of viable dissociated cells are usually obtained from fetal or neonatal tissues. The tissue sample is typically treated with proteolytic enzymes (such as trypsin and collagenase) to digest proteins in the extracellular matrix and with agents (such as ethylenediaminetetraacetic acid, or EDTA) that bind, or chelate, the Ca^{2+} on which cell–cell adhesion depends. The tissue can then be teased apart into single living cells by gentle agitation.

Several approaches are used to separate the different cell types from a mixed cell suspension. One exploits differences in physical properties. Large cells can be separated from small cells and dense cells from light cells by centrifugation, for example. These techniques will be described below in connection with the separation of organelles and macromolecules, for which they were originally developed. Another approach is based on the tendency of some cell types to adhere strongly to glass or plastic, which allows them to be separated from cells that adhere less strongly.

An important refinement of this last technique depends on the specific binding properties of antibodies. Antibodies that bind specifically to the surface of only one cell type in a tissue can be coupled to various matrices—such as collagen, polysaccharide beads, or plastic—to form an affinity surface to which only cells recognized by the antibodies can adhere. The bound cells are then recovered by gentle shaking, by treatment with trypsin to digest the proteins that mediate the adhesion, or, in the case of a digestible matrix (such as collagen), by degrading the matrix itself with enzymes (such as collagenase).

One of the most sophisticated cell-separation technique uses an antibody coupled to a fluorescent dye to label specific cells. The labeled cells can then be separated from the unlabeled ones in an electronic *fluorescence-activated cell sorter*. In this remarkable machine, individual cells traveling single file in a fine stream pass through a laser beam and the fluorescence of each cell is rapidly measured. A vibrating nozzle generates tiny droplets, most containing either one cell or no cells. The droplets containing a single cell are automatically given

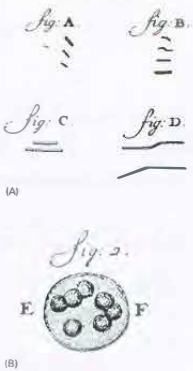


Figure 8–1 Microscopic life. A sample of "diverse animalcules" seen by van Leeuwenhoek using his simple microscope. (A) Bacteria seen in material he excavated from between his teeth. Those in fig. B he described as "swimming first forward and then backwards" (1692). (B) The eucaryotic green alga *Volvox* (1700). (Courtesy of the John Innes Foundation.)

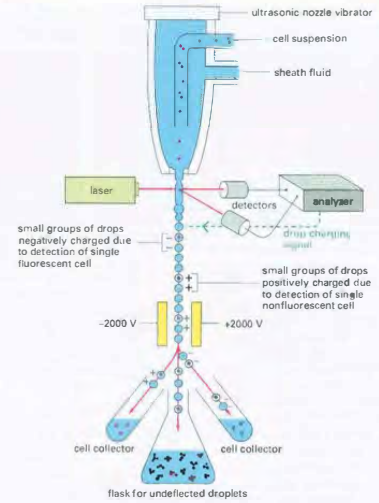


Figure 8–2 A fluorescence-activated cell sorter. A cell passing through the laser beam is monitored for fluorescence. Droplets containing single cells are given a negative or positive charge, depending on whether the cell is fluorescent or not. The droplets are then deflected by an electric field into collection tubes according to their charge. Note that the cell concentration must be adjusted so that most droplets contain no cells and flow to a waste container together with any cell clumps.

a positive or a negative charge at the moment of formation, depending on whether the cell they contain is fluorescent; they are then deflected by a strong electric field into an appropriate container. Occasional clumps of cells, detected by their increased light scattering, are left uncharged and are discarded into a waste container. Such machines can accurately select 1 fluorescent cell from a pool of 1000 unlabeled cells and sort several thousand cells each second (Figure 8–2).

Selected cells can also be obtained by carefully dissecting them from thin tissue slices that have been prepared for microscopic examination (discussed in Chapter 9). In one approach, a tissue section is coated with a thin plastic film and a region containing the cells of interest is irradiated with a focused pulse from an infrared laser. This light pulse melts a small circle of the film, binding the cells underneath. These captured cells are then removed for further analysis. The technique, called *laser capture microdissection*, can be used to separate and analyze cells from different areas of a tumor, allowing their properties to be compared. A related method uses a laser beam to directly cut out a group of cells and catapult them into an appropriate container for future analysis (Figure 8–3).

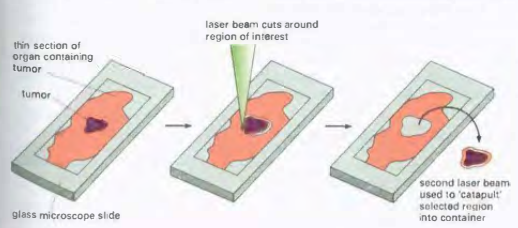


Figure 8–3 Microdissection techniques allow selected cells to be isolated from tissue slices. This method uses a laser beam to excise a region of interest and eject it into a container, and it permits the isolation of even a single cell from a tissue sample.

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 25 of 53

Once a uniform population of cells has been obtained—by microdissection or by any of the separation methods just described—it can be used directly for biochemical analysis. A homogeneous cell sample also provides a starting material for cell culture, thereby allowing the number of cells to be greatly increased and their complex behavior to be studied under the strictly defined conditions of a culture dish.

Cells Can Be Grown in a Culture Dish

Given appropriate surroundings, most plant and animal cells can live, multiply, and even express differentiated properties in a tissue-culture dish. The cells can be watched continuously under the microscope or analyzed biochemically, and the effects of adding or removing specific molecules, such as hormones or growth factors, can be explored. In addition, by mixing two cell types, the interactions between one cell type and another can be studied. Experiments performed on cultured cells are sometimes said to be carried out *in vitro* (literally, "in glass") to contrast them with experiments using intact organisms, which are said to be carried out *in vivo* (literally, "in the living organism"). These terms can be confusing, however, because they are often used in a very different sense by biochemists. In the biochemistry lab, *in vitro* refers to reactions carried out in a test tube in the absence of living cells, whereas *in vivo* refers to any reaction taking place inside a living cell (even cells that are growing in culture).

Tissue culture began in 1907 with an experiment designed to settle a controversy in neurobiology. The hypothesis under examination was known as the neuronal doctrine, which states that each nerve fiber is the outgrowth of a single nerve cell and not the product of the fusion of many cells. To test this contention, small pieces of spinal cord were placed on clotted tissue fluid in a warm, moist chamber and observed at regular intervals under the microscope. After a day or so, individual nerve cells could be seen extending long, thin filaments into the clot. Thus the neuronal doctrine received strong support, and the foundations for the cell-culture revolution were laid.

The original experiments on nerve fibers used cultures of small tissue fragments called explants. Today, cultures are more commonly made from suspensions of cells dissociated from tissues using the methods described earlier. Unlike bacteria, most tissue cells are not adapted to living in suspension and require a solid surface on which to grow and divide. For cell cultures, this support is usually provided by the surface of a plastic tissue-culture dish. Cells vary in their requirements, however, and many do not grow or differentiate unless the culture dish is coated with specific extracellular matrix components, such as collagen or laminin.

Cultures prepared directly from the tissues of an organism, that is, without cell proliferation *in vitro*, are called *primary cultures*. These can be made with or without an initial fractionation step to separate different cell types. In most cases, cells in primary cultures can be removed from the culture dish and made to proliferate to form a large number of so-called secondary cultures; in this way, they may be repeatedly subcultured for weeks or months. Such cells often display many of the differentiated properties appropriate to their origin: fibroblasts continue to secrete collagen; cells derived from embryonic skeletal muscle fuse to form muscle fibers that contract spontaneously in the culture dish; nerve cells extend axons that are electrically excitable and make synapses with other nerve cells; and epithelial cells form extensive sheets with many of the properties of an intact epithelium (Figure 8–4). Because these phenomena occur in culture, they are accessible to study in ways that are often not possible in intact tissues.

Serum-free, Chemically Defined Media Permit Identification of Specific Growth Factors

Until the early 1970s tissue culture seemed a blend of science and witchcraft. Although fluid clots were replaced by dishes of liquid media containing specified quantities of small molecules such as salts, glucose, amino acids, and vitamins,

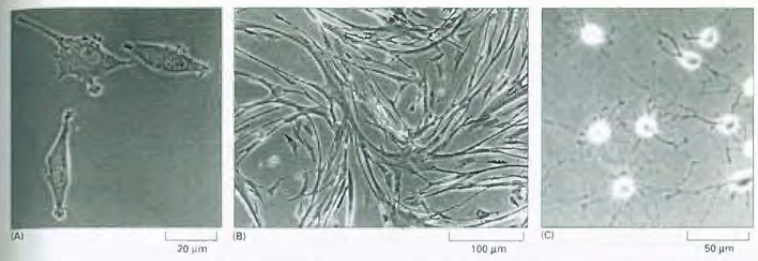




Figure 8–4 Cells in culture. (A) Phase-contrast micrograph of fibroblasts in culture. (B) Micrograph of myoblasts in culture shows cells fusing to form multinucleate muscle cells. (C) Oligodendrocyte precursor cells in culture. (D) Tobacco cells, from a fast-growing immortal cell line called BY2, in liquid culture. Nuclei and vacuoles can be seen in these cells. (A, courtesy of Daniel Zicha; B, courtesy of Rosalind Zalin; C, from D.G. Tang et al., *J. Cell Biol.* 148:971–984, 2000. © The Rockefeller University Press; D, courtesy of Gethin Roberts.)

most media also included either a poorly defined mixture of macromolecules in the form of horse or fetal calf serum, or a crude extract made from chick embryos. Such media are still used today for most routine cell culture (Table 8–1), but they make it difficult for the investigator to know which specific macromolecules a particular type of cell requires to thrive and to function normally.

This difficulty led to the development of various serum-free, chemically defined media. In addition to the usual small molecules, such defined media contain one or more specific proteins that the cells require to survive and proliferate in culture. These added proteins include growth factors, which stimulate cell proliferation, and transferrin, which carries iron into cells. Many of the extracellular protein signaling molecules essential for the survival, development, and

TABLE 8–1 Composition of a Typical Medium Suitable for the Cultivation of Mammalian Cells

AMINO ACIDS	VITAMINS	SALTS	MISCELLANEOUS	PROTEINS (REQUIRED IN SERUM-FREE, CHEMICALLY DEFINED MEDIA)
Arginine	biotin	NaCl	glucose	insulin
Cystine	choline	KCl	penicillin	transferrin
Glutamine	folate	NaH_2PO_4	streptomycin	specific growth factors
Histidine	nicotinamide	$NaHCO_3$	phenol red	
Isoleucine	pantothenate	$CaCl_2$	whole serum	
Leucine	pyridoxal	$MgCl_2$		
Lysine	thiamine			
Methionine	riboflavin			
Phenylalanine				
Threonine				
Trytophan				
Tyrosine				
Valine				

Glucose is used at a concentration of 5–10 mM. The amino acids are all in the L form and, with one or two exceptions, are used at concentrations of 0.1 or 0.2 mM; vitamins are used at a 100-fold lower concentration, that is, about 1 μM. Serum, which is usually from horse or calf, is added to make up 10% of the total volume. Penicillin and streptomycin are antibiotics added to suppress the growth of bacteria. Phenol red is a pH indicator dye whose color is monitored to ensure a pH of about 7.4.
Cultures are usually grown in a plastic or glass container with a suitably prepared surface that allows the attachment of cells. The containers are kept in an incubator at 37°C in an atmosphere of 5% CO_2, 95% air.

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 26 of 53

proliferation of specific cell types were discovered by studies seeking minimal conditions under which the cell type behaved properly in culture. Thus, the search for new signaling molecules has been made much easier by the availability of chemically defined media.

Eucaryotic Cell Lines Are a Widely Used Source of Homogeneous Cells

Most vertebrate cells stop dividing after a finite number of cell divisions in culture, a process called *cell senescence* (discussed in Chapter 17). Normal human fibroblasts, for example, typically divide only 25–40 times in culture before they stop. In these cells, the limited proliferation capacity reflects a progressive shortening of the cell's telomeres, the repetitive DNA sequences and associated proteins that cap the ends of each chomosome (discussed in Chapter 5). Human somatic cells have turned off the enzyme, called *telomerase*, that normally maintains the telomeres, which is why their telomeres shorten with each cell division. Human fibroblasts can be coaxed to proliferate indefinitely by providing them with the gene that encodes the catalytic subunit of telomerase; they can then be propagated as an "immortalized" cell line.

Some human cells, however, are not immortalized by this trick. Although their telomeres remain long, they still stop dividing after a limited number of divisions because the culture conditions activate cell-cycle *checkpoint mechanisms* (discussed in Chapter 17) that arrest the cell cycle. In order to immortalize these cells, one has to do more than introduce telomerase. One must also inactivate the checkpoint mechanisms, which can be done by introducing certain cancer-promoting oncogenes derived from tumor viruses (discussed in Chapter 23). Unlike human cells, most rodent cells do not turn off telomerase and therefore their telomeres do not shorten with each cell division. In addition, rodent cells can undergo genetic changes in culture that inactivate their checkpoint mechanisms, thereby spontaneously producing immortalized cell lines.

Cell lines can often be most easily generated from cancer cells, but these cells differ from those prepared from normal cells in several ways. Cancer cell lines often grow without attaching to a surface, for example, and they can proliferate to a very much higher density in a culture dish. Similar properties can be induced experimentally in normal cells by transforming them with a tumor-inducing virus or chemical. The resulting transformed cell lines, in reciprocal fashion, can often cause tumors if injected into a susceptible animal. Both transformed and immortal cell lines are extremely useful in cell research as sources of very large numbers of cells of a uniform type, especially since they can be stored in liquid nitrogen at −196°C for an indefinite period and retain their viability when thawed. It is important to keep in mind, however, that the cells in both types of cell lines nearly always differ in important ways from their normal progenitors in the tissues from which they were derived. Some widely used cell lines are listed in Table 8–2.

Among the most promising cell cultures to be developed—from a medical point of view—are the human embryonic stem (ES) cell lines. These cells, harvested from the inner cell mass of the early embryo, can proliferate indefinitely while retaining the ability to give rise to any part of the body (discussed in Chapter 21). ES cells could potentially revolutionize medicine by providing a source of cells capable of replacing or repairing tissues that have been damaged by injury or disease.

Although all the cells in a cell line are very similar, they are often not identical. The genetic uniformity of a cell line can be improved by cell cloning, in which a single cell is isolated and allowed to proliferate to form a large colony. In such a colony, or clone, all the cells are descendants of a single ancestor cell. One of the most important uses of cell cloning has been the isolation of mutant cell lines with defects in specific genes. Studying cells that are defective in a specific protein often reveals valuable information about the function of that protein in normal cells.

Some important steps in the development of cell culture are listed in Table 8–3.

TABLE 8–2 Some Commonly Used Cell Lines

CELL LINE*	CELL TYPE AND ORIGIN
3T3	fibroblast (mouse)
BHK21	fibroblast (Syrian hamster)
MDCK	epithelial cell (dog)
HeLa	epithelial cell (human)
PtK1	epithelial cell (rat kangaroo)
L6	myoblast (rat)
PC12	chromaffin cell (rat)
SP2	plasma cell (mouse)
COS	kidney (monkey)
293	kidney (human); transformed with adenovirus
CHO	ovary (chinese hamster)
DT40	lymphoma cell for efficient targeted recombination (chick)
R1	embryonic stem cells (mouse)
E14.1	embryonic stem cells (mouse)
H1, H9	embryonic stem cells (human)
S2	macrophage-like cells (*Drosophila*)
BY2	undifferentiated meristematic cells (tobacco)

*Many of these cell lines were derived from tumors. All of them are capable of indefinite replication in culture and express at least some of the special characteristics of their cell of origin. BHK21 cells, HeLa cells, and SP2 cells are capable of efficient growth in suspension; most of the other cell lines require a solid culture substratum in order to multiply.

Cells Can Be Fused Together to Form Hybrid Cells

It is possible to fuse one cell with another to form a **heterocaryon**, a combined cell with two separate nuclei. Typically, a suspension of cells is treated with certain inactivated viruses or with polyethylene glycol, each of which alters the plasma membranes of cells in a way that induces them to fuse. Heterocaryons provide a way of mixing the components of two separate cells in order to study their interactions. The inert nucleus of a chicken red blood cell, for example, is reactivated to make RNA and eventually to replicate its DNA when it is exposed to the cytoplasm of a growing tissue-culture cell by fusion. The first direct evidence that membrane proteins are able to move in the plane of the plasma membrane (discussed in Chapter 10) came from an experiment in which mouse cells and human cells were fused: although the mouse and human cell-surface proteins were initially confined to their own halves of the heterocaryon plasma membrane, they quickly diffused and mixed over the entire surface of the cell.

Eventually, a heterocaryon proceeds to mitosis and produces a hybrid cell in which the two separate nuclear envelopes have been disassembled, allowing all the chromosomes to be brought together in a single large nucleus (Figure 8–5). Although such hybrid cells can be cloned to produce hybrid cell lines, the cells tend to lose chromosomes and are therefore genetically unstable. For unknown reasons, mouse–human hybrid cells predominantly lose human chromosomes. These chromosomes are lost at random, giving rise to a variety of mouse–human hybrid cell lines, each of which contains only one or a few human chromosomes. This phenomenon has been put to good use in mapping the locations of genes in the human genome: only hybrid cells containing human chromosome 11, for example, synthesize human insulin, indicating that the gene encoding insulin is located on chromosome 11. The same hybrid cells are also used as a source of human DNA for preparing chromosome-specific human DNA libraries.

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 27 of 53

TABLE 8–3 Some Landmarks in the Development of Tissue and Cell Culture

1885	**Roux** shows that embryonic chick cells can be maintained alive in a saline solution outside the animal body.
1907	**Harrison** cultivates amphibian spinal cord in a lymph clot, thereby demonstrating that axons are produced as extensions of single nerve cells.
1910	**Rous** induces a tumor by using a filtered extract of chicken tumor cells, later shown to contain an RNA virus (Rous sarcoma virus).
1913	**Carrel** shows that cells can grow for long periods in culture provided they are fed regularly under aseptic conditions.
1948	**Earle** and colleagues isolate single cells of the L cell line and show that they form clones of cells in tissue culture.
1952	**Gey** and colleagues establish a continuous line of cells derived from a human cervical carcinoma, which later become the well-known HeLa cell line.
1954	**Levi-Montalcini** and associates show that nerve growth factor (NGF) stimulates the growth of axons in tissue culture.
1955	**Eagle** makes the first systematic investigation of the essential nutritional requirements of cells in culture and finds that animal cells can propagate in a defined mixture of small molecules supplemented with a small proportion of serum proteins.
1956	**Puck** and associates select mutants with altered growth requirements from cultures of HeLa cells.
1958	**Temin and Rubin** develop a quantitative assay for the infection of chick cells in culture by purified Rous sarcoma virus. In the following decade the characteristics of this and other types of viral transformation are established by **Stoker, Dulbecco, Green,** and other virologists.
1961	**Hayflick and Moorhead** show that human fibroblasts die after a finite number of divisions in culture.
1964	**Littlefield** introduces HAT medium for the selective growth of somatic cell hybrids. Together with the technique of cell fusion, this makes somatic-cell genetics accessible.
	Kato and Takeuchi obtain a complete carrot plant from a single carrot root cell in tissue culture.
1965	**Ham** introduces a defined, serum-free medium able to support the clonal growth of certain mammalian cells.
	Harris and Watkins produce the first heterocaryons of mammalian cells by the virus-induced fusion of human and mouse cells.
1968	**Augusti-Tocco and Sato** adapt a mouse nerve cell tumor (neuroblastoma) to tissue culture and isolate clones that are electrically excitable and that extend nerve processes. A number of other differentiated cell lines are isolated at about this time, including skeletal muscle and liver cell lines.
1975	**Köhler and Milstein** produce the first monoclonal antibody-secreting hybridoma cell lines.
1976	**Sato** and associates publish the first of a series of papers showing that different cell lines require different mixtures of hormones and growth factors to grow in serum-free medium.
1977	**Wigler and Axel** and their associates develop an efficient method for introducing single-copy mammalian genes into cultured cells, adapting an earlier method developed by **Graham and van der Eb**.
1986	**Martin** and **Evans** and colleagues isolate and culture pluripotent embryonic stem cells from mouse.
1998	**Thomson** and **Gearhart** and their associates isolate human embryonic stem cells.

Hybridoma Cell Lines Provide a Permanent Source of Monoclonal Antibodies

In 1975 the development of a special type of hybrid cell line revolutionized the production of antibodies for use as tools in cell biology. The technique involves propagating a clone of cells from a single antibody-secreting B lymphocyte so that a homogeneous preparation of antibodies can be obtained in large quantities. The practical problem, however, is that B lymphocytes normally have a limited life-span in culture. To overcome this limitation, individual antibody-producing B lymphocytes from an immunized mouse or rat are fused with cells derived from an "immortal" B lymphocyte tumor. From the resulting heterogeneous mixture of hybrid cells, those hybrids that have both the ability to make a particular antibody and the ability to multiply indefinitely in culture are selected. These **hybridomas** are propagated as individual clones, each of which provides a permanent and stable source of a single type of **monoclonal antibody** (Figure 8–6). This antibody recognizes a single type of antigenic site—for example, a particular cluster of five or six amino acid side chains on the surface of a protein. Their uniform specificity makes monoclonal antibodies much more useful for most purposes than conventional antisera, which generally contain a mixture of

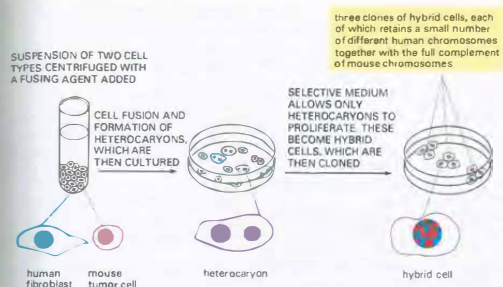


Figure 8–5 The production of hybrid cells. Human cells and mouse cells are fused to produce heterocaryons (each with two or more nuclei), which eventually form hybrid cells (each with one fused nucleus). These particular hybrid cells are useful for mapping human genes on specific human chromosomes because most of the human chromosomes are quickly lost in a random manner, leaving clones that retain only one or a few. The hybrid cells produced by fusing other types of cells often retain most of their chromosomes.

antibodies that recognize a variety of different antigenic sites on a macromolecule.

The most important advantage of the hybridoma technique is that monoclonal antibodies can be made against molecules that constitute only a minor component of a complex mixture. In an ordinary antiserum made against such a mixture, the proportion of antibody molecules that recognize the minor

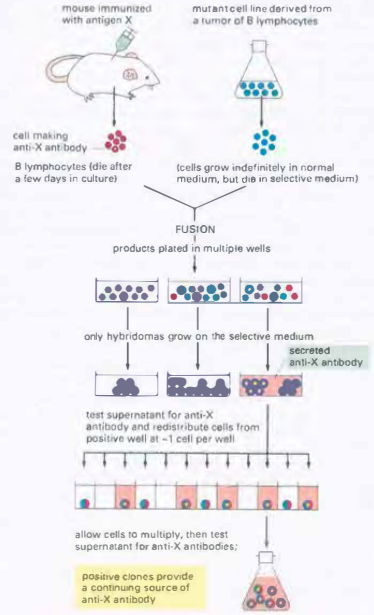


Figure 8–6 Preparation of hybridomas that secrete monoclonal antibodies against a particular antigen. Here the antigen of interest is designated as "antigen X." The selective growth medium used after the cell fusion step contains an inhibitor (aminopterin) that blocks the normal biosynthetic pathways by which nucleotides are made. The cells must therefore use a bypass pathway to synthesize their nucleic acids. This pathway is defective in the mutant cell line derived from the tumor, but it is intact in the cells obtained from the immunized mouse. Because neither cell type used for the initial fusion can grow on its own, only the hybrid cells survive

NORRIS MEDICAL LIBRARY
USC UNIVERSITY OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 28 of 53

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.

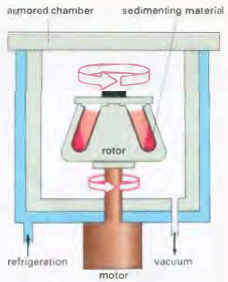
FRACTIONATION OF CELLS

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.

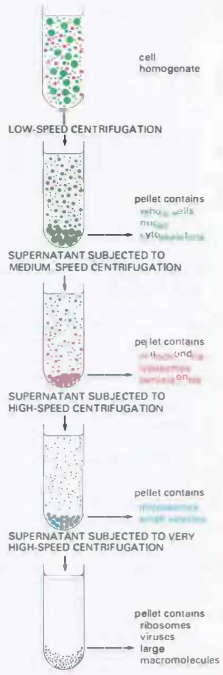
Centrifugation is the first step in most fractionations, but it separates only components that differ greatly in size. A finer degree of separation can be achieved by layering the homogenate in a thin band on top of a dilute salt solution that fills a centrifuge tube. When centrifuged, the various components 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



NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 29 of 53

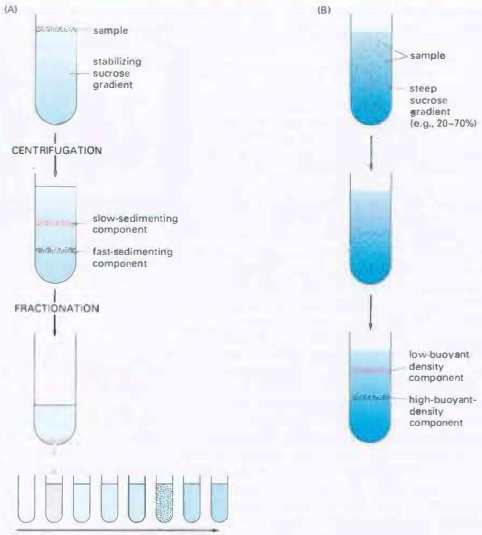


Figure 8–9 Comparison of velocity sedimentation and equilibrium sedimentation. In velocity sedimentation (A) subcellular components sediment at different speeds according to their size and shape when layered over a dilute solution containing sucrose. To stabilize the sedimenting bands against convective 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.

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 30 of 53

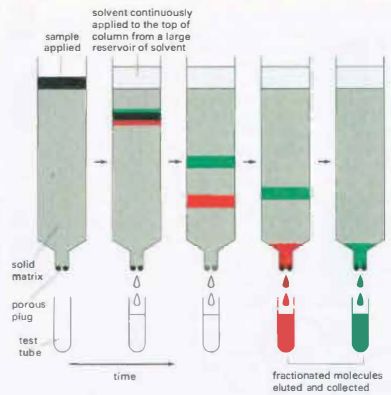


Figure 8–10 The separation of molecules by column chromatography. The sample, a mixture of different molecules, is applied to the top of a cylindrical glass or plastic column filled with a permeable solid matrix, such as cellulose, immersed in solvent. A large amount of solvent is then pumped slowly 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

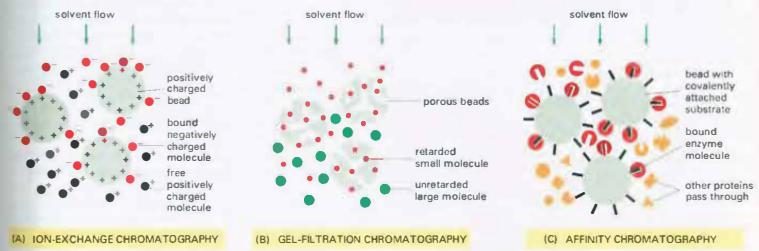


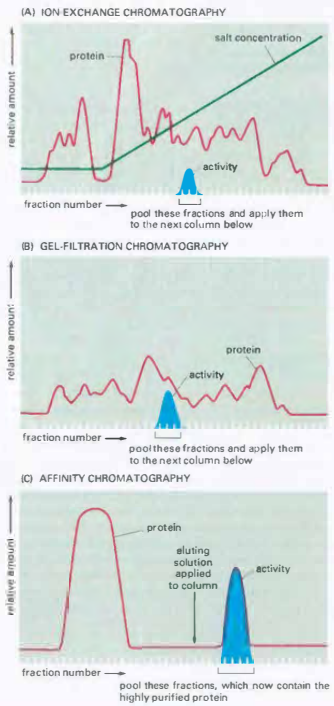
Figure 8–11 Three types of matrices used for chromatography. In ion-exchange chromatography (A) the insoluble matrix carries ionic charges that retard the movement of molecules of opposite charge. Matrices used for separating proteins include diethylaminoethylcellulose (DEAE-cellulose), which is positively 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.

NORRIS MEDICAL LIBRARY
UNIV OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 31 of 53

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.



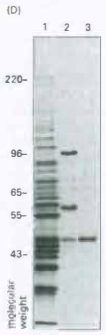


Figure 8–12 Protein purification by chromatography. Typical results obtained when three different chromatographic steps are used in succession to purify a protein. In this example a homogenate of cells was first fractionated by allowing it to percolate through an ion-exchange resin packed into a column (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.

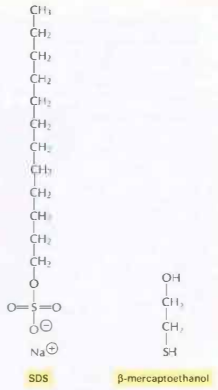


Figure 8–13 The detergent sodium dodecyl sulfate (SDS) and the reducing agent β-mercaptoethanol. These two chemicals are used to solubilize proteins for SDS polyacrylamide-gel electrophoresis. The SDS is shown here in its ionized form.

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

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 32 of 53

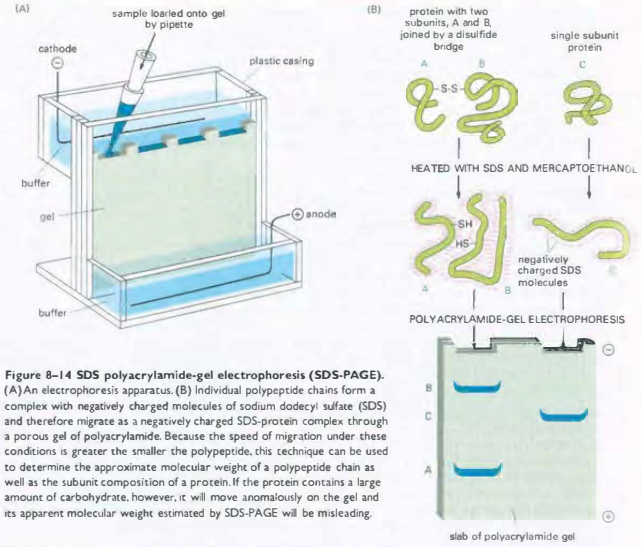


Figure 8–14 SDS polyacrylamide-gel electrophoresis (SDS-PAGE). (A) An electrophoresis apparatus. (B) Individual polypeptide chains form a complex with negatively charged molecules of sodium dodecyl sulfate (SDS) and therefore migrate as a negatively charged SDS-protein complex through a porous gel of 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.)

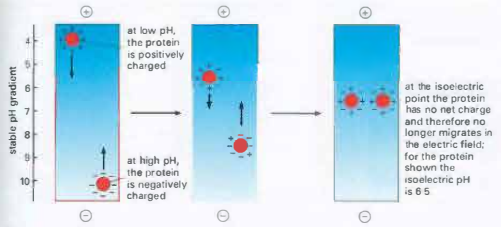


Figure 8–16 Separation of protein molecules by isoelectric focusing. At low *pH* (high H^+ concentration) the carboxylic acid groups of proteins tend to be uncharged (–COOH) and their nitrogen-containing basic groups fully charged (for example, $-NH_3^+$), giving most proteins a net positive charge. 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.)

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 33 of 53

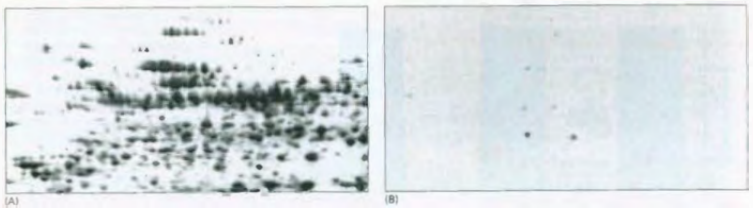


Figure 8–18 Western blotting. The total proteins from dividing tobacco cells in culture are first separated by two-dimensional polyacrylamide-gel electrophoresis and in (A) their positions are revealed by a sensitive protein stain. In (B) the separated proteins on an identical gel were then transferred 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

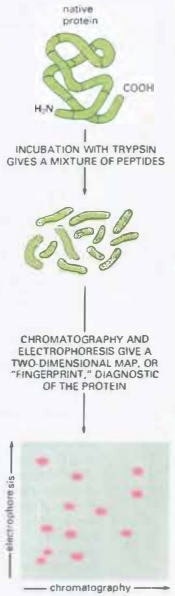


Figure 8–19 Production of a peptide map, or fingerprint, of a protein. Here, the protein was digested with trypsin to generate a mixture of polypeptide fragments, which was then fractionated in two dimensions by electrophoresis and partition chromatography. The latter technique separates peptides on 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.

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 34 of 53

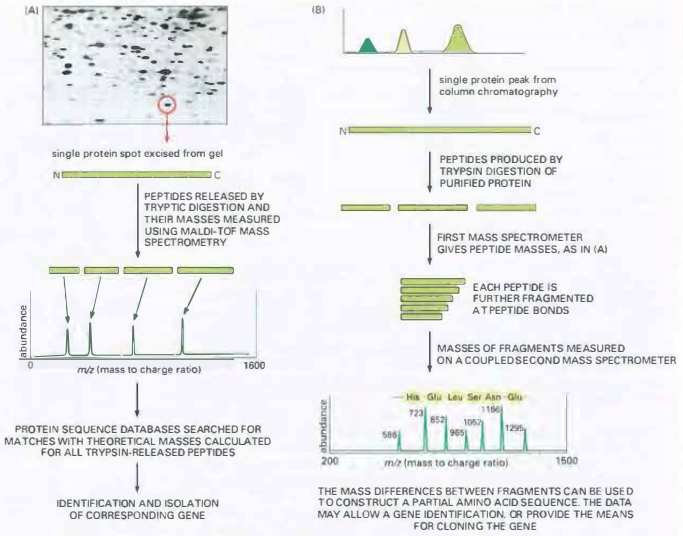


Figure 8–20 Mass-spectrometric approaches to identify proteins and sequence peptides. (A) Mass spectrometry can be used to identify proteins by determining their precise masses, and the masses of peptides derived from them, and using that information to search a genomic database for the corresponding 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,

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 35 of 53

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

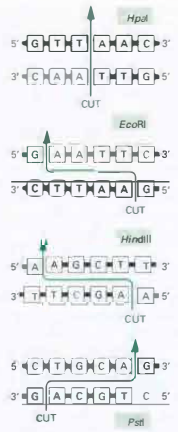
Recombinant DNA technology comprises a mixture of techniques, some new and some borrowed from other fields such as microbial genetics (Table 8–7). Central to the technology are the following key techniques:

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

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

1869	**Miescher** first isolates DNA from white blood cells harvested from pus-soaked bandages obtained from a nearby hospital.
1944	**Avery** provides evidence that DNA, rather than protein, carries the genetic information during bacterial transformation.
1953	**Watson and Crick** propose the double-helix model for DNA structure based on x-ray results of **Franklin and Wilkins**.
1955	**Kornberg** discovers DNA polymerase, the enzyme now used to produce labeled DNA probes.
1961	**Marmur and Doty** discover DNA renaturation, establishing the specificity and feasibility of nucleic acid hybridization reactions.
1962	**Arber** provides the first evidence for the existence of DNA restriction nucleases, leading to their purification and use in DNA sequence characterization by **Nathans and H. Smith**.
1966	**Nirenberg, Ochoa, and Khorana** elucidate the genetic code.
1967	**Gellert** discovers DNA ligase, the enzyme used to join DNA fragments together.
1972–1973	DNA cloning techniques are developed by the laboratories of **Boyer, Cohen, Berg**, and their colleagues at Stanford University and the University of California at San Francisco.
1975	**Southern** develops gel-transfer hybridization for the detection of specific DNA sequences.
1975–1977	**Sanger and Barrell and Maxam and Gilbert** develop rapid DNA-sequencing methods.
1981–1982	**Palmiter and Brinster** produce transgenic mice; **Spradling and Rubin** produce transgenic fruit flies.
1982	**GenBank**, NIH's public genetic sequence database, is established at Los Alamos National Laboratory.
1985	**Mullis** and co-workers invent the polymerase chain reaction (PCR).
1987	**Capecchi** and **Smithies** introduce methods for performing targeted gene replacement in mouse embryonic stem cells.
1989	**Fields and Song** develop the yeast two-hybrid system for identifying and studying protein interactions
1989	**Olson** and colleagues describe sequence-tagged sites, unique stretches of DNA that are used to make physical maps of human chromosomes.
1990	**Lipman** and colleagues release BLAST, an algorithm used to search for homology between DNA and protein sequences.
1990	**Simon** and colleagues study how to efficiently use bacterial artificial chromosomes, BACs, to carry large pieces of cloned human DNA for sequencing.
1991	**Hood and Hunkapillar** introduce new automated DNA sequence technology.
1995	**Venter** and colleagues sequence the first complete genome, that of the bacterium *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 *HpaI*, the cleavage leaves blunt ends; for others, such as *EcoRI*, *HindIII*, and *PstI*, the cleavage is staggered and creates cohesive ends. Restriction nucleases are obtained from various species of bacteria: *HpaI* is from *Hemophilus parainfluenzae*, *EcoRI* is from *Escherichia coli*, *HindIII* is from *Hemophilus influenzae*, and *PstI* is from *Providencia stuartii*.



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

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

Large DNA Molecules Are Cut into Fragments by Restriction Nucleases

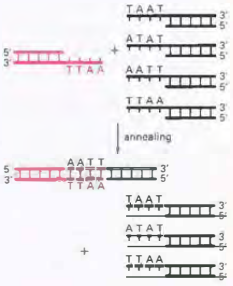
Unlike a protein, a gene does not exist as a discrete entity in cells, but rather as a small region of a much longer DNA molecule. Although the DNA molecules in a cell can be randomly broken into small pieces by mechanical force, a fragment containing a single gene in a mammalian genome would still be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

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

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

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

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



NORRIS MEDICAL LIBRARY
USC SCHOOL OF MEDICINE

Illumina Exhibit 1031, Page 36 of 53

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 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).

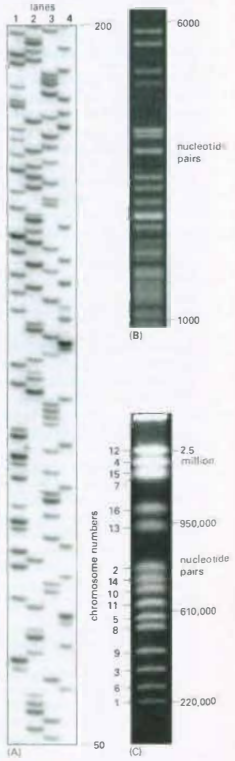


Figure 8–23 Gel electrophoresis techniques for separating DNA molecules by size. In the three examples shown, electrophoresis is from top to bottom, so that the largest—and thus slowest-moving—DNA molecules are near the top of the gel. In (A) a polyacrylamide gel with small pores is used to fractionate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

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

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 37 of 53

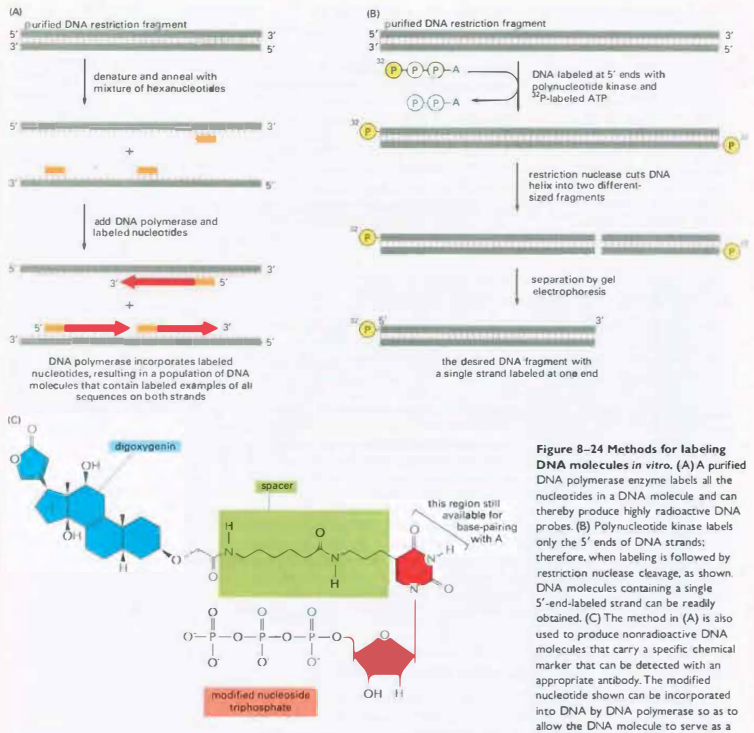


Figure 8–24 Methods for labeling DNA molecules *in vitro*. (A) A purified DNA polymerase enzyme labels all the nucleotides in a DNA molecule and can thereby produce highly radioactive DNA probes. (B) Polynucleotide kinase labels only the 5′ ends of DNA strands; therefore, when labeling is followed by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

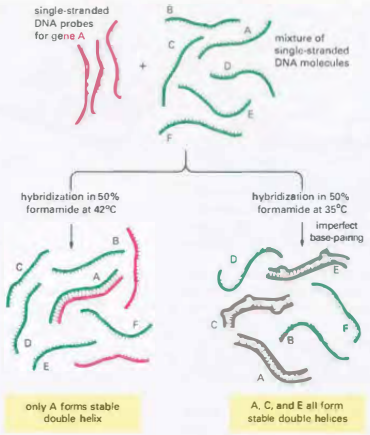


Figure 8–25 Different hybridization conditions allow less than perfect DNA matching. When only an identical match with a DNA probe is desired, the hybridization reaction is kept just a few degrees below the temperature at which a perfect DNA helix denatures in the solvent used (its *melting temperature*),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

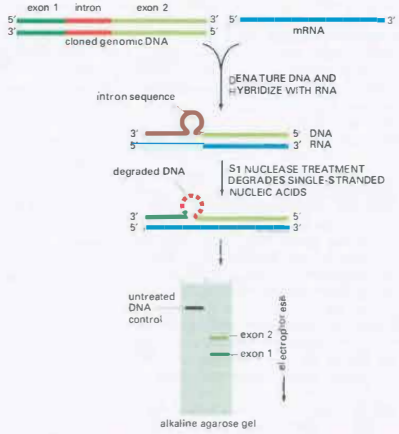


Figure 8–26 The use of nucleic acid hybridization to determine the region of a cloned DNA fragment that is present in an mRNA molecule. The method shown requires a nuclease that cuts the DNA chain only where it is not base-paired to a complementary RNA chain. The positions of the introns in eucaryotic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

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 38 of 53

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

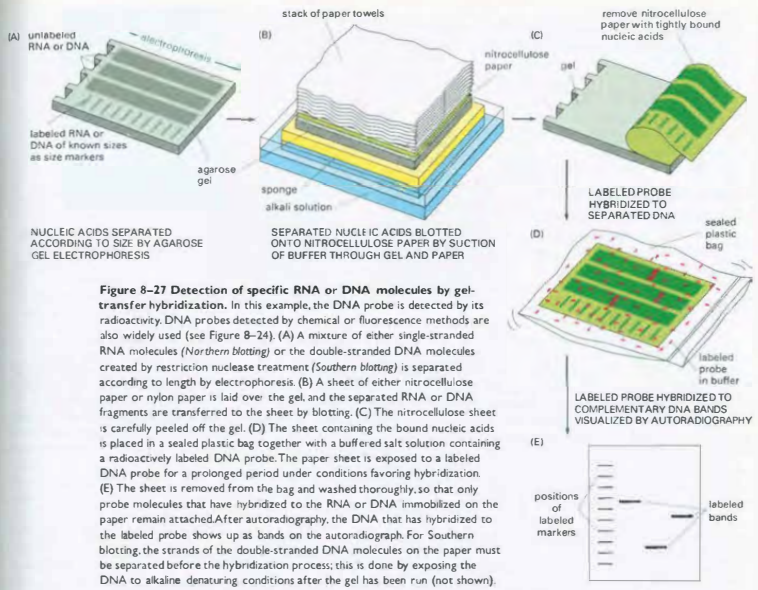


Figure 8–27 Detection of specific RNA or DNA molecules by gel-transfer hybridization. In this example, the DNA probe is detected by its radioactivity. DNA probes detected by chemical or fluorescence methods are also widely used (see Figure 8–24). (A) A mixture of either single-stranded RNA molecules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

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 39 of 53

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

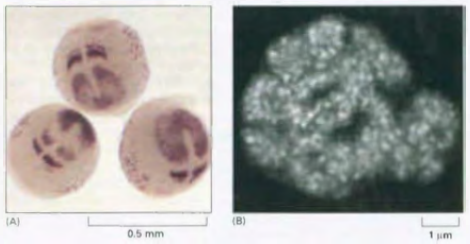




Figure 8–28 *In situ* hybridization to locate specific genes on chromosomes. Here, six different DNA probes have been used to mark the location of their respective nucleotide sequences on human chromosome 5 at metaphase. The probes have been chemically labeled and detected with fluorescent antibodies.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

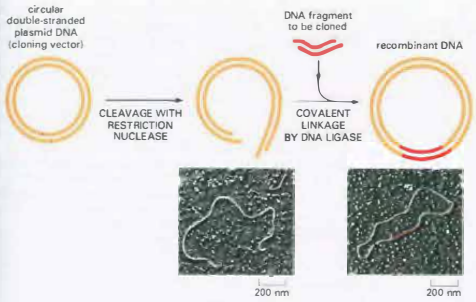


Figure 8–30 The insertion of a DNA fragment into a bacterial plasmid with the enzyme DNA ligase. The plasmid is cut open with a restriction nuclease (in this case one that produces cohesive ends) and is mixed with the DNA fragment to be cloned (which has been prepared with the same restriction nuclease),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

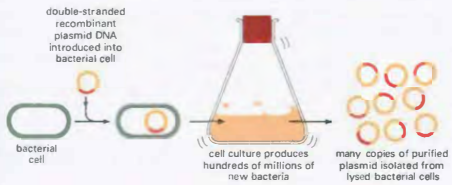


Figure 8–31 Purification and amplification of a specific DNA sequence by DNA cloning in a bacterium. To produce many copies of a particular DNA sequence, the fragment is first inserted into a plasmid vector, as shown in Figure 8–30. The resulting recombinant plasmid DNA is then introduced into a bacterium,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

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 40 of 53

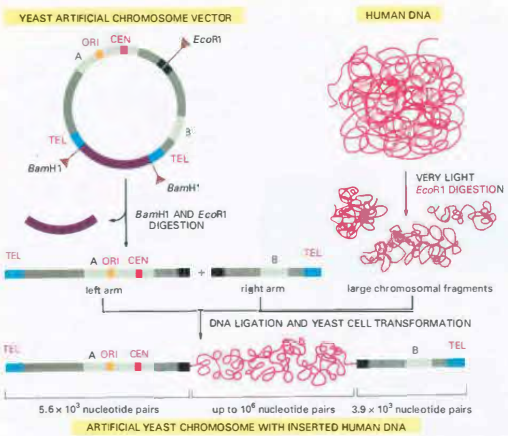


Figure 8–32 The making of a yeast artificial chromosome (YAC). A YAC vector allows the cloning of very large DNA molecules. TEL, CEN, and ORI are the telomere, centromere, and origin of replication sequences, respectively, for the yeast *Saccharomyces cerevisiae*. *Bam*H1 and *Eco*R1 are sites where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

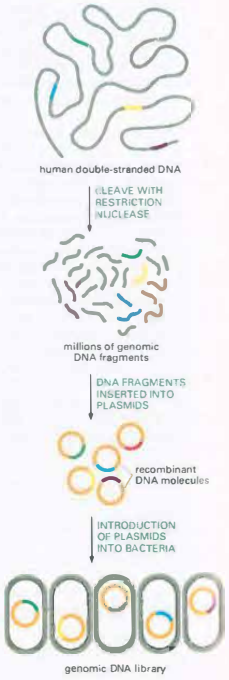


Figure 8–33 Construction of a human genomic DNA library. A genomic library is usually stored as a set of bacteria, each carrying a different fragment of human DNA. For simplicity, cloning of just a few representative fragments (*colored*) is shown. In reality, all of the *gray* DNA fragments would also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

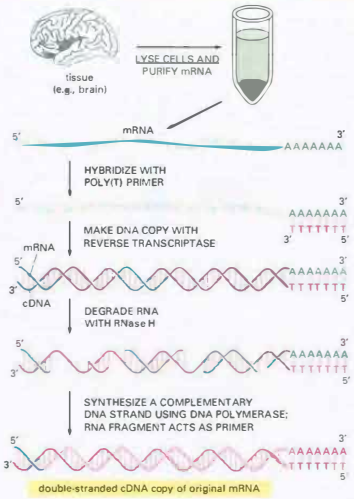


Figure 8–34 The synthesis of cDNA. Total mRNA is extracted from a particular tissue, and DNA copies (cDNA) of the mRNA molecules are produced by the enzyme reverse transcriptase (see p. 289). For simplicity, the copying of just one of these mRNAs into cDNA is illustrated. A short oligonucleotide complementary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

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 41 of 53

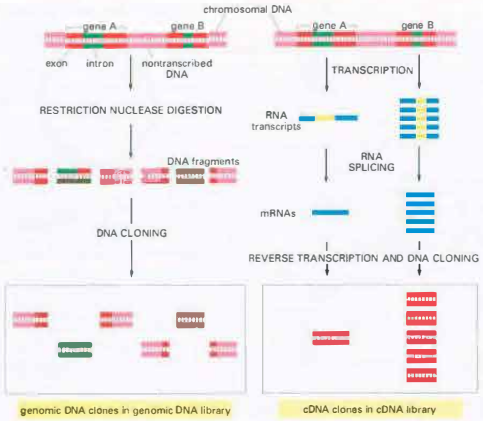


Figure 8–35 The differences between cDNA clones and genomic DNA clones derived from the same region of DNA. In this example gene A is infrequently transcribed, whereas gene B is frequently transcribed, and both genes contain introns *(green)*. In the genomic DNA library, both the introns and the nontranscribed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

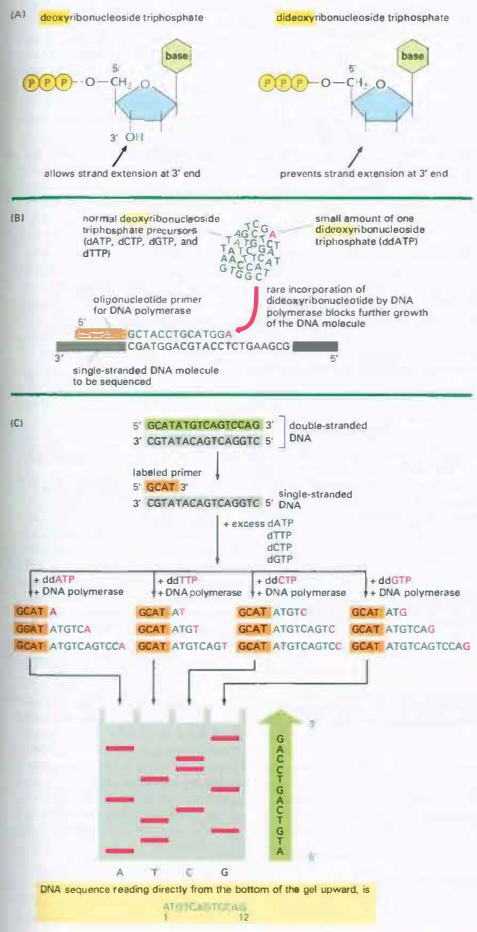


Figure 8–36 The enzymatic—or dideoxy—method of sequencing DNA. (A) This method relies on the use of dideoxyribonucleoside triphosphates, derivatives of the normal deoxyribonucleoside triphosphates that lack the 3′ hydroxyl group. (B) Purified DNA is synthesized *in vitro* in a mixture that contains single-stranded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

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 42 of 53

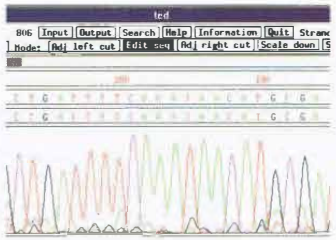


Figure 8–37 Automated DNA sequencing. Shown here is a tiny part of the data from an automated DNA-sequencing run as it appears on the computer screen. Each colored peak represents a nucleotide in the DNA sequence—a clear stretch of nucleotide sequence can be read here between positions 173 and 194 from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

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 43 of 53

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

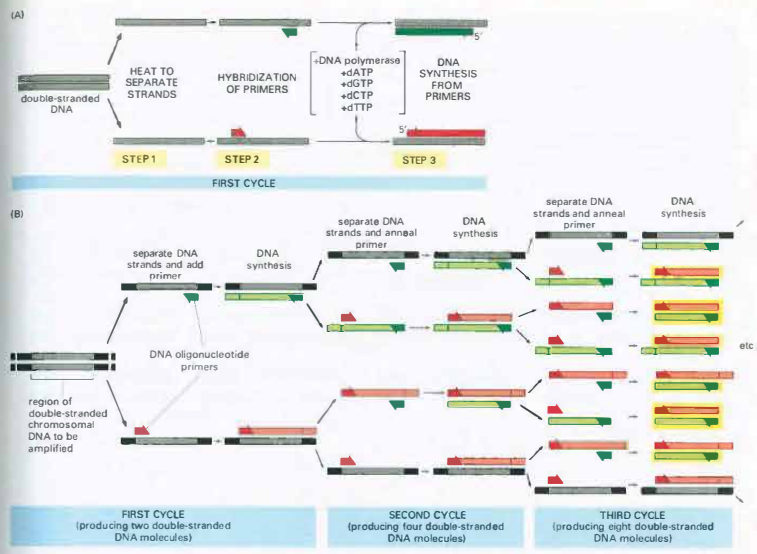


Figure 8–39 Amplification of DNA using the PCR technique. Knowledge of the DNA sequence to be amplified is used to design two synthetic DNA oligonucleotides, each complementary to the sequence on one strand of the DNA double helix at opposite ends of the region to be amplified. These oligonucleotides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

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 44 of 53

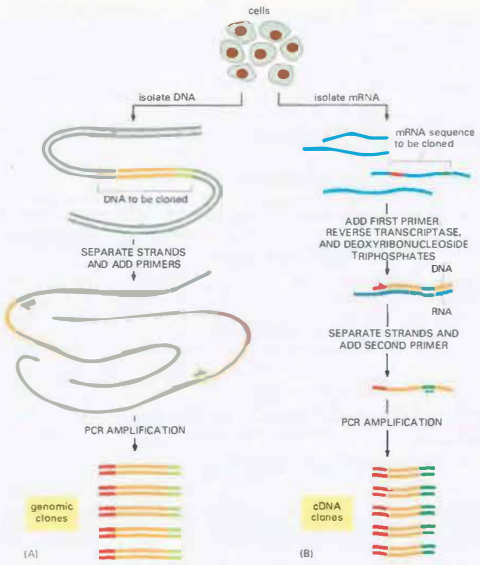


Figure 8–40 Use of PCR to obtain a genomic or cDNA clone. (A) To obtain a genomic clone by using PCR, chromosomal DNA is first purified from cells. PCR primers that flank the stretch of DNA to be cloned are added, and many cycles of the reaction are completed (see Figure 8–39). Since only the DNA between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

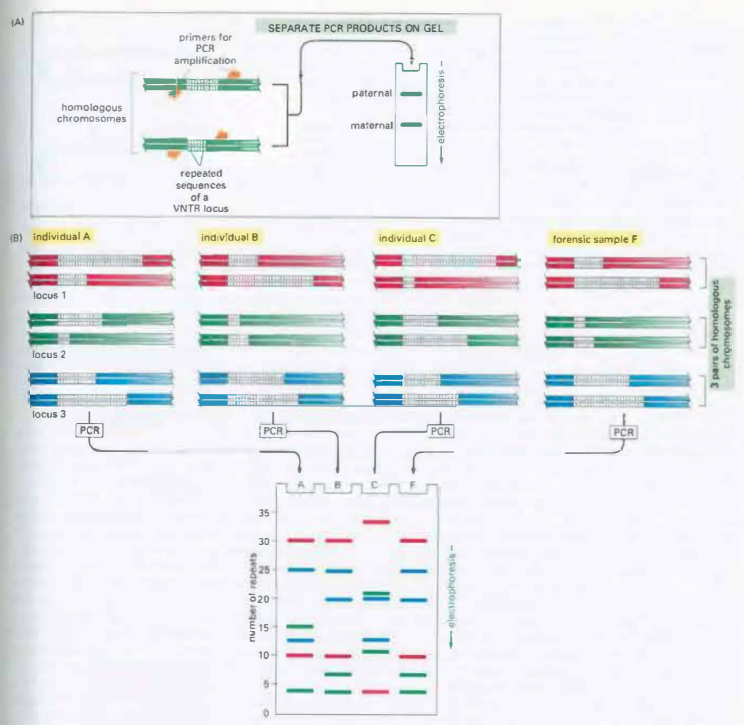


Figure 8–41 How PCR is used in forensic science. (A) The DNA sequences that create the variability used in this analysis contain runs of short, repeated sequences, such as CACACA . . . , which are found in various positions (loci) in the human genome. The number of repeats in each run can be highly variable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

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 45 of 53

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

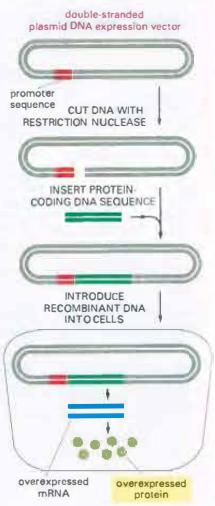
proteins—including hemoglobin, trypsin, immunoglobulin, and lysozyme—were analyzed in this way.

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

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

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

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



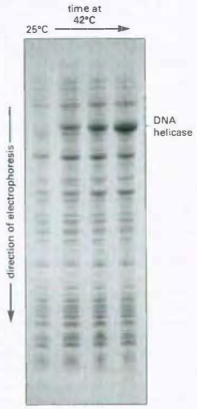


Figure 8–43 Production of large amounts of a protein by using a plasmid expression vector. In this example, bacterial cells have been transfected with the coding sequence for an enzyme, DNA helicase; transcription from this coding sequence is under the control of a viral promoter that becomes active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

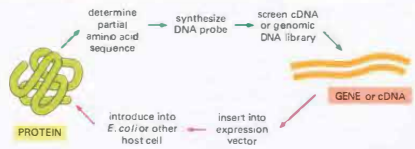


Figure 8–44 Knowledge of the molecular biology of cells makes it possible to experimentally move from gene to protein and from protein to gene. A small quantity of a purified protein is used to obtain a partial amino acid sequence. This provides sequence information that enables the corresponding gene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

Summary

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

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

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

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

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

ANALYZING PROTEIN STRUCTURE AND FUNCTION

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

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 46 of 53

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

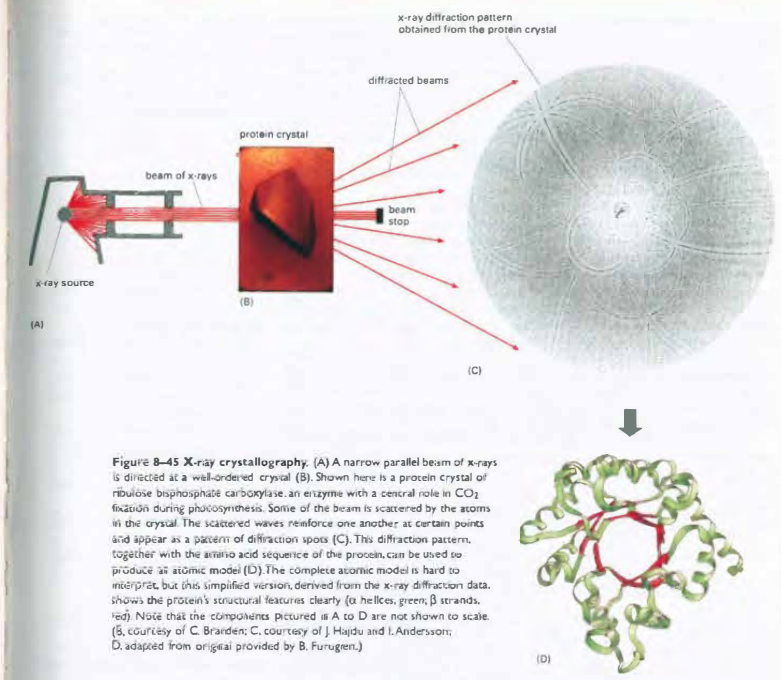


Figure 8–45 X-ray crystallography. (A) A narrow parallel beam of x-rays is directed at a well-ordered crystal (B). Shown here is a protein crystal of ribulose bisphosphate carboxylase, an enzyme with a central role in CO_2 fixation during photosynthesis. Some of the beam is scattered by the atoms in 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

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 47 of 53

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

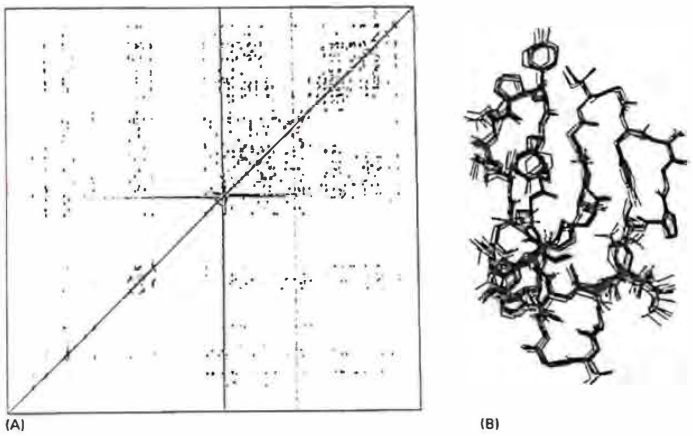


Figure 8–46 NMR spectroscopy. (A) An example of the data from an NMR machine. This two-dimensional NMR spectrum is derived from the C-terminal domain of the enzyme cellulase. The spots represent interactions between hydrogen atoms that are near neighbors in the protein and hence reflects the distance 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.)

Illumina Exhibit 1031, Page 48 of 53

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 p on either a

Illumina Exhibit 1031, Page 49 of 53

```
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  PNIVKLEDVIHTENKLYLVFEFLHQDLKKFMDASALTGIPLPLIKSYLFQLLQGLAFCHS 176
             NIV+L DV+H+E ++YLVFE+L  DLKKFMD+          LIKSYL+Q+L G+A+CHS
Sbjct: 61   GNIVRLEDVVHSEKRLYLVFEYLDLDLKKFMDSCPEFAKNPTLIKSYLYQILHGVAYCHS 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.

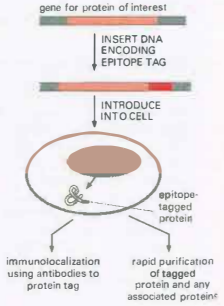


Figure 8–48 Epitope tagging allows the localization or purification of proteins. Using standard genetic engineering techniques, a short epitope tag can be added to a protein of interest. The resulting protein contains the protein being analyzed plus a short peptide that can be recognized by commercially 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.

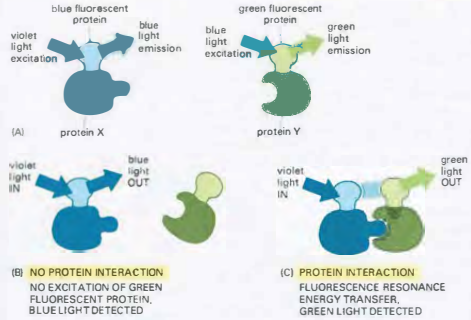


Figure 8–49 Fluorescence resonance energy transfer (FRET). To determine whether (and when) two proteins interact inside the cell, the proteins are first produced as fusion proteins attached to different variants of GFP. (A) In this example, protein X is coupled to a blue fluorescent protein, which is 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).

NORRIS MEDICAL LIBRARY

Illumina Exhibit 1031, Page 50 of 53

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.

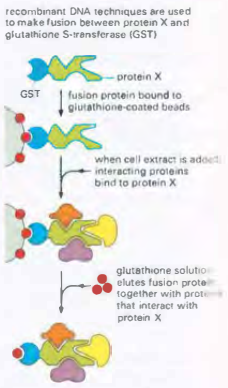


Figure 8–50 Purification of protein complexes using a GST-tagged fusion protein. GST fusion proteins, generated by standard recombinant DNA techniques, can be captured on an affinity column containing beads coated with glutathione. To look for proteins that bind to protein X, cell extracts can be passed 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).

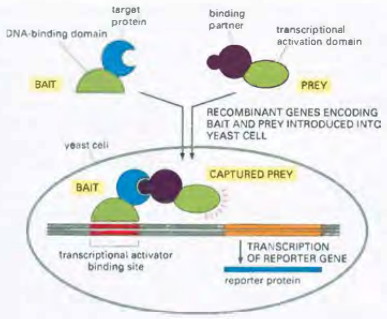


Figure 8–51 The yeast two-hybrid system for detecting protein–protein interactions. The target protein is fused to a DNA-binding domain that localizes it to the regulatory region of a reporter gene as "bait." When this target protein binds to another specially designed protein in the cell nucleus ("prey"), 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

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 51 of 53

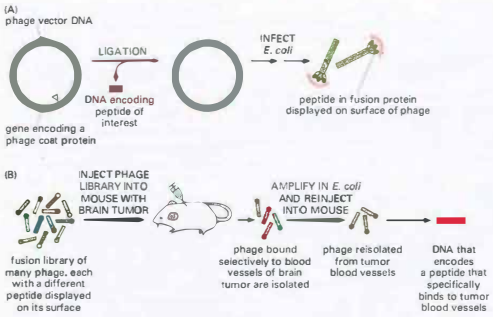


Figure 8–52 The phage display method for investigating protein interactions. (A) Preparation of the bacteriophage. DNA encoding the desired peptide is ligated into the phage vector, fused with the gene encoding the viral protein coat. The engineered phage are then introduced into *E. coli*, which produce 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.

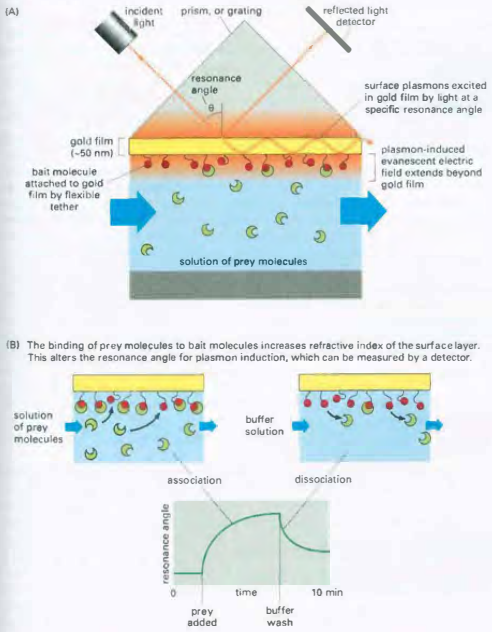


Figure 8–53 Surface plasmon resonance. (A) SPR can detect binding interactions by monitoring the reflection of a beam of light off the interface between an aqueous solution of potential binding molecules *(green)* and a biosensor surface coated with an immobilized bait protein *(red)*. (B) A solution 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.

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 52 of 53

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.

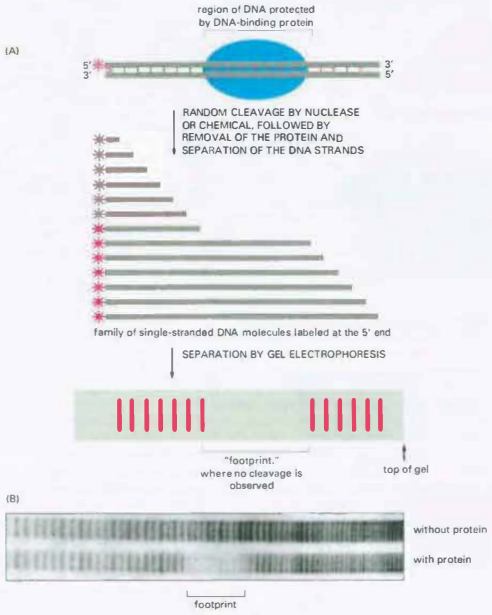


Figure 8–54 The DNA footprinting technique. (A) This technique requires a DNA molecule that has been labeled at one end (see Figure 8–24B). The protein shown binds tightly to a specific DNA sequence that is seven nucleotides long, thereby protecting these seven nucleotides from the cleaving agent. If 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.

NORRIS MEDICAL LIBRARY
UNIV. OF SOUTHERN CALIFORNIA

Illumina Exhibit 1031, Page 53 of 53