Tools and techniques in molecular biology

INTRODUCTION

Scientific discovery is driven by individuals who are able to look at a problem in a new way and have the right tools at their disposal. The Experimental approach panels throughout this book describe the process of discovery that has illuminated our understanding of some of the molecular processes and components described in the preceding chapters. Here we describe the tools and experimental methods that are used by biologists today in their continuing quest to understand how molecules in each cell carry out their many amazing functions.

Molecular biology became defined as a discipline around the middle of the twentieth century, when genetics and biochemistry were converging to provide an explanation for how genetic information was stored and harnessed by the cell. Although the use of model organisms in genetics was long established, we have since learned that many fundamental processes are so highly conserved from single-celled bacteria to multicellular eukaryotes that the study of these model organisms yields information relevant to all domains of life. With the development of techniques to manipulate genetic material and reintroduce it into cells, it has become possible to study biochemical pathways as never before.

The term "molecular biology" was coined to place emphasis on the molecules that carry out biological processes. To understand how something takes place in the cell, one must identify the molecules that participate in the process and determine their functions. A powerful approach to determining the components required for a given process, as well as the function of each macromolecule, is to replicate the process with purified components ***in vitro***, meaning in an artificial vessel (literally "within the glass"). An mRNA copy of a gene, for example, can be readily synthesized in the test tube by mixing the DNA with purified RNA polymerase, nucleoside triphosphates (NTPs), magnesium, and the appropriate buffers. The fact that RNA synthesis will now occur without any other cellular components present tells us that the molecules in the test tube mixture comprise the minimal requirements for RNA synthesis in the cell. *In vitro* assays also make possible quantitative studies of binding constants, reaction rates, and the way in which particular mutations affect one or more steps in the process.

However, *in vitro* studies have their limitations; when a particular process is reconstituted in the test tube, some of the proteins that help to regulate the process in the cell might still be missing, or an important post-translational modification might not be found in a recombinant protein used in the experiment. Despite these caveats, much of what is known in molecular biology derives from *in vitro* experiments.

The advent of whole genome sequencing and the resulting explosion of information about the nucleic acids and proteins in each organism have made possible new types of

experiments that exploit both cutting-edge techniques and the availability of vast amounts of information from thousands of organisms. The process of discovery has also been aided by other technological advances, including the use of robotics, the printing of slides containing thousands of different DNA or protein molecules, and the improvement in microscopy-based techniques. These methods have greatly increased the pace of discovery and have yielded profound insights into the inner workings of the cell. It is hard to believe that DNA was first shown to be the carrier of genetic information in cells less than 75 years ago.

We begin this chapter by describing the organisms that have served as models for the study of biology in all types of cells and explore how cells and viruses can be grown in culture. From there, we explore the amplification and cloning of nucleic acids and how such cloning, known as **recombinant DNA** technology, can be used both to identify new genes and to construct modified genes and chromosomes. We then present a large array of techniques used to study individual molecules, beginning with their purification from whole cells to modern methods for identifying particular proteins and nucleic acids. This is followed by a discussion of how we study interactions among molecules. The chapter concludes with a discussion of the techniques used to image whole cells and study processes ***in vivo*** (in the living organism) on the one hand, and individual macromolecules on the other hand. The application of all of these techniques in combination with new approaches yet to be developed will continue to expand our understanding of genome function.

19.1 MODEL ORGANISMS

Although millions of species exist in nature, only a few dozen, called model organisms, are extensively used in research. Many model organisms are studied not for their intrinsic importance to the economy or because they cause disease, but because they are easy to propagate and yet still can provide general insights into key molecular processes and components—insights which can be extrapolated to other organisms, including humans, and the diseases they face. This is because many molecules and molecular processes are evolutionarily conserved. For example, *Arabidopsis* is not a horticulturally important flowering plant and *Caenorhabditis elegans* is not an agriculturally or medically important roundworm, but both organisms serve as laboratory models for the understanding of basic biological processes. Likewise, most strains of *Escherichia coli* are not human pathogens, yet they are similar in many ways to bacteria that are.

Model organisms have certain shared properties

The most widely used model organisms share several biological properties that have made them the standards for investigation. A model organism must have fairly simple and, ideally, very defined nutritional requirements with a well-understood life cycle that allows it to be easily and safely grown in large quantities in the laboratory. A short generation time, small size, and the ability to store or maintain stocks of the organisms for years are also advantages.

Nearly all important model organisms have been extensively studied by genetic analysis. This means that mutations which reduce or alter the function of a gene can be readily introduced, propagated, identified, and characterized. Since the advent of recombinant DNA technology, a further requirement is the ability to introduce modified genes into the organism, so that the effect of any type of mutation can be studied. The genomes of model organisms have been sequenced, along with

many related species, and the communities of investigators studying these organisms have established databases containing genome sequence information, descriptions of mutant strains and their phenotypes, and the latest information about gene expression, enzyme activity, and protein–protein interactions. The availability of these data has greatly accelerated the pace of discovery and made possible new approaches to studying biology, as we will see throughout this chapter.

It is important to note that the nomenclature for gene and protein names varies from organism to organism. Refer back to page ii for a summary of the nomenclature rules adopted for some of the main model organisms discussed in this book.

Different model organisms have distinct advantages

Model organisms come from different kingdoms and phyla, and represent different evolutionary and biological properties, as depicted in Figure 19.1. Well-studied models include bacteria, single-celled eukaryotes, small animals, flowering plants, and lower vertebrates and mammals. How, then, does a scientist choose an experimental model in which to study a biological process? The answer depends on the question being asked, as well as the experimental tools that are available at the time. For example, much of our knowledge about DNA replication, transcription, and translation was originally gained from studies of bacteria. In part, this is because bacteria were among the first model organisms available, since it is easy to obtain large quantities of bacterial cells and they are easy to mutate. We now know that many aspects of the processes characterized in bacteria are universal to all organisms, validating the use of this model.

Phenomena that are unique to eukaryotes must, of course, be studied in eukaryotic cells. The inner workings of the nucleus of a cell can be readily studied in single-celled eukaryotes such as yeast, while the mechanisms by which a multicellular organism develops from a fertilized egg must be studied in a higher eukaryote, such as the fruit fly or the nematode. Similarly, phenomena unique to

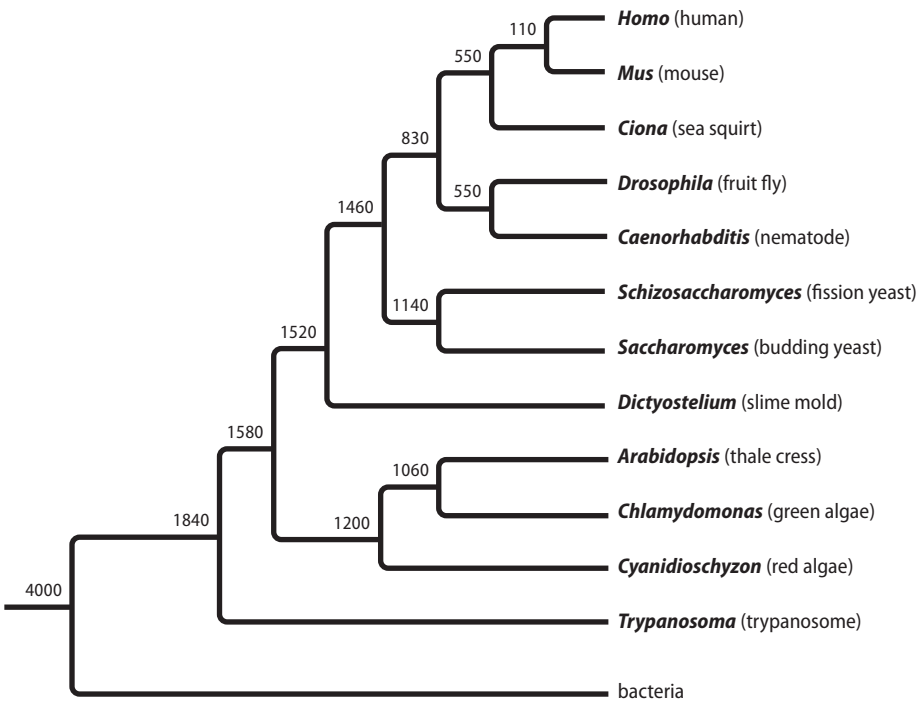


Figure 19.1 Phylogenetic tree of model organisms. The time of divergence, in millions of years, is indicated at the nodes of the tree (note that branch length is not proportional to time). For each organism, both the scientific name (italicized) and the common name are provided.

Adapted from Hedges SB, *Nature Reviews Genetics*, 2002;**3**:838–849.

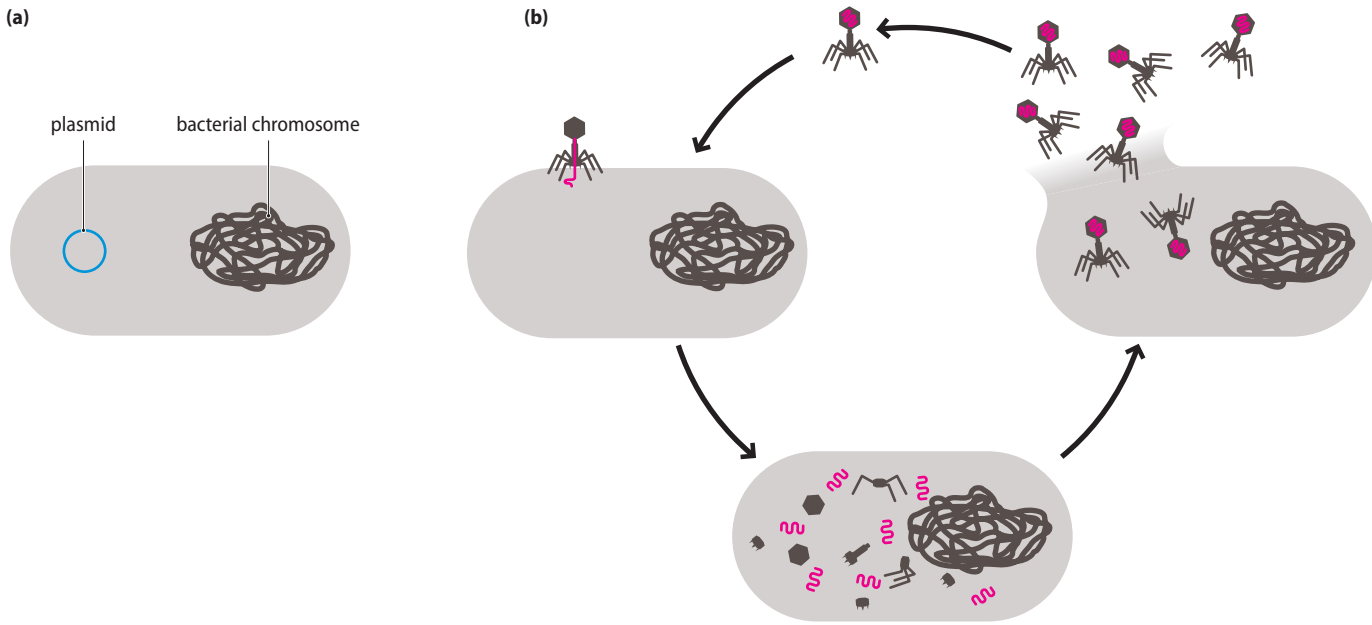


Figure 19.2 Bacterial cell with plasmid and phage. (a) Bacterial cell harboring a plasmid (in blue), a circular double-stranded DNA molecule that is separate from the bacterial chromosome. Plasmids are typically much smaller than the chromosome. (b) The life cycle of bacteriophage T2. A T2 bacteriophage infects an *E. coli* bacterium by injecting its DNA (pink) into the bacterium. The bacteriophage genes are expressed in the bacterium, generating the various bacteriophage proteins and allowing the bacteriophage to make multiple copies of itself. Finally, the new bacteriophage particles are released by cell lysis, after which they can infect new cells.

vertebrates or to mammals must be studied in a representative of that class of organisms. Often, the choice of the appropriate model organism has been instrumental in making rapid experimental progress. The organisms mentioned in the following text are just a selected subset of the important and informative model organisms that have been used.

E. coli, *Caulobacter crescentus*, and *Bacillus subtilis* are commonly studied bacteria

The gut bacterium *E. coli* is the best studied bacterium and quite possibly the best studied species. Many of the basic principles of molecular biology that are described throughout this book were first investigated in *E. coli*. This bacterium can be grown easily in the laboratory and divides about every 20–30 minutes under optimal conditions. Importantly, genes can be introduced into *E. coli* using extrachromosomal DNA elements known as **plasmids**, which are illustrated in Figure 19.2a. These double-stranded circular DNA molecules typically contain a DNA sequence that allows them to replicate and hence be stably maintained inside the bacterium, as well as a marker gene that allows selection for cells carrying the plasmid. A plasmid can harbor multiple inserted genes and be readily introduced into bacterial cells. Plasmids can also be easily manipulated by genetic engineering, making it straightforward to introduce mutant genes into bacteria and subsequently to study their effects. As we shall see later, plasmids can be used in eukaryotes as well.

E. coli cells can also be infected by a variety of viruses such as bacteriophage T2 and lambda, as depicted in Figure 19.2b. These bacteriophages provide another means by which to introduce genes into the cell. The dependence of bacteriophages on host cell function for their propagation has helped to provide insights into basic cellular processes.

Many other bacteria have interesting or important biological properties that make them useful as model organisms. *B. subtilis* can form stress-resistant spores, and its sporulation program is a thoroughly studied example of a developmental process. Sporulation involves sequential transcriptional activation of certain key genes, including changes in the sigma factor subunit of RNA polymerase. Thus, *B. subtilis* has given us insights into the regulation of transcription initiation.

The bacterium *C. crescentus* is another example of a bacterium with an interesting developmental process—it can adopt either of two different morphological forms. One form is the highly motile swarming cell that moves by means of a flagellum. The other form is a stalk cell, a stationary cell with a very different morphology. The switch between swarming and stalk formation is an example of a bi-stable biological switch. As we have seen in Chapters 5 and 7, *Caulobacter* has also been used to study bacterial cell division.

Saccharomyces cerevisiae, *Schizosaccharomyces pombe*, and *Dictyostelium discoideum* are commonly studied single-celled eukaryotes

Despite the enormous differences evident between single-celled eukaryotes and highly complex eukaryotes, such as humans, most fundamental pathways are remarkably conserved across all eukaryotes. Single-celled eukaryotes are therefore very attractive model systems for the study of eukaryotic molecular biology because of the ease with which they can be grown in culture and genetically manipulated. In addition, their short doubling time—about 90 minutes per generation for budding yeast—greatly enhances the rate at which studies of the effects of mutations can be performed.

The budding yeast *S. cerevisiae*, which is of widespread practical use in brewing, baking, and the production of biofuels, is also the best understood model for a eukaryotic cell. Budding yeast has been a key model organism for understanding the fundamental processes of transcription, meiosis, recombination, and cell division, among others. Yeast has a fairly compact genome, with about 6000 genes in a genome of 12 million base pairs (Mbp). This type of yeast is distinct from other eukaryotes, and even from some other yeasts, in its manner of cell division—rather than simply dividing into two cells of equal size, budding yeast, as its name implies, reproduces by budding, in which the daughter cell first appears as a small bud on the mother cell, as depicted in a microscopic image of budding yeast in Figure 19.3a and a diagram in Figure 19.3b. This small bud eventually grows into a daughter cell that separates from the mother cell and generates a bud of its own. Each *S. cerevisiae* cell can bud 20–30 times, depending on the strain.

The budding yeast life cycle, which is illustrated in Figure 19.4, lends itself to rapid genetic manipulation. As we learned in Chapter 8, this yeast can exist in both haploid and diploid forms. There are two haploid yeast cell types: **a** cells and α cells. A haploid yeast cell can only mate with a cell of opposite mating type to produce an **a**/α diploid cell. Both haploid and diploid cells can be propagated indefinitely, but under conditions of starvation, diploid cells undergo meiosis to form four haploid cells: two **a** cells and two α cells.

The ability to work with a haploid organism makes the analysis of mutant phenotypes much easier, as only one gene copy has to be mutated (unlike the situation in diploid cells, where both gene copies have to be mutated for recessive phenotypes to emerge). Moreover, different mutations can be combined by mating, followed by meiosis and sporulation. Finally, homologous recombination in budding yeast is extremely efficient, allowing the generation of chromosomal

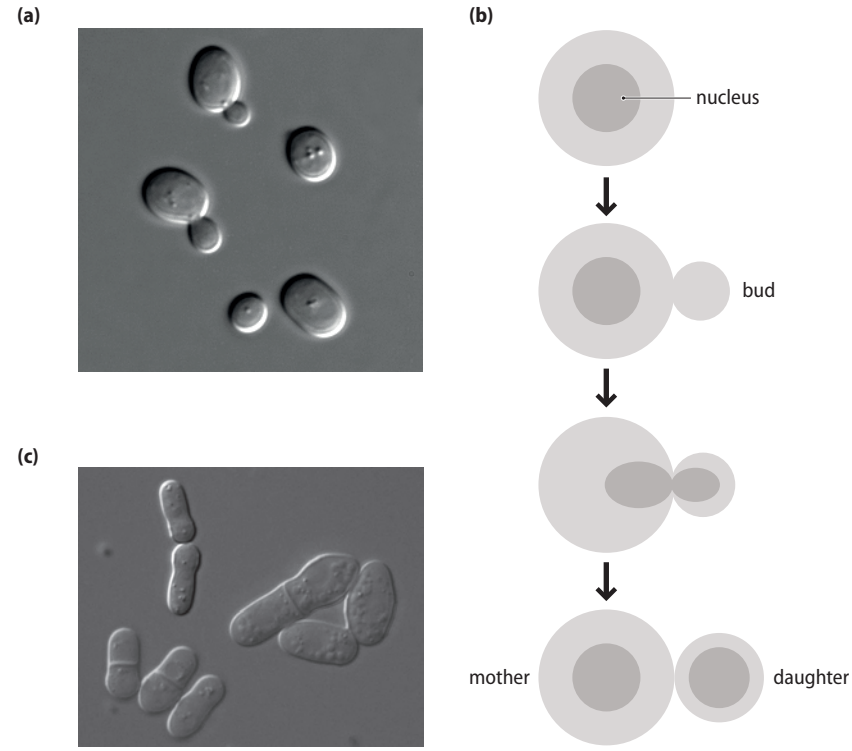


Figure 19.3 Budding and fission yeasts. (a) A microscopy image of budding yeast. (b) Diagram of the asymmetrical division of budding yeast. (c) A microscopy image of fission yeast.

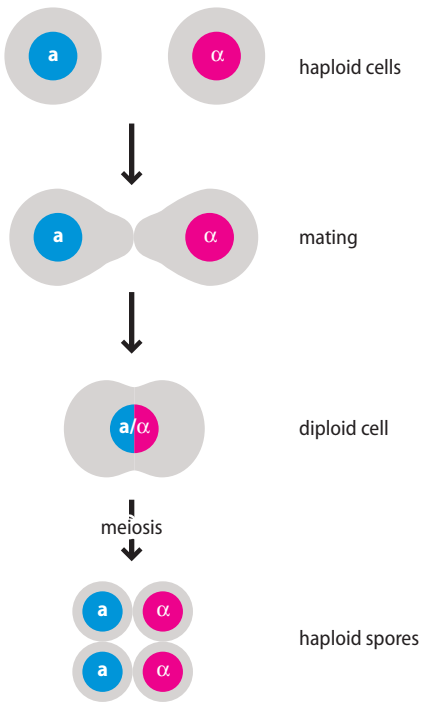


Figure 19.4 Life cycle of budding yeast. Yeast can exist as either haploid or diploid cells. Two haploid cells of opposite mating type, **a** and α, can mate with each other and fuse to form an **a**/α diploid cell. Under conditions of starvation, the diploid cell undergoes meiosis to produce four haploid spores: two **a** and two α.

gene fusions and gene deletions with relative ease. For this reason, yeast collections with all known protein-coding genes fused to a useful tag such as the green fluorescent protein (GFP), or where each of the known non-essential genes is deleted, are commercially available.

Some phenomena common to higher eukaryotes that are absent from budding yeast can be studied in its evolutionarily distant cousin *S. pombe*. This yeast reproduces by symmetrical cell division and is therefore called a fission yeast. A microscopy image of a fission yeast is shown in Figure 19.3c. Like budding yeast, fission yeast can be either haploid or diploid. However, DNA silencing and DNA structures of the centromeres in fission yeast are more similar to those found in multicellular eukaryotes than those of budding yeast. Moreover, because this yeast divides by fission, it has been used extensively in studies to elucidate the basis of spatial cues, such as the cell's middle. The similarities and contrasts between the two yeasts are often used to provide a broader perspective on the fundamental properties of eukaryotic cells.

Most single-celled yeasts lack extensive extracellular signaling or multicellular organization. In contrast, the slime mold *D. discoideum* is often used to provide insights into such intercellular communication and signaling. *Dictyostelium* cells live as motile solitary cells for much of their life cycle. However, in response to a secreted signal from one cell, the nearby cells stream together to form a colony of functionally differentiated cells called a slug. The slug moves by the coordinated action of individual cells and can attach itself to a surface, whereupon it differentiates into a complex cellular structure called the fruiting body. Figure 19.5 illustrates how the fruiting body is composed of a stalk and a structure containing spores. The secretion of the signal and the response of the cells to a signaling **gradient** for slug and stalk formation have been important in understanding the role of calcium in extracellular signaling.

Drosophila melanogaster, *C. elegans*, and *Arabidopsis thaliana* are commonly studied multicellular eukaryotes

Studies of single-celled eukaryotes, while tremendously important for understanding the inner workings of all eukaryotic cells, provide limited insights into how multicellular organisms develop from a single fertilized egg into an organism with distinct tissue types and morphology. The common fruit fly *D. melanogaster* has been one of the most important experimental systems for studying development since it was first established as a model organism over 100 years ago. Indeed, *Drosophila* mutants, such as the one shown in Figure 19.6a, have been instrumental in studying almost every process in animals.

The entire development cycle in *Drosophila* is very rapid, with a fertilized egg taking just about a week to develop into a mature fly. A single female can lay thousands of eggs, thereby producing an enormous number of progeny. Random mutations can be induced by feeding flies chemicals that induce mutations or by subjecting flies to ionizing radiation, after which the many progeny can be screened for new and interesting phenotypes. Indeed, modern recombinant DNA techniques make it possible to engineer specific mutations into flies. By analyzing the effects of mutations on the mature organism, as well as on the different stages of development, genes that control development and differentiation, as well as behavior, can be identified. A number of genes that control development in all metazoans were first identified in the fruit fly.

A particularly interesting feature of the fly is its salivary glands. In order to produce large amounts of the material secreted by these glands, cells of the salivary gland undergo multiple rounds of DNA replication without undergoing mitosis. This results in the formation of the polytene chromosomes shown in Figure 19.6b, which have been useful in studying chromosome structure and gene expression.

The nematode (or roundworm) *C. elegans*, which has a life cycle of about three days, was developed as a genetic model organism because it has many fewer cells than other multicellular animals despite carrying out complex biological processes and behaviors. An adult worm has only 909 cells (excluding the germline), of which 302 are neurons. In addition, these cells arise from a pattern of precise and largely invariant cell divisions, which can be tracked by virtue of the worm's transparency. The cell lineage patterns, which were established by the early 1980s, were instrumental in understanding the important role of apoptosis in organismal growth and development. *C. elegans* also has unique features that make it useful for genetic studies—the worms are either hermaphrodites, depicted in Figure 19.7a, or males and can reproduce by self-fertilization (hermaphrodites, as shown in Figure 19.7b) or by mating (a hermaphrodite with a male).

An unanticipated benefit of developing *C. elegans* as an experimental model was the relatively recent discovery of the RNA interference (RNAi) machinery, which grew out of attempts to manipulate gene expression in worms using

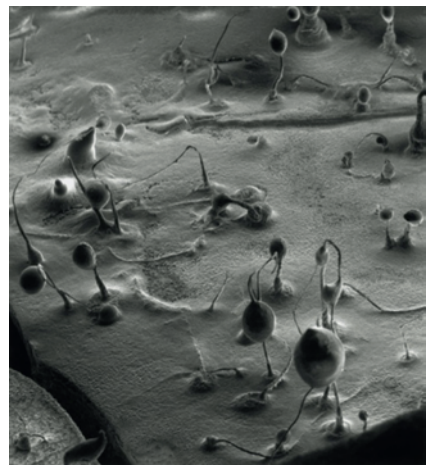


Figure 19.5 **Slime mold *D. discoideum*.** Scanning electron micrograph of fruiting bodies formed by a colony of *D. discoideum* cells.
From David Scharf/Science Photo Library.

(a)

(b)

Figure 19.6 **The fruit fly.** (a) Wild-type (top) and mutant (bottom) fruit flies. The mutation in the mutant fly causes the third thoracic segment to develop wings, instead of halteres. (b) Polytene chromosomes from salivary glands. The light and dark bands correspond to transcribed and non-transcribed regions, respectively. The numbers refer to the different chromosomes (L and R refer to the left and right arms, respectively, of chromosomes 2 and 3).

(a) Akbari et al. Unraveling cis-regulatory mechanisms at the abdominal-A and Abdominal-B genes in the Drosophila bithorax complex. *Developmental Biology* **293**;2:294–304. (b) © Brian Stavely.

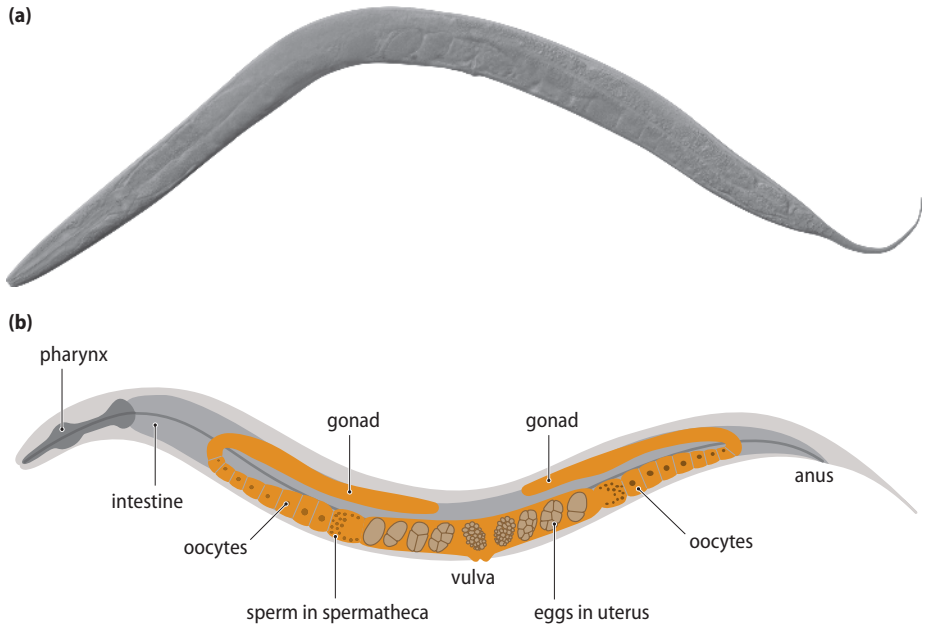


Figure 19.7 **The *C. elegans* hermaphrodite.** (a) Microscopy image of a hermaphrodite worm. (b) The hermaphrodite worm has two gonads that meet at a single vulva. The worm first produces sperm, which is stored in the two spermatheca, and then switches to producing oocytes. The oocytes are fertilized as they pass through the spermatheca.

See Experimental approach 9.3 to learn more about the fruit fly's salivary glands.

See Experimental approach 13.2 to learn more about the discovery of RNAi.

antisense RNA. Once it was realized that double-stranded RNA was causing targeted genes to be turned off, standard methods of worm genetics were rapidly used to identify components of the RNAi machinery. In fact, genes in *C. elegans* can be turned off simply by feeding the worm bacteria expressing the desired double-stranded RNA.

The most prominent model organism from the plant world is the flowering plant *A. thaliana*, the wild-type of which is shown in Figure 19.8a. Many other plants have served as model organisms, including maize, pea, tobacco, petunias, and snapdragons, but *Arabidopsis* plants are much smaller and have a shorter generation time. *Arabidopsis* is also a diploid and has a much smaller genome than agriculturally or horticulturally important plants (for example, some wheat cultivars are either tetraploid or hexaploid), making the isolation of mutants relatively straightforward. As we will see later, genetic manipulation of *Arabidopsis* is also aided by the fact that foreign genes can be introduced by infecting the plants with a soil bacterium, *Agrobacterium tumefaciens*. This soil bacterium harbors a tumor-inducing plasmid (Ti) into which genes can be cloned and via which they can be introduced into the plant cells.

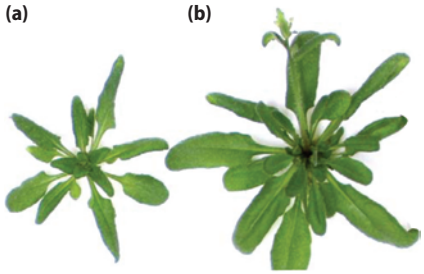


Figure 19.8 **The plant model *Arabidopsis*.** (a) Wild-type *A. thaliana*. (b) An *Arabidopsis* carrying a mutation that causes the plant to become bigger. Mutations such as this may be useful in the development of commercially valuable plants.

The zebrafish, frog, and mouse are model vertebrates

Several organisms serve as model vertebrates—that is, animals having a backbone and spinal column; these include reptiles, birds, and mammals. The small zebrafish (*Danio rerio*) commonly found in pet stores has many features of other model organisms that make it amenable to study. Zebrafish are easy to raise and breed in the laboratory, with each female producing hundreds of eggs in a single clutch. Although the generation time of zebrafish is similar to that of the mouse (3–4 months), the embryos develop in a matter of days. Importantly, embryo development occurs externally to the mother and the embryos are transparent, making it possible to observe the different stages of vertebrate embryogenesis and organ formation. A variety of techniques are available for genetic

manipulations in zebrafish, including the ability to generate transgenic fish, as shown in Figure 19.9. Such tools make it possible to even conduct large-scale screens of mutant fish.

Another vertebrate, *Xenopus laevis*, the African clawed frog, is very widely used in molecular biological and biochemical studies. Since oocytes and eggs of this frog are very large, it is possible to inject DNA and other molecules (and even whole nuclei) into them and examine the expression and function of these components. In addition, *Xenopus* eggs are a rich source for biochemical studies because the large quantities of concentrated cytoplasm can be easily isolated for biochemical reactions, as depicted in Figure 19.10. The extracts can be fractionated to identify the important components for a given reaction. Cell cycle and DNA replication studies, in particular, have used such *Xenopus* extracts to identify many biological molecules. *Xenopus* is not, however, extensively used for genetic analysis because it is a partial tetraploid, making genetic manipulations more difficult because there are four copies of each gene. The genome of the closely related, but diploid, frog *Xenopus tropicalis* has been sequenced, allowing genetic approaches to be more readily used in the study of this important vertebrate model system.

Some biological processes in humans can only be analyzed in other mammals. For example, the house mouse *M. musculus* has provided invaluable insights into human biology and disease. It is often the experimental organism of choice for comparisons with humans because it too is a mammal, and yet can be cultivated for laboratory research. The mouse genome is similar to the human genome in size, and there is extensive sequence homology between genes. Mice offer an excellent system to study the effects of genes associated with human disease; they can be used to study both single-gene traits and, increasingly, complex traits. Our understanding of the complexities of the mammalian immune system, for example, was largely derived from experimental studies in mice.

Multiple approaches use mice to study those genes implicated in human disease. For example, it is possible to generate genetically modified mice by manipulating the genes in mouse embryos. The techniques for generating genetically modified mice are discussed in Section 19.6 The ability to disrupt single genes, thereby creating a "knockout mouse," has been an invaluable tool in studying gene function in mammals and in understanding the molecular basis of an increasing number of genetic human diseases. Figure 19.11 illustrates just one such example, in which a mouse model was used to study obesity.

Some genes cannot be studied in this way because embryos in which both copies of the gene (on the two homologous chromosomes) have been knocked out fail to complete development and reach adulthood. In many of these cases, **conditional mutation** is used. These mutants make it possible to generate animals in which essential genes are deleted in only certain tissues. As the gene is not deleted in the entire mouse, embryonic development can occur. It is also possible to produce strains of transgenic mice, that is, mice whose cells express a foreign or mutant form of a gene. Despite the powerful tools available for mice, it should be noted that there are still important differences between rodents and humans, for example in metabolism, that preclude extrapolation for all diseases.

Rats (*Rattus norvegicus*) are also used for studying molecular mechanisms of disease, but to date the genetic tools for rat research are less developed than in the mouse.

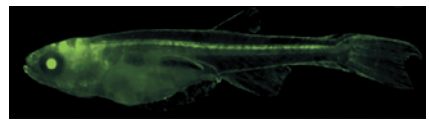


Figure 19.9 Zebrafish. The image shows a transgenic zebrafish expressing GFP under the control of regulatory elements of the *inhibitor of differentiation 1* (*id1*) gene, which is expressed in the bones, skin, retina, pineal, optic tectum, and cerebellum. From http://www.sars.no/research/BeckerGrp.php.

Courtesy of Mary Laplante. See also Kikuta et al. *Genome Research* **17**: 545–555, 2007.

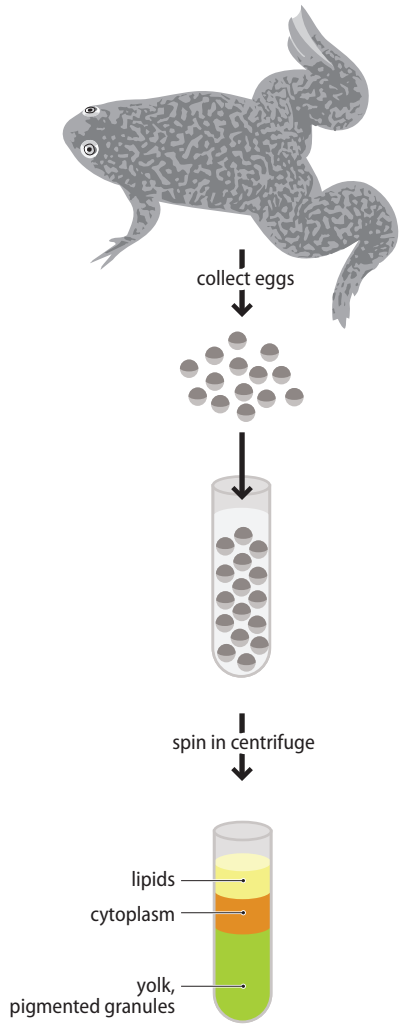


Figure 19.10 Preparation of *Xenopus* egg extracts. Each frog can lay hundreds of eggs. The eggs are collected in a centrifuge tube and spun to crush the eggs and separate the various egg components. The cytoplasmic fraction is often used for biochemical studies.

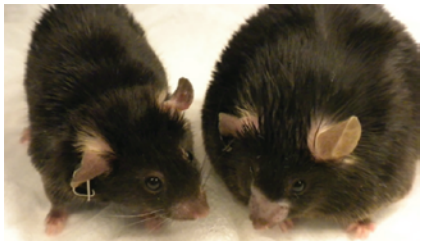


Figure 19.11 A mouse model for studying obesity. The mouse on the left is wild-type, while the mouse on the right carries a mutation in the *leptin* gene that causes it to be obese.
Courtesy of Shannon Reilly and the University of Michigan.

Additional model organisms such as *Tetrahymena* are being exploited for their special properties

Many of the model organisms described earlier in this section are representatives of particular phylogenic branches and together give a general overview of biological systems. Some model organisms, however, are chosen not for their common properties, but rather for some unique trait that can give insights into specific biological processes. A prime example is the ciliated protozoan *Tetrahymena thermophila*, shown in Figure 19.12a. *Tetrahymena* is a single-celled eukaryote that is found in freshwater ponds. It is classified as a ciliate because it is covered with cilia, small appendages that allow it to move through the water. This single-celled eukaryote is unusual in that it has two nuclei—a diploid micronucleus, which serves as a germline "vessel," and a polyploid somatic macronucleus, from which gene expression occurs.

The macronucleus is derived from a copy of the diploid micronucleus. This takes place in a developmental process that involves both gene amplification and the rearrangement of some regions of DNA and the elimination of other sections, as depicted in Figure 19.12b. As a result of the amplification, a very high copy number of small chromosomes arises, with many additional copies of the ribosomal RNA genes. These features led to the discovery of both self-splicing rRNA and telomeres. The presence of both a transcriptionally silent micronucleus and a transcriptionally active macronucleus in the same cell allowed investigation of the role of histone modification in transcriptional regulation. These discoveries are described in Experimental approaches 4.2 and 10.2.

Even though its unique genome gymnastics make it seem unusual, studies in *Tetrahymena* have uncovered important processes and principles that are broadly applicable. Given the conservation of fundamental mechanisms in biology, what is learned in other unusual organisms can shed light on processes across biological systems. For this reason, the exploration of diverse organisms to study specific biological processes makes it possible to gain new fundamental insights. With the increased ease of genome sequencing and ever improving tools for genome manipulation, such as the CRISPR–Cas9 system discussed in Section 19.6, we are witnessing more and more exploitation of organisms with special properties and are able to draw ever more fruitful comparisons between organisms.

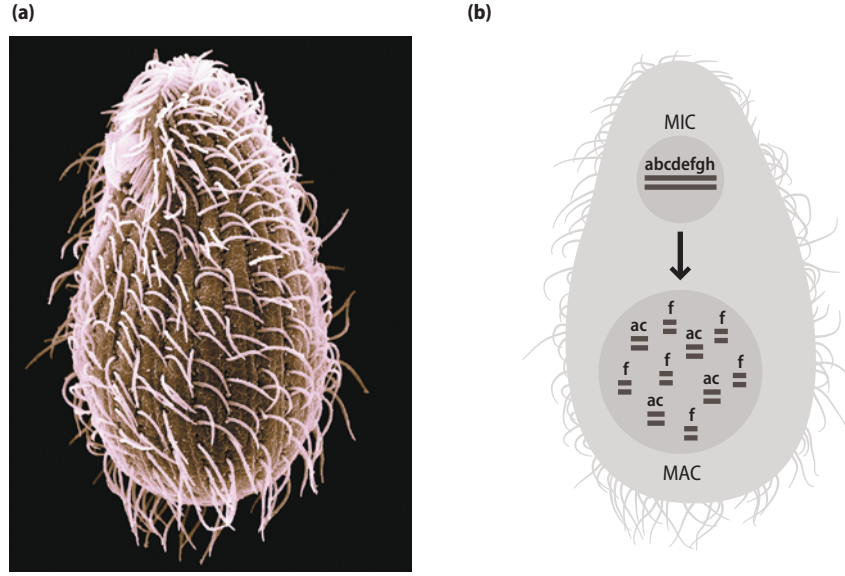


Figure 19.12 The ciliated protozoan *T. thermophila*. (a) Electron micrograph of *T. thermophila*. (b) *Tetrahymena* has two nuclei—a diploid micronucleus (MIC) that serves as a germline, and a polyploid somatic macronucleus (MAC). Gene expression occurs from the macronucleus, which is derived from a copy of the diploid micronucleus in a developmental process involving multiple DNA rearrangements.
(a) Courtesy of Aaron J. Bell.

19.2 CULTURED CELLS AND VIRUSES

Detailed biochemical studies require samples of macromolecules that have been isolated from a cell. It is generally important to purify the molecule or molecules of interest from a homogeneous population of cells, in order to avoid variations in sequence or post-translational modifications that may be found among different strains or tissue types. Although proteins and nucleic acids can be isolated from relatively homogeneous large organs, such as a bovine heart or liver, most of the experiments underlying the information in this book were carried out with proteins and nucleic acids isolated from a single type of cell grown while suspended in liquid or immobilized on agar, plastic, or glass surfaces. We refer to this type of cell growth as **cell culture**. Some of the different types of cells that are grown in culture and the advantages and limitations of cell culture are the focus of this section.

Single-celled organisms can be grown in liquid culture or on agar plates

Single-celled organisms are most readily grown in culture. Model organisms, such as the bacterium *E. coli* and the budding yeast *S. cerevisiae*, can be grown in liquid medium consisting of a mixture of nutrients that is buffered at a pH necessary for cell growth; such a culture is depicted in Figure 19.13a. The cells can also be grown on solid medium containing agar. (Agar is a polysaccharide derived from algae that is liquid at high temperatures but which solidifies at lower temperatures; agar can be added to culture medium to form a gel on which microorganisms can grow.) When cells are spread on agar at low density, each cell grows and divides, giving rise to single colonies, as seen in Figure 19.13b. When cells are spread at high density, the proliferating cells coat the surface of the agar, giving rise to a lawn.

An advantage of culturing cells is that conditions under which the cells are growing can be controlled. For example, the cells can be continuously supplied with air by bubbling in gas or by vigorous stirring or shaking of liquid medium, or they can be grown completely in the absence of oxygen. Similarly, the cells can be maintained at different temperatures or grown with different nutrient sources.

Many species of bacteria, algae, fungi, and some archaea can also be grown in liquid culture or on solid medium. Many other organisms, such as those that live in unusual environments, and even some that grow in soil, cannot be grown in the laboratory because we do not yet know how to create culture conditions under which they can grow and thrive. Our detailed knowledge of biochemical processes within cells has therefore been biased toward organisms that can be grown in culture, though increasing information about the non-culturable organisms is being derived from the sequencing of their genomes.

Many differentiated cells from multicellular organisms need to undergo modifications to be maintained in culture

Eukaryotic cells from multicellular organisms can also be grown in culture. These cells, which can be derived by dissociating cells from tissues and are known as **primary cells**, are typically grown in a plastic culture bottle, as shown in Figure 19.14, or on a plate under conditions that allow the cell to adhere to the surface. Some cell types, such as blood cells, can be grown for short periods while suspended in liquid culture.

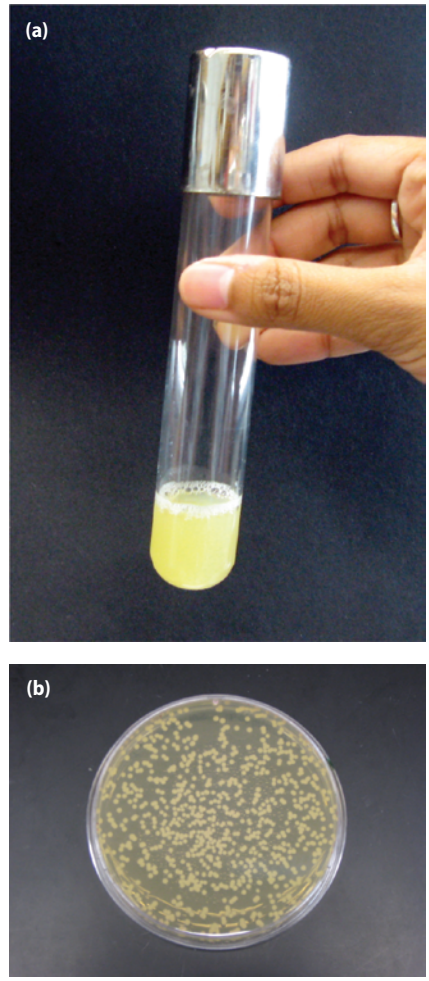


Figure 19.13 Culturing of single-celled organisms. (a) Growth of the yeast *S. cerevisiae* in a glass tube. The solution is turbid because of the presence of yeast cells at about 10^8 cells/ml. The nutrient solution itself is yellow. (b) Growth of the yeast *S. cerevisiae* on an agar plate. A solution containing several hundred yeast cells was spread on the surface of a nutrient-containing agar plate; the agar is yellow because of the nutrients. On incubation, each cell gives rise to a single colony containing several million progeny cells.

Photos courtesy of Susan Michaelis, Johns Hopkins University School of Medicine.

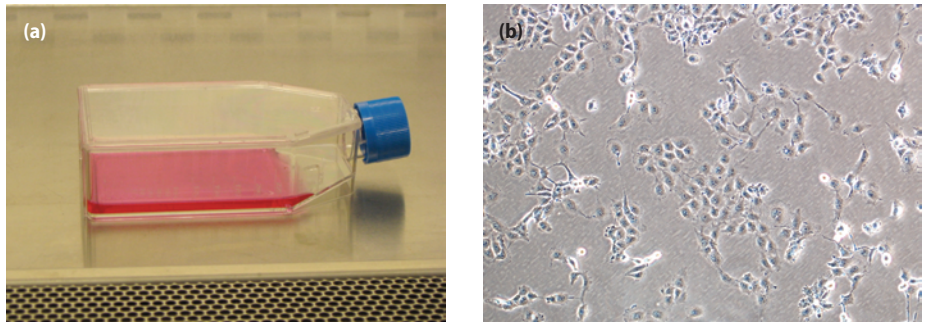


Figure 19.14 Eukaryotic cells grown in tissue culture. (a) Plastic flask containing mammalian cells in culture. The growth media is pink because of a pH indicator dye. (b) Layer of single tissue culture cells, as viewed by light microscopy.

(b) kindly provided by Sin Urban, Johns Hopkins University.

The growth of cells under these tissue culture conditions is often limited by a number of factors. When grown on a solid surface, contact between cells may trigger cell cycle arrest so that cells cease to grow, a phenomenon known as **contact inhibition**. In addition, most primary cells can only divide a limited number of times due to certain intrinsic properties of the cells themselves. For example, growth is limited in many cases by telomere shortening. Since telomeres shorten at every division, exceedingly short telomeres eventually cause **senescence**, during which cell division ceases. To avoid these problems, researchers often use special strains of cells, such as cancer cells, which can bypass senescence. These cultured cells, referred to as **cell lines**, are often derived from tumors that actively express telomerase and have undergone other genetic changes that remove the normal inhibitions to continued cell division. Cells can also be made to grow continuously in culture by infecting with a cancer-causing virus.

Some commonly used cell lines are listed in Figure 19.15. These cell lines have been central to the study of eukaryotic cells. However, some caution is warranted. The very changes that permit their continuous growth in culture also mean that these cells differ in fundamental ways from cells in intact tissues. Additionally, as cell lines are grown in culture, the genomes can acquire mutations or rearrangements. Errors in the distribution of the cell lines among researchers have also led to situations where cell lines received differ from what was expected. Thus, it is prudent to be vigilant about the status of the cells being characterized and to

Some commonly used cell lines

cell line	cell type	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	adrenal gland cell	rat
SP2	plasma cell	mouse
COS	kidney cell	monkey
293	kidney cell, transformed with adenovirus	human
CHO	ovary	Chinese hamster
DT40	lymphoma cell	chick
R1	embryonic stem cell	mouse
E14.1	embryonic stem cell	mouse
H1, H9	embryonic stem cell	human
S2	late embryonic cell	*Drosophila*
BY2	undifferentiated meristematic cell	tobacco

Figure 19.15 Commonly used cell lines.

Adapted from Alberts et al. *The Molecular Biology of the Cell*, 4th edition. Garland Science.

be aware that differences, compared to intact tissues, as well as uncharacterized genome changes, can affect the outcomes of experiments.

Stem cells are the precursor cells to all cell types

Stem cells are a type of cell that can give rise to many different cell lineages. **Embryonic stem (ES) cells** are derived from early embryos and are considered to be totipotent cells, meaning that they have the potential to give rise to every cell type of the body. Other stem cells can generate a more limited repertoire of cell types; these are called **pluripotent cells**, as they give rise to a restricted set of cell lineages. Such pluripotent stem cells are present during development and are also found in fully developed animals. Adult stem cells, which are derived from a fully mature adult organism, are usually pluripotent, and the number of different cell types they can generate depends on the specific cell type. Stem cells from skin, hair follicles, and the intestine have been studied, but the best characterized stem cells are those from blood.

Stem cells typically have an unlimited capacity to divide, in stark contrast to differentiated cells, which either do not divide or whose capacity to divide is limited. When stem cells divide, they do so in a special manner—one of the two daughters forms a new stem cell, while the other daughter differentiates into a particular cell type, as illustrated in Figure 19.16. Under some circumstances, the differentiated daughter can also give rise to other cell types and is referred to as a progenitor cell. Progenitor cells may have a more restricted set of cell fates than stem cells and often have a more limited cell division capacity.

The best way to demonstrate that a cell is indeed a stem cell is to isolate the cell and implant it into a recipient animal. A stem cell will give rise to the entire set of differentiated cells in this new setting. This type of transplantation is also the basis of bone marrow transplants that are carried out for the treatment of blood diseases and cancer. Blood cells have a limited lifespan, with some types lasting only a day. Blood cells must therefore be continually renewed throughout the life of the organism. In bone marrow transplants, stem cells that generate blood cells are isolated from the bone marrow, where they normally are found, and transferred to another individual. These blood precursor cells are referred to as **hematopoietic stem cells**, as **hematopoiesis** is the process of regenerating blood cells that occurs every day in adults. The differentiation of hematopoietic stem cells into specific blood cell lineages is one of the best understood hierarchies of stem cell differentiation and is illustrated in Figure 19.17.

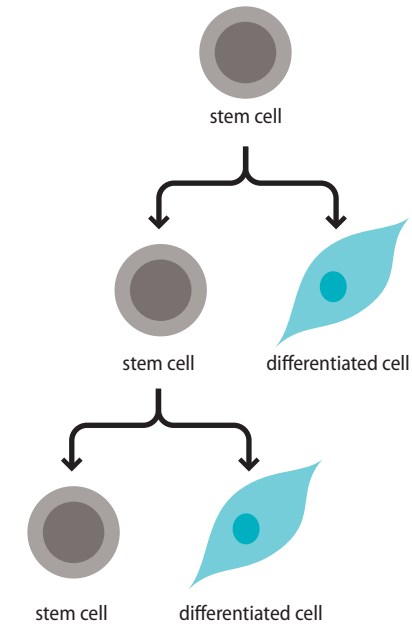


Figure 19.16 Stem cells are capable of self-renewal and differentiation. Stem cells divide in a special manner, yielding one daughter cell that is a stem cell and another daughter cell that can differentiate into a particular cell type.

ES cells can give rise to cell types needed to treat disease

ES cells have received a great deal of attention for their potential medical usage. These totipotent cells have been manipulated in the laboratory to generate specific cell lineages that might be used to treat specific diseases. For example, ES cells treated with certain growth factors in culture can cause >99% of the cells to differentiate into neurons. It is hoped that these neurons could be transplanted into patients with Parkinson's disease as a way to treat the loss of function of specific neurons, a characteristic of this disease.

One of the dangers of the use of ES cells, however, is the potential to generate tumors. For example, if less than 100% of the cells fully differentiate, the remaining undifferentiated cells could continue to grow as ES cells and later differentiate randomly to form a tumor. Indeed, direct injection of mouse ES cells into

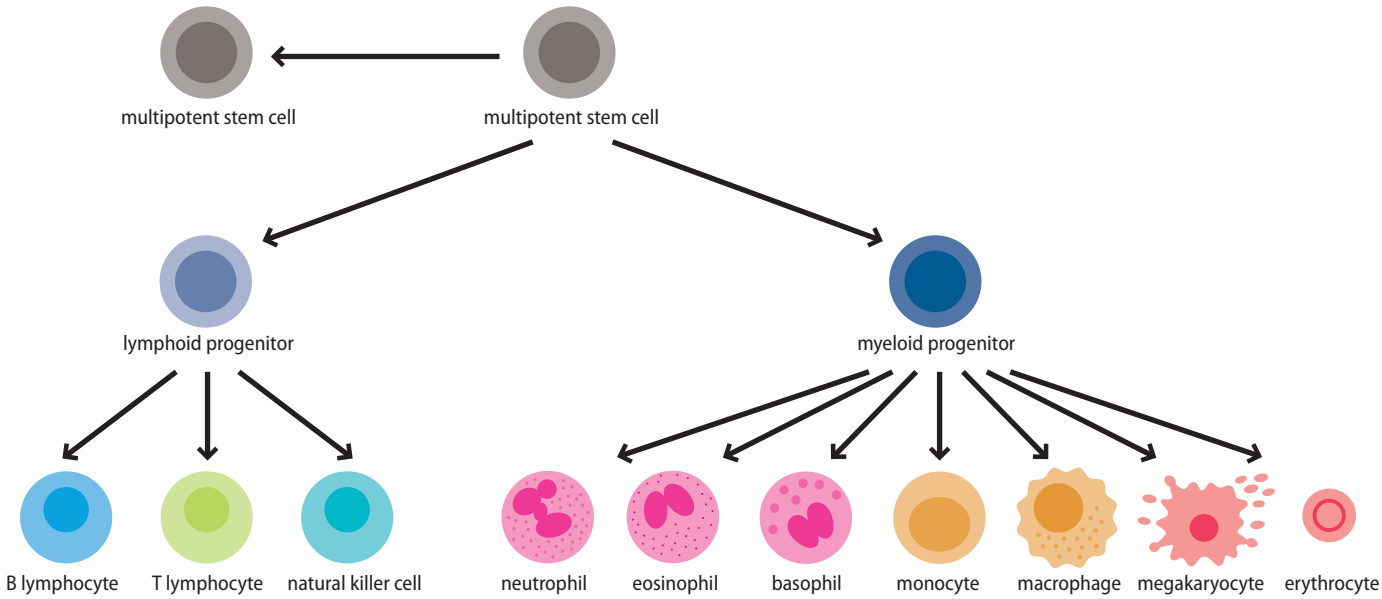


Figure 19.17 The differentiation of hematopoietic stem cells into specific blood cell lineages. Hematopoietic stem cells are isolated from blood or the bone marrow. They are capable of self-renewal (that is, they can produce more stem cells) and also produce progenitor cells that can differentiate into many other cell types.

mice gives rise to a tumor known as a teratoma, which comprises a mixture of many different differentiated cell types.

Induced pluripotent stem cells are a special form of stem cells derived from differentiated cells

In recent years, a new method has been developed to form totipotent cells for medical research. These cells are termed induced pluripotent stem (iPS) cells. Researchers found that introducing four specific transcription factors into fully differentiated skin cells would "reprogram" the skin cells, turning them into stem cells—that is, the transcription factors induce the expression of a set of genes that is needed in early embryonic development. iPS cells seem to be, in many ways, equivalent to ES cells.

Importantly, this new approach makes it possible to treat an individual with differentiated cells derived from their own tissue. Called autologous transplants, these cells are less likely to be rejected by the host's immune system. Moreover, the ability to form iPS cells from adult tissue bypasses the ethical issues surrounding the use of ES cells derived from human embryos.

Viruses must be cultured in living cells

As we saw in earlier chapters, studies of viruses that infect bacteria or eukaryotic cells have provided important insights into basic cellular mechanisms. Viruses are also a valuable means of introducing genes into cells. For example, the introduction of the specific transcription factors required to form iPS cells is often done with viral vectors. In some viruses, the genome is encoded by a DNA molecule, while other viruses (such as retroviruses) use an RNA molecule as their genome. Regardless, however, none of the viruses are free-living life forms, and they must be propagated within cells.

See Section 17.6 for a description of the retrovirus life cycle.

Most viruses can infect just a few types of cells at most, so the ability to grow a particular virus depends on the ability to culture the type of cell in which it grows. Perhaps the most important advance in developing a vaccine against the crippling disease polio was the development of methods to grow poliovirus in cultured human cells.

Viruses that have been propagated in cultured cells can be isolated in several ways. Since viruses infect a cell and harness the cell's machinery to produce many virus particles, one isolation method is to infect cultured cells with the virus and harvest the virus-producing cells that result. The cell membrane can then be disrupted with detergents to release the virus particles, which can be separated from the cell debris by a variety of methods. With lytic viruses, which cause the cell membrane to burst in order to release the virus particles, cells can be infected and grown until the cells burst and release virus particles into the culture medium, or onto the surface of a Petri dish where lysis of cells in a lawn produces a clear circle called a plaque, as depicted in Figure 19.18. The virus particles present in a plaque or shed into the culture medium can then be purified.

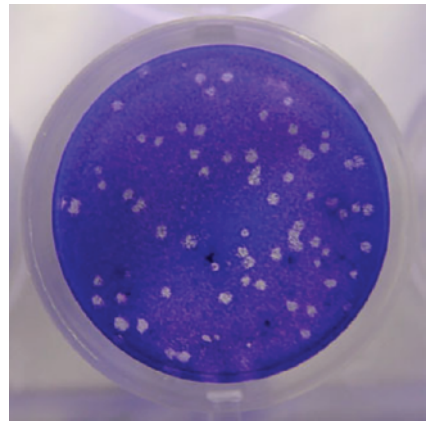


Figure 19.18 Viral plaque assay. To assay for a virus, appropriate cells are grown on a plate and then infected with the possible virus-containing sample. As a virus infects a cell, it produces many virus copies that, after cell lysis, infect neighboring cells. As a result, a zone of dead cells is formed. In the example shown, live cells take up a purple dye, and the white areas are plaques where cells died due to viral infection. The number of plaques is correlated with the number of live viruses in the sample.

19.3 AMPLIFICATION OF DNA AND RNA SEQUENCES

The ability to manipulate DNA to generate many copies, or clones, of wild-type and mutant derivatives and the ability to determine the sequence of a particular DNA or RNA fragment have been central to the advancement of molecular biology. Both molecular cloning and sequencing of DNA and RNA rely upon the generation of many copies of the DNA or RNA of interest from a small amount of starting material. In this section, we discuss commonly used methods for manipulating and amplifying DNA and RNA sequences.

The polymerase chain reaction can be used to amplify a specific DNA sequence

The polymerase chain reaction, commonly referred to as PCR, can be used to generate thousands or even millions of copies of a particular sequence of double-stranded DNA, beginning with just a few copies of starting material. The method relies upon repeated cycles of DNA strand separation, followed by replication of each DNA strand. This doubling of the number of DNA molecules with every cycle leads to rapid amplification of the DNA sequence. Virtually any DNA segment can be amplified by PCR.

The boundaries of the DNA sequence to be amplified are set by two synthetic DNA oligonucleotides, each complementary to the 3′ end of one of the DNA strands in the DNA fragment to be amplified. PCR itself then proceeds as illustrated schematically in Figure 19.19. In the first cycle of PCR, the two strands of a DNA template are first separated by heat denaturation, followed by cooling of the sample. As the temperature drops, the two oligonucleotides (also called primers) hybridize to the denatured DNA, each complementary to a different 3′ end of the sequence to be amplified. DNA polymerase, which is present in the reaction along with dNTPs, utilizes the oligonucleotides as primers to replicate the complementary strand. This procedure produces two copies of the desired DNA sequence. The cycle is then repeated, as once again the DNA is denatured so that primers can hybridize and be extended.

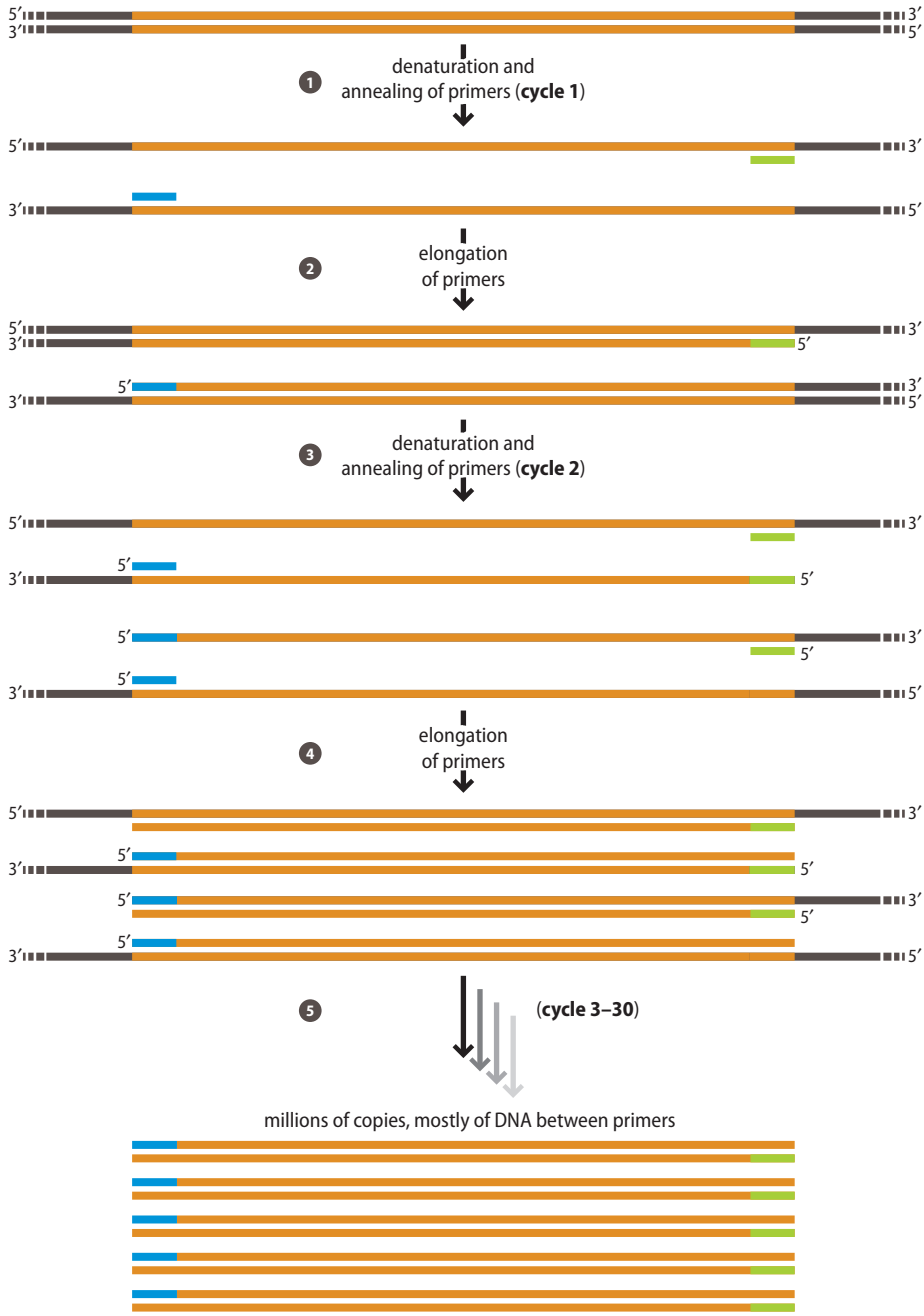


Figure 19.19 The use of PCR to amplify a specific DNA fragment, using multiple rounds of primer annealing and extension. Oligonucleotides (in blue and green) are designed to be complementary to each end of the fragment that is to be amplified (in orange). Such oligonucleotides, also known as primers, are typically around 20 bases in length. In the first PCR cycle, the DNA is denatured and the primers anneal to the complementary site on the single-stranded DNA (step 1). These primers then serve to prime DNA replication, usually by a heat-resistant DNA polymerase from the bacterium *Thermus aquaticus* (step 2). After the DNA is fully replicated, a second PCR cycle begins where the DNA is again denatured and a second round of primer annealing and extension takes place (steps 3 and 4). With every successive round (step 5), the amount of new DNA present in the reaction mixture is doubled such that the DNA region between the primers is preferentially amplified.

With each cycle, the number of template strands doubles. It is this amplification that makes PCR such a powerful experimental tool—it allows over one million copies of a single double-stranded template to be generated in just 20 cycles. PCR is usually done in an automated manner in a machine that cycles between different temperatures, using a DNA polymerase that is stable at high temperatures so that repeated cycles of heat denaturation do not inactivate it.

Although the PCR method is extremely powerful, a few caveats should be noted. Oligonucleotide primers can sometimes hybridize to sites other than those intended due to partial sequence complementarity. When this happens, the wrong DNA sequence could be amplified. Thus, primers need to be designed with care, and there are specialized computer programs to assist with this.

In addition, misincorporation of bases by the polymerase can introduce random mutations during DNA synthesis. If such a mutation is introduced in an early cycle, the mutation will be amplified at each cycle. Finally, amplification of long DNA segments, typically over 4000–5000 base pairs (bp), may require specialized DNA polymerases and conditions.

While PCR is most commonly used to produce faithful copies of a DNA sequence, it can also be used to introduce random mutations by carrying out the PCR reaction under special conditions. For example, altering the salt concentration affects the fidelity of the DNA polymerase, while altering the ratios between the four deoxyribonucleotides in the reaction favors the misincorporation of bases. The resulting population of DNA fragments containing random mutants can then be introduced into cells through a variety of methods and used in genetic screens for altered protein function.

PCR can be used to add sequences that are not in the template DNA

PCR can be applied in various ways. For example, it is sometimes useful to add additional nucleotides that are not found in the template DNA to either end of the amplified DNA fragment. These additional sequences are particularly useful for subsequent steps in cloning (see Section 19.4), as well as in a variety of other applications. Additional nucleotides can be added to the amplified DNA fragment by synthesizing a primer that contains additional nucleotides of the desired sequence at its 5′ end. As a result, the 5′ end of the primer is not complementary to the template DNA sequence. However, as the DNA is amplified over successive PCR cycles, the sequences of the primer become part of the new DNA template and are incorporated into the PCR products, as illustrated in Figure 19.20. The additional DNA that has been added to the amplified fragment can then be used in subsequent manipulations, as we shall see in Section 19.4.

RNA amplification depends on a DNA intermediate

In some cases, investigators wish to generate copies of a particular RNA molecule or obtain the full-length sequence of an RNA transcript. Since RNA cannot be directly cloned, a complementary DNA (cDNA) copy must be generated and then used in rounds of DNA amplification similar to PCR. In contrast with PCR amplification of a DNA sequence, though, the ends of a given RNA molecule of interest may not be known, for example, because of alternative splicing or there being multiple possible promoters from which the RNA may be transcribed.

This problem is overcome by a modification of the PCR technique for DNA cloning called RACE (for rapid amplification of cDNA ends) that makes it possible to amplify an RNA sequence even when its precise 3′ and 5′ boundaries are not known. The RACE method can be used to obtain the nucleotide sequence of an RNA transcript from a short known sequence within the transcript to either the 3′ or 5′ end. By combining these two approaches, known as 3′-RACE and 5′-RACE, one can determine the complete sequence of the RNA transcript.

For the purpose of this discussion, we will consider the amplification of an mRNA molecule. Since eukaryotic mRNAs typically have a poly(A) tail at their 3′ ends, the presence of this characteristic stretch of nucleotides can be exploited in 3′-RACE, as illustrated in Figure 19.21a. A primer is designed such that its 3′ end is a poly(T) sequence—a sequence that is complementary to the poly(A) tail.

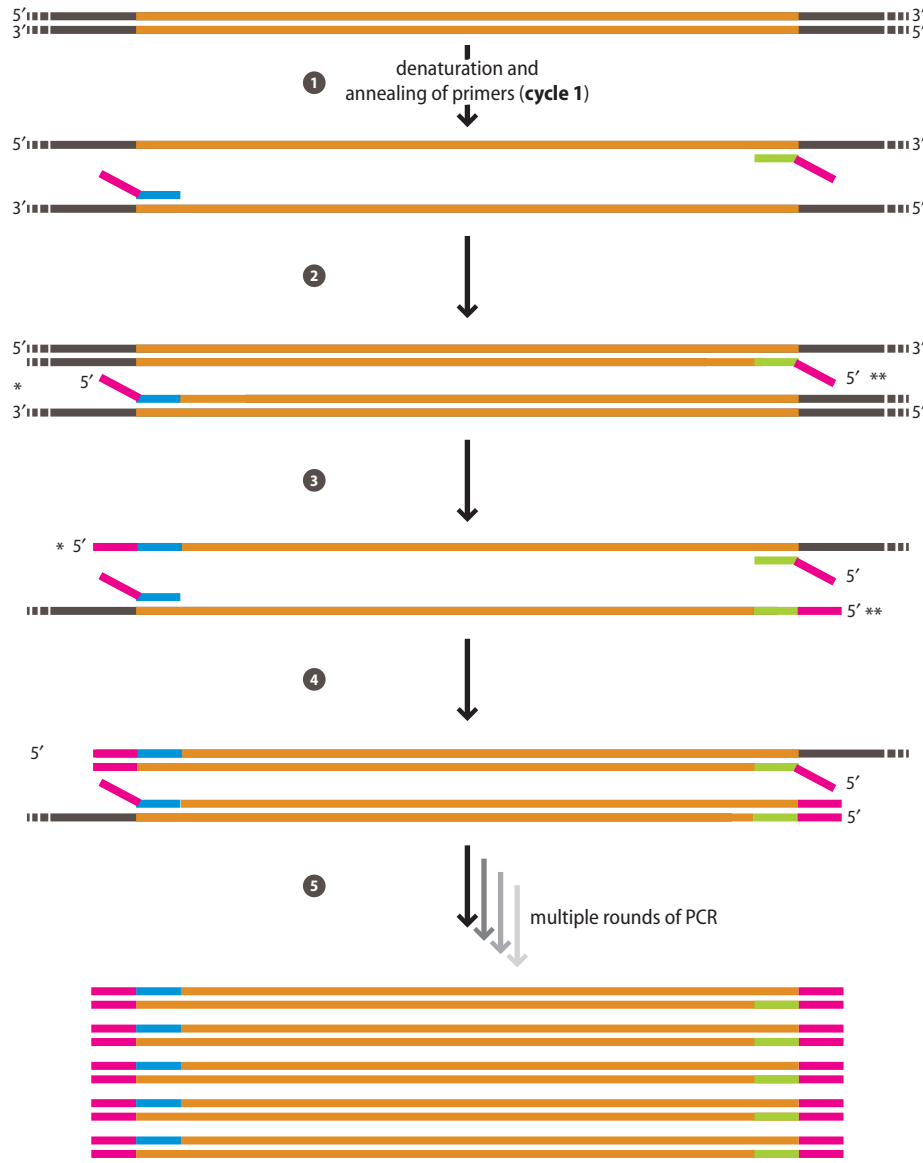


Figure 19.20 Adding specific sequences to DNA that is amplified by PCR. To add DNA sequences to the ends of a DNA fragment, one or both primers have unique sequences on the 5′ ends (pink), in addition to a region of complementarity to the template DNA (green or blue). During the first round of PCR, only the complementary sequences will anneal to the template DNA (step 1). However, once these primers are extended by DNA polymerase (step 2), the new sequences become part of the new template DNA (* or **). In the next round of amplification (steps 3 and 4; for simplicity, only the newly synthesized DNA is shown), the template DNA that was made in the first round will contain the added sequences, which will be incorporated into the amplified DNA in each successive amplification cycle (step 5).

Additional nucleotides may be added to the 5′ end of the primer to facilitate subsequent manipulations, as described earlier for DNA PCR.

In the first step, a cDNA copy of the RNA is made using a retroviral DNA polymerase called reverse transcriptase (see Section 6.4), which uses the RNA molecule as a template to synthesize a complementary DNA strand. The poly(T) primer, called an anchor primer, is annealed to the RNA and then extended by reverse transcriptase to form an RNA–DNA hybrid. If a mixture of mRNAs is used as template, as would be the case if the mRNA is purified from a cell, this process will result in many different RNA–DNA hybrids, one for each type of polyadenylated RNA present.

Next, a second primer is used, this time using a sequence that is internal to the mRNA of interest (which would be complementary to the newly synthesized DNA). The RNA–DNA hybrid is denatured, the internal primer is allowed to anneal to the DNA strand, and the reaction now proceeds in a manner similar to PCR. Multiple rounds of amplification will generate a cDNA product, one strand of which is identical to the portion of the original RNA molecule that lies between the two primers used in these reactions.

The process is somewhat more complicated if one wants to convert the 5′ end of the RNA to DNA. In this method, called 5′-RACE, the sequence at the 5′ end of the RNA molecule is unknown. To overcome this problem, an artificial 5′ end is generated. As illustrated in Figure 19.21b, the first step of 5′-RACE is to anneal a primer that is complementary to an internal sequence within the RNA. This antisense primer is extended by reverse transcriptase, which synthesizes a DNA copy of the RNA molecule from the primer all the way to the 5′ end of the RNA molecule. A poly(A) tail is then added to the 3′end of the complementary DNA strand using an enzyme called terminal transferase. The subsequent steps are similar to 3′-RACE; the DNA–RNA hybrid is denatured, and an anchor primer complementary to the poly(A) sequence is used to prime synthesis of the cDNA. With two complementary DNA strands, one a copy of the original RNA and one the antisense strand, the DNA is amplified, using the original primer to the internal sequence and the primer complementary to the added poly(A) tail.

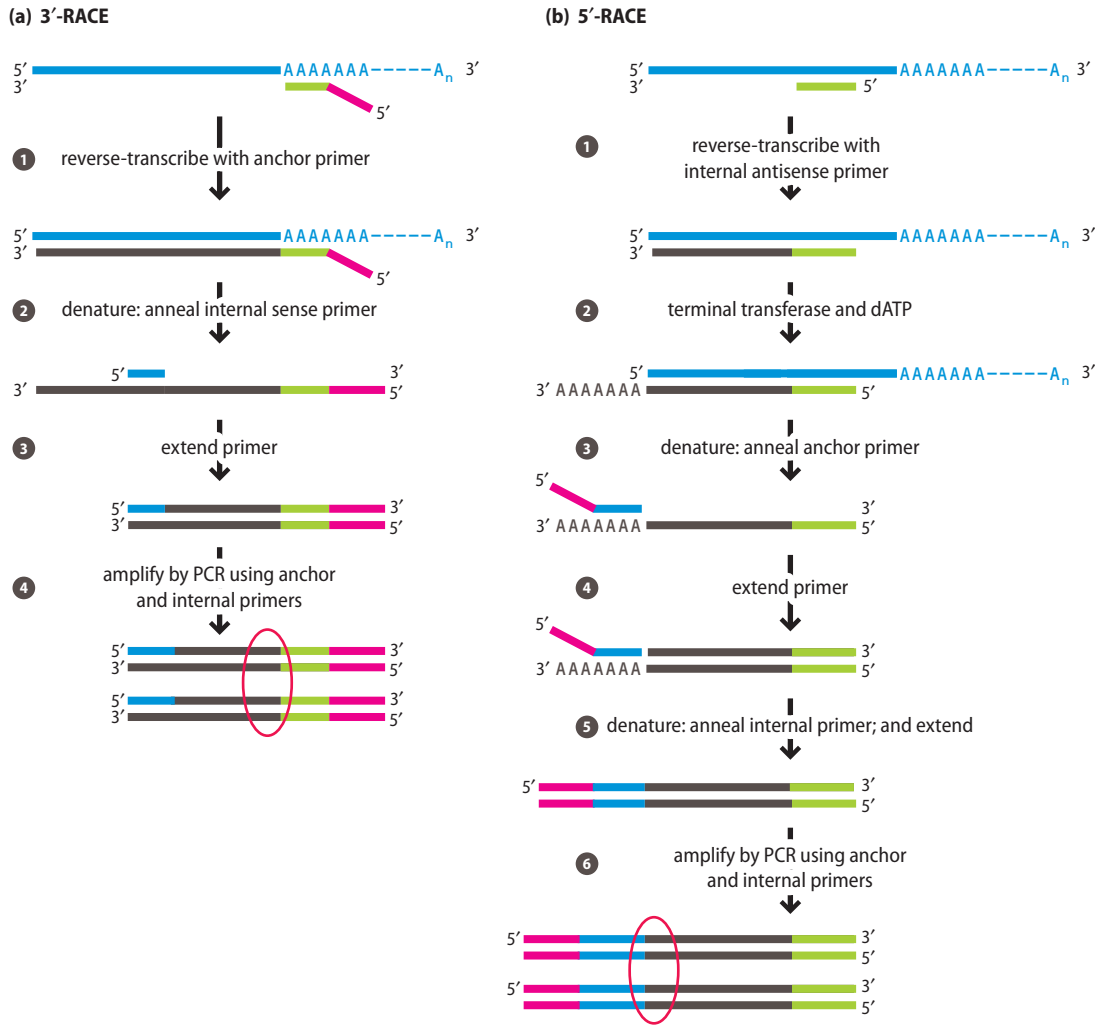


Figure 19.21 3′- and 5′-RACE. (a) 3′-RACE. The mRNA (in blue) is annealed to an anchor primer that is made of a sequence complementary to the mRNA's poly(A) tail (poly(T), in green) and a sequence of choice (pink). Once annealed, the primer is extended by reverse transcriptase, which creates a DNA molecule that is complementary to the mRNA (step 1). The DNA–RNA hybrid is denatured, and the DNA strand is annealed to a primer (blue) with an internal mRNA sequence. The primer is extended by a DNA polymerase (step 2) to create a double-stranded DNA molecule. This DNA can then be amplified by PCR (step 3) that includes the anchor primer (pink) and the internal primer (blue). The end result (step 4) is a DNA molecule, one of the strands of which is identical in sequence to the original mRNA. (b) 5′-RACE. This reaction is similar in principle to 3′-RACE, except that the anchor primer is on the 5′ end. The reaction begins with an internal primer (green) that is complementary to the RNA sequence, which is extended by reverse transcriptase (step 1). The 3′ end of the newly synthesized DNA is then extended by terminal transferase, which adds a poly(A) tail (step 2). This DNA strand can then be copied with an anchor primer that is made of a sequence complementary to the added poly(A) sequence (blue) and a sequence of choice (pink) (step 3) and further amplified by PCR (steps 4–6), as in 3′-RACE.

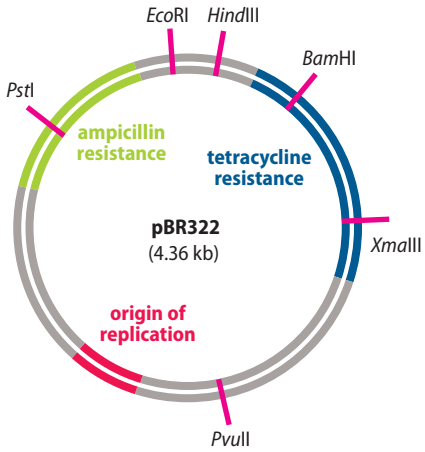


Figure 19.22 Plasmid vector. The bacterial plasmid pBR322 contains an origin of replication that allows propagation in *E. coli* cells and selectable marker genes that grow on ampicillin and tetracycline, as well as several unique restriction sites.

19.4 DNA CLONING

The ability to propagate and re-engineer specific DNA fragments in cells has been central to the study of molecular biology in the past few decades. Thanks to these methods, scientists are able to change the sequence and expression of any gene or DNA region of interest. In this section, we discuss methods for cloning DNA sequences.

Specific DNA fragments can be amplified and modified by cloning into plasmid or viral vectors

The process of isolating and propagating specific DNA regions is called **DNA cloning** because many copies of the same DNA sequence are generated from the original copy. The first step in cloning is to introduce the DNA fragment into an appropriate DNA vector that can be used to propagate the desired DNA fragment independently of the host cell chromosome. Since the final product is a DNA molecule containing both the inserted fragment and the vector DNA, the methods for generating and manipulating the cloned gene are referred to as **recombinant DNA** techniques.

Vectors for cloning can be plasmids, which are circular double-stranded DNA molecules that can replicate independently of the chromosome, or viruses. These vectors contain specific elements that make them suitable for cloning. For example, as illustrated in the schematic of the commonly used pBR322 vector in Figure 19.22, a vector contains a specific DNA sequence that allows it to replicate within its target host; plasmids that are propagated in *E. coli* contain an origin of replication that is recognized by the bacterium's DNA replication machinery. These vectors usually also carry a gene known as a **selectable marker** that allows only the host cells containing the vector to grow under the specific culture conditions. For example, vectors can carry an antibiotic resistance gene that allows only the vector-containing cells to grow in the presence of an antibiotic drug that would otherwise kill the host, or a gene encoding a metabolic enzyme that allows the cells to grow in specific medium that would otherwise not sustain cell growth. DNA cloning is often carried out using bacteria, although other host organisms such as yeast or cultured cells are also used.

Vectors used for cloning have a range of useful properties

The bacterium *E. coli* is the organism of choice for most routine cloning. A wide range of vectors can be used to propagate genes in *E. coli*; the vector chosen depends on the experiment to be undertaken. Different vectors can accommodate inserts of different size; typical bacterial plasmids might have inserts of up to 15 kilobases (kb), whereas a specialized kind of bacterial vector, called a bacterial artificial chromosome (BAC), can have inserts that are hundreds of kilobases in size.

Many vectors can replicate in several different organisms and are therefore called shuttle vectors. These vectors contain two or more origins of replication: one for replicating in *E. coli*, and the others for replicating in a different organism of choice. These vectors may include two or more selectable markers: one for *E. coli*, and the other(s) for another organism. In this way, the cloning steps can be done in *E. coli*, after which the vector can be introduced into cells of the organism of choice.

Vectors also vary widely in copy number within the cell. Some vectors are present in one or two copies per cell, whereas others are present in more than 50

copies. This is important to consider because the expression of a gene is often proportional to the number of copies present.

Once a DNA segment is cloned into a vector, it can be further manipulated according to the investigator's needs; a cloned DNA fragment is more easily manipulated than the equivalent DNA sequence situated in its "native" chromosomal location. For example, a cloned gene can be introduced into a cell containing a mutation or deletion of a particular gene, and one can then test whether the introduced gene complements the mutation (that is, restores a particular gene activity lost by a given mutation). It is also possible to fuse the gene of interest to a **reporter gene**, such as a gene coding for GFP (see Section 19.13), or to a short peptide that facilitates detection or purification of the protein. Finally, by placing the gene next to a strong promoter, it is possible to direct high levels of gene expression, thus making it easier to purify the protein of interest.

The first step in cloning typically involves using PCR, as described in Section 19.3, to make many copies of the DNA fragment of interest. Once the desired DNA is amplified, there are two main strategies for cloning a DNA fragment:

- ligation—in which the two ends of the DNA fragment are cut with restriction enzymes and the fragment is then "glued" to the vector DNA by a DNA ligase
- recombination—in which the DNA fragment is inserted into the vector by recombination enzymes by virtue of sequence homology.

We begin by describing the use of restriction enzymes, which are an important tool for cloning by ligation.

Restriction enzymes cleave DNA at specific sites

One of the earliest tools developed for manipulating DNA was the use of bacterial proteins, known as restriction enzymes, that recognize and cleave double-stranded DNA at specific sequences, as illustrated in Figure 19.23a. These sequences, known as restriction sites, are typically palindromic sequences of 4–8 bp in length. Bacteria use these enzymes as part of a defense mechanism against foreign DNA, but scientists have adapted them as tools for manipulating DNA in the test tube.

There are hundreds of restriction enzymes that differ from one another in the DNA sequence that they recognize and in the location of the cleavage sites relative to the restriction site. Depending upon the restriction enzyme, the DNA cleavage reaction may leave an end that is blunt or that has a small region of 5′ or 3′ overhanging bases, as shown in Figure 19.23b. The overhangs are useful during cloning, as we will see next. Restriction enzymes are also used for diagnostic purposes (see Section 19.11).

Cloning by ligation requires restriction enzymes and DNA ligase

In cloning by ligation, a linear DNA fragment of interest is enzymatically joined with a larger DNA vector to yield one double-stranded DNA molecule, as shown in Figure 19.24a. The inserted DNA fragment, which we will refer to as the insert, may be excised from a larger region of DNA using a restriction enzyme, as illustrated in Figure 19.23a. More commonly, the insert is amplified by PCR, as described in Section 19.3. A PCR-amplified DNA fragment may be obtained using PCR primers engineered to include particular restriction sites at both its 3′ and 5′ ends.

Figure 19.24b illustrates how the insert and vector are prepared. The insert is prepared by digesting it with the appropriate restriction enzymes, which often

(b)

microorganisms	restriction enzymes	cleavage sites	cleavage products
Escherichia coli	*Eco*RI	5′-G↓AATTC-3′ 3′-CTTAA↑G-5′	5′-G AATTC-3′ 5′-CTTAA G-3′
Haemophilus influenzae	*Hind*III	5′-A↓AGCTT-3′ 3′-TTCGA↑A-5′	5′-A GACTT-3′ 5′-TTCGA A-3′
Haemophilus parainfluenzae	*Hpa*I	5′-GTT↓AAC-3′ 3′-CAA↑TTG-5′	5′-GTT AAC-3′ 5′-CAA TTG-3′
Klebsiella pneumoniae	*Kpn*I	5′-GGTAC↓C-3′ 3′-C↑CATGG-5′	5′-GGTAC C-3′ 5′-C CATGG-3′

Figure 19.23 Restriction enzymes.
(a) Restriction enzymes recognize a specific DNA sequence (that is, specific combinations of adenosine (A), cytidine (C), guanosine (G), and thymidine (T)) to cleave the double-stranded DNA. In the example shown, the restriction enzyme *Eco*RI recognizes the GAATTC sequence and cleaves after the G on both DNA strands. As a result, the new DNA ends now each contain a 5′ overhang. Since restriction enzymes of this type recognize palindromic sequences, the overhangs are complementary to each other and can be used in a reverse reaction, in which these ends are ligated (see Figure 19.24). (b) Examples of restriction enzymes produced by different bacteria. Note that these enzymes recognize different sequences and leave different types of overhangs.
From Stem Cell Information, 2001. http://stemcells.nih.gov/info/scireport/chapter5.asp.

generate single-stranded overhanging ends. The vector is prepared by digesting it with the same restriction enzymes to generate overhanging ends that are complementary to the overhangs of the insert. The DNA insert and vector DNA are then incubated together, so that the compatible ends of the insert and vector can base-pair with each other. A DNA ligase is used to covalently join the adjacent 5′ and 3′ ends, thereby connecting the DNA backbone.

The ligated plasmid is introduced into bacteria using a process called **transformation** (see Figure 19.24a), in which the receiving cells usually need to be treated with salts or exposed to an electric field before they are competent to take up the DNA. (Cells able to take up DNA are referred to as competent cells.) The presence of a selectable marker in the plasmid—for example, a gene that confers antibiotic resistance—is then used to select for cells that have received the plasmid. This is done by taking cells that have undergone the transformation reaction and plating them on solid growth medium containing that antibiotic, so that only the cells that contain the ligated plasmid with the resistance gene will grow.

Gibson cloning allows for ligation of multiple DNA fragments

A variation on cloning by ligation can be used to join several different DNA fragments in a single ligation reaction, accelerating the process of obtaining clones. In this approach, called Gibson cloning, which is depicted in Figure 19.25, adjacent DNA fragments to be combined share regions of homology that can be introduced by PCR, as described earlier. The DNA fragments, typically a vector and one or more inserts, are mixed in a test tube and treated with an exonuclease that leaves a 5′ overhang at each DNA end. Complementary single-stranded regions base-pair,

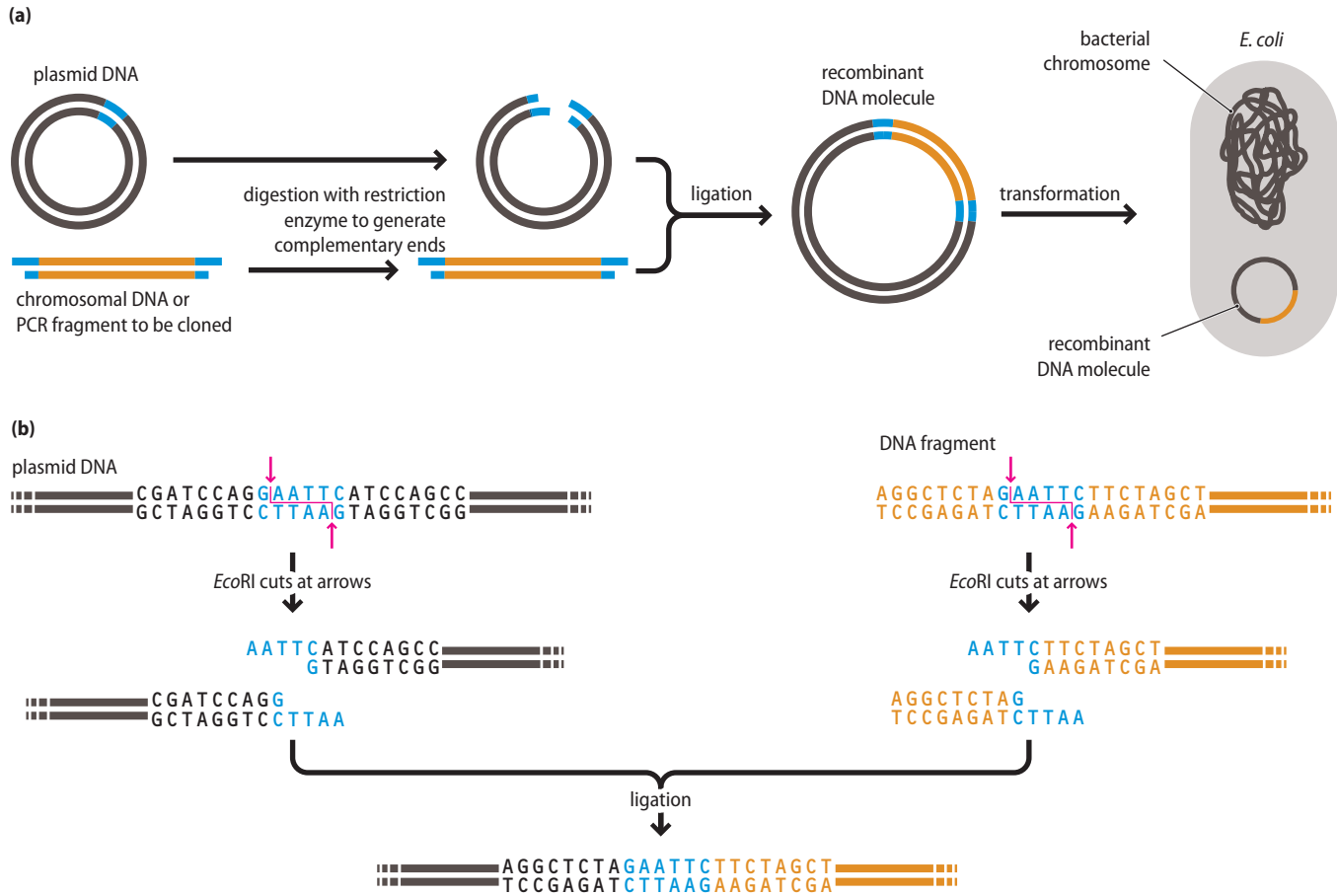


Figure 19.24 Cloning using restriction endonucleases. (a) In this cloning scheme, a plasmid is cut with a restriction endonuclease to create two ends (the restriction sites on both plasmid and fragment DNA are shown in blue). The fragment DNA to be cloned is cut with the same restriction endonuclease, making the single-stranded ends of the plasmid and the DNA fragment complementary, such that they can anneal. To clone the fragment DNA into the plasmid, the cut plasmid and the DNA fragment are mixed and allowed to anneal, and then the nicks are sealed by DNA ligase. The new plasmid is then transformed into a host, most commonly *E. coli*, where it can be propagated. (b) An example of DNA annealing following cleavage by the restriction endonuclease *Eco*RI is depicted (again the restriction site is in blue).

thus directing the order in which the DNA fragments and vector assemble. DNA polymerase fills in the gapped single-stranded regions, and DNA ligase seals the nicks. The mixture is then introduced into bacteria, which are plated on media with antibiotics to select for cells with complete plasmids. An advantage of this method is that it does not require restriction enzymes or the presence of restriction sites at desired locations.

Cloning by recombination requires regions of DNA homology at the ends of the DNA fragment

The ability to amplify DNA by PCR and to introduce desired additional sequences to the ends of the DNA fragment has allowed investigators to introduce fragments into vectors by recombination. Recombination can occur when there is sufficient homology between the ends of the linear DNA and the target site. This method was initially used for research in yeast where homologous recombination between a linear DNA fragment and either the chromosome or a plasmid is very efficient. More recently, methods have been developed to use recombination to insert a gene

The modification of the chromosome by recombination is discussed further in Section 19.6.

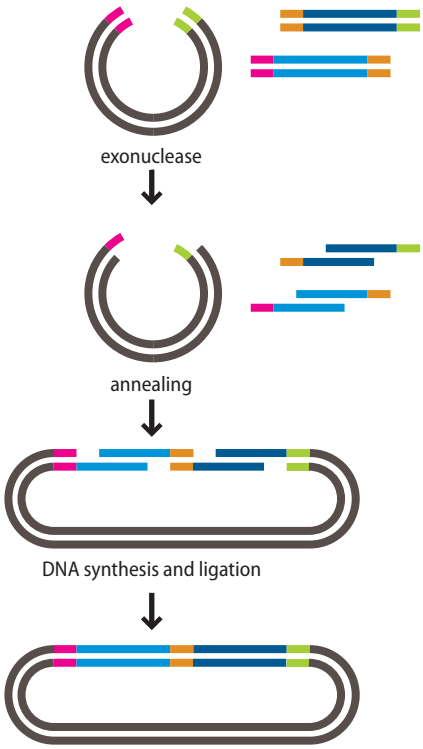


Figure 19.25 Gibson cloning. The DNA fragments to be cloned are designed to be flanked by DNA sequences of 20–40 bp in length that are identical to either the ends of the linearized vector (pink and green regions) or to the ends of the other fragment (orange regions). All DNA fragments are mixed and subjected to an exonuclease that digests DNA from 3′ to 5′, leaving a 5′ overhang. The DNA fragments are then allowed to anneal to each other, leaving single-stranded gaps. These gaps are filled in by a DNA polymerase and then sealed by ligase. Intact plasmids are selected by transformation into bacteria, as in the ligation-mediated cloning described in Figure 19.24.

into bacterial chromosomes or plasmids in cells or into purified DNA in a test tube using purified recombination enzymes. Such an approach is illustrated schematically in Figure 19.26. The advantage of these methods is that they bypass the need for restriction sites at specific locations.

To clone a fragment of DNA by recombination, the DNA is first amplified with primers that introduce an additional sequence at each end of the amplified fragment that is homologous to the target sites in the vector. If the recombination reaction is to happen *in vivo*, a linearized vector and the fragment are transformed into cells and the subsequent recombination reaction produces a recombinant plasmid. The plasmid will also code for a selectable marker that makes it possible to select for cells in which the recombination event took place.

Certain types of recombination reactions can also be performed in a test tube. In this case, the DNA insert and target vector are incubated *in vitro* with purified recombination enzymes, so that the DNA fragment becomes incorporated into the vector. Bacterial cells are then transformed with the plasmid, as in ligation cloning.

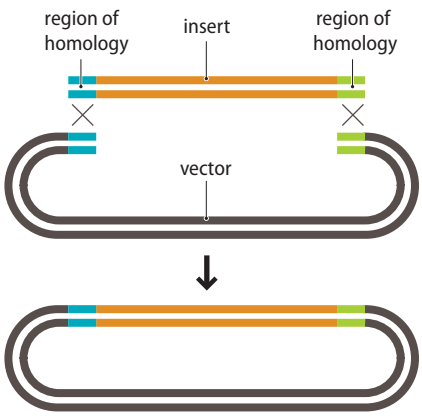


Figure 19.26 Cloning by recombination. DNA is amplified with a set of primers that contain, at their 5′ ends, sequences that are homologous to the vector (in turquoise and green). The recombination reaction can take place in cells, such as yeast cells or bacteria, via a process in which a linearized vector and the fragment are transformed into the cells and the recombination reaction (indicated by X) occurs *in vivo* to produce a recombinant plasmid. The plasmid will also code for a selectable marker (not shown), allowing the identification of cells in which the recombination event took place.

Specific mutations can be introduced by site-directed mutagenesis or PCR

In the examples described previously, the purpose of cloning was to generate exact copies of the DNA or RNA of interest. However, it is also often helpful to change the function of a gene or to inactivate it by introducing mutations at specific sites. These mutations can be introduced after cloning by a method known as **site-directed mutagenesis**, or they can be introduced by a specialized PCR process, known as PCR sewing, and then cloned into the vector of choice.

In site-directed mutagenesis of a cloned gene, sense and antisense oligonucleotides that are complementary to the DNA sequence of interest, with the exception of one or more nucleotide changes corresponding to the desired mutation, are synthesized, as depicted in Figure 19.27. The plasmid is denatured, and each oligonucleotide is hybridized to one of the two strands of the denatured plasmid. These oligonucleotides serve as a primer for DNA polymerase, which synthesizes the remaining DNA strand. The process is then repeated; some plasmids will contain a strand of template DNA and a newly synthesized strand, while some will have two new strands. The reaction mixture will also include some of the original DNA.

To eliminate plasmids that do not contain the mutation on both strands, the mixture is incubated with an enzyme that cleaves plasmids containing one or both of the original DNA strands (that is, strands that were not synthesized in the test tube). The original DNA strands are methylated at the adenines of GATC sites, thanks to an enzyme present in many *E. coli* strains. This methylation is absent from the newly synthesized strands because the methylase responsible for this modification is not present in the reaction mixture. Plasmids containing the

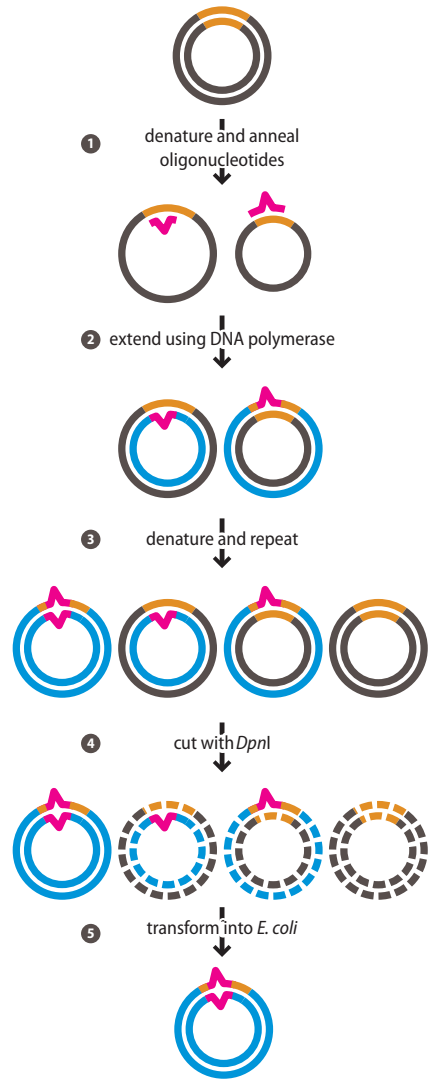


Figure 19.27 Introduction of specific mutations by site-directed mutagenesis. In this scheme, oligonucleotides are synthesized whose sequence contains the desired mutations (depicted as pink Vs in the DNA) but are otherwise complementary to the flanking sequence (step 1). These oligonucleotides are annealed to the denatured plasmid and serve as primers for DNA replication, which extends the primers and replicates the entire plasmid (step 2; new strands are shown in blue). This process is then repeated to generate a population of plasmids containing the mutation on both strands (step 3). The plasmids are then digested with the enzyme *Dpn*I, which cleaves plasmids containing one or both strands of the parental DNA (in black), which are methylated at the sequence GATC (step 4). When bacteria are then transformed with the reaction mixture, the intact circular plasmid will preferentially promote colony formation on selective media (step 5).

methylated bases can then be digested by the restriction enzyme *Dpn*I, which specifically cleaves DNA that contains GATC sequences in which the adenine is methylated on either one or both strands. After *Dpn*I digestion, the only intact plasmids are the ones in which both strands are newly synthesized and so contain the mutation. Since circular, uncut plasmids transform bacteria with high efficiency, the circular plasmids containing the mutation preferentially will be propagated.

PCR can also be used to introduce a specific change into the amplified DNA, as illustrated in Figure 19.28. In this instance, the PCR primer carries the desired mutation such that the products of the PCR will also predominantly carry the mutation. The resulting product can be used in a subsequent PCR reaction in which two PCR fragments are "sewn" together by annealing the ends to each other. The products can then be cloned. In all cases, the resulting clones are sequenced to verify that the desired change has been introduced. We discuss methods for sequencing in Section 19.9.

A library of clones can be made from genomic sequences, cDNA copies of mRNAs, or synthetic DNA sequences

In molecular biology, a **library** is a collection of vectors (usually plasmids or viruses), each containing a different cloned sequence. For example, one can create a library of plasmids containing DNA fragments that, together, span the entire genome of an organism. Libraries are often used in genetic studies—for example, one could isolate a mutant yeast strain with an interesting phenotype without knowing which gene is mutated. By transforming the mutant strain with a library containing all the yeast genes, one can then identify a cloned gene that reverses the phenotype, thus revealing the identity of the altered gene in the mutant yeast strain. Such complementation of a mutant phenotype by a library was used in Experimental approaches 4.3 and 5.1.

A library can be made of total genomic DNA, of expressed genes only, or of a subset of expressed genes (for example, genes that are expressed only under a certain condition). For a genomic library, a genome is fragmented (for example, by mechanical shearing or by restriction enzyme digestion), and each fragment is independently ligated into the plasmid vector so that each plasmid carries a different fragment of the genome.

For a library made of expressed genes, a complementary DNA fragment of each mRNA, known as cDNA, is first made using the reverse transcriptase enzyme, as described earlier in this section. A **cDNA library** featuring all mRNAs expressed in a certain cell type or tissue can be produced; such libraries make it possible to identify and study genes expressed in certain biological contexts—for example, from a particular human tissue.

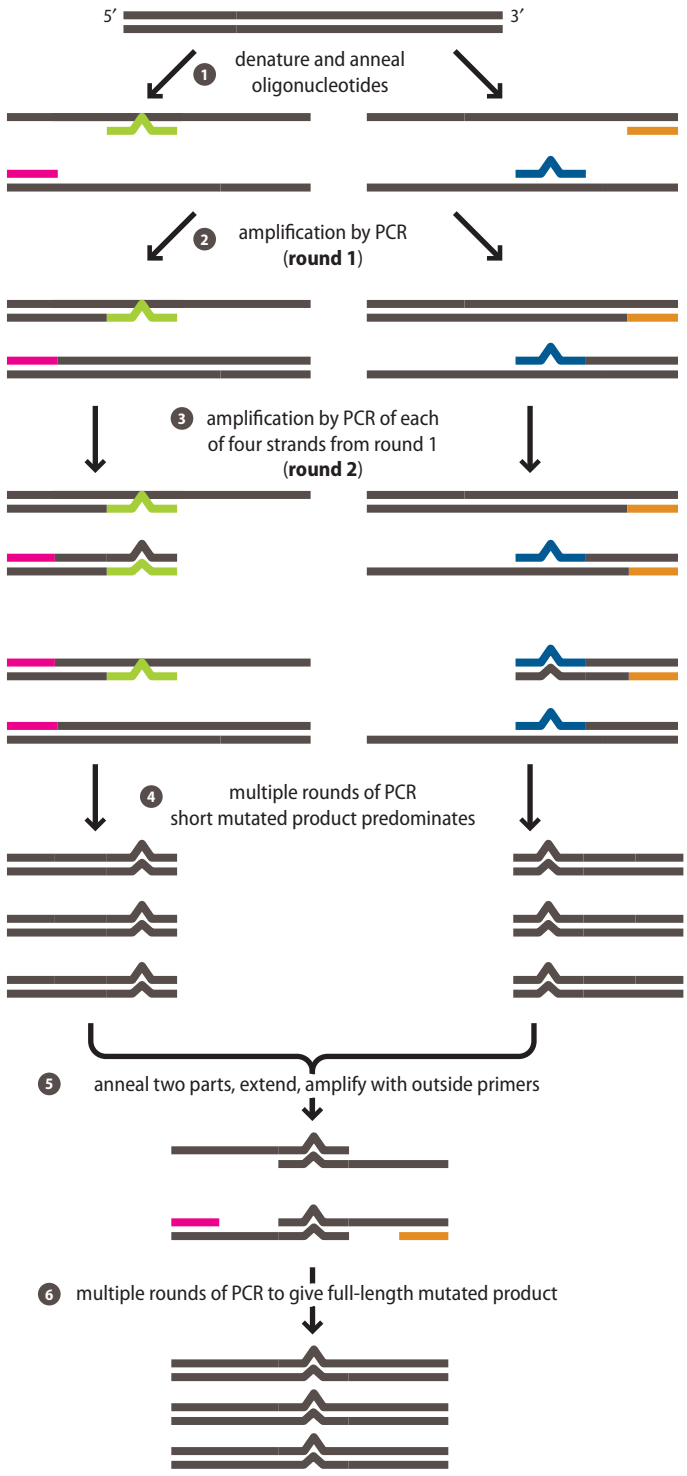


Figure 19.28 Introduction of specific mutations by PCR. In this scheme, the mutagenesis reaction takes place as part of a PCR amplification process. The first step of the mutagenesis reaction takes place in two PCR reactions, one amplifying the left half of the DNA fragment using a terminal (pink) primer and an interior mutagenic primer (green) and the other amplifying the right half of the DNA fragment using an interior mutagenic primer (dark blue) and a terminal (orange) primer (steps 1 to 4). The green and blue primers are complementary to each other and carry a mutation in the same base pair. By incorporating the mutation into the oligonucleotide used in the PCR reaction, the majority of the amplified products will eventually carry the mutation in both strands, similar to adding sequences to the end of PCR products described in Figure 19.20. Once the two half-reactions are completed, the products are used in a "sewing" PCR reaction, as shown (step 5); because the blue and green oligonucleotides were complementary to each other, the ends of the two PCR products can anneal as shown and the DNA can be amplified using the terminal primers (orange and pink) (step 6). The end result is a DNA fragment that is like the starting material, except that it carries a mutation at the desired position.

More recently, investigators have designed libraries that express small hairpin RNAs (shRNAs), whose function is to inactivate gene expression in mammalian cells via RNAi pathways. These libraries contain hundreds of thousands of clones, each expressing a short RNA that can fold onto itself and direct the RNAi machinery to degrade the complementary mRNA. These libraries are generated using many different short palindromic synthetic oligonucleotides (of 40–50 bases) that are cloned into plasmids or viral vectors. Special computer programs are used to

ensure that these libraries cover a certain portion of the genome, and the libraries typically contain several different clones for each gene.

Researchers have used shRNA libraries to screen for shRNAs that, when present, lead to a particular phenotype—for example, resistance to HIV infection. The researcher can then identify the sequence on the clone that conferred resistance, which is then used to identify the gene that was targeted by the clone. The use of this kind of library is described in Experimental approach 17.3.

We learn more about RNAi pathways in Chapter 13.

19.5 UNDIRECTED GENOME MANIPULATION

In Section 19.4, we learned how mutations are introduced into cloned regions of DNA. However, when a gene is expressed from a plasmid or a virus, its expression level is often different from its normal level. In addition, to best eliminate gene function, one must inactivate the chromosomal copy (or copies, in a diploid organism) of the gene. Scientists therefore often want to introduce mutations into a genome.

Mutations in the genome were traditionally identified by an approach now referred to as **forward genetics**, in which the researcher generates a mutant that is defective in a particular process (or uses an existing mutant) and then seeks to identify and understand the gene(s) responsible for the observed phenotype or disease. For example, to study the process of DNA repair, researchers predicted that mutants defective in this process would be sensitive to radiation. Cells were treated in a way that produced random mutations in the genome and screened for mutants that were sensitive to radiation. Regardless of the phenotype, once mutants with the desired phenotype are in hand, one of a variety of methods could be used to identify the mutated genes—for example, by complementing the phenotype with a genomic library.

We discuss genome libraries in Section 19.4.

Another approach to studying gene function is **reverse genetics**. In this case, the gene is already in hand, and the researcher seeks to understand its function by altering the gene's sequence and analyzing the phenotypic changes that result when the altered sequence is expressed *in vivo*. For example, a previously unstudied protein may be identified through its physical interaction with a known protein. To study the new protein, a researcher can delete the specific gene coding for this protein and examine how this affects cell function.

Sometimes it is also desirable to introduce a specific gene or specific mutations into a genome in a non-random fashion. The introduction of specific genes into the genome of an organism is denoted **transgenesis** and results in the generation of a **transgenic organism**.

In this section, we will discuss several different methods for generating random mutations, including chemical mutagenesis and transposition. In Section 19.6, we will discuss methods for eliminating the function of, or mutating, specific genes.

Mutations can be introduced by radiation or chemicals that damage DNA

As discussed in Chapter 15, mutations can be introduced more or less at random by the action of radiation or DNA-damaging chemicals. Thus, one of the simplest ways of introducing mutations is to irradiate with ultraviolet (UV) light or to expose the organism to agents such as ethyl methanesulfonate or ethylnitrosourea. The advantage of these types of mutagens is that they often introduce point mutations that can change the function of the gene product, instead of simply inactivating the gene (for example, due to a large deletion), which could result in cell death.

This type of mutagenesis brings with it the challenge of identifying the DNA sequence change in the context of a large genome (for example, that of the mouse). However, this limitation is rapidly being overcome by advances in DNA sequencing technology that make it possible to identify the mutation by sequencing the whole genome, making it likely that mutations induced by radiation or chemicals will remain a widely used tool in genetic screens.

Transposons can be exploited for mutagenesis and transgenesis

We learn more about transposons in Chapter 17.

Transposons were discovered because of the way their movement resulted in mutagenesis, since the insertion of a transposon into a coding sequence disrupts gene function. This property of transposons has been exploited in several genetically tractable model organisms to carry out insertional mutagenesis.

Transposon mutagenesis is most often carried out using a two-component system, where one component is a regulated source of transposase and the other component is a "mini" transposon consisting of a selectable marker gene flanked by the transposon end sequences necessary for transposition, as illustrated in Figure 19.29. These elements are typically placed on separate plasmids that lack sequence elements necessary for efficient DNA replication or segregation and are thus eventually lost over multiple generations. Once the transposable element inserts into the target DNA, the plasmid carrying the transposase is eventually lost, at which point the inserted element is stably integrated and cannot transpose again.

The plasmids or viruses containing the transposon are introduced into the cell by transformation or infection. Cells containing the transposon are isolated by screening for the selectable marker encoded by the integrated transposon, which may be an antibiotic resistance gene or a visual marker such as a fluorescent protein.

Transposons are a particularly useful experimental tool because the DNA that flanks the transposon can be isolated and sequenced, with the transposon providing a molecular tag to identify the site of the insertion mutation. Some of the transposons that are the most widely used for mutagenesis are listed in Figure 19.30.

In addition to generating mutations, transposon insertions can be used to obtain information about the activity of specific promoters or enhancers, or to artificially activate a gene. For example, a transposon can contain a reporter gene whose expression can easily be assayed, as depicted in Figure 19.31. When the

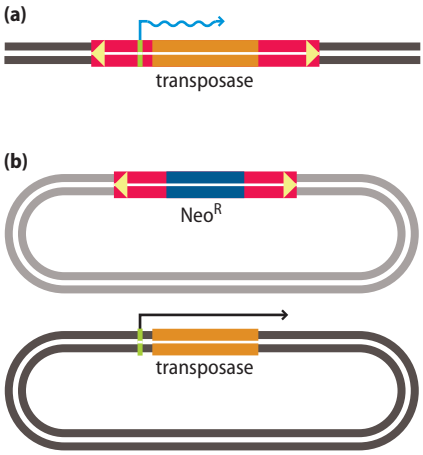


Figure 19.29 Two-component transposon system. (a) The structure of a transposon as found in nature, with its terminal recombination sequences (yellow triangles) and transposase gene (orange). (b) In a two-component transposition system, one component is a "mini" transposon made of transposon ends flanking a cargo DNA segment, here the gene encoding Neo^R (which provides resistance to the drug neomycin). The other component is the transposase gene under the control of a regulatable promoter. When a cell contains both components, induction of transposase expression mobilizes the mini transposon.

Some transposons commonly used as tools

organism	commonly used transposons
bacteria	*Tn5*
plants	*Ac*
insects	P elements, *piggyBac*, *hobo*
fish	*Tol2*
mammals	*Sleeping Beauty*, *Tol2*, *piggyBac*, *Minos*

Figure 19.30 Transposons commonly used as tools.

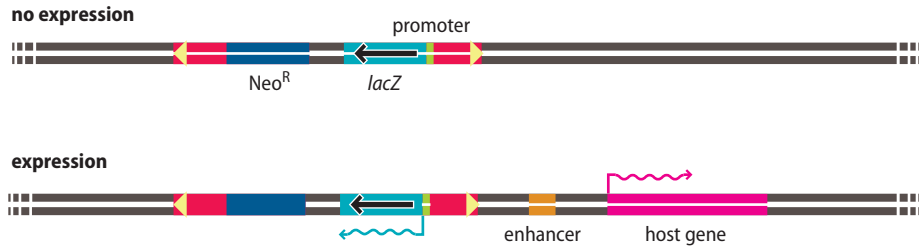


Figure 19.31 An enhancer–sniffer transposon. The transposon contains a selectable marker (to confirm transposon insertion into the genome) and a reporter gene (*lacZ*) with only a weak basal promoter such that it is not expressed in the absence of an enhancer. If the transposon inserts at a chromosomal region lacking a nearby enhancer (top panel), the *lacZ* reporter will not be expressed. If, however, the transposon inserts adjacent to the enhancer of a host gene (bottom panel), the reporter is expressed similarly to the host gene.

transposon is inserted in the vicinity of a promoter or enhancer element, the expression of the reporter gene is under the control of the neighboring promoter or enhancer. By assaying the relative levels of the reporter gene product, it is possible to obtain a measure of the level at which the endogenous gene is expressed. This approach is sometimes referred to as an **enhancer trap**.

A converse strategy is to activate endogenous genes using a transposon that contains a strong promoter. When the transposon containing the promoter is inserted into the chromosome, transcription of nearby genes can be controlled by the inserted promoter, rather than by the endogenous promoter.

Transposons can be used not only to disrupt genes, as we have just discussed, but also as vectors for inserting genes of choice into cells or organisms. Transposons are thus being tested as vehicles for so-called "gene therapy", in which a wild-type copy of a gene is introduced into the host chromosome in an effort to cure disease.

We explore mutagenesis of the human genome by an endogenous transposable element in Experimental approach 17.1.

Plant genomes can also be modified by mobile genetic elements

Mobile DNA elements have been the major tool for manipulating the genomes of those plants that can be infected by the bacterium *A. tumefaciens*. Such infection results in the transfer of a DNA segment called the T-DNA from a specialized Ti plasmid into the plant genome, as shown in Figure 19.32. The virulence region of the Ti plasmid encodes the proteins that act on sequences at the ends of the T-DNA segment to transfer the DNA from the bacterium to the plant.

In nature, the T-segment encodes plant hormones that stimulate the plant cells to grow in unregulated fashion as "crown gall tumors;" these "tumors" produce compounds that the bacterium can metabolize for growth. For genome engineering, the tumor-inducing genes inside the T-DNA are replaced with a selectable marker, such as an antibiotic resistance gene, and any other DNA of interest (for example, a gene expressing a plant protein fused to GFP). Another useful strategy is to use the T-DNA to deliver a transposon into the plant genomic DNA, which can be subsequently remobilized for further mutagenesis, or to express small inhibitory RNAs (siRNAs) to inhibit targeted genes.

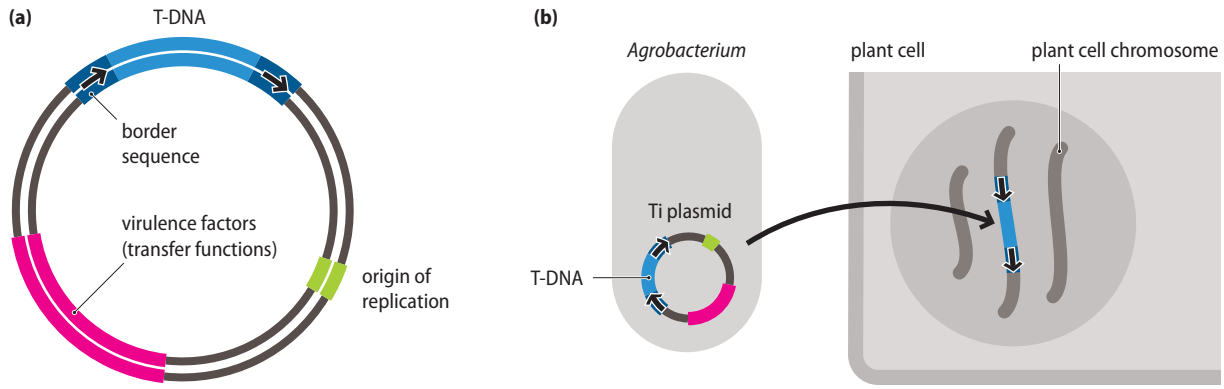


Figure 19.32 Use of T-DNA for modification of plant genomes. (a) The Ti plasmid from *A. tumefaciens* includes a bacterial origin of replication (green), the T-DNA, which is composed of two border sequences (dark blue) and the DNA that lies in between (light blue), and the virulence region (pink) that codes for proteins that copy and move the T-DNA to the plant cell. (b) Insertion of the T-DNA segment into the plant genome occurs when the proteins coded by the bacterial virulence region copy the T-DNA segment and move it to the plant cell, where it is integrated into the genome. When used for creating transgenic plants, the natural T-DNA segment is replaced with the DNA of interest.

19.6 DIRECTED GENOME MANIPULATION

We discuss homologous recombination in more detail in Chapter 16 and encounter its use when discussing cloning in Section 19.4.

In addition to randomly mutagenizing a genome, researchers often want to modify the function of a gene or the genome in a precise manner. For example, an experiment might call for a specific gene function to be eliminated, for a specific genomic sequence to be replaced, or for an exogenously supplied piece of DNA to be added. In some cells and organisms, scientists turn to approaches such as RNAi for "knocking down" the expression of specific genes, sometimes using the libraries described in Section 19.4. However, it is becoming increasingly common to delete or modify specific regions of the genome by homologous recombination. We will discuss these approaches for directed manipulation of gene expression and genomes in this section, particularly the use of CRISPR systems to target recombinases or other protein complexes to specific sequences. Throughout the section, we need to be cognizant that none of these approaches are 100% accurate.

The expression of genes can be disrupted by siRNAs

We learn about the double-stranded RNA fragments called siRNAs (21–22 bp), which can trigger the degradation of mRNA containing a region of complementarity to the siRNA, in Chapter 13.

Rather than directly disrupting a gene at the DNA level to study its function, the RNAi pathway discussed in Chapter 13 can be exploited to target particular mRNAs for degradation, thereby reducing expression of the targeted gene. In Section 19.4, we discussed the use of an shRNA library to disrupt gene function. shRNAs are small synthetic RNAs that form hairpins and are processed by the cell's RNAi machinery. It is also possible to use synthetic siRNAs (short double-stranded RNAs that do not require processing) to accomplish the same goal. While shRNAs can be engineered to be stably expressed within a cell or an organism, siRNAs are mostly used in tissue culture where they are transfected into cells and transiently inhibit expression. Both siRNA and shRNA libraries are now commercially available for every gene in the human and mouse genome.

A caveat for using RNAi to study gene function is the possibility of "off-target" effects, whereby genes that are related in sequence are also targeted by the same RNAi construct. If the expression of more than one gene is affected by a given RNAi construct, it can be difficult to determine which inactivated gene contributes to the observed phenotype. Sequences used for siRNA have to be designed carefully to avoid regions that are homologous with other genes. Additionally, if several different siRNAs are used to target the same gene and yield similar results, confidence that the gene does, in fact, influence the phenotype of interest is increased since it is unlikely that the independent siRNAs target the same off-target genes. Finally, the best control is to reintroduce an siRNA-"resistant" version of the targeted gene, in which the RNA sequence that was targeted by the siRNA is altered without changing the protein sequence, into the targeted cells. If this complements the observed phenotype, it is a good indication that the specific gene plays a role in the phenotype.

Specific sequences or gene disruptions can be introduced by homologous recombination

A more permanent way of altering gene expression is by homologous recombination whereby a linear fragment of DNA that has homology to a genomic sequence on both of its ends is introduced into cells and becomes incorporated into the host cell genome by homologous recombination. This process is illustrated in Figure 19.33. This process may be mediated by proteins that are normally found

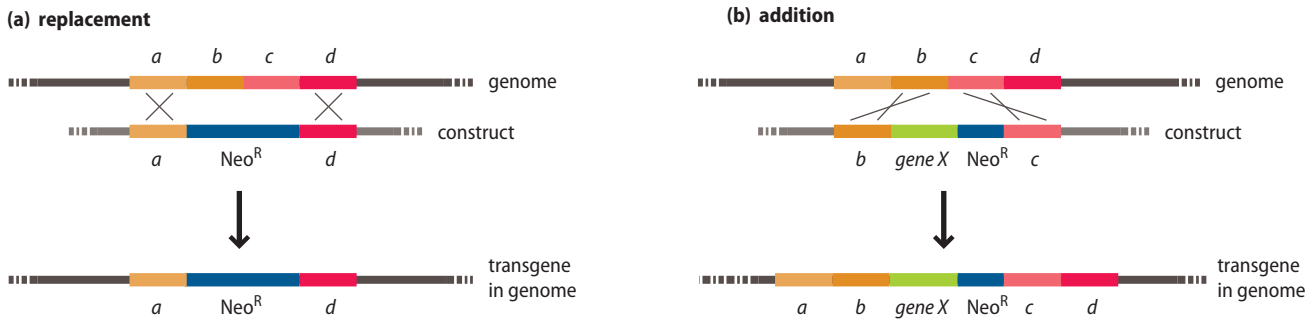


Figure 19.33 Gene replacement and addition by homologous recombination. (a) Gene replacement by homologous recombination. To delete the sequences between the *a* and *d* DNA segments, a linear DNA containing the gene encoding Neo^R (conferring resistance to the drug neomycin) is flanked by sequences that are identical to *a* on one side, and *d* on the other. When this linear DNA is introduced into cells, it recombines with the genomic *a* and *d* sequences, resulting in the substitution of the genomic *b* and *c* regions with the *Neo*^{R} gene. (b) Gene addition by homologous recombination. Let us assume that a researcher wants to express *gene X* in a particular cell type, without removing any cellular DNA sequences. To do so, a linear DNA segment containing *gene X* and a selectable marker (Neo^R) is flanked by sequences homologous to the genomic insertion site (in this case, sequences *b* and *c*). Homologous recombination between the *b* and *c* regions of the linear DNA and the *b* and *c* regions of the genomic DNA results in the insertion of *gene X* and Neo^R into the genome.

in the host cell, or the cell may be engineered to increase the efficiency of the recombination reaction—for example, by introducing components of the CRISPR system, as described later in this section. In some organisms, such as *B. subtilis*, *S. cerevisiae*, and mammalian cells, the ends of the linear targeting DNA are sufficiently recombinogenic that recombination can occur in a wild-type strain. By contrast, other organisms such as *E. coli* need to be modified—for example, by decreasing the nuclease activity of RecBCD to reduce the resection of the linear DNA, and introducing the recombination system from bacteriophage lambda to increase the efficiency of recombination.

RecBCD is discussed in Chapter 16 and in Experimental approach 16.1.

A DNA fragment that is incorporated by recombination can be engineered to contain specific sequences. For example, the DNA fragment often contains a marker gene that allows for selection of cells in which recombination has taken place. The DNA fragment can also contain a unique sequence, often referred to as a bar code, which makes it easy to identify a particular gene replacement. Researchers have taken advantage of this approach to substitute each non-essential yeast gene with the same selectable marker gene, but with a different unique barcode sequence, as illustrated in Figure 19.34. The barcode sequences make it possible to identify each different knockout strain uniquely in a pool of mutants, now most commonly by DNA sequencing.

Homologous gene replacements have also been quite useful for examining the function of genes in mammalian cells in culture. In these cells, DNA can be integrated by both homologous recombination and non-homologous end joining (NHEJ). Often it is important to distinguish between cells in which homologous targeted integration has occurred and those in which NHEJ has occurred. A common strategy is to include two markers on the targeting DNA. As depicted in Figure 19.35, in this strategy, one marker will be retained and the other will be lost when the desired homologous recombination occurs, while both markers are retained upon random integration.

NHEJ is described in Chapter 16.

Knockout mice are generated using homologous recombination

Mice in which one or more genes have been disrupted, known as knockout mice, have become important tools for studying the role of particular genes in mammals. Knockout mice are also generated through a procedure that relies on homologous

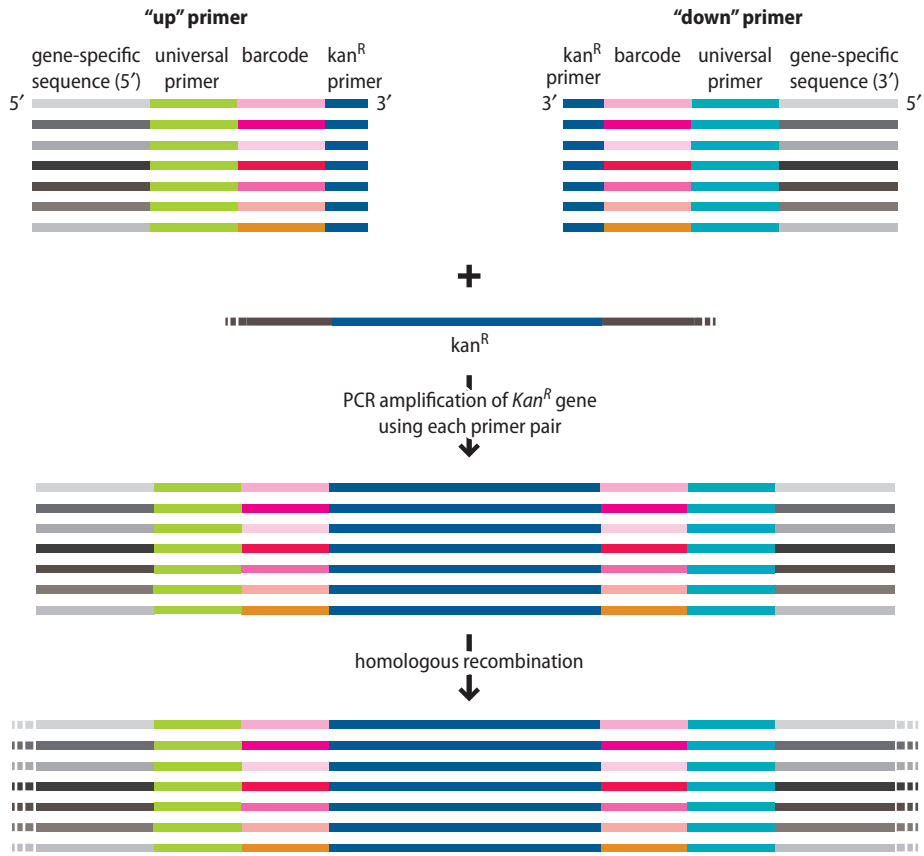


Figure 19.34 Gene disruptions (knockouts) of *S. cerevisiae* genes. The disruption of each non-essential gene in the yeast *S. cerevisiae* deletion collection requires a pair of DNA primers, referred to here as an "up" primer and a "down" primer. Each primer contains (starting from the 5′ end): a gene-specific sequence that is immediately upstream (for the up primer) or downstream (for the down primer) of the coding region (shades of gray); a universal primer, one for all "up" primers and another for all "down" primers (light green and light blue); a barcode sequence that is unique to each primer (shades of pink); and primers to amplify the *kan*^{R} gene (in blue). Each pair of primers is used in a PCR reaction to amplify the *kan*^{R} gene (conferring resistance to kanamycin), creating a collection of DNA fragments, each homologous at its ends to the DNA regions flanking the coding sequence of a specific gene. Homologous recombination between these gene-specific sequences and the chromosomal sequences results in the replacement of the gene with a *kan*^{R} fragment flanked by universal primers and two barcodes specific to the targeted gene. The strains of interest (for example, those that are able to survive a particular treatment) can then be identified by virtue of the unique barcodes associated with their *kan*^{R} cassette.

From Meneely, P. (2009). *Advanced Genetic Analysis: Genes, Genomes, and Networks in Eukaryotes*. Oxford: Oxford University Press.

recombination to re-engineer mouse embryos, as illustrated in Figure 19.36a. The gene of interest is first replaced with a selectable marker by homologous recombination in mouse ES cells in culture, as described earlier.

For the example shown in Figure 19.36, we will assume the ES cell is derived from a white mouse. The ES cells are placed in a selective medium so that only the cells containing a correctly targeted selectable marker will survive. The transformed cells are then injected into a mouse embryo from a black-coated mouse, which is at the blastocyst stage of development and hence contains several hundred cells. The ES cells become incorporated into the blastocyst and will contribute to the tissues of the animal that develops. The blastocyst becomes implanted in the uterus of a black female mouse. Blastocysts that are naturally in the mother will give rise to black mice, while transplanted blastocysts that contain engineered ES cells will give rise to chimeric mice with a mixed black-and-white coat color, depending on which cells in the injected blastocyst contributed to a particular skin patch. Such a chimeric mouse is shown in Figure 19.36b.

The use of mice with different coat colors to generate the ES cells and to act as a source for the blastocysts makes it possible to identify progeny mice in which the ES cells contribute to the tissues. If the engineered ES cells contribute to the germ line, these mice can be crossed, first to a wild-type mouse to generate heterozygous mice, and then to themselves to generate homozygous mice in which all cells contained the gene knockout.

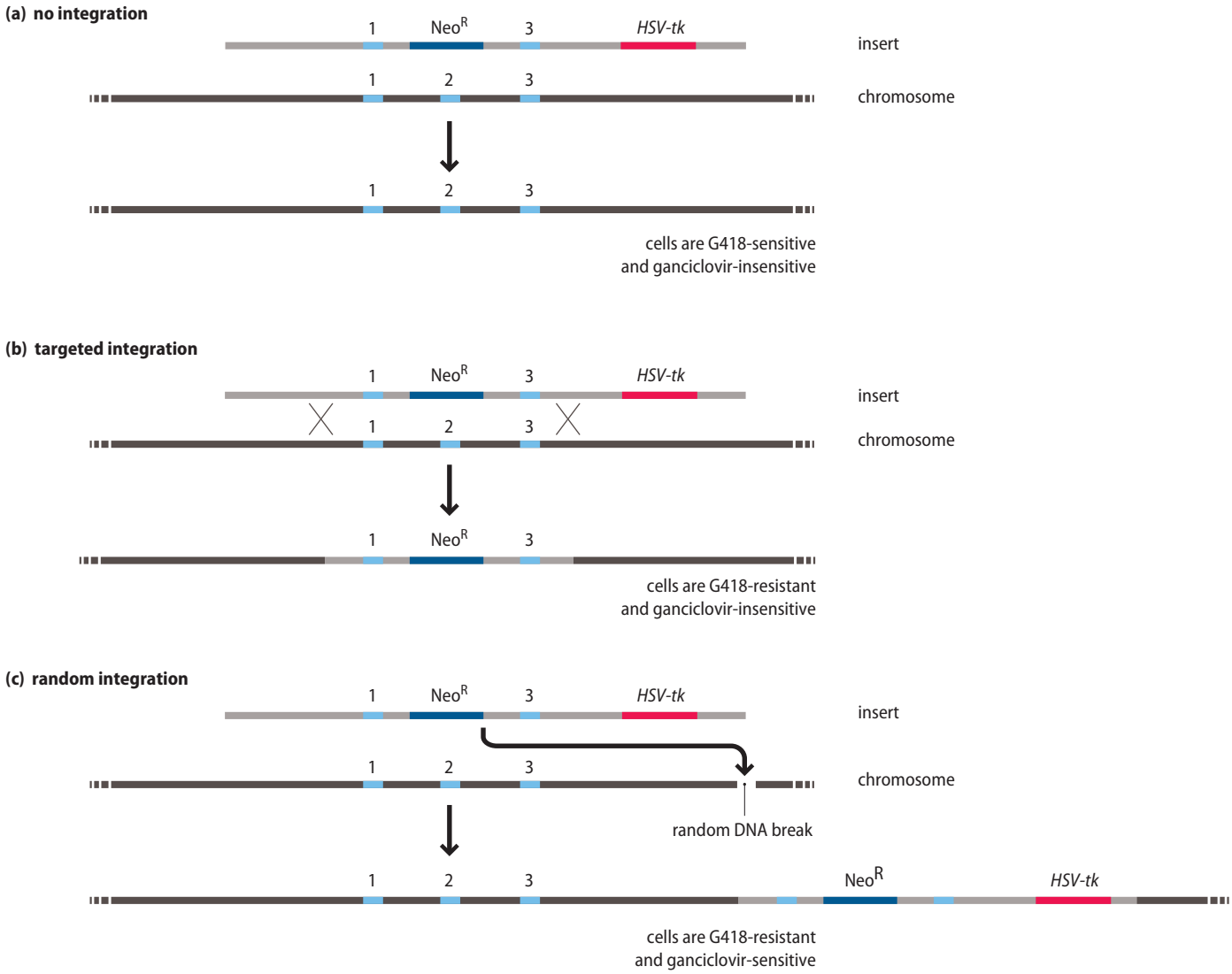


Figure 19.35 Characterization of gene integration in mammalian cells. The insert used in this example contains the gene for Neo^R, which confers resistance to G418, and *HSV-tk*, which confers sensitivity to the drug ganciclovir. The purpose here is to replace chromosomal DNA segment "2" with the *Neo*^{R} gene. (a) When the construct is not integrated, cells remain sensitive to neomycin and insensitive to ganciclovir. (b) Integration by homologous recombination at the desired site results in cells that are resistant to neomycin and insensitive to ganciclovir. (c) Random integration of the insert by NHEJ results in cells that are neomycin-resistant and sensitive to ganciclovir.

From Meneely, P. (2009). *Advanced Genetic Analysis: Genes, Genomes, and Networks in Eukaryotes*. Oxford: Oxford University Press.

Cre recombinase allows mutations to be generated at specific times or in specific cells

If a gene plays a critical role in cell viability or during development, it may not be possible to study its function in the mature animal by generating a knockout in an ES cell. An example is a gene knockout that causes the embryo to die or cease developing, thereby making it impossible to assess the function of the gene in the mature organism. In this case, a useful alternative strategy is to make a conditional allele of the gene of interest, which allows the gene to be deleted at a chosen time or in a specific cell type. A conditional allele can also be useful in studying the effects of a knockout in just a subset of tissue types, rather than in the whole organism.

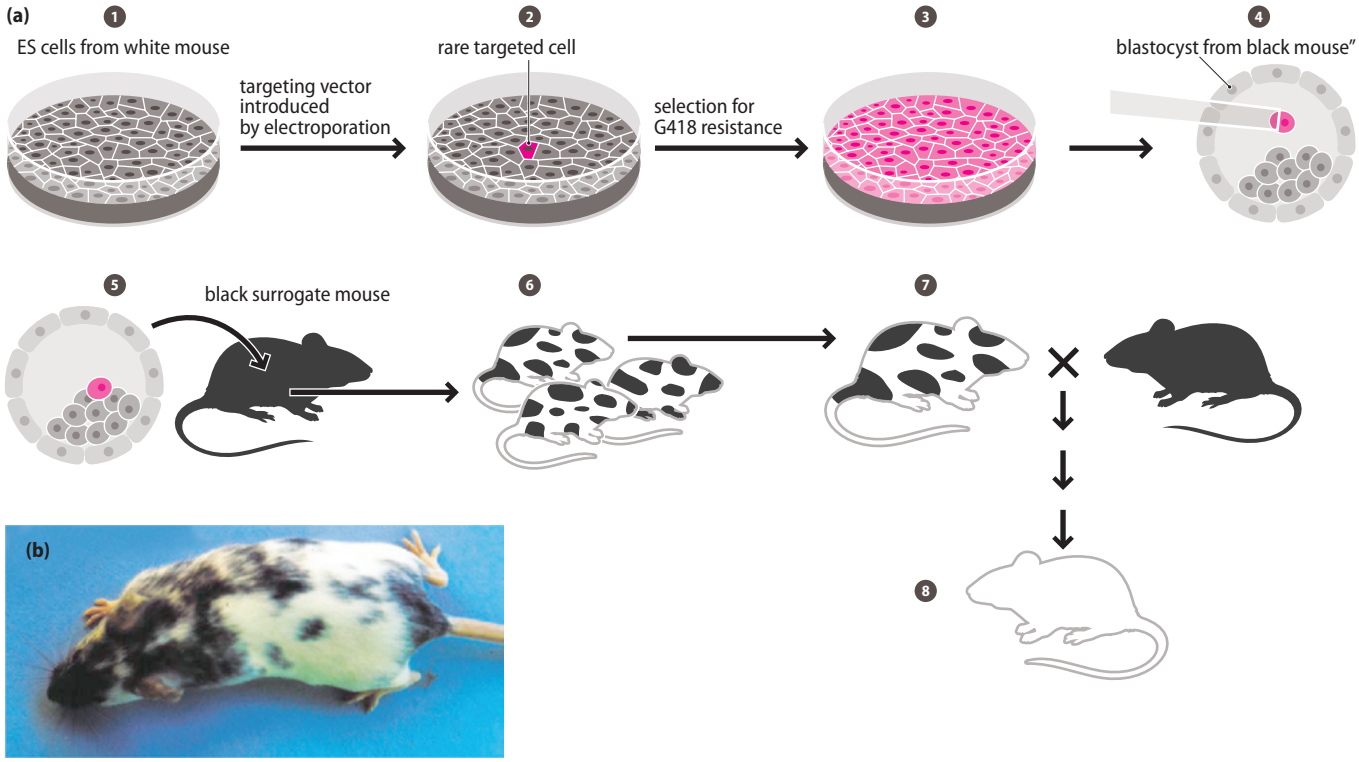


Figure 19.36 Generation of a transgenic mouse. (a) Scheme for generating a transgenic mouse: (1) ES cells from a mouse with white coat color are grown in culture. (2) The targeting vector is introduced into the ES cells by selection for a marker that confers resistance to the drug G418 (pink cell). (3) Cells in which the targeting vector has been correctly incorporated are grown as a pure population by selection in media containing G418. (4) The ES cells are introduced into a blastocyst from a mouse with a black coat color. (5) The blastocyst containing the targeted ES cells is injected into a female surrogate black mouse with black coat color. (6) Chimeric pups that contain white fur from the ES cells and black coat color from the mother are born. (7) Chimeric mice are backcrossed to a mouse with black fur in an effort to identify mice in which the ES cells gave rise to the germ line. (8) White coat color mice derived entirely from the ES cells are obtained by crossing chimeric pups with the mutations in the germ line to each other. (b) Photograph of a chimeric mouse, with patches of both white and black fur.
From Meneely, P. (2009). *Advanced Genetic Analysis: Genes, Genomes, and Networks in Eukaryotes*. Oxford: Oxford University Press.

A common method for producing conditional knockouts relies upon the site-specific Cre recombinase. As we learned in Section 17.13, the Cre recombinase mediates recombination between DNA sequences called *loxP* sites. When the *loxP* sites are incorporated in a direct orientation (meaning oriented in the same way in the DNA), recombination between these sites results in deletion of DNA that lies between the *loxP* sites, as depicted in Figure 19.37a. By controlling whether or not Cre recombinase is expressed at a particular time or in particular tissues, it is possible to control when and where the gene is deleted.

To generate a conditional knockout mouse, it is necessary to generate two different transgenic mice: one bearing the gene of interest flanked by *lox* sites (this allele is referred to as "floxed", for "flanked by lox"), and the second carrying the *Cre* gene under an inducible or a tissue-specific promoter. First, a transgenic mouse carrying the floxed allele in the desired position is constructed, as described previously (see Figure 19.36). The mouse that is homozygous for the floxed allele is then crossed with a transgenic mouse carrying the *Cre* gene under an inducible promoter (for example, a heat shock promoter) or under a tissue-specific promoter; this strategy is depicted in Figure 19.37b. The floxed gene will thus be deleted only in tissues expressing Cre recombinase.

The use of the Cre-*loxP* system is not limited to mice and has been used in other model organisms such as yeast, *Drosophila*, and *C. elegans*.

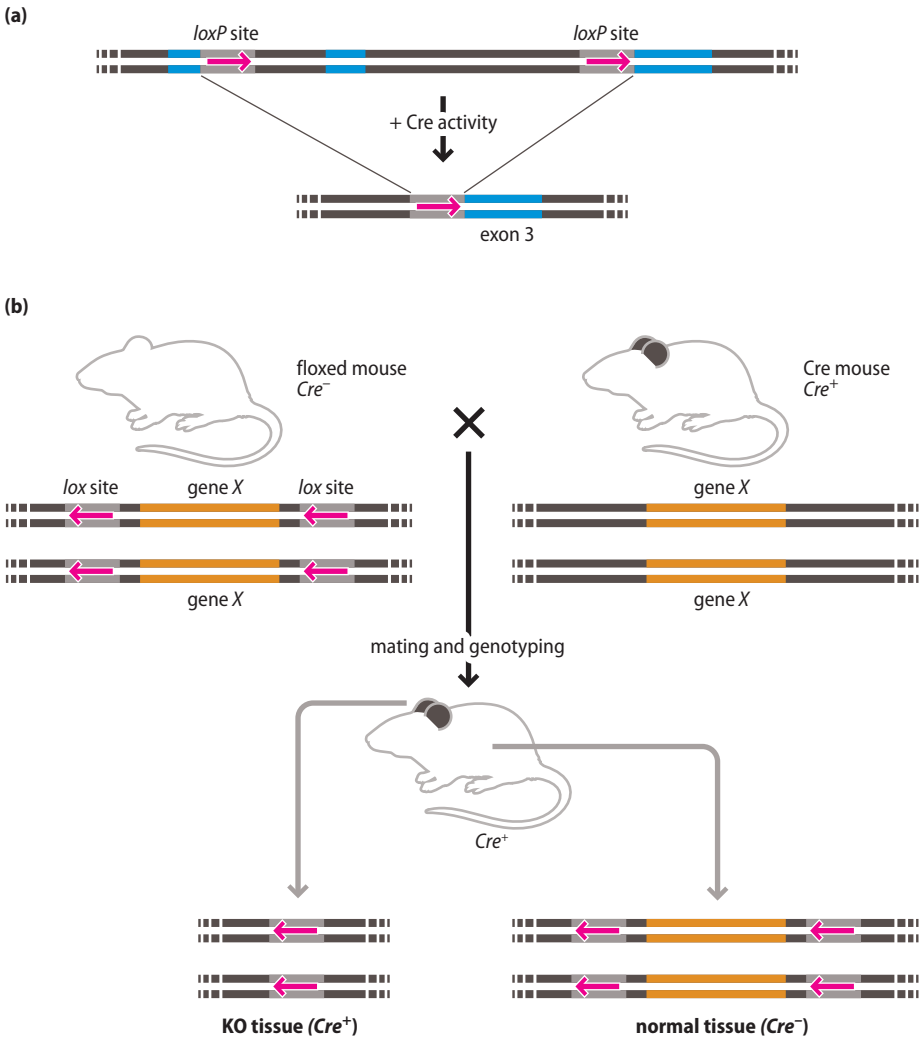


Figure 19.37 Using Cre-*lox* recombination. (a) The principles of Cre-*lox* recombination. The substrate DNA contains a *loxP* site (pink arrows) in the intron between exons 1 and 2 and another *loxP* site in the intron between exons 2 and 3. Cre-mediated recombination results in deletion of exon 2. (b) Scheme for making a conditional deletion using Cre-*lox* recombination. A homozygous transgenic mouse carrying gene *X* flanked by *loxP* sites (floxed gene *X*), but not expressing the *Cre* gene, is crossed with a mouse expressing Cre in a tissue-specific manner (in the ears in the example shown), but carrying wild-type copies of gene *X*. After a couple of rounds of mating, a mouse homozygous for floxed gene *X* and the *Cre* gene is generated. Note that this mouse will have intact copies of gene *X* throughout its body, except where Cre is expressed (in the ears). In this way, it is possible to study the role of gene *X* in a specific tissue, while the rest of the mouse has a wild-type genotype.

From Meneely, P. (2009). *Advanced Genetic Analysis: Genes, Genomes, and Networks in Eukaryotes*. Oxford: Oxford University Press.

CRISPR systems can create double-strand breaks to generate deletions and increase homologous recombination

Recently, the ability to efficiently introduce targeted DNA double-strand breaks has dramatically increased through the application of CRISPR (for clustered regulatory interspaced short palindromic repeats) systems. As we discussed in Section 13.7, CRISPR systems are part of the cellular defense against foreign nucleic acids in bacteria and archaea. In these systems, cells recognize foreign DNA and insert a piece of it as a spacer (initially termed a protospacer) into a cluster of palindromic repeat sequences in the chromosome that together comprise a CRISPR array. This array serves as a form of acquired memory of prior encounters with foreign DNA. When these segments are transcribed, the resulting RNA is processed into short crRNAs, which correspond to each of the spacers. These crRNAs then associate with a specific nuclease for targeting to foreign DNA (or, in a few cases, RNA) that shares homology with the spacer sequences.

A great variety of CRISPR systems have been found in different bacterial and archaeal species, some with multiple proteins required for RNA-directed targeting to the foreign nucleic acids and others where just one protein, such as Cas9,

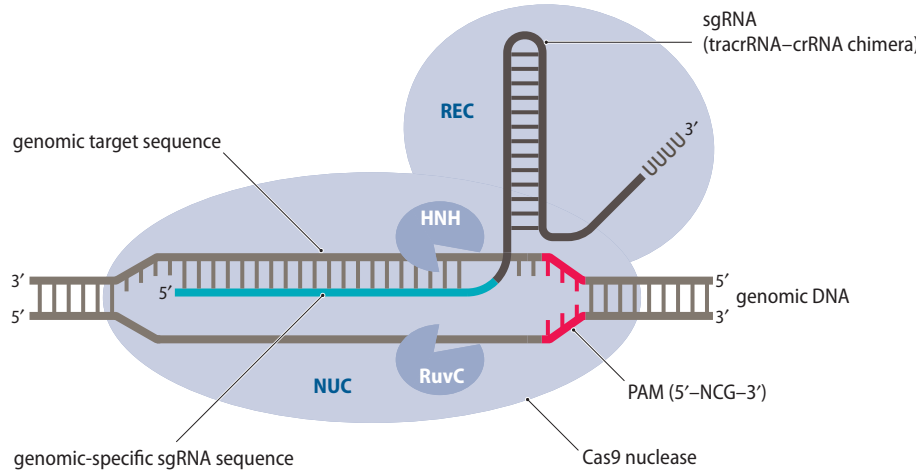


Figure 19.38 sgRNA:Cas9 complex generates double-strand DNA breaks. Model of sgRNA (which is a chimera of the tracrRNA and crRNA) base-paired with target DNA in complex with the Cas9 protein. For cleavage to occur, the target DNA sequence must have an adjacent PAM motif. The Cas9 protein has two nuclease domains denoted HNH and RuvC.

carries out all of the steps in targeting. Cas9 acts in conjunction with only two small RNAs; one is the crRNA and the other the tracrRNA, a *trans*-activating RNA that base-pairs with the 3′ end of the crRNA and is required for crRNA processing, as well as the binding and activation of the Cas9 nuclease. The Cas9 nuclease is fully active when the 3′ end of the crRNA is fused to the 5′ end of the tracrRNA to give a single guide RNA (sgRNA) and thus becomes a system that only requires one protein and one RNA, as shown in Figure 19.38. Researchers have taken advantage of the sgRNA:Cas9 and other systems with limited components to simply and effectively target specific DNA sites in heterologous genomes.

One additional feature required for sgRNA:Cas9-mediated cleavage of target DNA is a short 2- to 6-base pair sequence termed the PAM (protospacer adjacent motif) sequence that must be adjacent to the genomic DNA being targeted. For example, in *Streptococcus pyogenes*, the PAM is NGG. So, practically speaking, when choosing a site for cutting in the genome, one must design an sgRNA that lies adjacent to a PAM sequence in the genome. (That PAM sequence must be NGG if *S. pyogenes* Cas9 is employed.) As shown in Figure 19.39a, when the sgRNA:Cas9 complex binds to the DNA complementary to the crRNA portion of the sgRNA with an adjacent PAM sequence, Cas9 generates a double-strand DNA break.

In the absence of any added DNA, the double-strand break will be repaired by NHEJ. This process often involves the loss of a few nucleotides at the break site, thus generating mutations at a desired location. If, however, linear DNA with homology to sequences on either side of the break is present, repair may take place by homologous recombination, thereby resulting in the insertion of the DNA sequence of choice at a specific chromosomal location. This CRISPR–Cas9 system has been used to introduce double-strand breaks in a wide variety of organisms, ranging from bacteria to plants, mice, and humans.

Given the ease of designing sgRNAs capable of hybridizing to almost any desired sequence, a variety of applications have been developed using sgRNAs and Cas9 as tools. For example, two different sgRNAs capable of targeting different chromosomal regions can be expressed simultaneously, leading to double-strand breaks in different regions of the chromosome whose repair can lead to chromosomal rearrangements or large deletions, as illustrated in Figures 19.39b and c.

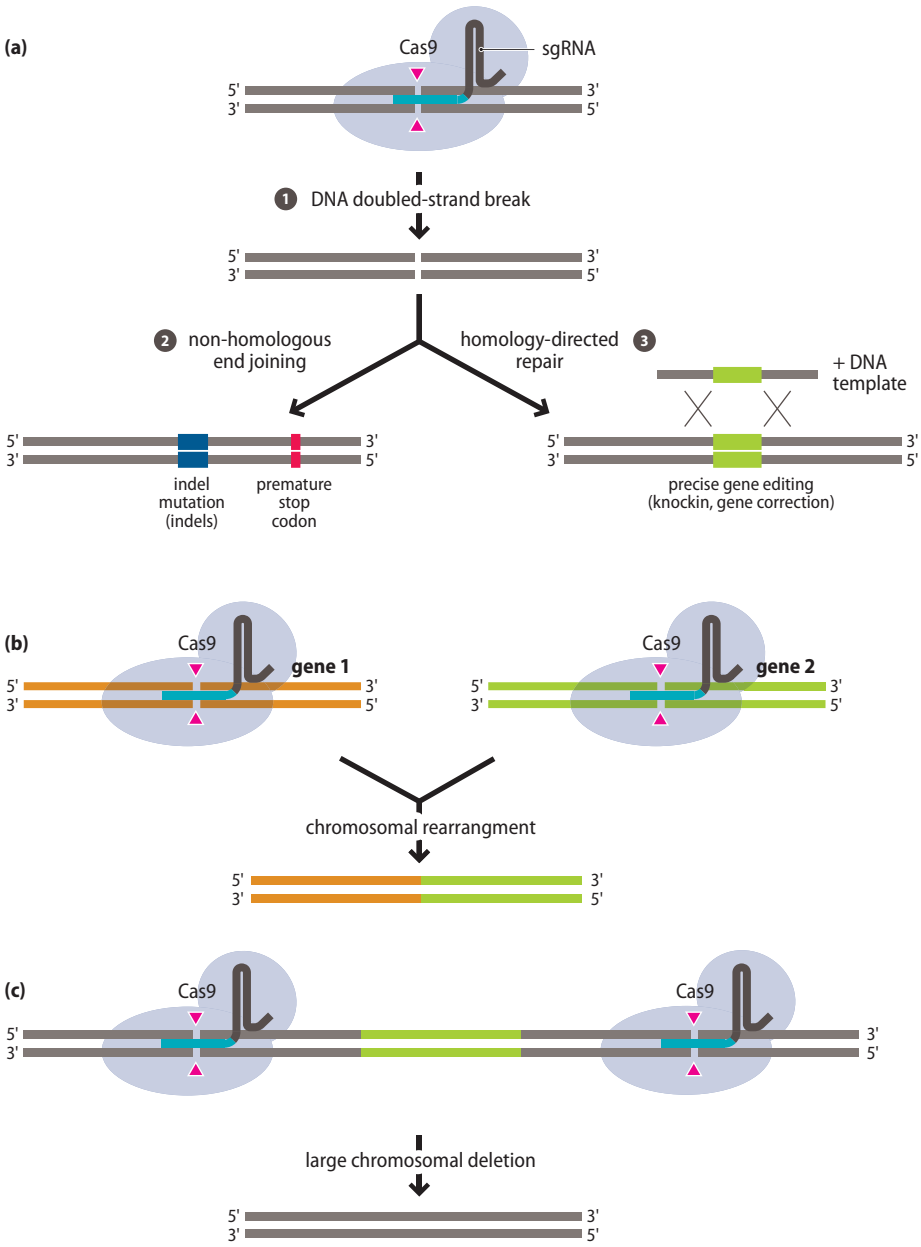


Figure 19.39 Using CRISPR–Cas9 to generate double-strand DNA breaks. (a) When the sgRNA:Cas9 complex recognizes the DNA site via base-pairing between the sgRNA and the complementary DNA sequence, the Cas9 protein introduces a double-strand DNA break (step 1). The double-strand break can be repaired by either the more mutageneic NHEJ (step 2) or by homologous recombination (indicated by the two Xs), if a DNA fragment with homology to either side of the break is present (step 3). If two sgRNAs with homology to different regions of a chromosome are introduced, sgRNA:Cas9-directed cleavage can lead to chromosomal rearrangements (b) or large deletions (c).
From Heidenreich and Zhang 2016 *Nat. Rev. Neurosci* **17**:36.

As in the case of shRNA libraries, libraries of sgRNAs capable of targeting thousands of different genes have been generated. These libraries can be used to select for specific cellular phenotypes that arise when individual genes are knocked out. For example, one could target 100 000 genes in a pool of cells where each cell has one gene knocked out. If the entire pool of cells is sequenced before and after exposure to the desired condition, one can determine which mutations confer a growth advantage (these mutations would be enriched) and which mutations affect genes that are required (these mutations would be lost) under the imposed set of conditions. sgRNAs can also be multiplexed to target multiple genomic regions in a cell simultaneously. In this case, Cas9 would be expressed together with several different sgRNAs such that it is possible to change 5–10 different genomic sites in one cell at the same time.

We discuss the use of genome-wide screens using CRISPR in Experimental Approaches 18.1 and 18.2.

Cleavage-defective CRISPR systems can be used to block transcription or to target proteins to specific genomic regions

Cas9 generates a double-strand break in DNA by utilizing two independent nuclease domains (the HNH and RuvC domains indicated in Figure 19.38), each of which cuts one strand of the double helix. Cas9 proteins that inactivate one or both of these nuclease sites have been employed as tools. If only one of the nuclease sites is mutated, the partially active Cas9 can serve as a site-specific nickase that only cleaves one strand. Two single-strand nicks near each other can stimulate NHEJ in that region of the chromosome. Because the adjacent nicks have to be in the same cell, the introduction of partially active Cas9, together with two sgRNAs complementary to nearby regions, has been used as a way to increase targeting specificity. Catalytically dead Cas9 (dCas9) proteins that lack both nuclease activities are also useful as sequence-specific binding proteins capable of delivering specific effectors to specific regions of the genome. As illustrated in Figure 19.40, dCas9 has been used as a road block to stop transcription, has been fused to transcription activator and repressor proteins, as well as chromatin-modifying factors, to modulate transcription and chromatin structure, and has been fused to fluorescent proteins to allow the localization of a specific region of DNA in live cells by microscopy.

Some limitations of CRISPR systems need to be considered

The list of applications for CRISPR-associated proteins and targeting RNAs is continuing to grow. However, again, some caution is warranted. As for siRNAs, CRISPR systems and other targeted recombination approaches can have "off-target" effects. Significant effort is being put into improving the specificity of CRISPR systems, but it is prudent for scientists to verify phenotypes associated with gene mutations by independent experiments and to complement deletions to ensure that the reintroduced wild-type gene restores the original phenotype. It may also be useful to turn off the activity of Cas proteins after the initial cleavage. The desire to turn off Cas activities is a strong factor in driving the characterization of viral anti-CRISPR proteins, which can be powerful tools in inhibiting Cas proteins.

anti-CRISPR proteins are discussed in Section 13.7.

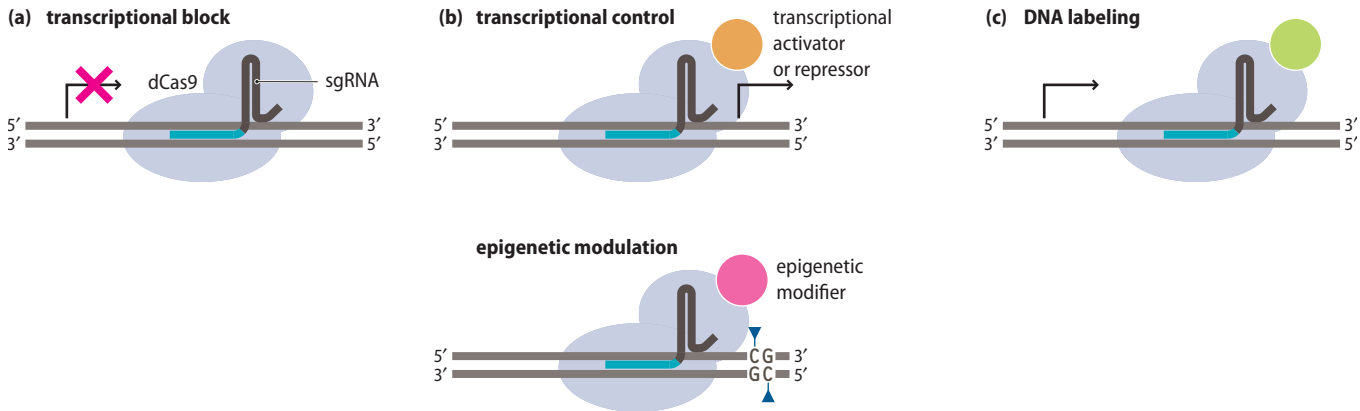


Figure 19.40 Uses of catalytically inactive dCas9. (a) Catalytically inactive dCas9 can serve as a road block to transcription. (b) When fused to a transcription activator or repressor or chromatin-modifying activity, dCas9 can be used to modulate nearby transcription or chromatin structure. (c) dCas9 fused to a fluorescent protein can be used to identify the location of a specific region of DNA in the cell.

Modified from Heidenreich and Zhang 2016 *Nat. Rev. Neurosci* **17**:36.

Another consideration with respect to the application of CRISPR technology is how the system is delivered. For organisms that can be transformed or transfected, the sgRNA and Cas9 (or similar protein) can be expressed in the organism. A parallel approach is to transiently deliver the sgRNA:Cas9 complex. Since the complex is lost from the cell, the transient approach appears to decrease off-target effects and increase the efficiency of gene editing. The situation is more challenging in organisms for which there are barriers to the introduction of DNA or protein complexes. Some approaches being explored for delivery, such as viral infections, can have undesirable secondary consequences.

Finally, the increasing ease with which targeted genome modifications can now be carried out raises ethical questions requiring ongoing discussion about whether these modifications should be carried out in all cells in all organisms.

19.7 DETECTION OF BIOLOGICAL MOLECULES

In order to study or purify cellular components, methods are needed to detect the presence of specific biological molecules and distinguish them from others in a mixture. Some detection methods take advantage of the intrinsic properties of a molecule, such as its biochemical activity or its intrinsic spectroscopic properties. Molecules can also be detected by attaching to them other molecules with particular spectral, chemical, or fluorescent properties that are then readily detectable. Many of these detection methods can also be used to determine the **amount** of a molecule that is present. Such measurements are essential for quantitative studies of the thermodynamic and kinetic properties of molecules. In this section, we review some of the basic methods currently in use for detecting the presence of certain biological molecules. Methods aimed at detecting specific DNA and RNA sequences or specific proteins are discussed in the coming sections.

We note that molecules are most generally studied and followed as a population of molecules and less often as single molecules. Thus, when we refer to the properties of "a protein" or "a DNA fragment," what we most typically mean is the properties of a whole population of proteins or DNA fragments of a certain type.

Cellular components can be monitored by their spectroscopic properties

Biological molecules absorb electromagnetic radiation over a range of wavelengths specific to the molecule of interest; this property can be exploited to identify and quantify molecules in solution. The amount of radiation absorbed over a range of wavelengths is called the **absorption spectrum**. Several key biological molecules have spectroscopic properties that allow their presence to be readily monitored by virtue of the characteristic absorption spectra associated with them. For example, Figure 19.41a shows how DNA and proteins have distinct absorption spectra in the UV range of the electromagnetic spectrum. An absorption spectrum typically has one or more peaks of absorbance that can be used as a chemical "signature" to monitor the presence of a particular species. DNA and RNA have an absorbance maximum at ~260 nm that reflects the presence of heterocyclic bases (A, C, G, and T/U), whereas a typical protein with aromatic residues (tryptophan, tyrosine, or phenylalanine) usually has an absorbance peak at ~280 nm. Other peaks in

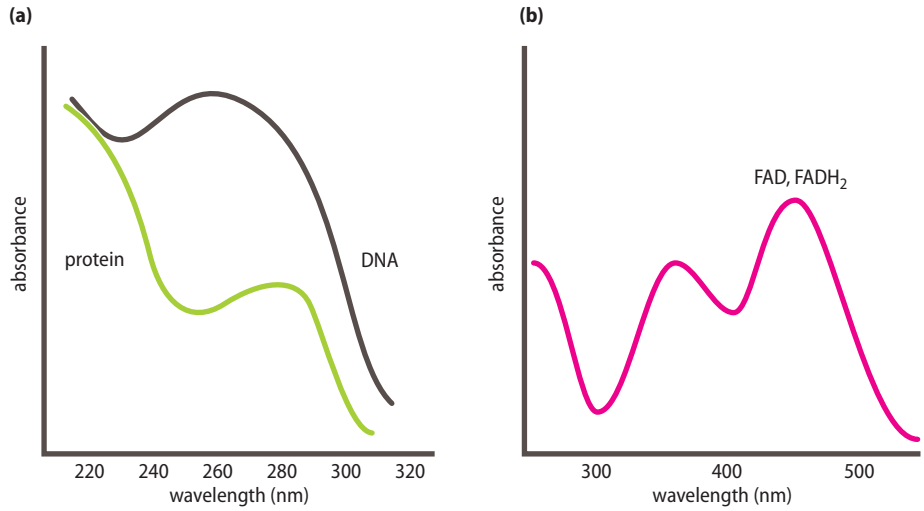


Figure 19.41 Representative absorption spectra. (a) Absorption spectra of DNA and protein at equal concentrations. At a wavelength of 260 nm, or A_{260}, a 50 μg/ml solution of double-stranded DNA or a 40 μg/ml solution of single-stranded RNA has an absorbance of ~1.0. At A_{280}, a 1 mg/ml solution of an "average" protein has an absorbance of ~1.0, although this value depends on amino acid composition. (b) The absorption spectrum of FAD and $FADH_2$. Both FAD and $FADH_2$ exhibit two absorption peaks, at 380 and 450 nm.
From Harm, W. (1980). *Biological Effects of Ultraviolet Radiation*. Cambridge: Cambridge University Press.

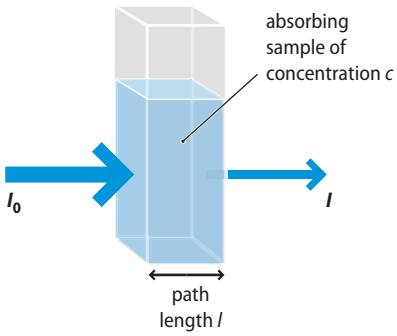


Figure 19.42 Absorption spectroscopy. The relative amount of light absorbed by a sample (I/I_0) is determined by the type of molecule it contains, the concentration of the molecule, and the distance the light traverses through the liquid. These parameters are related to one another according to the Beer–Lambert Law: $A_\lambda = \varepsilon_\lambda\, l\, c$, where A is the measured absorbance at wavelength λ, ε is the molar absorptivity or extinction coefficient (constant for each substance, units of $M^{-1}\ cm^{-1}$ where M is molar), l is the path length (in cm), and c is the concentration (in M). The key term is ε, which varies with wavelength and is a property of the particular molecule being evaluated.

the DNA, RNA, and protein absorption spectra at shorter wavelengths reflect the properties of the peptide and nucleic acid backbones.

Many small molecules in the cell also have characteristic absorption spectra that make them relatively easy to visualize (and identify). As shown in Figure 19.41b, the nucleotide cofactor flavin adenine dinucleotide (FAD) has a characteristic absorption spectrum with peaks at 380 and 450 nm. Proteins that are complexed with FAD (or its reduced derivative $FADH_2$) can therefore be directly monitored by measuring their absorbance at these wavelengths. Cofactors that cause proteins to absorb light in the visible portion of the electromagnetic spectrum (from 400-nm to 700-nm wavelength) cause otherwise colorless proteins to appear colored and are therefore called **chromophores** (from the Greek "chromo," which means color, and "phore," which means carrying). For example, the red color of hemoglobin, which gives blood its characteristic color, comes from the oxygenated heme prosthetic group that is complexed with globin peptide chains.

Optical absorbance is measured in an instrument called a **spectrophotometer**. Typically, a liquid sample is placed in a transparent vessel called a cuvette (illustrated in Figure 19.42), which is placed in the spectrophotometer. The intensity of a light beam that has passed through the sample is compared to the intensity of a beam that has passed through air or through a sample containing water or buffer. The relative amount of light absorbed by the sample is determined by three things: the concentration of the molecule, the distance the light traverses through the liquid, and the type of molecule the sample contains.

Each molecule has intrinsic properties that affect the absorbance of light and are indicated by a parameter called the **extinction coefficient**, denoted by the Greek letter ε. The extinction coefficient of a particular protein at 280 nm, for example, is dictated primarily by the total number of tryptophan and tyrosine residues in the protein. The light absorbed by the sample at a given wavelength λ, denoted A_λ, will then be the product of the extinction coefficient (ε_λ), the concentration of the molecule (c), and the distance the light traverses through the fluid (l). This relationship, known as the Beer–Lambert Law, can be represented by the following equation: $A_\lambda = \varepsilon_\lambda\, l\, c$.

When the extinction coefficient of the molecule under study is known, the absorbance of a sample can be used to determine the concentration of a given molecule (particularly if no other species are present that could contribute to the

overall absorbance at that wavelength). In addition, the purity of a sample can sometimes be evaluated by asking whether the ratios of absorbance at different wavelengths match those of a pure sample. For example, the ratio of the absorbance at 260 and 280 nm is typically ~2 for a DNA population, whereas the ratio is ~0.6 for a typical protein preparation. Therefore, if the A_{260}/A_{280} ratio for a DNA sample is significantly less than 2, it is likely that it is contaminated with protein.

Cellular components can be monitored by binding or reaction with fluorochromes

In addition to using their intrinsic spectral properties, molecules can be detected by their binding or attachment to a specific molecule that either fluoresces or has a distinct color or spectroscopic property. The detectable molecules can be naturally occurring, as in the FAD/$FAHD_2$ example mentioned earlier, or they can be developed specifically for the purpose of detecting cellular components. For example, a method for determining protein concentration, known as the Bradford assay, relies on the binding of a dye, called Coomassie Brilliant Blue, to proteins. On its own, the dye has very little absorbance at 595 nm. However, when bound to protein, the dye changes conformation and absorbs light at 595 nm. Since the amount of bound dye is proportional to the amount of protein in a sample, the absorbance of the protein–dye mixture at 595 nm can be used to determine the concentration of the protein in the sample.

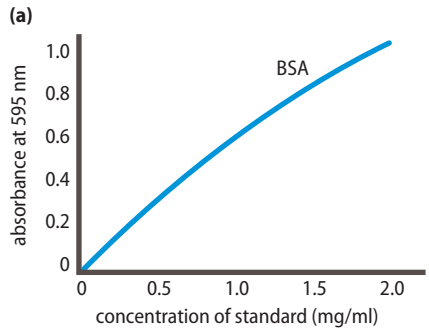
To use this assay, a standard curve is first generated by incubating a series of protein standards of known concentration with the dye and measuring their absorbance at 595 nm; an example of a standard curve is shown in Figure 19.43a. The protein sample in question is then similarly incubated with the dye and its absorbance compared to the standard curve to estimate the protein concentration in the sample (see Figure 19.43b). A drawback to this assay is that it assumes that the unknown protein binds about as much dye per milligram of protein as the protein used as the standard, which is not always the case. Coomassie Brilliant Blue is also commonly used to stain proteins immobilized in gels after separation by electrophoresis, as we shall see in Section 19.8.

A commonly used molecule for the detection of DNA, and sometimes RNA, is ethidium bromide. This molecule inserts (intercalates) between the DNA or RNA bases, and the intercalated molecules fluoresce when irradiated with UV light. Methylene blue dye is also commonly used to stain nucleic acids. Although not as sensitive as ethidium bromide, methylene blue is less toxic and less likely to interfere with subsequent manipulations (such as restriction enzyme digestion or annealing to complementary sequences) because it does not intercalate in the nucleic acid. Both reagents are commonly used to stain nucleic acid samples that have been resolved on acrylamide or agarose gels.

Radioactive labeling can be used to detect molecules

A powerful method for detecting very small amounts of a molecule is to incorporate radioactive atoms into them. Isotopes such as ^{32}P or ^{33}P, ^{35}S, ^{14}C, or ^{3}H are unstable, unlike their more common isotopes ^{31}P, ^{32}S, ^{12}C, or ^{1}H. This instability means that they break down over time, emitting ionizing radiation that can be detected.

Various approaches are available for incorporating radioactive isotopes into larger molecules, including proteins, nucleic acids, sugars, or lipids. Radioactive isotopes can be incorporated biosynthetically, for example, by growing cells in



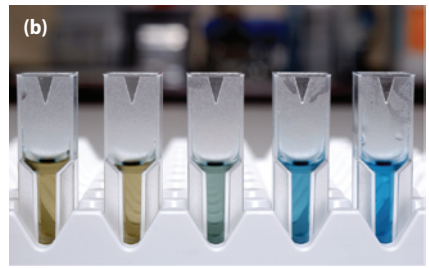


Figure 19.43 Typical standard curve for a Bradford assay. (a) A standard curve is constructed by performing the assay with known quantities of protein. Note that the assay becomes somewhat non-linear at higher protein concentrations. Samples are typically diluted until they fall within the linear range of the assay. (b) Visual examples of Bradford assay cuvettes with increasing concentrations of the protein bovine serum albumin (BSA).
From www.eiroforum.org/media/gallery/embl.php

We learn more about the use of acrylamide and agarose gels when we discuss electrophoresis in Section 19.8.

Methods for detecting radioactivity
Geiger-Muller survey meter (Geiger counter)
Typically handheld gas-filled monitor that detects ions formed upon radioactive decay.
liquid scintillation counting
A scintillation counter converts radioactive emissions into visible light. Liquid scintillation fluid is an organic solvent containing a fluor (a compound that fluoresces in the presence of radioactivity). Scintillation fluid is mixed with the radioactive sample (which may be a small amount of fluid, a gel slice, or a small piece of filter paper) in a vial which is then placed inside the counter in the dark, so that the light emissions can be detected and counted.
film autoradiography
Photographic emulsions contain silver halide crystals that form silver atoms in the presence of radioactivity, leaving an image on the film after processing. X-ray film containing silver halide crystals is the format usually used in the laboratory and is exposed by pressing the film flat against a gel or paper that contains the radioactivity. (Gel separation and autoradiography will be discussed further in Section 19.8).
phosphorimager autoradiography
A phosphorimager screen is coated with a phosphor compound whose atoms are excited when they encounter radioactivity. When the screen is exposed to radioactivity, for example by being pressed against a gel, a latent image of excited atoms forms in the screen. When the screen is scanned with a laser, the stored energy in the excited atoms is released in the form of light, which can be measured.

Figure 19.44 Methods for the detection of radioactivity.

the presence of ^{35}S-labeled methionine, which becomes incorporated into newly translated proteins. Molecules can be labeled *in vitro* using enzymes that transfer a radiolabeled moiety to either nucleic acids or proteins. For example, adenosine triphosphate (ATP) synthesized with ^{32}P at the γ phosphate (the outermost phosphate) can be used as a donor to transfer the radioactive phosphate to a different molecule. To label nucleic acids, T4 polynucleotide kinase is used to transfer the γ phosphate of ATP to the substrate (for example, the 5′ end of a strand of DNA or RNA). In an analogous fashion, protein kinases typically transfer the γ phosphate of ATP to particular serine, threonine, or tyrosine residues in a protein.

Radioactively labeled molecules can be detected by a range of techniques, as shown in Figure 19.44. In many cases, it is possible to measure the amount of radioactivity present and thereby determine the quantity of a radioactively labeled molecular species present in a sample.

Cellular components can be monitored by activity assays

Experimental approach 14.2 provides an example of an activity assay.

In order to detect the presence of an enzyme and study how the rate of the catalyzed reaction changes under different conditions or as the result of mutation, it is necessary to have a method to monitor the particular activity of the protein. For example, a protein with kinase activity can be assayed by monitoring the addition of a radioactive phosphate to the target of the kinase. The advantage of an activity assay is that it can, in many cases, be sensitive enough to detect minute quantities of the species that might not be detectable by any other means, especially if the sample in question is impure. This is, in part, because the product of an enzyme-catalyzed reaction may be easier to detect in small quantities than the enzyme itself, or because enough product may accumulate that it becomes more abundant than the enzyme itself.

The products of some enzymatic reactions cannot be easily monitored directly. However, in some cases, it is possible and desirable to couple the reaction of interest to a second enzymatic reaction whose products can be monitored more readily. In the example shown in Figure 19.45a, the reaction of interest is the hydrolysis of ATP by topoisomerase II, which yields ADP and inorganic phosphate. Although methods are available to measure ATP and ADP levels, a simple approach is to use a

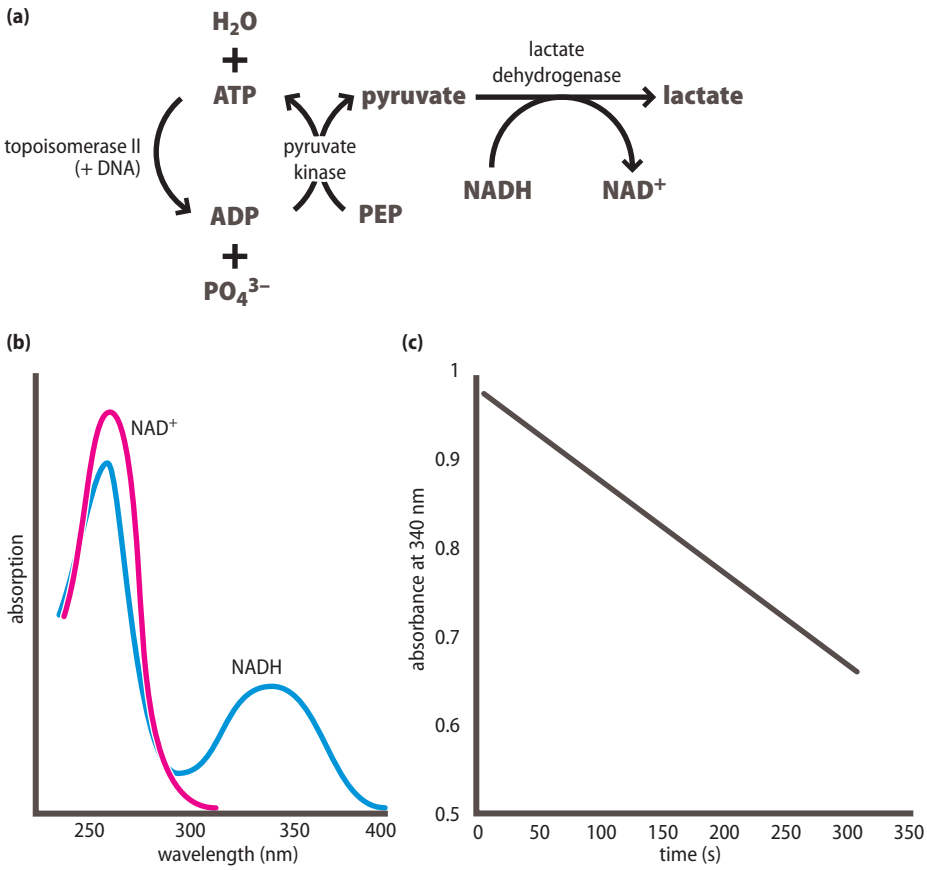


Figure 19.45 Measurement of enzymatic activity by coupling several reactions. ATP hydrolysis by DNA topoisomerase II can be followed by coupling the hydrolysis reaction to the oxidation of the reduced form of nicotinamide adenosine dinucleotide (NADH) to its oxidized form (NAD^+). (a) The reaction scheme. Topoisomerase II hydrolyzes ATP to adenosine diphosphate (ADP). This hydrolysis can be measured indirectly using two subsequent coupled reactions. First, the ADP that was generated by topoisomerase is converted back to ATP by pyruvate kinase, which uses phosphoenolpyruvate (PEP) as a phosphate donor. This converts PEP to pyruvate. Pyruvate, in turn, is a substrate for lactate dehydrogenase, which converts pyruvate into lactate, using NADH as a cofactor; for each pyruvate that is converted to lactate, one NADH molecule is converted to NAD^+. Thus, NADH concentration is directly correlated with ATP hydrolysis by topoisomerase. (b) Absorption spectra of NAD^+ and NADH. The spectra of these molecules are markedly different—NADH absorbs light at 340 nm, whereas NAD^+ does not. (c) When ATP is hydrolyzed by topoisomerase, the concentration of NADH decreases as it is converted to NAD^+ and the absorption at 340 nm decreases in tandem.

product of the reaction—in this case, ADP—to drive additional enzymatic reactions whose products can be easily detected using a spectrophotometer.

In this example, the coupled reactions take advantage of the fact that, as shown in Figure 19.45b, NADH has a peak absorption at 340 nm while NAD^+ does not absorb light at this wavelength. NADH, phosphoenolpyruvate, and two additional enzymes are added to the reaction, along with ATP and topoisomerase II. As ADP from the topoisomerase II reaction accumulates, it is used by pyruvate kinase to generate pyruvate, while regenerating ATP. The pyruvate is then converted into lactate by lactate dehydrogenase in a reaction that consumes NADH and releases NAD^+. Since one molecule of NADH is consumed for every molecule of ATP hydrolyzed by topoisomerase II, one can follow the progress of the topoisomerase reaction by using a spectrophotometer to monitor the absorbance of NADH at 340 nm. As the reaction progresses, NADH is converted to NAD^+ and the absorbance at 340 nm decreases, as depicted in Figure 19.45c.

Activity assays also make it possible to look for, and isolate, a biological molecule whose existence is suspected, but whose identity is unknown. For example, researchers were able to carry out RNA polymerase II-dependent transcription in a test tube using a nuclear extract when the identities of the particular proteins in the mixture that were required to facilitate transcription by RNA polymerase II were unknown. These proteins were identified by fractionating nuclear extracts and assaying the different fractions for their ability to promote transcription *in vitro*. Once a sufficiently pure sample with the desired activity was obtained, the proteins could be identified through a variety of methods, many of which will be described in the following sections.

19.8 SEPARATION AND ISOLATION OF BIOLOGICAL MOLECULES

Much of the information in this book derives from experiments done in the test tube using purified components. In order to study the behavior and properties of a specific molecule or complex in solution, one must be able to separate it from other cellular components. This can be done by taking advantage of the different physical and chemical characteristics that distinguish one molecule from another—for example, mass, shape, overall charge, or the ability to bind to a specific reagent. It is also possible to introduce a tag with known properties into a protein or nucleic acid molecule that can then be used to isolate the tagged molecule. In this section, we will discuss some of the ways for isolating cellular structures and molecules.

Organelles and molecules can be separated by centrifugation

The first step in purifying biological molecules from cells often involves isolating the different organelles or separating soluble and membrane-associated molecules from one another. This process is known as **cell fractionation** and can be done with whole tissues or with cells grown in culture. To begin this process, the intact cells must be disrupted to release their contents—but the treatment must not be so harsh as to disrupt the organelle or the molecular structures being studied. Figure 19.46 illustrates a variety of methods that can be used to disrupt cells; these include grinding or the use of ultrasonic waves, high pressure, or detergents.

Once cells have been disrupted, the individual components can be separated. This is commonly done using a **centrifuge**, an instrument used to spin samples in test tubes at high speed. The centrifuge creates centripetal force perpendicular to the rotation axis. Test tubes (or other vessels) are placed in a central holder in the centrifuge called a **rotor**, which either allows the tubes to swing outward as the rotor spins or holds the test tubes at a fixed angle; these differences are illustrated in Figure 19.47. In either case, the centripetal force that results when samples are accelerated at upward of 100 000 *g* (where *g* denotes the acceleration due to gravity) causes organelles or large molecular complexes to move toward the bottom of the tube and form a pellet.

The rate at which a particle descends through the tube depends upon its size and the speed (angular velocity) at which the centrifuge rotates—the faster the

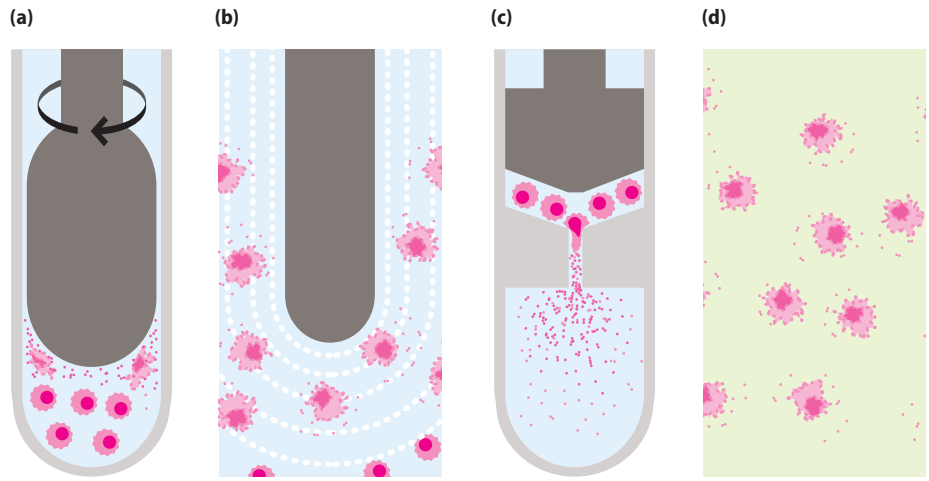


Figure 19.46 Methods for disrupting cells. Various methods are used to disrupt cells, including: (a) grinding cells—for example, with a pestle that fits tightly in a test tube; (b) subjecting cells to ultrasonic waves using an instrument called a sonicator (the tip of the sonicator is shown); (c) passing cells through a small opening by applying very high pressure; and (d) treating cells with mild detergents that generate holes in membranes.

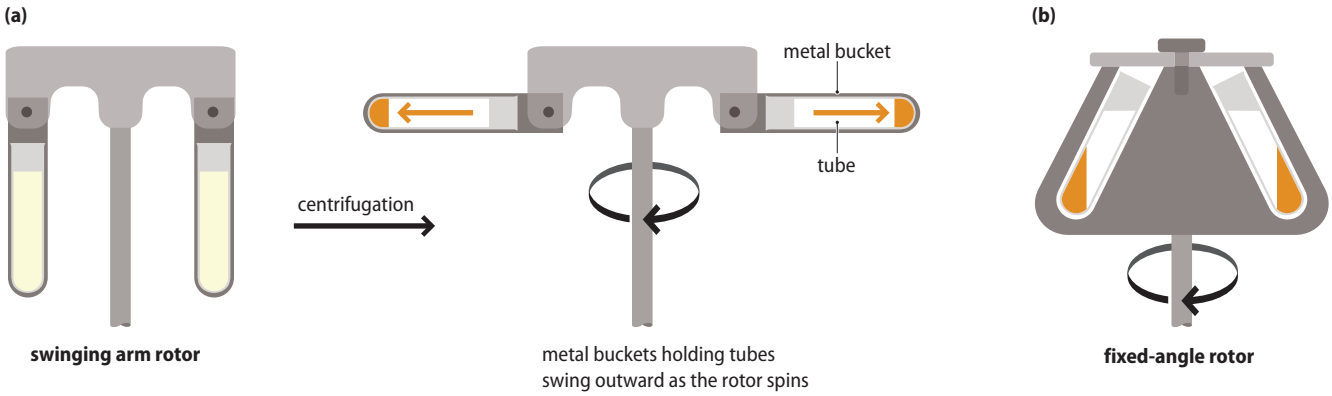


Figure 19.47 Centrifugation. Solutions can be subjected to centripetal force in a centrifuge rotor, which is spun at very high speeds (in the order of tens of thousands of revolutions per minute). This causes larger complexes and organelles to sediment to the bottom of the test tube. (a) In a swinging bucket rotor, the test tubes are placed in metal buckets that can swing outward as the rotor begins to spin. The centripetal force causes large complexes and organelles to sediment and collect at the bottom of the tube. (b) In a fixed-angle rotor, the test tube bottom slants away from the central axis of rotation and remains at a fixed angle. The centripetal force acts perpendicular to the axis of rotation. As the rotor slows and stops, the layers of increasing density can reorient so that they align along the vertical axis of the tube—that is, the densest layer is at the bottom and the least dense layer is at the top.

rotor turns, the greater the acceleration, and thus the force that is exerted on the particle. By choosing a particular speed and amount of time, the desired particle can be collected at the bottom of the test tube (see Figure 19.48). For example, the cytosolic components of a cell that has been broken or lysed can be readily separated from the heavier membrane-bound organelles (or organelle fragments) by centrifugation. At a different speed of rotation, membranes will settle to the bottom of the tube, while the cytosolic components will stay in solution.

Better separation between different cellular components can be achieved by centrifuging the fractions through a gradient of different sugar or salt concentrations. When sucrose (a sugar) or cesium chloride (CsCl; a salt) is dissolved in water, the resulting solution is significantly denser than water or dilute buffer. It is possible to create a **gradient** of concentrations in which the most concentrated, and hence the densest, solution is at the bottom of the centrifuge tube and the least dense (and least concentrated) solution is at the top; such density gradients are depicted in Figure 19.48. The gradient can be established prior to adding the sample, or it can be allowed to develop during centrifugation as a result of the centripetal force on the solute molecules.

A density gradient can separate different cellular components in one of two ways. In velocity sedimentation (see Figure 19.48a), the sample is layered on top of the gradient, and the sample is centrifuged for a fixed amount of time. The rates

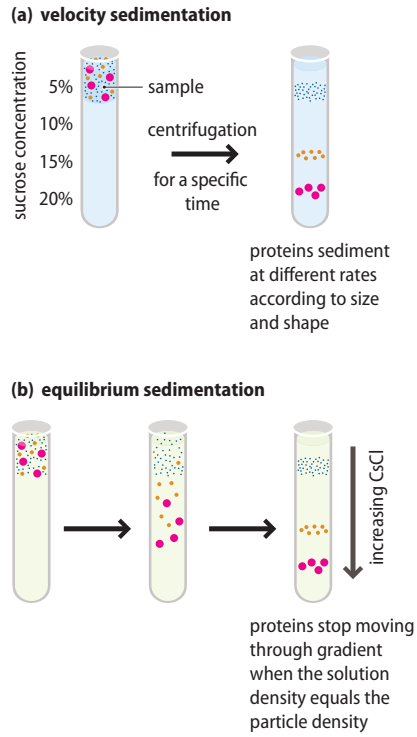


Figure 19.48 Density gradient sedimentation. Cellular components can be separated from one another by centrifugation through a solution whose density is highest at the bottom of the test tube and lowest at the top. (a) In a sedimentation velocity experiment, particles are separated by the rate of their sedimentation. The gradient (for example, of sucrose) is first created by layering solutions of decreasing density in the test tube and then introducing the sample at the top of the tube. The rate at which particles sediment through this gradient is determined by their mass and shape. In the example shown, the red particles are larger and denser than the dark blue and orange particles; thus, during centrifugation, they move faster through the gradient and end up at a denser part of the gradient. (b) In an equilibrium sedimentation experiment, particles are separated by density. A solution containing the sample and a high concentration of a substance such as the salt CsCl is centrifuged at high speeds for long periods. Eventually, a density gradient forms under the centripetal force. The different particles in the sample will be found at a position in the tube where the solution density matches the particle density.

We see an example of the use of CsCl gradient centrifugation to separate DNA of different densities in Experimental approach 6.1.

at which different molecules or organelles migrate toward the bottom of the tube depend on both their size and their shape. After centrifuging the sample for a fixed amount of time, samples are collected from either the top or the bottom of the tube. (This can be done by punching a hole in the bottom of the tube and allowing the sample to drip out.) In equilibrium sedimentation, the sample is spun for a long period of time until the molecules reach equilibrium and are no longer moving within the gradient (see Figure 19.48b). A molecule or complex will stop moving through the gradient when it reaches a position in the tube where the density of the gradient matches the density of the molecule.

Information about the size and shape of a macromolecule or complex can be obtained by a more sophisticated instrument known as an analytical ultracentrifuge, which can both spin samples at very high speeds and simultaneously record the concentration of the macromolecule at different radial positions in the ultracentrifuge vessel (known as a cell). This method uses a standard buffer, rather than a salt gradient. The rate at which a molecule or complex of molecules sediments in the centrifuge cell is denoted by a unit called a Svedberg (or S). The S value of a given complex is governed by both its molecular weight and its shape. The different subunits of the ribosome, which we learned about in Chapter 11, are denoted by the S value that characterizes their sedimentation behavior.

Macromolecules can be separated based on their solubility

The initial extract prepared from disrupted cells is often very complex, containing thousands of different proteins and complexes. These macromolecules differ in size and shape, as well as in other properties, including overall charge and the relative distribution of hydrophobic and polar residues on their surfaces. Each of these physical properties distinguishes one macromolecule or complex from the next and can be exploited to separate molecules.

A simple method for separating macromolecules from one another is to fractionate them based on their solubility, typically in a salt solution such as ammonium sulfate. Most proteins can be induced to become insoluble and precipitate at sufficiently high concentrations of ammonium sulfate, although the precise concentration at which a protein becomes insoluble varies widely among macromolecular species. Thus, by using the appropriate salt concentration, one can separate the protein of interest (and other proteins or complexes with similar solubility properties) from many other proteins. Either the solution containing the soluble protein of interest can be isolated or the precipitate containing the protein of interest can be resuspended in buffer at a low-salt concentration, thus enabling the protein to re-enter the solution. Other types of precipitating agents—for example, polyethylene imine—are also used to precipitate both proteins and nucleic acids.

Proteins and protein complexes can be purified by column chromatography

A common method for purifying proteins is column chromatography, which uses a glass or plastic tube filled with material (the matrix) that separates proteins based on their physical properties as they flow through the column. The protein sample is first applied to the top of the column, as shown in Figure 19.49. As a buffer solution is flowed continuously through the column, proteins in the sample migrate through the column at different rates, depending on the nature of the column matrix and

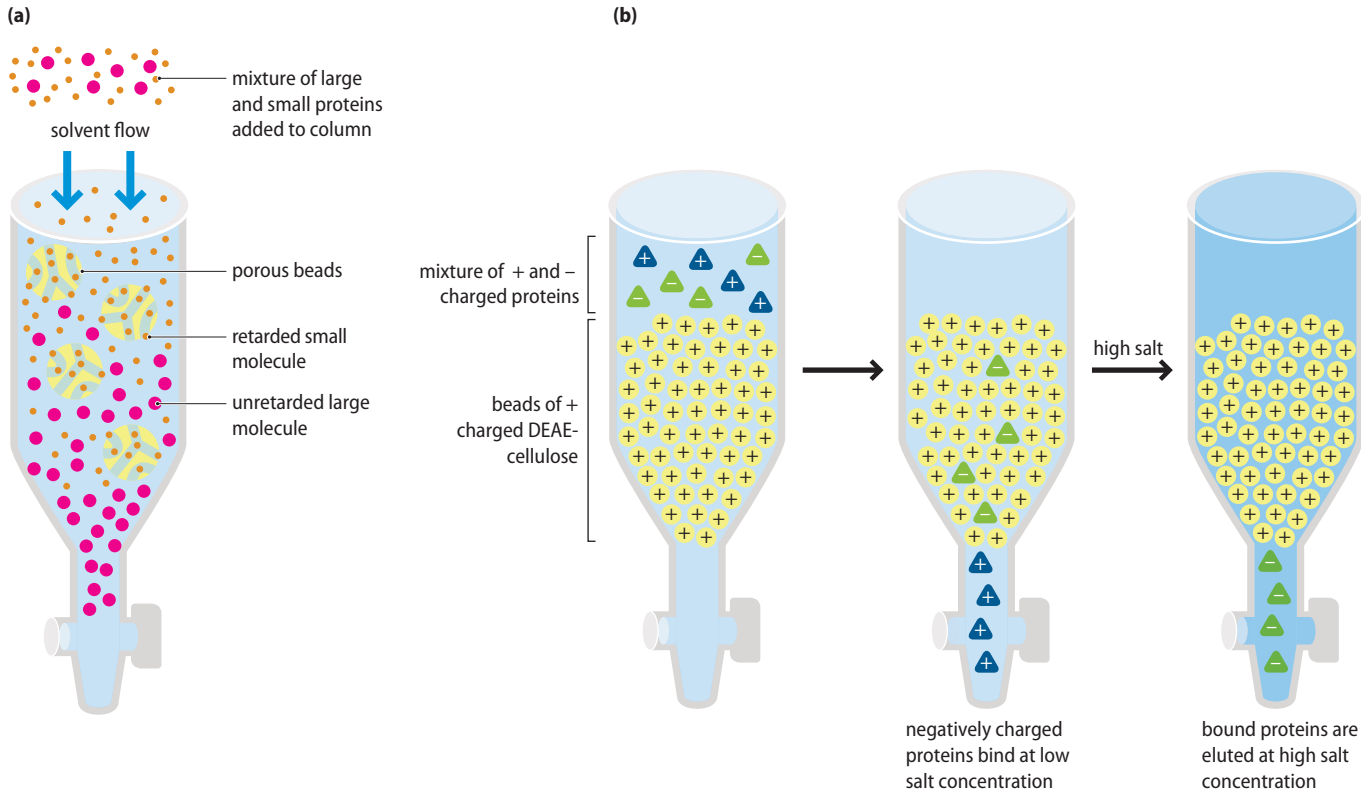


Figure 19.49 Chromatographic separation by size or charge. (a) Size exclusion chromatography, also known as gel filtration, separates molecules based on their size and shape. The column is filled with porous beads (in yellow). Large molecules (red) that cannot fit into the pores in the beads pass quickly between the beads, while smaller molecules (orange) that can enter the beads take more time to pass through the column. (b) Ion exchange chromatography separates molecules based on charge. If the beads in the column are positively charged, negatively charged proteins (green) will bind to the column matrix at low salt concentrations, as a result of ionic interactions, while positively charged proteins (blue) will flow through the column. The negatively charged proteins can be induced to dissociate from the beads by increasing the salt concentration.

the physical and chemical properties of the proteins. Different types of columns are used to separate proteins, based on size, charge, hydrophobicity, and other properties.

In **size exclusion chromatography**, protein molecules and complexes can be separated on the basis of size and shape on a **gel filtration** column (see Figure 19.49a). The column matrix consists of microscopic beads, each containing many tiny pores. Small molecules are more likely to go through the pores of the matrix and thus travel a greater distance and move through the column more slowly. Larger molecules, on the other hand, are excluded from entering the pores in the beads and can therefore only pass through the spaces between the beads. As a result, they travel a shorter distance overall and pass through the column more quickly.

Ion exchange chromatography columns separate proteins on the basis of surface charge (see Figure 19.49b). These columns contain a matrix bearing either positively or negatively charged chemical groups. In low-salt solutions, proteins with a negative surface charge will bind to the positively charged matrix in anion exchange columns, while proteins with a positive surface charge will bind to the negatively charged matrix in cation exchange columns. The bound proteins can then be washed off (or "eluted") in buffer containing higher salt concentrations, which competes for the charged chemical groups on the matrix.

Another property that can be used to purify proteins is the ability of the protein to bind selectively to a particular molecule or chemical group. This approach is called **affinity chromatography**, since it relies on the protein binding to a molecule for which it has specific affinity. For example, the heparin polymer mimics the sugar–phosphate backbone of DNA and RNA and, when attached to a matrix, can be used to purify proteins that bind nucleic acids. The dye Cibacron blue is used in an analogous way to purify proteins that bind to NAD^+.

Not all proteins have a known substance to which they bind, however. Thus, it is often most convenient to use cloning and recombinant DNA techniques (see Section 19.4) to add a protein or nucleic acid "tag" that binds a known substance, as we discuss next.

Proteins and nucleic acids can be engineered to contain tags that facilitate purification

Instead of simply relying on the native properties of a protein to separate it from other proteins in the cell, the recombinant DNA technology described in Section 19.4 can be used to add an affinity tag that can be exploited to purify a protein using affinity chromatography. The tag generally consists of amino acids added to the N- or C-terminus, thus extending the polypeptide chain (see Figure 19.50a). Some commonly used tags are a histidine-rich peptide that binds tightly to a matrix containing nickel ions; a peptide that binds to a specific antibody immobilized on a column; or an entire protein domain such as glutathione-*S*-transferase (GST) that binds to a matrix containing immobilized glutathione. More than one affinity tag can be incorporated into a single polypeptide chain, making it possible to use different types of affinity purification sequentially.

How are tagged proteins purified? The engineered protein of interest fused to the tag is expressed from a plasmid or the chromosome. The fusion protein is then isolated by lysing the cells and passing the resulting extract on the appropriate affinity column, as illustrated in Figure 19.50b. The fusion protein binds to the column, while the other components in the extract pass through. The fusion protein is then eluted from the column with a molecule that competes with the fusion protein's interaction with the matrix. Imidazole, for example, is used to elute histidine-tagged proteins from a nickel column (since it competes with the imidazole ring of histidine side chains, which coordinate the metal), whereas glutathione is used to elute proteins from a GST column.

An important caveat is that affinity tags have the potential to alter the activity of a protein. For systems where the activity of the protein can be assayed *in vivo* or *in vitro*, it is important to check that the tag has not altered the activity. Tagged proteins can also be engineered to contain intein sequences, which we learned about in Chapter 14, or a protease recognition site between the tag and the protein sequence of interest. These can then be exploited to remove the tag, either by changing the solution conditions to promote intein self-splicing or by adding a protease.

DNA or RNAs of interest similarly can be fused to tags that allow them to be separated from a pool of other molecules by affinity chromatography. Examples of such tags include sequences that are tightly bound by specific DNA- or RNA-binding proteins. These proteins can be attached to a solid matrix (such as in affinity purification) or be isolated together with the tagged nucleic acid using one of the methodologies described earlier.

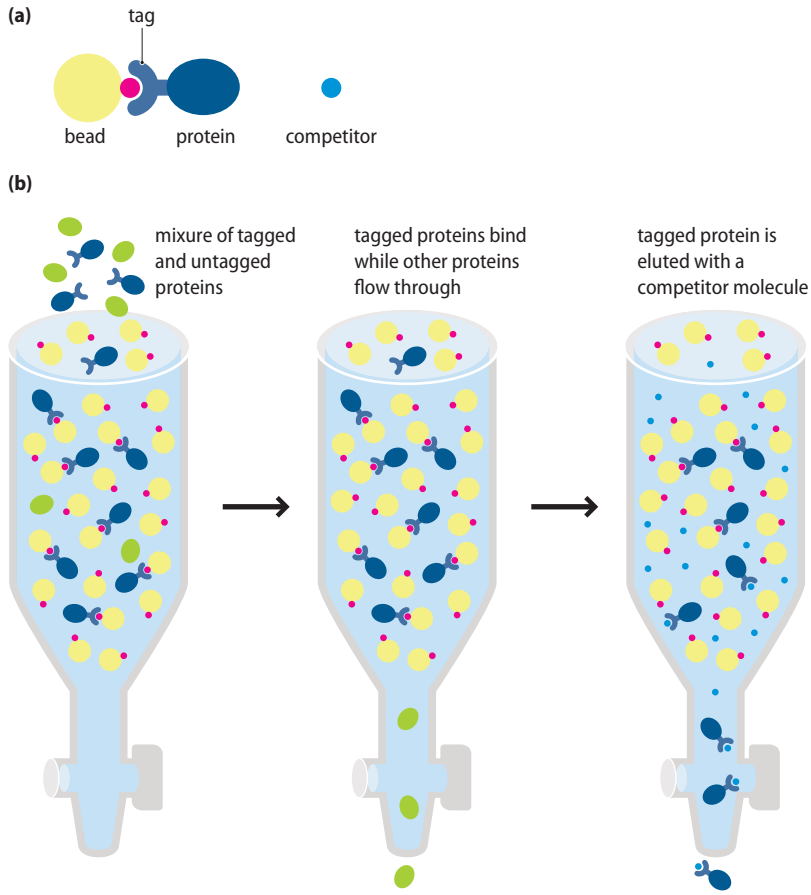


Figure 19.50 Using tags to purify proteins. (a) A fusion protein containing an additional protein tag can be engineered. The tag can bind to a specific molecule (pink) on the beads in the column. The competitor (light blue) has the same structure as, or a similar structure to, that of the resin-bound molecule. (b) Affinity purification of a tagged protein. The protein tag binds to the molecule that is immobilized on a column, causing the fusion protein (dark blue) to be retained while other proteins (green) flow through. The protein can be eluted by adding a molecule (light blue) that competes with the fusion domain for binding to the column matrix.

High-performance liquid chromatography and thin-layer chromatography can be used to separate small molecules

The method most commonly used to separate small peptides, nucleic acids, and small molecules is called **high-pressure or high-performance liquid chromatography (HPLC)**. In a common application of this method known as reverse-phase chromatography, the column matrix consists of small beads derivatized with hydrophobic alkyl chains (such as $C_{18}H_{37}$ or C_8H_{17}). Small molecules bind to the beads via hydrophobic interactions and are then eluted with increasing concentrations of organic solvent. The liquid is passed through the column at high pressure, which increases the resolution.

Instead of using a column matrix, small molecules can also be separated on a flat support covered by a thin layer of an absorbent such as silica or cellulose. Termed **thin-layer chromatography (TLC)**, this method of separation is also based on the principle that different molecules will travel at different rates through the absorbent. The sample is applied to the bottom of the plate, whose end is then dipped in a suitable solvent. The solvent containing the different molecules then migrates through the absorbent by capillary action. The distance the molecules travel in a given time is determined by their chemistry, which affects how they interact with the plate and with the solvent. Different types of molecules, such as particular lipids or nucleotide phosphates—for example, ATP, ADP, and adenosine monophosphate (AMP)—can then be distinguished by their different locations on the TLC plate. The location of the different molecules

along the plate can be determined by their ability to bind a specific dye or, if they are radiolabeled, by using methods for detecting radioactivity, as described in Section 19.7.

RNA and DNA molecules can be separated on the basis of size on agarose and acrylamide gels

As we recall from Chapter 2, DNA and RNA are negatively charged. Thus, when nucleic acids are subjected to an electrical field, they migrate toward the positive pole. This property is used to separate DNA and RNA molecules of different sizes by a method called **gel electrophoresis**, an apparatus for which is depicted in Figure 19.51. This method relies on a porous gel made of agarose or polyacrylamide through which nucleic acids migrate when the gel slab is immersed in an aqueous solution and an electric field is applied.

The gel slab is prepared either by heating a solution containing the polysaccharide agarose, thus causing it to liquefy, and then allowing it to solidify in a mold or by polymerizing acrylamide (which can be induced by adding a cross-linking chemical) between two panes of glass or plastic. When the agarose or acrylamide solidifies, the individual molecules become polymers that form a mesh through which nucleic acids can pass. The pore size in the gel mesh can be adjusted by altering the concentration of agarose or acrylamide. The nucleic acid sample is applied to the electronegative negative end of the gel, as illustrated in Figure 19.51a, and the sample migrates through the gel toward the positive pole at the opposite end. The rate at which nucleic acids migrate through the gel depends on their length, with smaller molecules migrating faster than longer ones. The DNA or RNA molecules can then be visualized by staining with dye or by autoradiography, as we learned in Section 19.7.

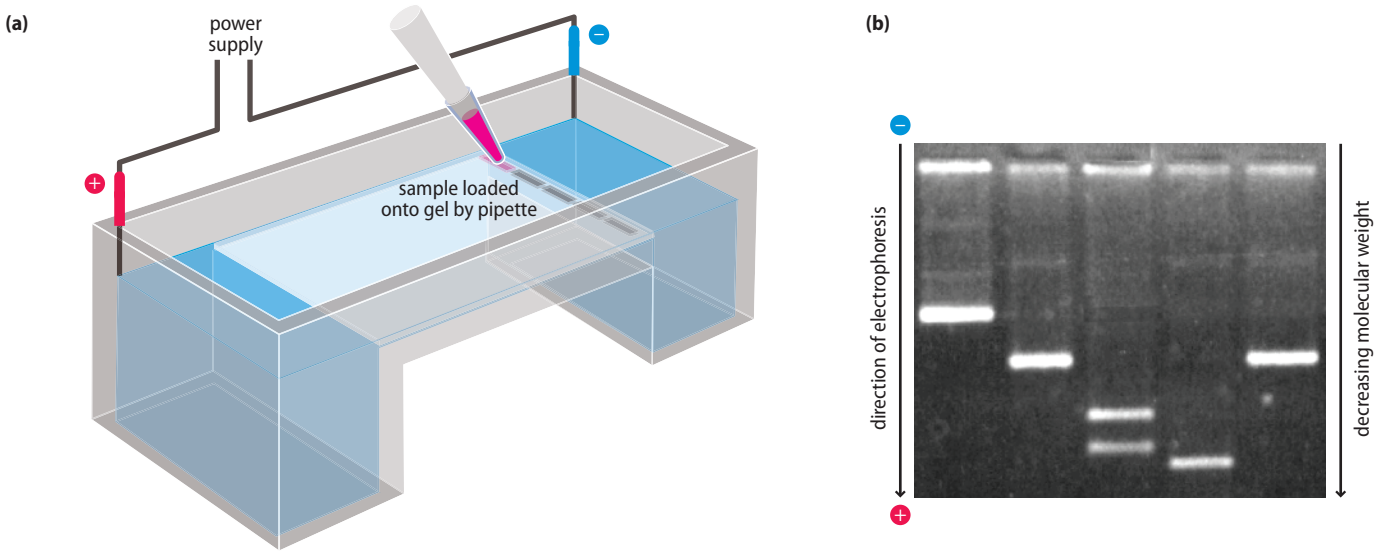


Figure 19.51 Agarose gel electrophoresis. (a) Apparatus showing an agarose slab gel (in light blue) immersed in buffer and connected to a power supply. The DNA (or RNA) sample (in pink) is introduced into the slots (gray) at the negative end of the gel, and a voltage is applied, causing the DNA fragments to migrate toward the positive pole. (b) Example of DNA fragments separated on an agarose gel. After electrophoresis, the gel was immersed in a solution containing a dye that binds to DNA (ethidium bromide). The DNA bands (in white) were detected by exposing the gel to UV light.

DNA molecules of longer than approximately 20 000 bases are not well resolved or separated by conventional electrophoresis on agarose gels, which are typically used to separate DNA fragments of more than a few hundred base pairs in length. Instead, chromosomes and large DNA fragments must be separated using a variant of the gel method, termed pulsed-field gel electrophoresis. In these gels, electric fields are applied in a series of orthogonally oriented pulses (meaning at right angles to one another). Every time the orientation of the electric field is changed, the DNA molecules are reoriented. Larger DNA molecules move more slowly through the gel because they take longer to reorient each time the electrical field is shifted, allowing for separation of very large fragments on the basis of size.

A variant of this method is the field inversion gel (FIG), in which the charges at the two ends of the gel are periodically reversed in pulses of several seconds. This switching or inversion of the electric field serves the same purpose as the pulsed-field gels—to reorient the large DNA molecules. A FIG is simpler to manufacture and use, although separation by FIG is typically not as good as it is for pulsed-field gels.

Proteins can be separated by size and charge on one- and two-dimensional polyacrylamide gels

Gel electrophoresis can also be used to analyze the composition of a protein mixture or complex by separating the different polypeptides based on their molecular weight. Unlike nucleic acids, however, proteins vary in charge and shape and thus do not migrate in a gel in a predictable way. To separate proteins by molecular weight using gel electrophoresis, the proteins must first be denatured and then coated with a negatively charged detergent. This is done by adding the strong ionic detergent sodium dodecyl sulfate (SDS) and usually also by heating the sample to induce the protein to unfold and allow the SDS molecules to bind to the hydrophobic regions of the protein. In the presence of SDS, proteins behave as unstructured polymers coated with a negative charge, especially if all disulfide bonds are eliminated by adding a reducing agent. Proteins that have been coated with SDS can be separated by SDS-polyacrylamide gel electrophoresis (**SDS-PAGE**).

SDS-PAGE gels differ from gels used to separate nucleic acids in that they are usually discontinuous. The sample first migrates through a small, slightly acidic "stacking" gel with a low acrylamide concentration, followed by a longer, mildly basic "resolving" gel containing a higher concentration of acrylamide; the stacking and resolving gels are depicted in Figure 19.52a.

The running buffer contains glycine, a zwitterion that is neutral in the acidic stacking gel and becomes negatively charged in the basic resolving gel. This change in ionization, along with the differing acrylamide concentrations in the stacking and resolving gels, helps the protein sample to become focused into a tight band in the stacking gel. Once the proteins reach the resolving gel, they migrate according to their molecular weights, with small proteins separating into faster migrating bands and larger proteins migrating more slowly. The mixture of proteins is thus separated into discrete bands according to size. The percentage of acrylamide in the resolving gel can be adjusted to maximize the resolution of individual bands over a desired molecular weight range. The proteins can then be visualized by a variety of means, including the binding of the dye Coomassie Brilliant Blue, as described in Section 19.7 and shown in Figure 19.52b, or by silver staining whereby silver ions bind to the amino acid side chains (in particular, the

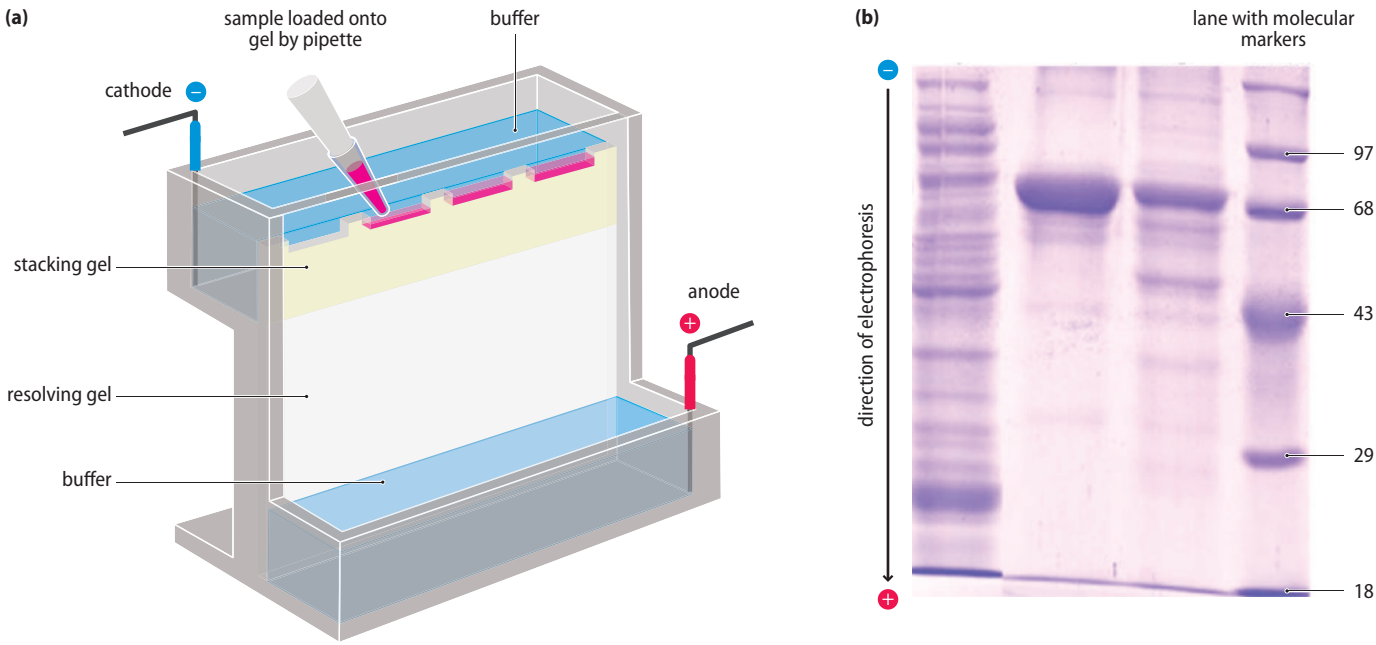


Figure 19.52 SDS-PAGE. (a) This type of gel is used to separate proteins by molecular weight after they have been first denatured in the presence of the detergent SDS. The gel is discontinuous, with the sample first migrating through a stacking gel (in yellow), which helps to compact the sample into a single tight band, followed by separation on a resolving gel, which separates proteins according to their size. (b) Depiction of an SDS-PAGE gel that has been stained with Coomassie Brilliant Blue to visualize different proteins. The right lane contains a mixture of proteins of known mass, to indicate the approximate size of proteins of similar mass on the gel (numbers are in kilodaltons (kDa)).

sulfhydryl and carboxyl groups) and are then chemically reduced to metallic silver, which is easily visualized.

Sometimes it is useful to separate proteins on the basis of other properties, such as their isoelectric point. The net charge of a protein depends upon its amino acid composition and relative number of acidic and basic residues. The isoelectric point pI is the pH at which a protein has no overall charge. Proteins are positively charged in solutions at a pH below the pI and are negatively charged at a pH above the pI.

An **isoelectric focusing (IEF) gel** is a type of polyacrylamide gel that has a stable pH gradient across the gel, as depicted in Figure 19.53a. When a native protein is loaded onto the high pH end of the gel and an electric field is applied, the protein will begin to migrate toward the positive pole of the gel because the protein has an overall negative charge (assuming the pI is lower than the highest pH in the IEF gel, which is typically around 9 or 10). The protein will continue to migrate through the gradient of decreasing pH until it reaches a zone in the gel where the pH corresponds to the pI of the protein. At this point, the protein is no longer charged and ceases to migrate in the gel.

IEF can be used in combination with SDS-PAGE to separate proteins by both pI and molecular weight. This method is called **two-dimensional gel electrophoresis** and is depicted in Figure 19.53b. Two-dimensional gels are particularly useful for separating proteins that have a similar size or pI and are thus difficult to separate by SDS-PAGE or IEF alone. Proteins can be first separated on the basis of pI using IEF and separated subsequently on the basis of size using SDS-PAGE (see Figure 19.53c).

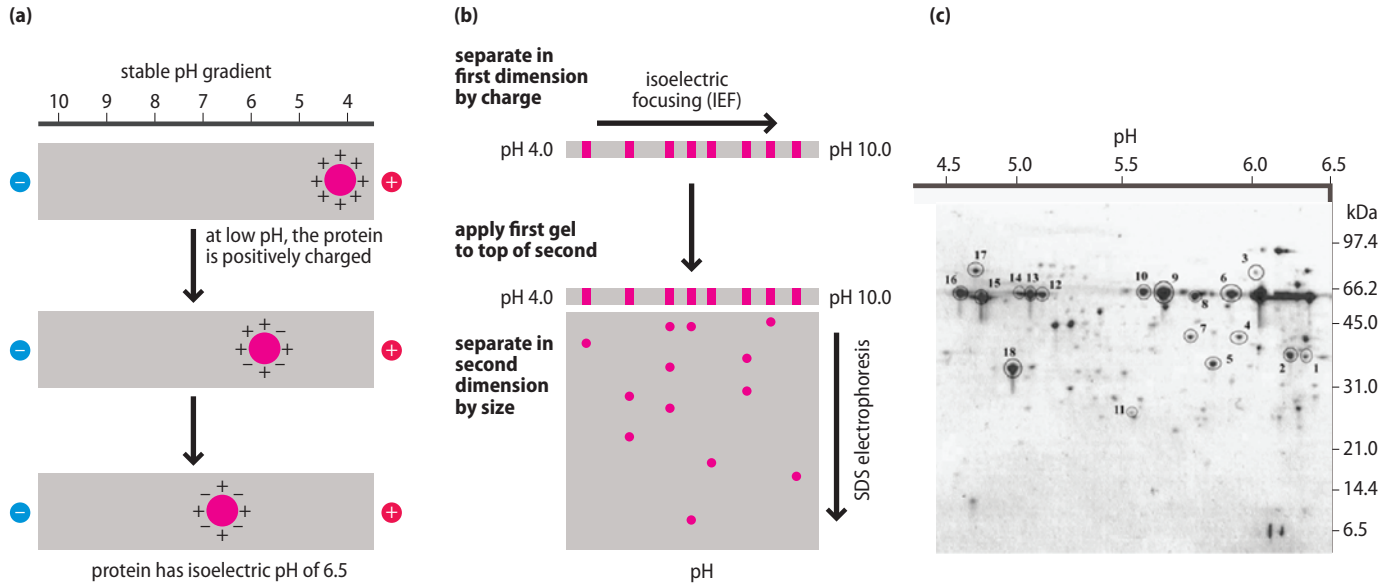


Figure 19.53 Separation by isoelectric point. (a) An IEF gel contains a pH gradient. The protein shown (pink) is positively charged when its surrounding pH is acidic. When an electric field is applied, the protein migrates through the gel until it reaches a pH that matches its pI, at which point it is no longer charged and therefore stops migrating. (b) Two-dimensional gel electrophoresis. Proteins are first separated along one dimension by their pI, and then along the second dimension by molecular weight. In practice, the first separation by IEF is done in a thin tube of polymerized acrylamide containing the desired pH gradient. After the proteins reach their pI, the gel is removed from the tube and placed across the top of the type of acrylamide gel shown in Figure 19.52. (c) Example of a two-dimensional gel of *D. melanogaster* sperm. The purified sperm proteins are run on a pH 4–7 isoelectric focusing strip and run in the second dimension on a 12.5% SDS-PAGE gel. This gel demonstrates the resulting separation of proteins, which allows one to excise selected proteins from the gel (circled) for further analysis.

From Karr, T.L. (2008). Application of proteomics to ecology and population biology. *Heredity* **100**: 200–206.

19.9 IDENTIFYING THE COMPOSITION OF BIOLOGICAL MOLECULES

Once a cellular component of interest has been purified away from other components, it is usually desirable to determine its composition. For example, investigators often wish to determine the nucleotide sequence of a DNA or RNA molecule, or the amino acid sequence of a protein. In this section, we will discuss various methods for determining the sequence of these biological molecules. Many of the technologies described in this section are not sensitive enough to determine the sequence of a single molecule, and thus require the isolation of sufficient material for successful analysis.

The sequence of both RNA and DNA molecules can be determined by primer extension in the presence of dideoxynucleotides (ddNTPs)

RNA and DNA molecules can both be sequenced by generating DNA copies of the molecules of interest in the presence of chain-terminating nucleotides in a procedure referred to as **chain-termination dideoxy sequencing** or Sanger sequencing. In this approach, illustrated in Figure 19.54, an oligonucleotide primer that specifically hybridizes to the RNA or DNA is annealed. (If the sequence of the RNA or DNA is completely unknown, the RNA can be converted first to cDNA, as described in Section 19.3, and oligonucleotides of known sequence can be ligated to the ends of the cDNA or DNA molecule of interest.) The primer is then extended with an enzyme—a DNA polymerase in the case of DNA or reverse transcriptase

Figure 19.54 Chain-termination DNA sequencing using ddNTPs. (a) A DNA primer (turquoise) is annealed to one of the DNA strands whose sequence is being determined; DNA synthesis is carried out by DNA polymerase I (pol I; or reverse transcriptase if RNA is being sequenced), together with dNTPs, in the presence of limiting amounts of ddNTPs. If radioactive detection is used (as in panel c), one or more of the dNTPs is radioactively labeled. If detection is by fluorescence (as in panel d), each ddNTP is uniquely fluorescently labeled. (b) Products of DNA synthesis, each terminated by one of the ddNTPs. (c) An example of radioactive reaction products separated on a polyacrylamide gel and detected by autoradiography. (Note that this sequence is unrelated to the example shown in panel b). When radioactivity is used, each sequencing reaction contains a different ddNTP and each reaction is loaded on a separate lane. The sequence can then be read from bottom to top (GCACAATGTCA …). (d) The products of DNA synthesis using fluorescently labeled ddNTPs are separated by capillary gel electrophoresis, which is similar to regular gel electrophoresis (see Section 19.8), except that the gel is made in a capillary and it separates the molecules of a single sample. When fluorescently labeled ddNTPs are used, a single sequencing reaction can contain all four ddNTPs. The color of the fluorescent label at the 3′ end of the terminating nucleotide is detected, as the corresponding fragment migrates past the laser. (e) An example of a sequencing reaction using fluorescent ddNTPs displayed on a chromatogram that reveals the sequence of the newly synthesized strand.

(c) Fig 2 from Sanger et al. DNA sequencing with chain-terminating inhibitors. 1977, **74**;12: 5463–5467. credit: MRC Laboratory of Molecular Biology.

in the case of RNA. If all four deoxynucleotide triphosphates (dNTPs) are supplied, a full-length DNA copy of the template strand will be synthesized. If, however, a small amount of ddNTPs (which are nucleotides that lack a 3′ hydroxyl group) is added, chain elongation terminates when that nucleotide is incorporated. This occurs because the chain cannot be extended by the polymerase when there is no 3′ OH group on the last added nucleotide.

For the sequencing reactions, four independent chain elongation reactions are performed, each containing the four dNTPs, together with a small amount of a single ddNTP (ddA, ddC, ddG, or ddT). When supplied in an appropriate dNTP-to-ddNTP ratio, the ddNTP will be incorporated at random throughout the synthesized DNA strand, generating a pool of extension products terminating at each position that the given nucleotide occurs. The sequence of the DNA being made can be read by comparing the pattern of extension products of the four different reactions.

Typically, radioactively labeled dNTPs or fluorescently labeled ddNTPs are included in the reactions to make it possible to detect the extension products when they are resolved by regular gel electrophoresis or by capillary gel electrophoresis. An example of regular gel electrophoresis using radioactive labeling is shown in Figure 19.54c; an example of capillary gel electrophoresis using fluorescent labeling is shown in Figure 19.54d.

A single reaction can typically reveal the sequence of up to 900 bases. As we will see in Section 19.10, automated DNA sequencing and newer techniques are used when determining complete genome sequences of individual species, mixtures of species, and even individual organisms.

The sequence of RNA polymers can be determined by reaction with specific nucleases or chemicals

The sequence of an RNA molecule can be determined with the aid of enzymes or chemical reagents that cleave RNA molecules after particular nucleotides. By carrying out these experiments under conditions in which a partial cleavage reaction is performed, a population of identical RNA molecules are cleaved at different places. The entire sequence can then be reconstructed based on the distance of each cleavage site from the end of the molecule. For this, the cleaved molecules need to be visualized.

One approach is to label either the 5′ or 3′ end of the unknown RNA, so that it can be visualized when separated on a polyacrylamide gel (see Section 19.8). RNA can be easily labeled at the 5′ end with radioactive phosphate using T4 polynucleotide kinase and γ-[^{32}P]-ATP. Alternatively, the 3′ end can be labeled with [^{32}P]-pCp (cytidine-3′, 5′-bis-phosphate) using T4 RNA ligase. The end-labeled samples are then exposed to sequence-specific nucleases or certain sequence-specific chemical modification reagents, resulting in strand cleavage. These fragments can then be resolved on denaturing polyacrylamide gels to directly read out the sequence of the polymer. By comparing the length of the cleaved molecules, it is possible to obtain information about the entire sequence, as illustrated in Figure 19.55a.

For example, the ribonuclease T1 preferentially cleaves after G residues. Thus, the positions of G residues can be determined by partially digesting a labeled RNA molecule with ribonuclease T1, which produces a variety of fragments, each ending in a G. Likewise, ribonuclease U2 cleaves after A residues; RNase PhyM cuts after U and A; RNase *B. cereus* cuts after U and C; and RNase CL cuts after C. The positions of each of these fragments on a gel are then compared with the

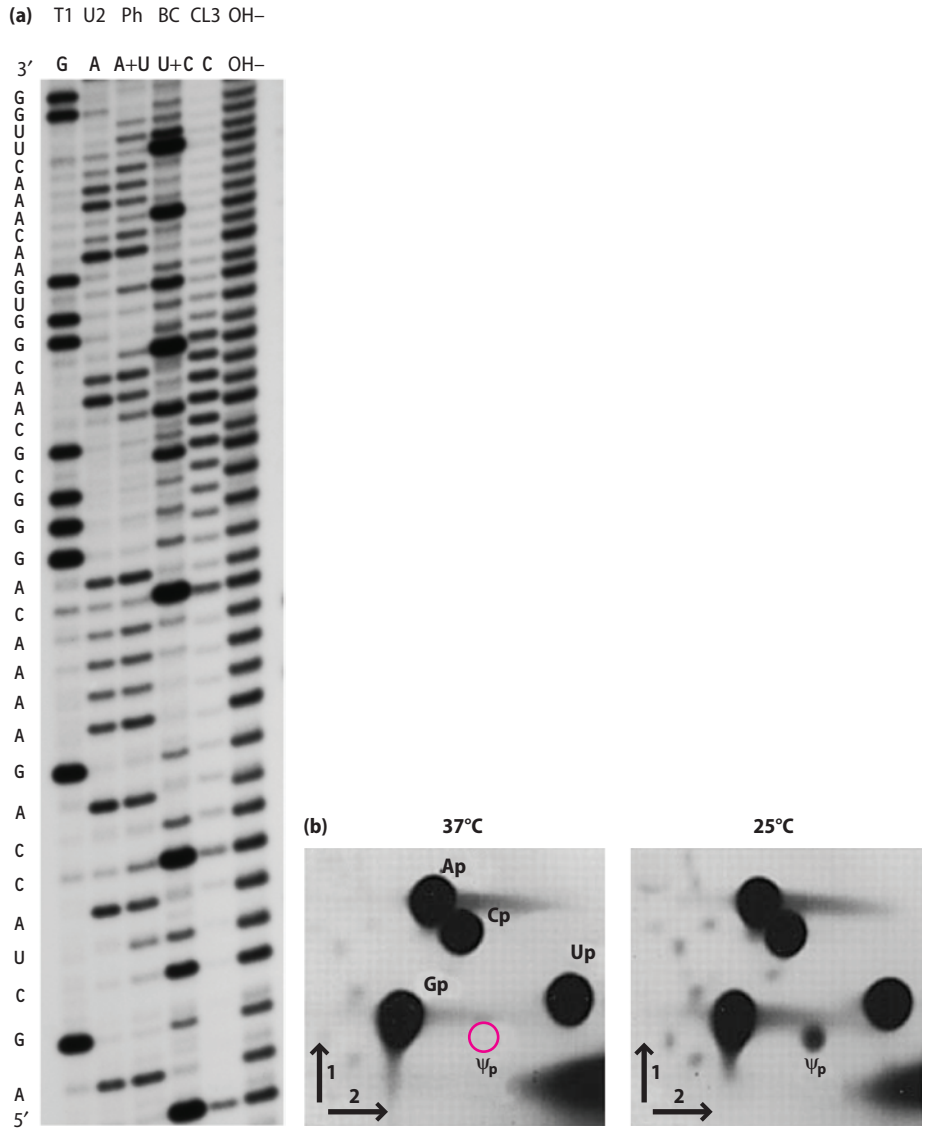


Figure 19.55 Sequencing RNA polymers. (a) Nuclease and chemical digestion of RNA. 5′ end-labeled RNA was prepared by *in vitro* transcription, subjected to partial digestion with ribonucleases T1, U2, PhyM (Ph), *B. cereus* (BC), CL3, or with base (OH^-), and analyzed by denaturing polyacrylamide gel electrophoresis. The treatment with base generates a ladder of products resulting from cleavage at each nucleotide position, whereas T1 only cleaves after guanosine (thus the incomplete banding pattern), U2 cuts after A, PhyM cuts after U and A, *B. cereus* after U and C, and CL3 cuts after C. The sequence of this RNA, as interpreted from these data, is shown on the left side. (b) Detection of a modified nucleotide by two-dimensional TLC. *In vivo* labeled ^{32}P pre-rRNA was isolated at 37°C or 25°C from yeast containing a temperature-sensitive allele in the pathway for RNA pseudouridylation. Thus, pseudouridylation is expected to happen at 25°C, but not at 37°C. The RNA was digested with RNase T2 and analyzed by two-dimensional cellulose TLC; the samples were run on a TLC plate with one solvent mixture, and the plate is then rotated in 90° and run in the second dimension with another solvent mixture. Arrows indicate the two directions of chromatography. Spots corresponding to the Ap, Gp, Cp, and Up residues are indicated. Pseudouridine 3′-monophosphate, whose expected position is indicated in each panel (Ψ_p), is not detectable in the heat-treated population (37°C, on left).

(a) from Carol Greider (b) from Bousquet-Antonelli, C., Henry, Y., Gélugne, J-P., Caizergues-Ferrer, M., and Kiss, T. (1997). A small nucleolar RNP protein is required for pseudouridylation of eukaryotic ribosomal RNAs. *EMBO Journal* **16**: 4770–4776.

positions of fragments resulting from a partial base (OH^-) digestion of the same labeled RNA fragment, where a band is seen for each consecutive nucleotide (see Figure 19.55a).

Another approach is to use chemicals that modify bases at particular positions and then to detect the location of the modified base. Dimethylsulfate, for example, modifies the N1 position of A residues, the N3 position of C residues, and the N7 position of G residues. When a G residue is modified at the N7 position, treatment of the RNA strand with the chemical aniline will cause strand cleavage at the modified base. The fragments can then be separated based on size, as described for the enzymatically cleaved RNA.

Another method of sequence analysis using chemical modification is primer extension. The RNA to be analyzed is annealed to a radiolabeled primer, which is then extended by reverse transcriptase. In the case of the modified A and C residues, where the modification happens on the face of the nucleotide involved in Watson–Crick base-pairing, the reverse transcriptase cannot place a nucleotide opposite the modified base and extension will be aborted. Since different RNA

molecules will be modified at different places, the lengths of the reverse transcriptase products, as analyzed by gel electrophoresis, will indicate the positions of the A and C residues. Other chemical reagents typically used for RNA modification include kethoxal (G-specific), 1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (CMCT) (U-specific), and *N*-methylisotoic anhydride (NMIA) (acylation at the ribose 2′ OH position).

RNA molecules intrinsically may contain both standard and modified ribonucleotides. Because post-transcriptionally modified nucleotides sometimes behave like standard nucleotides in typical sequencing methods, their detection may require the use of other methods such as TLC (see Section 19.8). This approach, for example, can be used to detect pseudouridine in an RNA molecule, as illustrated in Figure 19.55b.

The application of chemical probing to ribosome structural analysis is described in Experimental approach 11.1.

The N-terminal sequence of a protein can be determined by derivatization and cleavage

The N-terminal sequences of a protein can be determined by sequential derivatization, release, and identification of the N-terminal amino acid in a process called **Edman degradation**, a process depicted in Figure 19.56. In this procedure, the most N-terminal amino acid is specifically modified by phenylisothiocyanate (PITC). This modified amino acid is then specifically cleaved from the protein by mild acid treatment (which does not affect the remainder of the protein). The released derivatized amino acid is then identified by column chromatography, such as ion exchange chromatography (see Section 19.8), in which each amino acid has a characteristic elution profile. The cycle is then repeated with the next amino acid, and so on.

There are two limitations of this method: (i) some modified residues are resistant to PITC modification or acid hydrolysis; and (ii) the ability to identify an amino acid decreases with each round of cleavage (due to limitations in efficiency as the process is iterated). In practice, only a limited number of N-terminal amino acids (around 50–60 residues) can be identified through this approach, though, in many cases, this number of residues correctly identified may be sufficient for protein identification, especially in organisms whose entire genome sequence is known and protein sequences can be predicted. There are also methods for determining the C-terminal sequence of a protein based on the same principles as N-terminal sequencing, but these approaches are more problematic and are much less widely used.

Mass spectrometry can be used to identify most molecules

Mass spectrometry allows the identification of the widest spectrum of molecules, even if the molecules of interest are only present in small amounts. Mass spectrometry is used extensively to identify proteins following their separation in, and

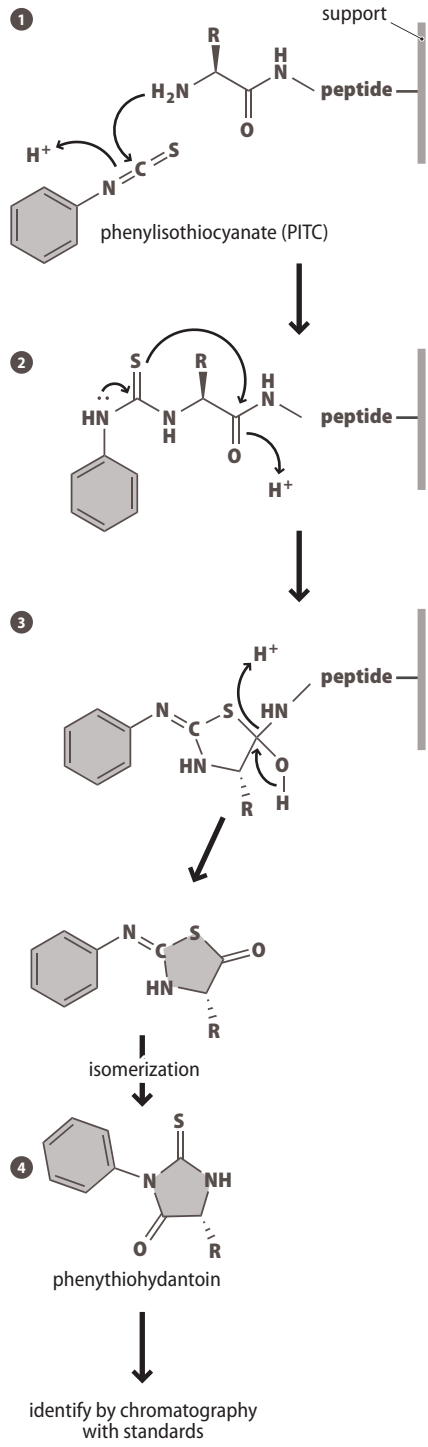
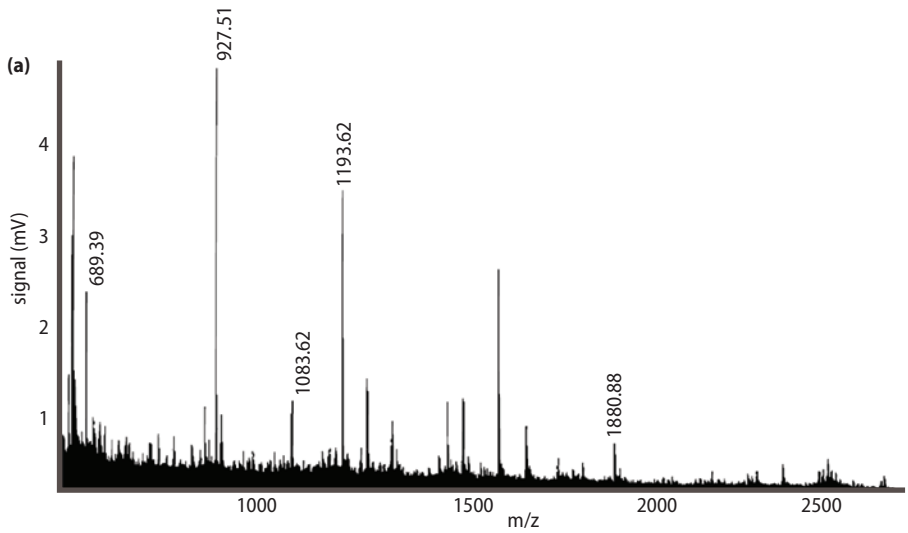


Figure 19.56 Steps in Edman protein degradation. The peptide is adsorbed to a solid support via its C-terminus and sequenced using sequential Edman degradation steps, in which one residue is removed from the N-terminus per cycle. Each cycle has the following steps: (1) PITC is reacted with the primary amine function of the peptide. (2) Under mildly alkaline conditions, a phenylthiocarbamoyl derivative is formed. (3) Mild acid treatment causes the terminal amino acid to be released, freeing the N-terminus of the peptide for another cycle of PITC reaction. (4) The released amino acid isomerizes to form a phenylthiohydantoin species that can be identified by chromatography or mass spectrometry.



(b) Calculated mass

amino acid	calculated mass (monoisomers)	sequence
212–217	688.37	AWSVAR
137–143	926.49	YLYEIAR
1–10	1192.59	DTHKSEIAHR
323–335	1566.74	DAFLGSFLYEYSR
484–499	1879.91	RPCFSALTPDETYVPK

Figure 19.57 Representative mass spectrometric analysis. (a) BSA was digested with trypsin, and a MALDI spectrum was generated. The mass/charge (*m*/*z*) ratios are indicated on the *x*-axis and the relative intensity is plotted on the *y*-axis. (b) Table showing the tryptic fragments of BSA that were detected and their calculated mass (accounting for the relative abundance of different amino acids).

From Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels. *Analytical Chemistry* **68**: 850–858.

excision from, a polyacrylamide gel; it can also be used to identify carbohydrates, nucleic acids, and small molecules.

The method of mass spectrometry makes it possible to determine the molecular weight of a molecule with very high accuracy by measuring its mass-to-charge ratio. A small amount of the material of interest must first be converted to gas-phase ions. This can be achieved by bombarding a sample embedded in a specific chemical matrix with a laser (matrix-assisted laser desorption/ionization or MALDI) or by generating a spray of the sample in the presence of a strong electric field (electrospray ionization or ESI). The ionized molecules are separated according to their mass-to-charge ratios, and the mass of the separated ions is then measured upon striking a detector (see Figure 19.57a). The mass of small biological molecules can be determined to within a single dalton (Da), while the mass of larger biological molecules can be determined with somewhat less accuracy.

Often, particularly in the case of larger molecules, the mass resolution is not sufficient to allow for unambiguous identification of a molecular species. In these cases, the molecules are usually digested or hydrolyzed into smaller fragments whose individual masses can be determined with greater accuracy. For example, proteins can be digested with specific proteases, such as trypsin, and the masses of each of the resulting fragments determined (see Figure 19.57b).

For organisms whose entire genome sequence is known, this information is usually sufficient to identify the protein, based on the predicted coding regions. If identification is not possible based on the fragment masses alone, however, the proteolytic fragments can be broken down further to the individual amino acids whose masses can then be determined. This is often referred to as tandem mass spectrometry (or MS/MS)—first, the mass of the proteolytic fragment is determined, and then the masses of the amino acids in the fragment are determined, thereby identifying the exact amino acid composition of the fragment.

Mass spectrometry has become an important tool in studying protein biology. Using this approach, it is possible to determine not only the sequence of a protein,

but also whether amino acids are modified—for example, by phosphorylation. Proteins can be identified not only after purification, but also in complex mixtures containing hundreds of proteins. For example, one can use mass spectrometry to determine the composition of a large multiprotein complex, or even how different growth conditions affect the abundance of different cellular proteins. This use of mass spectrometry will be discussed further in Section 19.13.

BLAST sequence alignment can be used to assign functions to genes

Once the sequence of DNA, RNA, or protein is obtained, it is possible to get some information about its function by examining regions of homology in other proteins or genes. As we discussed in Chapter 18, sequence homology implies that two sequences were derived at some point from a common ancestor. As such, one of the most common and powerful methods to infer the function of an unknown gene is to compare its sequence to that of genes with known functions.

The comparison of two or more sequences is called an **alignment**, in which the nucleotide or amino acid query sequence is compared with the nucleotide or amino acid sequence of known genes or proteins. After the alignment has been performed, it is assigned a score, based on whether the residues are identical, similar, or different. Such alignments make it possible to identify other proteins or RNA molecules that are homologs with a common evolutionary origin, and therefore are likely to carry out similar functions or activities. The region of sequence similarity may encompass an entire coding sequence, or it may be restricted to one or more protein domains (for example, kinase domains, zinc finger DNA-binding domains, or carbohydrate-binding domains). The presence of an identifiable domain whose function is known in other proteins often provides clues about the function of an uncharacterized protein.

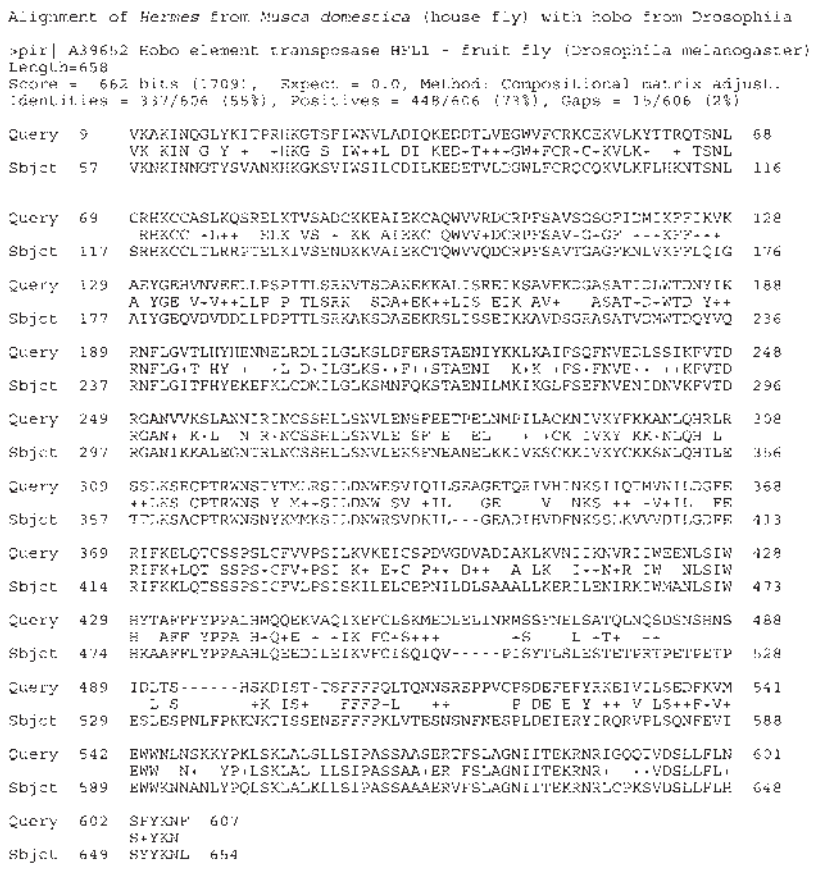
The most widely used method for sequence alignments is a suite of related programs known collectively as **BLAST** (Basic Local Alignment Sequence Tool), which are available from the National Center for Biotechnology Information (NCBI) website and on many other web servers worldwide.

You can find BLAST on the NCBI website at http://blast.ncbi.nlm.nih.gov/.

The simplest BLAST alignments are done as pairwise alignments (that is, comparisons between a pair of sequences) using the query sequence and each sequence in a chosen database. BLAST algorithms can align either DNA sequences (BLASTN) or protein sequences (BLASTP). For proteins, the program starts with a string of three amino acids and searches for homology; it then moves in a predefined "window" to examine adjacent sequences for additional homology. For DNA, an 11-nucleotide string is the initial sequence length that is probed.

After the alignment is optimized, it is given a score that provides a measure of how closely two sequences match each other. The score depends on the amount of similarity between the two sequences and the length of sequence that is similar. Gaps may be introduced into one sequence or the other to maximize the alignment of two sequences, although a penalty in the score is assigned for each gap. The score of the homology is expressed as an E value, which is a measure of the probability that the sequence alignment observed could have occurred by chance. The lower the E value, the better the homology, as the alignment in question is less likely to have occurred by chance.

An example of the output from a BLASTP search is given in Figure 19.58, which shows the results of a search for proteins similar in sequence to the Hermes transposase from the housefly *Musca domestica* (the query sequence). Aside from finding other version of Hermes transposase in the database, the BLAST search



```
Alignment of Hermes from Musca domestica (house fly) with Hobo from Drosophila

>pir|| A39652 Hobo element transposase HFL1 - fruit fly (Drosophila melanogaster)
Length=658
Score =  662 bits (1709),  Expect = 0.0, Method: Compositional matrix adjust..
Identities = 337/606 (55%), Positives = 448/606 (73%), Gaps = 15/606 (2%)

Query  9    VKAKINQGLYKITPRHKGTSFIWNVLADIQKEDDTLVEGWVFCRKCEKVLKYTTRQTSNL  68
            VK KIN G Y +  +HKG S IW++L DI KED-T+++GW+FCR-C-KVLK-  + TSNL
Sbjct  57   VKNKINNGTYSVANKHKGKSVIWSILCDILKEDETVLDGWLFCRQCQKVLKFLHKNTSNL  116


Query  69   CRHKCCASLKQSRELKTVSADCKKEAIEKCAQWVVRDCRPFSAVSQSCPIDMIKPFIKVK  128
             RHKCC +L++  ELK VS + KK AIEKC QWVV+DCRPFSAV-Q-CP  +++KPF+++
Sbjct  117  SRHKCCLILRRFTELKIVSENDKKVAIEKCTQWVVQDCRPFSAVTGAGPKNLVKPFLQIG  176

Query  129  ARYGRHVNVRELLPSPTTLSRKVTSDAKEKKALISREIKSAVEKDGASATIDLWTDNYIK  188
            A YGE V+V++LLP P TLSRK  SDA+EK++LIS EIK AV+   ASAT+D+WTD Y++
Sbjct  177  AIYGEQVDVDDLLPDPTTLSRKAKSDAEEKRSLISSEIKKAVDSGRASATVDMWTDQYVQ  236

Query  189  RNFLGVTLHYIENNELRDLILGLKSLDFERSTAENIYKKLKAIFSQFNVEDLSSIKFVTD  248
            RNFLG+T HY +  +L D+ILGLKS++F++STAENI  K+K +FS+FNVE++ ++KFVTD
Sbjct  237  RNFLGITFHYEKEFKLCDMILGLKSMNFQKSTAENILMKIKGLFSEFNVENIDNVKFVTD  296

Query  249  RGANVVKSLANNIRINCSSELLSNVLENSPEETPELNMPILACKNIVKYFKKANLQHRLR  308
            RGAN+ K+L  N R+NCSSELLSNVLE SP E  EL   + +CK IVKY KK+NLQH L
Sbjct  297  RGANIKKALECNTRLNCSSELLSNVLEKSPNEANELKKIVKSCKKIVKYCKKSNLQHTLE  356

Query  309  SSLKSECPTRWNSTYTMLRSTLDNWESVIQTLSRAGRTQEIVHTNKSTIQTMVNTLDGFE  368
            ++LKS CPTRWNS Y M+++STLDNW SV +TL   GR    V  NKS ++ -V+TL  FE
Sbjct  357  TTLKSACPTRWNSNYKMMKSTLDNWRSVDKTL---GRADIHVDFNKSSLKVVVDTLGDFE  413

Query  369  RIFKDLQTCSSPSLCFVVPSILKVKEICSPDVGDVADIAKLKVNIIKNVRIIWEENLSIW  428
            RIFK+LQT SSPS+CFV+PSI K+ E+C P++ D++  A LK  I++N+R IW  NLSIW
Sbjct  414  RIFKKLQTSSSPSICFVLPSISKILELCEPNILDLSAAALLKERILENIRKIWMANLSIW  473

Query  429  HYTAPFFYPPALHMQQEKVAQIKEFCLSKMEDLELINRMSSPNELSATQLNQSDSNSHNS  488
            H  AFF YPPA H+Q+E + +IK FC+S+++        +S    L +T+  ++
Sbjct  474  HKAAFFLYPPAAHLQEEDILEIKVFCISQLQV-----PISYTLSLESTETFRTPETPETP  528

Query  489  IDLTS------HSKDIST-TSFFFPQLTQNNSREPPVCPSDEFEFYRKEIVILSEDFKVM  541
              L S       +K IS+   FFFP-L   ++       P DE E Y ++ V LS++F+V+
Sbjct  529  ESLESPNLFPKKNKTISSENEFFFPKLVTESNSNFNESPLDEIERYIRQRVPLSQNFEVI  588

Query  542  EWWNLNSKKYPKLSKLALSLLSIPASSAASERTFSLAGNIITEKRNRIGQQTVDSLLFLN  601
            EWW  N+  YP+LSKLAL LLSIPASSAA+ER FSLAGNIITEKRNR+  ++VDSLLFL+
Sbjct  589  EWWKNNANLYPQLSKLALKLLSIPASSAAAERVFSLAGNIITEKRNRLCPKSVDSLLFLH  648

Query  602  SFYKNF  607
            S+YKN
Sbjct  649  SYYKNL  654
```

Figure 19.58 BLASTP search using Hermes transposase as a query. BLAST alignment of the protein sequence (amino acid one letter code) between the Hermes transposase (Query) and the closely related Hobo transposase from *Drosophila* (Sbjct = Subject). The image shown was taken from the BLAST website. The alignment starts with amino acid number 9 within the Hermes sequence and with amino acid 57 in Hobo, as shown by the numbers 9 and 57 to the left of the aligned sequences. The numbers in each row to the right indicate the position in the sequence of the last amino acid in the row to allow orientation of the sequences over the many lines. The amino acids that are identical between the two sequences are indicated by the identical letter placed at that position between the two sequences, and positions where the two proteins have similar amino acids are indicated with a plus sign. When gaps have to be introduced in one of the two sequences to optimize the alignment, dashed lines are shown between amino acids.

also found homology with the Hobo transposase, which is an orthologous protein from *Drosophila*.

While BLAST searches are the most commonly used searches for initially examining sequence alignments, there are a wealth of other programs that compare sequences and help identify specific domains, sites of protein modification, protein interaction domains, and other features. An increasing number of protein database entries incorporate multiple parameters, including, for example, structural information. As always, scientists need to be critical when interpreting the results of the alignment and domain-finding programs. For example, a protein of interest may share extensive homology, and even structural similarity, with a particular enzyme but have substitutions in critical active site residues, meaning that it would be flawed to conclude that the protein of interest has the particular enzymatic activity.

19.10 OBTAINING AND ANALYZING SEQUENCES ON A GENOMIC SCALE

The availability of complete genome sequences has revolutionized molecular biology. Obtaining and using sequence information require methods for both high-throughput DNA sequencing, as well as computational analyses of the sequences obtained. The first genome to be completely sequenced, that of *Haemophilus influenzae*, was completed in 1995. Less than 20 years later, the sequences of thousands

of genomes have been reported. The pace at which new genomes are sequenced continues to accelerate, driven by advances in DNA sequencing technology, as well as in the methods for their analysis. Here we explore the rapidly evolving methods used to generate genome sequence information.

The sequence of a whole genome can be determined by analyzing many overlapping DNA fragments

How does one obtain the complete sequence of a genome? We lack the technology to sequence an entire chromosome from one end to the other, so other approaches must be used. Sequencing methods are rapidly evolving, but they are typically unified by a common underlying approach. The general strategy of whole genome sequencing is to obtain a collection of many smaller overlapping DNA fragments that together represent the entire genome. As we will learn later, there are different methods for generating these DNA fragments; the key point is that, at the start, it is not necessary to know where each fragment came from in the genome. By determining the base sequence of each fragment and then comparing the sequences of many overlapping regions, the complete sequence of the genome can be assembled in the correct order with the aid of a computer, as illustrated in Figure 19.59. This strategy is known as "shotgun" sequencing because it begins with random fragmentation of the genome. When the human genome was first sequenced, the small overlapping DNA fragments were cloned into a library that was grown in *E. coli* to generate sufficient DNA for sequencing. Today, most protocols use PCR amplification to generate DNA for sequencing.

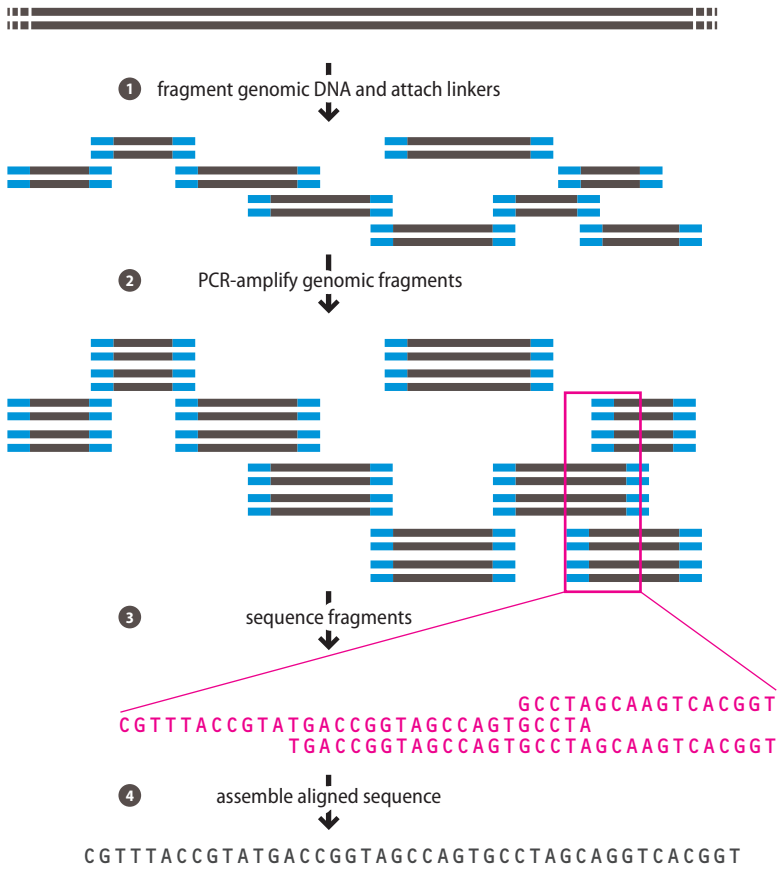


Figure 19.59 Assembly of a genome sequence from the sequences of multiple overlapping fragments. Genomic DNA is first fragmented by shearing or restriction enzyme digestion to generate a collection of thousands or millions of random overlapping fragments, after which linkers (blue) are ligated to all of the ends (step 1). The genomic fragments are amplified by PCR using primers complementary to the attached linkers (step 2). After sequencing of all of the fragments (step 3), computer programs can remove the linker sequence and assemble a complete genomic sequence by detecting overlaps between the many random fragments (step 4).

Genome sequencing requires over-sequencing

In order to obtain a complete genome sequence, it is necessary to sequence many more base pairs than the total length of the genome. There are three reasons for this. First, DNA sequencing is not completely error-free. These errors can be detected and eliminated by having multiple sequences of the same region. Second, since DNA fragments are sequenced randomly, many different overlapping fragments must be sequenced to ensure that all regions of the genome have been sequenced at least once. In theory, clones corresponding to eight genome equivalents need to be sequenced to ensure coverage of 99.9% of the genome. For example, to obtain a nearly complete sequence of the 4.6 Mbp *E. coli* chromosome using shotgun sequencing, it is necessary to sequence about 37 Mbp of DNA (eight genome equivalents). In practice, even higher coverage is often sought, as it has the additional benefit of increasing the accuracy of the final sequence. This is because an error in the sequence of a single fragment can be recognized and corrected when many independent clones corresponding to the same region are sequenced.

The third reason that a complete genome sequence requires over-sequencing is that a sufficient number of random overlapping DNA segments are needed in order to determine the correct order in which overlapping sequences must be aligned to assemble the complete genome. The presence of repetitive DNA in multiple genomic locations complicates the task because they make it difficult to uniquely align the fragments containing identical repetitive sequences. This task becomes nearly impossible if the regions of the genome containing many repeats are longer that the sequenced shotgun fragments. Computer sequence alignment will not assemble this part of the genome from overlapping sequences because the repeat will have homology in many different regions.

This inherent limitation in the shotgun technique is the reason that the first genome sequences obtained for many organisms, including humans, had regions that were missing. However, increasingly sophisticated computer assembly algorithms and the inclusion of genetic linkage data are facilitating the assembly of regions containing repetitive DNA sequences. In addition, new methods of sequencing are making it possible to sequence longer regions and thus connect the unique sequences that flank these repetitive regions.

DNA sequences can be obtained by a variety of methods

For many years, the primary method used to sequence whole genomes was an approach called chain-termination sequencing, which uses ddNTPs as chain terminators, as described in Section 19.9. This DNA sequencing method was originally performed by hand and analyzed by gel electrophoresis. Later, machines were developed that automated the sequencing and detection process.

A single high-throughput automated chain-termination sequencing machine can process about 1000 reactions each day, generating about 0.5 Mbp of total sequence. (Each sequencing reaction is termed a "read.") This process would have to be repeated about 75 times to sequence eight genome equivalents of the 4.6 Mbp *E. coli* genome (37 Mbp). To sequence the human genome (3×10^9 bp), about 70 million chain-termination sequencing reactions, each generating 500 bp of sequence, were needed to generate ten genome equivalents of sequence, or around

35×10^9 bp. This immense effort required 100 automatic sequencing machines generating 1000 sequencing reactions per day for about two years.

Recent advances in sequencing technology have greatly increased the speed of DNA sequencing, while at the same time dramatically reducing the cost. These new methods are collectively called NextGen sequencing or, more commonly, deep sequencing. These methods share certain features, most importantly that they are "massively parallel," meaning that huge numbers of unique DNA molecules can be sequenced simultaneously, thus generating a large number of unique reads that can then be aligned. In current deep sequencing methods, DNA fragments generated by PCR directly from genomic DNA are sequenced. New so-called single-molecule methods are currently being developed that will make it possible to sequence individual DNA molecules directly, eliminating the need for PCR and thereby avoiding the errors that can be introduced during PCR amplification.

While the pace of innovation in sequencing technology exceeds the pace of textbook writing, we will describe the details for one method, known as Illumina sequencing (after the company that developed the method), that is currently widely used. This method utilizes fluorescently labeled nucleotides that are chemically blocked at the 3′ end, so that they can be added by a polymerase, one at a time, to a growing DNA strand.

In the Illumina sequencing method, illustrated schematically in Figure 19.60, the genomic DNA is fragmented and oligonucleotide adapter segments are ligated to the ends of the resulting DNA fragments. The DNA strands are then separated, and the single strands are allowed to hybridize to a surface that is densely coated with a mixture of immobilized oligonucleotides corresponding to single strands of the adapters. Hybridization of both ends of each DNA fragment with a complementary adapter immobilized on the surface thus generates a bridge (see Figure 19.60). After PCR amplification, each spot on the array will contain a cluster of identical sequences. This high local concentration of each unique sequence is needed to produce a sufficiently robust fluorescent signal in the steps that follow.

The DNA in each spot is denatured to make it single-stranded, in preparation for sequencing. Fluorescently labeled nucleotides (A, T, G, and C, each labeled with a differently colored fluorophore to allow them to be distinguished) are then added, along with a primer and DNA polymerase. Since the nucleotides are also chemically modified such that the 3′ end is blocked, the polymerase can add only one base at a time. A laser excites the fluorophore, while a camera monitors whether the color of the emitted fluorescence corresponds to the addition of an A, T, G, or C base. The DNA is then chemically treated to remove the chemical block and the fluorophore, thus exposing the 3′ end and making it available for the next round of fluorescent nucleotide incorporation. By repeating this process for many cycles, hundreds of nucleotides can be sequenced rapidly.

The Illumina sequencing method, like other deep sequencing methods, typically produces short reads of 75–150 bp. Recall that, in whole genome shotgun sequencing, the genome sequence is assembled by aligning overlapping regions among individual sequence reads. The short reads from the Illumina method thus compound the problem of sequence assembly and make it necessary to have a greater depth of coverage of the genome to allow correct assembly of the final sequence.

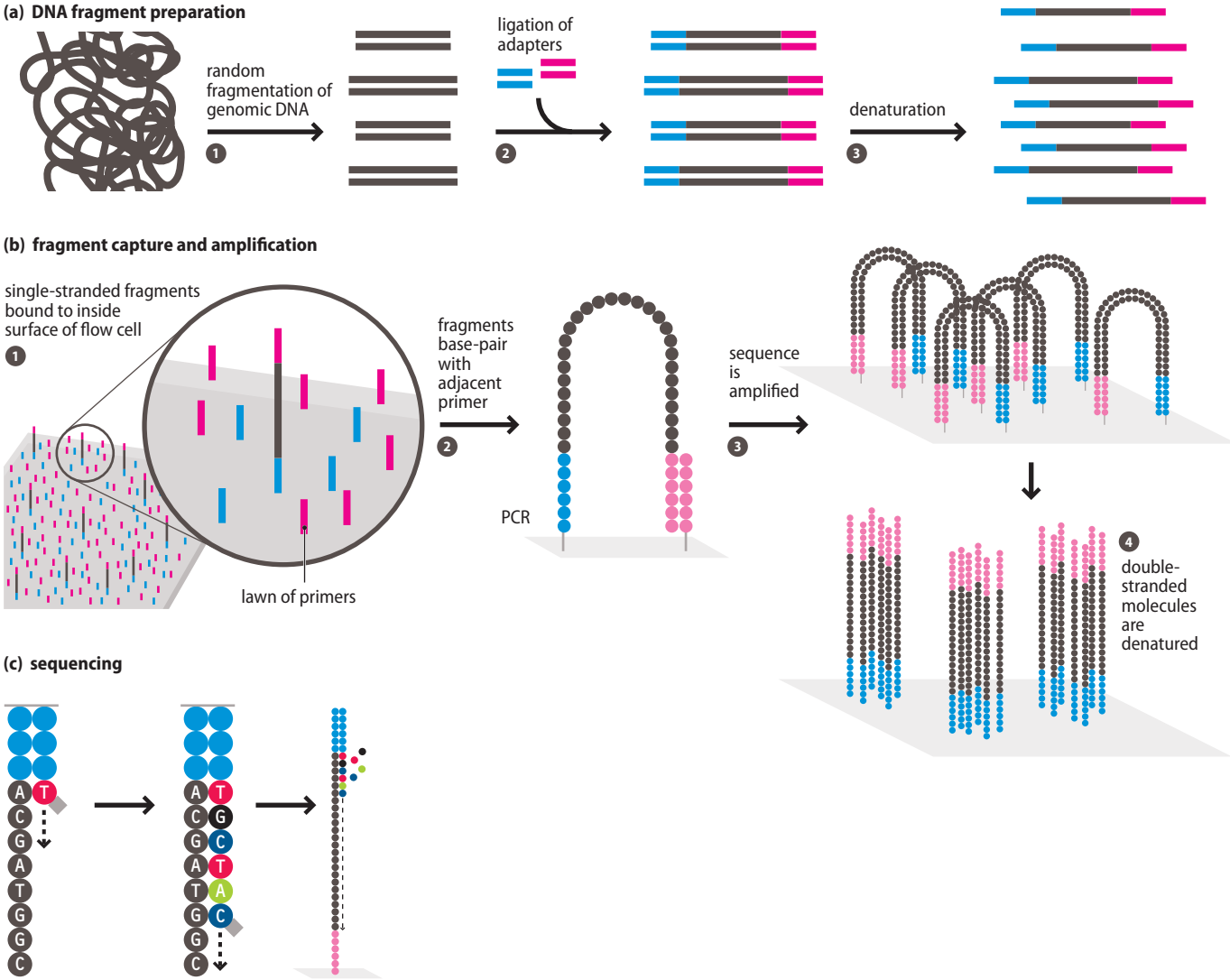


Figure 19.60 Illumina sequencing. (a) DNA fragment preparation. The DNA to be sequenced is broken into fragments by shearing or restriction digestion (step 1). Adapters are ligated to the ends of the fragments (step 2), after which the DNA is denatured (step 3). (b) Fragment capture and amplification. The fragments are attached to a surface decorated with a dense lawn of primers complementary to the adaptors (step 1). The attached fragments base-pair with the adjacent primers (step 2). The individual fragments are then amplified by unlabeled nucleotides using an adjacent primer to form a double-strand molecular bridge (step 3). Finally, the DNAs are denatured to yield clusters of identical fragments that are sequenced in parallel (step 4). (c) Sequencing showing a single fragment. A specific primer and four 3′ blocked nucleotides, each with a different fluorescent label, are added. After washing away the unincorporated nucleotides, laser excitation and optical detection identify the added nucleotide at each cluster position. The 3′ blocks are removed, and the cycle is repeated.

From Mardis (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics.* **9**:387–402.

At the time of writing, the Illumina method is the most popular because it is easy to automate and is highly accurate. However, many other methods are rapidly being developed. New methods to generate read lengths of several kilobases greatly simplify the alignment and assembly of genome sequence. Unlike the Illumina method, which relies on PCR amplification, some of the methods under development can directly sequence single molecules of genomic DNA without the need for PCR, as noted earlier. In addition, there has been significant innovation in the methods used to detect specific nucleotide incorporation. While the Illumina method uses a laser and camera to detect a fluorescent signal, some of the newer methods rely on changes in electrical properties or pH to detect the nucleotide added at a particular position, thus reducing the cost of the laser detector.

Whole genome sequencing has many uses

As we described in Chapter 18, deep sequencing is a powerful tool for identifying sequence variations in the genome that may play a role in human disease. Deep sequencing can also be used to rapidly re-sequence entire genomes and thereby identify naturally occurring genetic variation. In such applications, a reference genome sequence is used as a template to align short sequence reads into an already assembled complete genome, greatly reducing the difficulties in assembly. As sequencing becomes even less expensive, whole genome sequencing has become a powerful means of identifying disease-associated mutations.

With modifications to the DNA preparation protocol, other features of genomic DNA, such as nucleotide modifications and sequences bound by proteins, can be identified. For example, to detect cytosine methylation across a whole genome, the DNA can be treated with the chemical bisulfite, which reacts with unmodified cytosines such that they will be "read" as T, while methylated cytosines, which are not modified, will be "read" as C. As we will discuss in Section 19.15, proteins also can be cross-linked to the genomic DNA, and the corresponding sequences bound by the proteins determined by deep sequencing. This approach has even been extended to identifying DNA sequences that interact with each other in the compact chromatin that we learned about in Chapter 4.

Exome sequencing reduces the amount of genomic sequence information that needs to be analyzed

As we learned in in Chapter 1, protein-coding genes represent only 1.5% of the human genome. With the assumption that many of the physiological changes that underpin human disease are a consequence of changes within protein-coding genes, an approach termed "exome sequencing" was developed. This method seeks to determine the sequence of all protein-coding exons in the genome, rather than the genome sequence in its entirety.

Since sifting through trillions of base pairs to find those few that may be altered in those exhibiting a disorder or disease requires extensive computational analysis, the amount of sequence information that needs to be obtained and sorted can be significantly reduced by focusing solely on protein-coding genes. By only sequencing exons that are thought to be expressed as proteins, however, this method misses mutations in promoters and splice sites that might affect gene function. Neverthless, the advantages of only having to sequence and interpret nucleotide changes for 1.5% of the genome in many cases outweigh the potential disadvantage of missing a functionally important base change in non-coding regions. Exome sequencing has therefore become a common technique in recent years.

The sequencing step for exome sequencing is the same as for the deep sequencing methods we have just described, with one additional element—the enrichment for protein-coding genes. After genomic DNA fragments are amplified, the double-stranded DNA is denatured and hybridized with the RNA molecules that correspond to all known exons in the human genome. These RNA sequence probes are biotinylated, which allows them to be captured by magnetic beads coated with streptavidin, as shown in Figure 19.61. Since the RNA probes correspond only to exons, only DNA fragments containing protein-coding genes will hybridize with the RNA. Once the RNA–DNA hybrids are formed and precipitated with the streptavidin-coated beads, the remaining 98.5% of the genomic DNA is

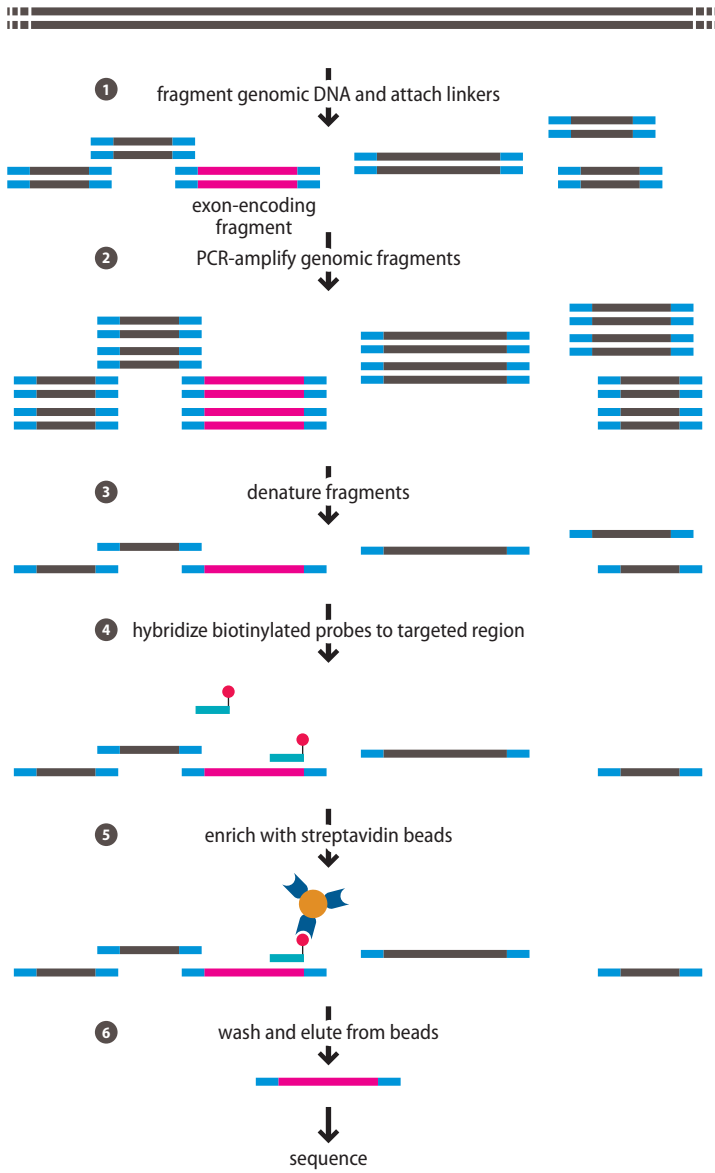


Figure 19.61 Exome sequencing. As for the sequencing of complete genomes, linkers are attached to fragmented genomic DNA (step 1), and the fragments are amplified by PCR (step 2). Subsequently, the fragments are denatured (step 3) and incubated with biotinylated RNA probes (aqua with red biotin) corresponding to all known exons shown in purple (pink) (step 4). The DNA fragments hybridizing to the biotinylated probes can be captured by precipitating with streptavidin-coated beads, which strongly bind the biotin at the end of the RNA probes (step 5). After the unbound DNA is washed away, the remaining exon-containing fragments can be released from the RNA probes (step 6) and sequenced.

washed away. The purified DNA is then eluted and sequenced by the Illumina or other deep sequencing methods. As the cost and speed of whole genome re-sequencing continue to decrease and the ability to analyze complex data sets increases, the use of exome sequencing is likely to decline in favor of whole genome sequencing.

The sequence of the complete transcriptome can be determined by deep sequencing

The cDNA copies of RNA sequences that we learned about in Section 19.3 can also be subject to deep sequencing, which allows us to determine the sequences of all RNAs in a cell, denoted the transcriptome, or specific subsets of RNAs. As we will discuss more in Section 19.12, this method, termed RNA-Seq, can be used to monitor the prevalence of transcripts in particular cells or tissues. Similar to

the experiments carried out for the detection of modified DNA bases, RNA can be treated with specific enzymes or chemicals that react with modified ribonucleotides in ways that can be detected by deep sequencing; such detection allows transcriptome-wide identification of the modifications. Finally, we will also learn in Section 19.15 that deep sequencing can be used to identify the RNA sequences bound by specific proteins or protein complexes transcriptome-wide in a manner similar to the genome-wide identification of protein-binding sites on DNA.

19.11 DETECTION OF SPECIFIC DNA SEQUENCES

The base-pairing ability inherent in a DNA sequence can be exploited to detect a specific DNA sequence amid a vast excess of unrelated sequences. In the previous sections, we discussed how the sequence of a specific DNA segment, or even the genome, is determined; in this section, we describe a number of methods by which a known DNA segment can be identified and distinguished from other DNA sequences in the context of many DNA fragments, a pool of DNA clones, or even chromosomes within a cell.

DNA molecules can be distinguished by their restriction enzyme digestion patterns

As discussed in Section 19.4, restriction enzymes are DNA endonucleases that cleave DNA at specific sequences. Different restriction enzymes recognize different sequences, as shown in Figure 19.23. These sequences are often palindromic and are between 4 and 8 bp in length. Restriction enzymes usually cleave DNA either at the restriction site or at a fixed distance from it. By choosing the right restriction enzyme (or enzymes), it is possible to distinguish between different DNA molecules and to learn something about their structure.

Consider, for example, a bacterial plasmid called pBR322, shown in Figure 19.62a. The sequence of this plasmid contains restriction sites for dozens of different restriction enzymes, of which only two are shown. Some restriction enzymes, such as *Eco*RI, have a single restriction site in the pBR322 sequence; some enzymes, such as *Acl*I, have multiple restriction sites, and some enzymes have none. The size of the DNA fragments generated by a restriction enzyme digest can be determined by separating the fragments on an agarose gel (described in Section 19.8), as illustrated in Figure 19.62b. As such, restriction enzymes can be used to verify whether a cloning procedure worked as planned. In the example shown in Figure 19.62b, digestion by *Acl*I was used to confirm that a 1-kb fragment was inserted at the *Eco*RI site of pBR322, resulting in an increase in size in one of the *Acl*I fragments.

Specific DNA fragments can be detected by Southern blot hybridization

One method to detect a specific DNA fragment among many DNA fragments is called **Southern blot analysis**, which is illustrated in Figure 19.63. This method is named after its inventor Edwin Southern, and hence the name carries an upper case "S." (Note that in the names of "northern blot" and "**western blot**" techniques, which we will learn about in Sections 19.12 and 19.13, respectively, the "n" and

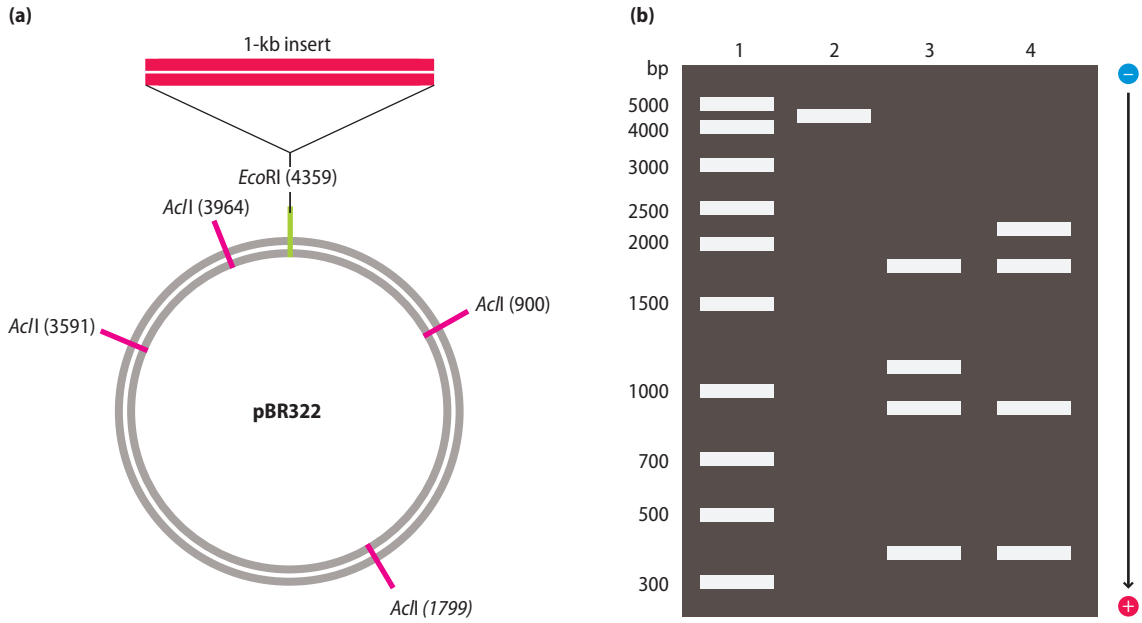


Figure 19.62 DNA analysis using restriction enzymes. (a) Plasmid pBR322 is 4361 bp in length. Shown is a map of the plasmid with the locations of the restriction sites for *Eco*RI (green) and *Acl*I (red). The numbers in parentheses indicate the position of the first base pair of the restriction site. The middle of the *Eco*RI site was designated as position 1. Also shown is the position of a 1-kb insert (cloned into *Eco*RI restriction site) examined in panel (b). (b) An illustration of an agarose gel used for separating different DNA molecules, which migrate, according to size, toward the positive pole. The gel is typically stained with a DNA-binding fluorescent dye (such as ethidium bromide) and visualized by exposing the gel to UV light. As a result, the DNA molecules appear as bright bands on a dark background. Lane 1 was used to separate a DNA ladder, which is a mixture of DNA fragments of known sizes (shown on the left). Plasmid pBR322 was digested with *Eco*RI (lane 2) to generate one DNA fragment, or with *Acl*I (lane 3) to generate four DNA fragments. Lane 4 shows the results of an *Acl*I digest of a pBR322 plasmid with a 1-kb fragment inserted at the *Eco*RI site. Notice that the band in lane 3 of ~1300 bp, corresponding to the fragment between the *Acl*I sites at 3964 and 900 bp, is missing from lane 4, and instead there is a band of ~2300 bp, consistent with a 1-kb insert at the *Eco*RI site.

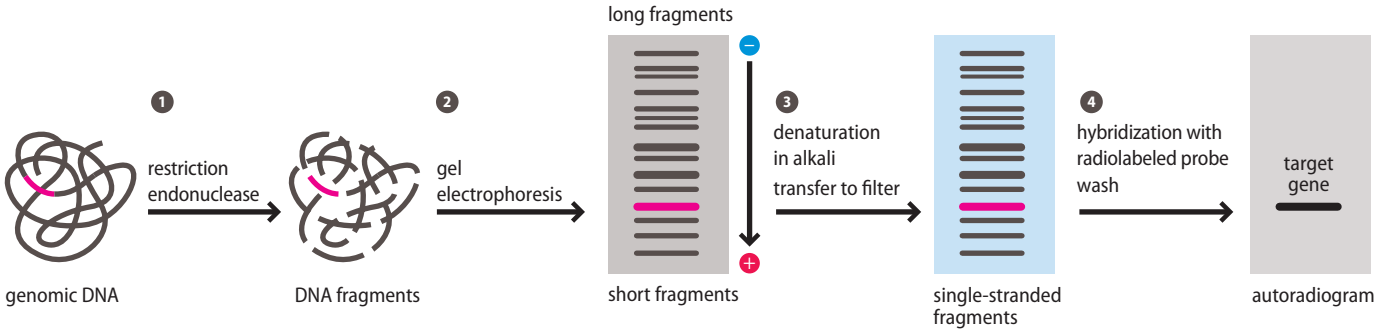


Figure 19.63 Southern blot analysis. In the example shown, the goal is to detect a region of interest (shown in pink) within the genomic DNA. The genomic DNA is first digested with restriction enzymes (step 1) and separated by gel electrophoresis (step 2). The DNA is then denatured and transferred to a filter (in blue, step 3). In parallel, a DNA fragment complementary to the region of interest is labeled, for example, with a radioactive isotope. The radioactive DNA probe is then hybridized to the filter (step 4). The probe will hybridize to its complementary DNA on the filter, which can be detected by exposing the filter to radiation-sensitive film.

"w" are not capitalized because neither term is derived from a person's name; rather, they are whimsical names given to these techniques as they are similar to the Southern blot technique.)

In overview, Southern blot analysis proceeds as follows. DNA fragments cut by specific restriction enzymes are separated on an acrylamide, agarose, or pulse-field gel, as described in Section 19.8. Without fragmentation, the DNA would be too large to enter the gel and it would not be possible to analyze specific genomic regions. The DNA fragments separated in the gel are denatured with alkali to expose the single strands for hybridization and are then transferred to a positively

charged membrane. The membrane is incubated with a single-stranded DNA probe that is complementary to the sequence to be identified and is labeled with either radioactivity, fluorescence, or a chemical tag. The labeled DNA anneals to its complementary strand, and the location of the target DNA fragment is identified by exposing the membrane to a device that can detect the probe (an x-ray film in the case of radioactivity) or to a machine that can detect fluorescence or color (in the case of fluorescent or chemical probes).

Using this method, one can determine whether a particular chromosomal locus is altered by the insertion or deletion of sequences, as the DNA fragment size will be either larger or smaller than expected. One can also learn if a particular locus is present in fewer or more copies in one sample, compared to another, because the intensity of the signal is somewhat proportional to the number of copies.

Experimental approach 17.1 describes how Southern blot analysis was used in the detection of a transposon inserted into the factor VIII gene that causes hemophilia.

Specific DNA clones can be identified in libraries by colony or plaque hybridization

We discuss the construction of libraries in Section 19.4.

A similar hybridization method can be used to identify a specific DNA clone in a library of many different clones. If the library is in a bacterial plasmid or BAC vector, colony hybridization can be used. Alternatively, if the library is represented in a bacteriophage vector, plaque hybridization can be used.

The process of colony hybridization is depicted in Figure 19.64. During this process, a collection of bacterial colonies, each of which arises from multiple divisions of a single cell containing one DNA clone, is transferred to a positively charged membrane, generating a copy of the plate on the membrane. The cells are then lysed *in situ* on the membrane.

In an analogous method called plaque hybridization, the bacteriophages that formed plaques on bacterial lawns are transferred to the membrane. The DNA from the colony or plaque is denatured with alkali and probed in the same way as for the Southern blots described earlier. Because the membrane is a copy of the original plate, the researcher can go back to the plate once a positive signal is obtained and isolate the bacterial colony or bacteriophage that gave rise to the positive signal indicative of the desired DNA fragment. The colony or bacteriophage can then be propagated. These hybridization approaches make it possible to identify and isolate a single clone that contains the specific sequence of interest from among thousands of clones in a library.

Chromosome abnormalities can be detected by chromosome spreads

Sometimes it is also of interest to detect larger regions of DNA, including whole chromosomes. The complete chromosome complement of an organism, known as a karyotype, is determined by visualizing stained chromosomes isolated from mitotic cells. One stain that is commonly used for this purpose is Giemsa stain,

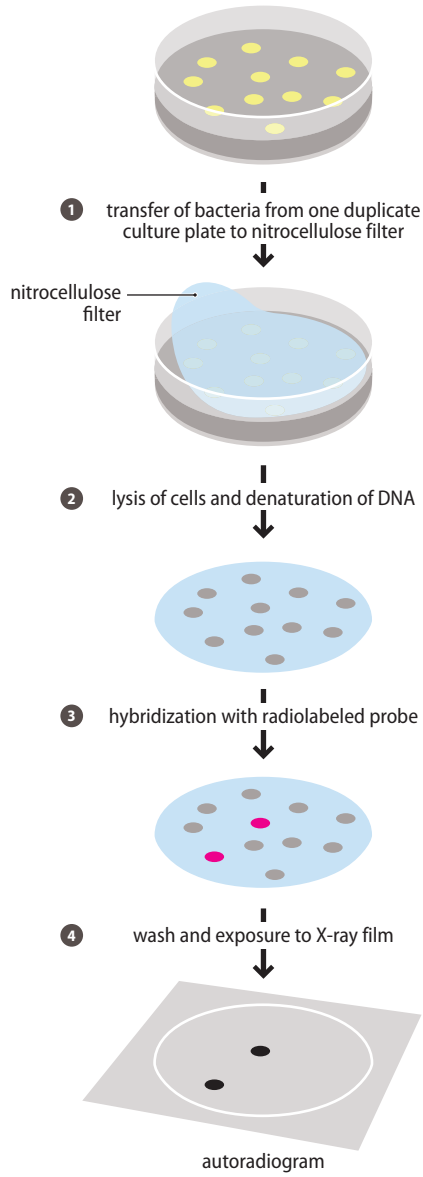


Figure 19.64 Colony hybridization. An outline of the method for probing bacterial colonies, each of which contains a unique plasmid, for the presence of a plasmid containing a DNA sequence of interest. The colonies are first transferred onto a filter that is pressed on the surface of the plate (step 1). The filter is processed in much the same way as that described for Southern blot analysis (step 2), then hybridized with a labeled DNA probe to the DNA region of interest (step 3). After washing, the filter hybridized with a radioactively labeled probe is exposed to film (step 4). If the colony carrying the relevant plasmid was present on the original plate, there will be a spot corresponding to its location on the film. The investigator can then isolate the corresponding colony from the original plate.

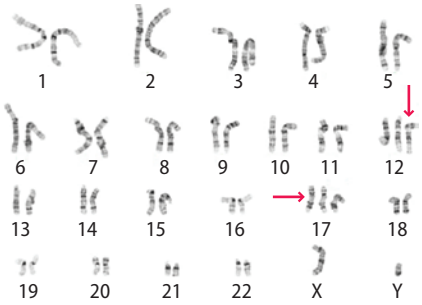


Figure 19.65 The karyotype of a human stem cell showing trisomy 12 and trisomy 17. The karyotype is determined by staining chromosomes of metaphase cells with Giemsa stain and identifying chromosomes based on their size and banding pattern. In the example shown, the cell has one X chromosome, one Y chromosome, and two copies of each autosome, with the abnormal exception of chromosomes 12 and 17, which are present in three copies each (marked by red arrows).

Adapted from Meisner and Johnson, *Methods* 2008; **45**: 133–141.

which generates a characteristic banding pattern by binding to gene-poor, A/T-rich regions, as shown in Figure 19.65. The stained chromosomes are then examined under the microscope to detect the presence of additional chromosomes, missing chromosomes, and gross deletions or rearrangements. This approach of visual inspection of chromosomes is called karyotype analysis.

Karyotype analysis is routinely used in medical diagnosis. Aneuploidy (extra or missing chromosomes) and other chromosomal aberrations, such as deletions and translocations, are frequently the underlying cause of birth defects, intellectual disability, miscarriages, and other syndromes. Thus, direct visualization of chromosomes can be used in prenatal screening procedures, such as amniocentesis. In this procedure, fluid containing cells shed from the fetus is collected from the amniotic sac. The cells are then cultured and used for karyotypic analysis. Chromosome abnormalities are similarly diagnosed in the case of malignancies and other diseases that may involve chromosomal rearrangements.

DNA sequences can be localized along a chromosome by fluorescence *in situ* hybridization

The examination of the overall structure of a chromosome makes it possible to detect gross chromosomal changes. However, it is often helpful to locate a specific DNA sequence **within** a chromosome. The method used to accomplish this is called **fluorescence *in situ* hybridization (FISH)**, as illustrated in Figure 19.66. This approach is also useful in diagnosing genetic diseases associated with chromosome abnormalities.

In this method, cells are fixed to a slide, permeabilized, and their DNA hybridized with a fluorescent DNA probe in a manner similar to the Southern blots described earlier. If more than one probe is used, molecules that fluoresce at different wavelengths are attached to the different probes. The chromosomes are also stained with a fluorescent dye, 4′-6′-diamidino-2-phenylindole (DAPI), described in Section 19.16, which binds specifically to DNA. The image that appears under a fluorescence microscope shows the DAPI-stained chromosomes in blue, with the regions of interest highlighted by the fluorescent probe or probes. In Figure 19.66, the probes are against regions on two different chromosomes, as shown in the example from a healthy individual (see Figure 19.66a). In a certain type of leukemia, a translocation occurs that brings these two regions in close proximity. As a result, the two probes overlap, as shown in Figure 19.66b.

The use of FISH to study chromosome segregation is described in Experimental approach 7.1.

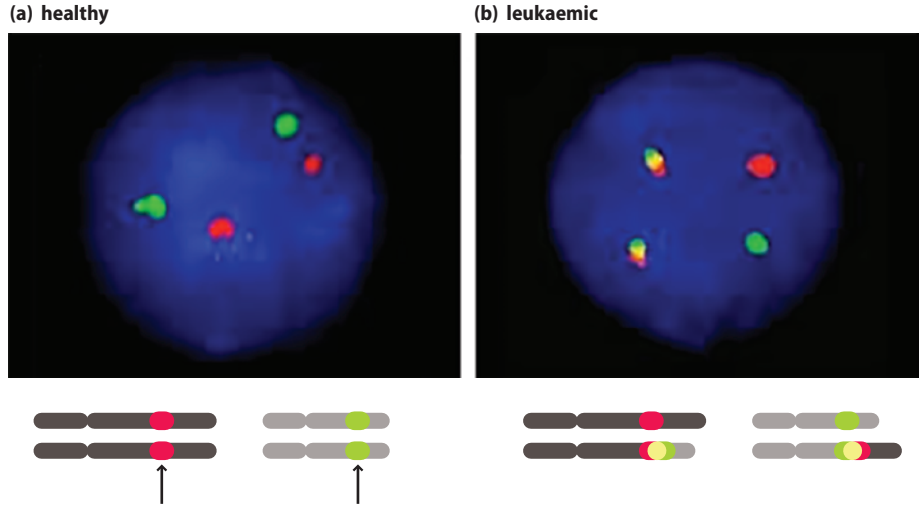


Figure 19.66 FISH. In the example shown, the assay was used to detect a reciprocal translocation between chromosomes 4 and 11, using probes that flank the breakpoint on both sides. (a) In a healthy individual, each probe gives rise to two spots, one for each homologous chromosome. Arrows point to the breakpoints that occurred in the leukemia patient shown in panel (b). (b) In a leukemia patient, there was a reciprocal translocation between one chromosome 4 and one chromosome 11. The unaffected chromosomes still give green or red signals, but the probes become very close to each other on the chromosomes that experienced the translocation, giving rise to merged signals (yellow).

Photo courtesy of Sabine Strehl, Children's Cancer Research Institute.

Two modifications of the FISH procedure, called **chromosome painting** and **spectral karyotyping (SKY)**, can be used to examine whole chromosomes. These methods are used to identify specific chromosomes or chromosomal rearrangements. For chromosome painting, the DNA for one entire chromosome is purified using flow cytometry, in which chromosomes are separated on the basis of size and base composition. The isolated chromosomes are then labeled with a fluorescent dye. SKY is similar to chromosome painting, except that, in this case, every chromosome has been separately purified by flow cytometry and each is labeled with a different and spectrally distinct ratio of five different fluorescent dyes. This set of labeled chromosome probes is then pooled, denatured, and used to probe a metaphase spread of chromosomes. Figure 19.67 illustrates that each pair of chromosomes in a normal karyotype has a unique color. If chromosome translocations or rearrangements have occurred, an abnormal karyotype will be seen. The chromosomes involved in the translocation are then known because the color uniquely identifies each chromosome.

Alterations in DNA from patients can be detected by array comparative genomic hybridization

Although FISH and SKY are very useful, they are not sensitive enough to detect small deletions or amplifications. In the case of FISH, we also need to have prior knowledge of which DNA region is to be analyzed. Array comparative genomic hybridization (aCGH) is a quantitative method for detecting changes in copy number (deletion, amplification) on a genome-wide basis. This method is based on the abundance of the DNA sequence in a test cell population being compared to the abundance of the same sequence in normal cells. In this way, it is possible to determine whether certain sequences are missing or amplified.

aCGH involves labeling genomic DNA from both normal cells and those from a patient with two different fluorophores. For example as shown in Figure 19.68a, one pool of DNA can be labeled with fluorescein, which emits green light when excited, while the other pool is labeled with rhodamine, which emits red light when excited. Instead of hybridizing these two labeled DNA pools to chromosomes, the

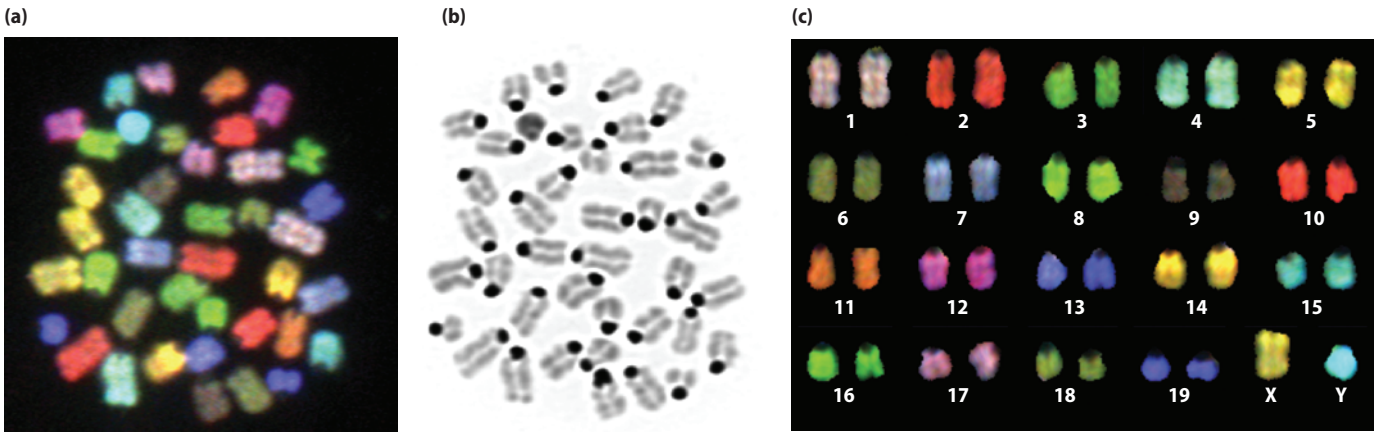


Figure 19.67 SKY. (a) DNA probes are generated for each individual chromosome by creating a unique combination of fluorophores, making each chromosome appear as a different color when these probes are hybridized to mitotic chromosome spreads. (b) In parallel, the chromosomes are also stained with a DNA dye that reveals the specific banding pattern of each chromosome. (c) An arranged SKY karyotype of the chromosomes can be generated from the chromosomes shown in (a).

Kindly provided by Margaret Strong, Johns Hopkins School of Medicine, unpublished.

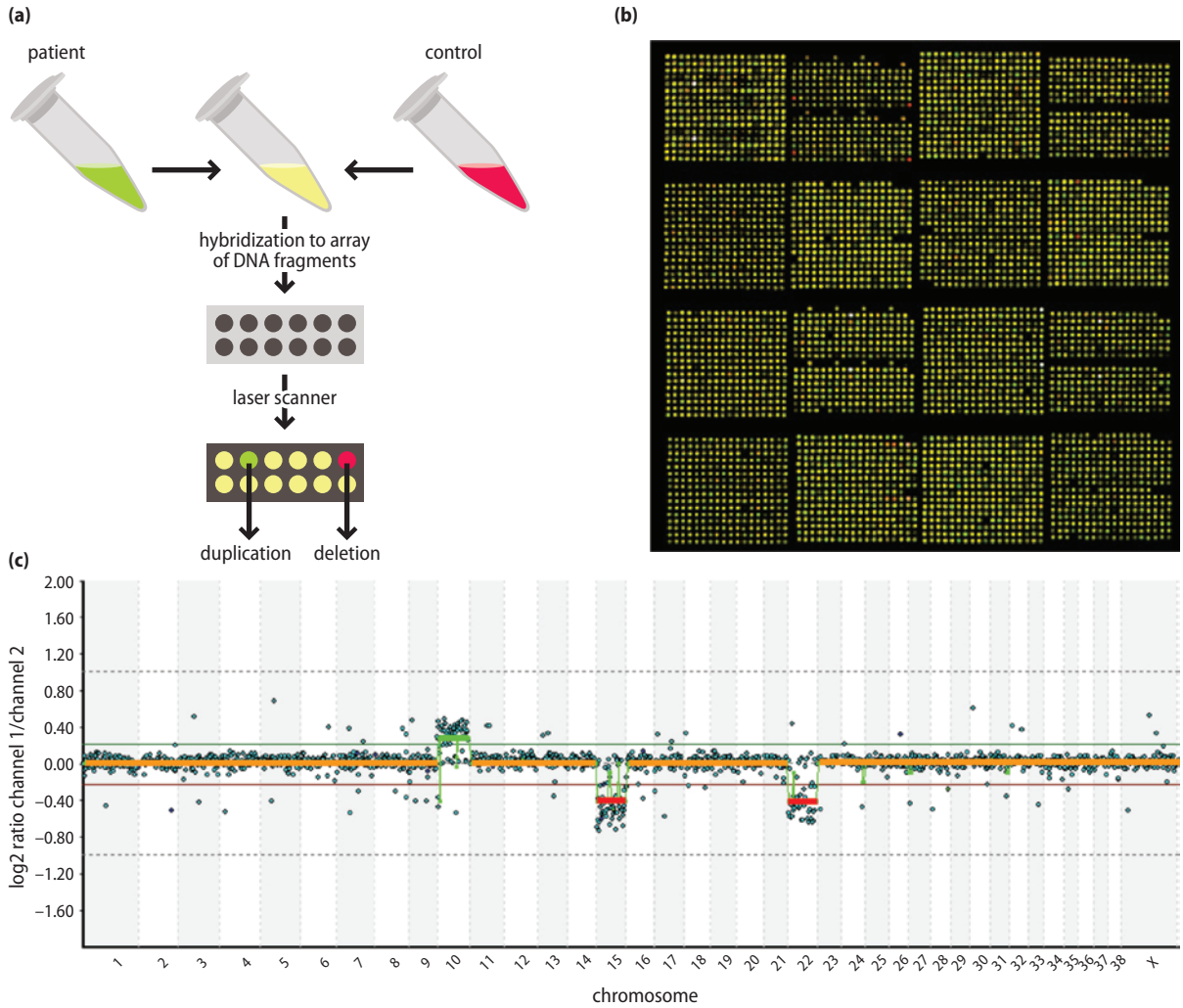


Figure 19.68 aCGH. (a) DNA from the sample to be tested is labeled with a green fluorescent dye and reference DNA is labeled with a red fluorescent dye. The two samples are mixed and co-hybridized to an array containing genomic DNA fragments that have been spotted on a glass slide. (b) The resulting ratio of fluorescence intensities is proportional to the ratio of the copy numbers of DNA sequences in the test and reference genomes. The green spots indicate extra chromosomal material (duplication) in the test sample relative to the reference sample at that particular region, while red spots indicate relatively less of the test DNA (deletion). The example shown is a small part of the entire array from a study on canine cancer. (c) The information generated by fluorescence intensities, such as in panel (b), is converted to the abundance of different chromosomal regions. In the example shown, the tested canine cell line had an extra chromosome 10 and only one copy of chromosomes 15 and 22. Note that this type of analysis can uncover deletions and amplification of short chromosomal domains.
(a) from Shinawi, M., and Cheung, S.W. (2008). *Drug Discovery Today* **13**: 760. (b), (c) copyright Matthew Breen.

We discuss the use of DNA microarrays to detect RNAs on a genome-wide level in Section 19.12.

labeled DNA is hybridized to a **DNA microarray** that contains DNA fragments from the entire genome, spotted in rows and columns on a microchip, as depicted in Figure 19.68b. If the copy number of a particular region is the same in the patient and in the control, the signal from the two fluorophores will be equal; it will therefore appear yellow because the equal amounts of red and green fluorochromes generate a yellow light. However, if the patient has a deletion or duplication, the signal will be red or green, respectively. (In this instance, the excess of one fluorophore will generate a signal that exceeds the signal from the second, less abundant fluorophore.) The microarray is scanned by a special fluorescence scanner that measures the signal from each microarray spot and plots it with respect to each chromosome, as illustrated in Figure 19.68c.

19.12 DETECTION OF SPECIFIC RNA MOLECULES

As we learned in Chapter 9, many genes are regulated at the level of transcription. It is therefore often desirable to study the expression of an individual gene or a collection of genes by assaying the amount of RNA that is present. As described in this section, many methods that are used to examine RNA levels rely on hybridization to specific probes, although the use of reporter genes can also provide information about timing and patterns of expression. It is also possible to survey the entire population of RNAs that is expressed in a given cell type or tissue, as we will discuss toward the end of this section.

Northern blot analysis and nuclease protection can be used to evaluate the expression of individual genes

Just as we learned in Section 19.11 that specific DNA fragments can be detected by Southern blot, we can detect RNA through an analogous approach known as northern blot. Specific RNA transcripts are generally detected by hybridization to specific DNA or RNA probes that have been labeled in some way.

So how is a northern blot performed? The procedure is summarized in Figure 19.69. First, RNAs are separated by size on denaturing acrylamide or agarose gels. (The denaturing step eliminates secondary structures in the RNA that would otherwise alter its migration in the gel and make it impossible to assess its true size.) The RNA is then transferred directly to a positively charged membrane. Unlike a Southern blot, where the DNA is fragmented prior to electrophoresis, RNA molecules are relatively small and thus will typically enter the gel and transfer efficiently to the filter without prior fragmentation. Also, RNA can be probed directly because it is single-stranded and therefore does not need to be further denatured once it is on the filter (unlike double-stranded DNA). At this stage, the membrane is "probed" with a radioactively or fluorescently labeled DNA (or RNA) probe, the unbound probe is washed off, and the hybridized labeled probe is detected.

Northern blot analysis has a number of benefits as it directly follows the signal of the actual transcript in the sample, in contrast to other techniques we describe later. As such, the size of a full-length transcript can be determined by comparing the position of the signal in the membrane to the positions of marker bands

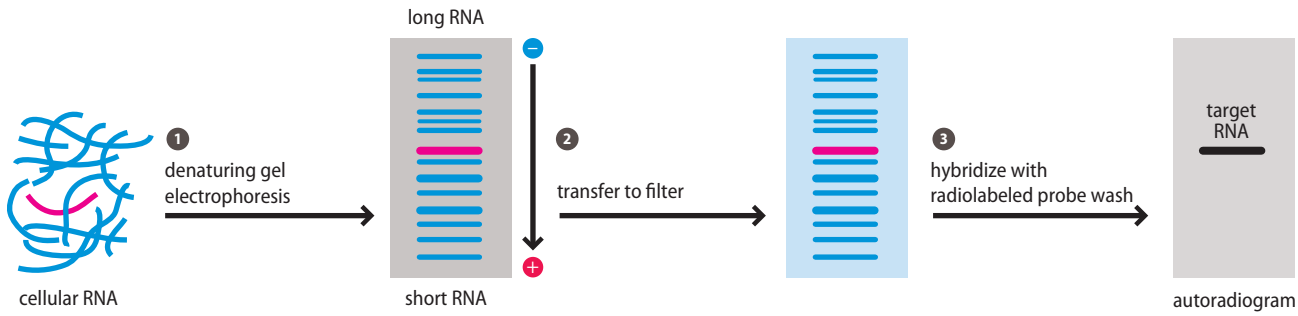


Figure 19.69 Analyzing RNA abundance by northern blot analysis. RNA is isolated from cells and separated by size on a denaturing gel (step 1). The RNA of interest is shown in pink. The RNA is then transferred to a filter (blue, step 2) and probed with a DNA or RNA probe that is coupled to a detectable molecule (for example, a radiolabeled isotope, step 3). The filter is then analyzed to determine the positions on the gel where the probe has hybridized (black band on the autoradiogram). In the case of a radioactive probe, the filter can be exposed to an x-ray film or scanned on a phosphorimager, which can quantify the amount of radioactive signal.

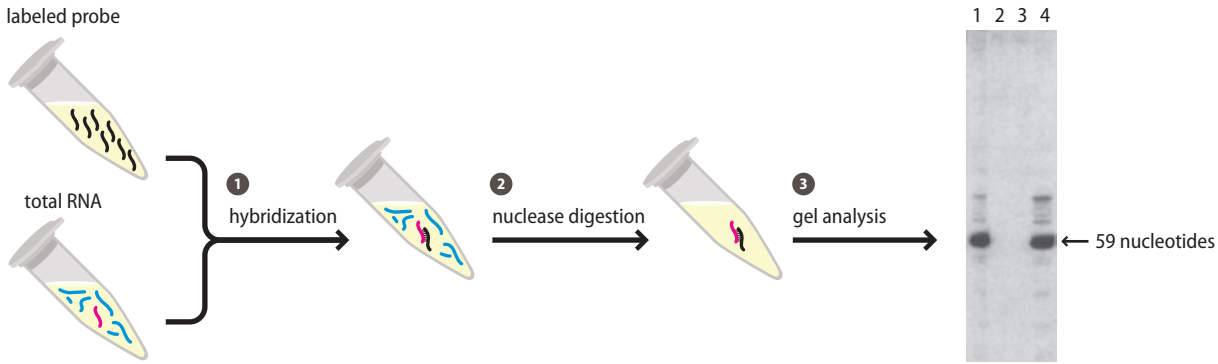


Figure 19.70 Analyzing RNA abundance by nuclease protection assay. RNA is isolated from cells and is hybridized in solution with a radioactive probe (DNA or RNA) targeting a specific sequence (step 1, the cellular sequence of interest is shown in pink). The mixture is then treated with nucleases that digest the non-target RNA and any probe that is not annealed, but not double-stranded RNA or DNA–RNA hybrids. Both the probe and the RNA are protected from nuclease digestion wherever the two are complementary to each other. The resulting mixture is subjected to gel electrophoresis; the amount of undigested probe visible in the gel is proportional to the abundance of the RNA of interest. In the example shown, mRNA was prepared from four cell lines (lanes 1–4), and the 59-bp probe used was against a segment of murine terminal transferase mRNA. Cell lines 1 and 4 express terminal transferase, whereas cell lines 2 and 3 do not.
Image adapted from Carey et al, *Cold Spring Harb Protoc* 2013 276–285.

of known size. Moreover, if there are multiple RNA species that hybridize to the selected probe (for example, due to alternatively processed versions of the transcript), each will be observed as an independent signal on the blot. Northern blot analysis can also be used to determine the relative abundance of a specific transcript under different growth conditions by comparing the intensity of the signals from samples isolated under different conditions.

Another approach that has traditionally been used to determine the amount of an RNA molecule is a **nuclease protection assay**, which is illustrated schematically in Figure 19.70. In this approach, a radioactively labeled DNA or RNA probe of a defined size is hybridized to the transcript of interest in solution. Nucleases such as S1 nuclease, which cleaves single-stranded nucleic acids, but not double-stranded DNA–RNA or RNA–RNA hybrids, are then used to digest the unhybridized RNA and thus reveal molecules of interest. These digestion-resistant products are then resolved by gel electrophoresis and visualized. The intensity of the band is proportional to the amount of target RNA in the sample.

The specific ends of RNA molecules can be determined by nuclease protection, primer extension, or 5′- or 3′-RACE

Sometimes it is desirable to determine the identity of the 5′ or 3′ end of an RNA molecule—for example, to identify the start of the transcript (and thus determine where the promoter of the gene was located). All methods for identifying end regions rely on the hybridization of DNA or RNA probes in the vicinity of the ends. For example, the nuclease protection assays we have just described can be used to map the ends of an RNA. Such end identification is performed by using a probe that extends beyond the end of the RNA of interest on one side and overlaps the end of the RNA on the other side, as shown in Figure 19.71a. During nuclease digestion, the part of the probe that extends beyond the transcript will be digested, and the length of the nuclease-resistant probe will reveal the position of the 5′ or 3′ end of the RNA of interest.

Primer extension assays can also be used to identify the 5′ end of an RNA molecule. As shown in Figure 19.71b, a labeled oligonucleotide complementary to sequences near the 5′ end of an RNA is extended by reverse transcriptase. Since

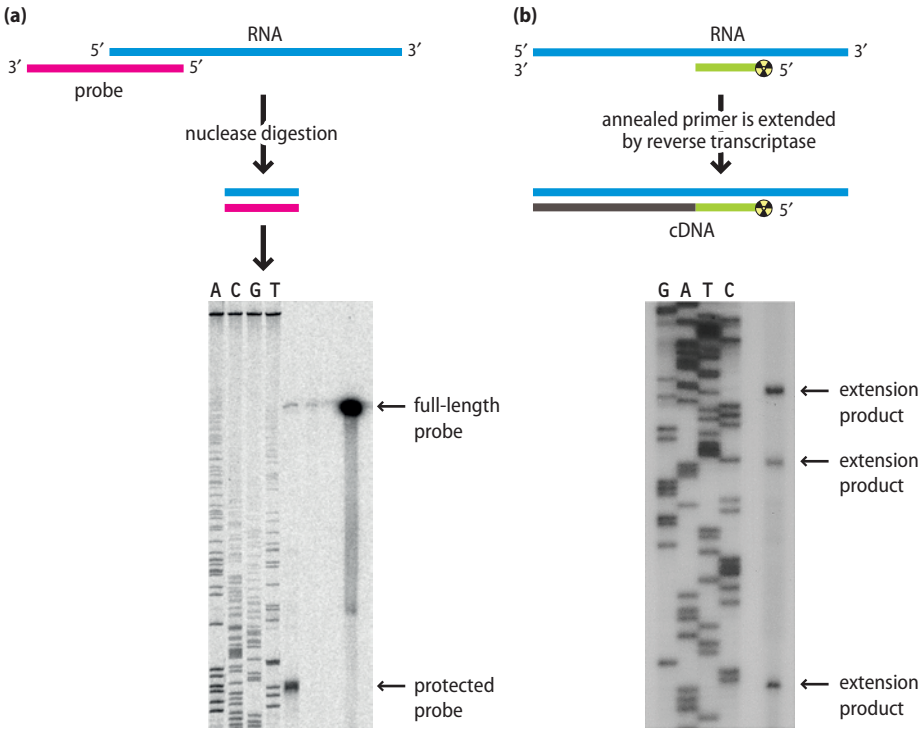


Figure 19.71 Methods to determine the identity of the ends of RNA molecules by nuclease protection and primer extension. (a) Identification of an RNA end by nuclease protection. The probe (in red) is designed to extend beyond the putative 5′ end of the RNA. After nuclease digestion, which removes any single-stranded nucleic acid, the 3′ end of the probe will lie at the 5′ end of the RNA. Determining the size of the protected probe will reveal where the transcript begins, relative to the 5′ end of the probe. (b) Identification of an RNA end by primer extension. A radioactive oligonucleotide (green) is hybridized to the RNA (blue) and then extended using reverse transcriptase to produce single-stranded cDNA (black). The resulting cDNA is subsequently analyzed on a denaturing polyacrylamide gel (by PAGE; bottom panel). A sequencing reaction is run next to the primer extension product to determine the size. In the example shown, there are three RNA forms: 59, 90, and 105 bp long.

(a) from Begnell, DRD, Tahlan, K, Colvin, KR, Jensen, SE, Leskiw, BK. Expression of ccaR, encoding the positive activator of cephamycin C and clavulanic acid production in *Streptomyces clavuligerus*, is dependent on *bldG*. *Antimicrobial Agents and Chemotherapy*, 2005;**49**:1529–1541.

(b) from Opdyke, JA, Kang, J-G, and Storz, G. GadY, a small-RNA regulator of acid response genes in *Escherichia coli*. *Journal of Bacteriology*, 2004;**186**:6698–6705.

the reverse transcriptase enzyme falls off when it reaches the 5′ end of the transcript, it is possible to determine the 5′ end precisely by comparing the extended product with a sequencing ladder.

Since reverse transcriptase will also fall off at many modified nucleotides in RNA, as already mentioned in the context of RNA sequencing, primer extension analysis can also be used to map the positions of such modified nucleotides. These modifications might occur naturally in the cell or may be introduced by treating cells or RNA with specific compounds such as dimethyl sulfate, which methylates adenosine and cytosine residues. Given that the ability of the adduct to modify a particular nucleotide depends on how accessible that nucleotide is in an RNA tertiary structure, the combination of RNA modification and reverse transcription can give insights into the structure of the RNA.

Another approach for identifying RNA ends is called **5′- or 3′-RACE**. We encountered this amplification method when we discussed the cloning of RNA molecules in Section 19.3 (see Figure 19.21). To identify a 5′ or 3′ end, an oligonucleotide of known sequence is ligated to the respective end of the RNA before the RNA is reverse-transcribed (or sometimes to the end of the reverse transcriptase product). Consequently, the resulting product has an additional string of nucleotides (whose sequence is known) at one end, which can then be used as a primer site for amplification by PCR. The products of PCR amplification are then sequenced, revealing the complete sequence of the original RNA molecule.

Quantitative reverse transcription-PCR can be used to evaluate levels of expression of individual genes

The PCR method described in Section 19.3 can also be exploited to detect the presence of a specific RNA and determine its relative quantities. This approach is commonly called quantitative reverse transcription-PCR (**qRT-PCR**). In this method, a

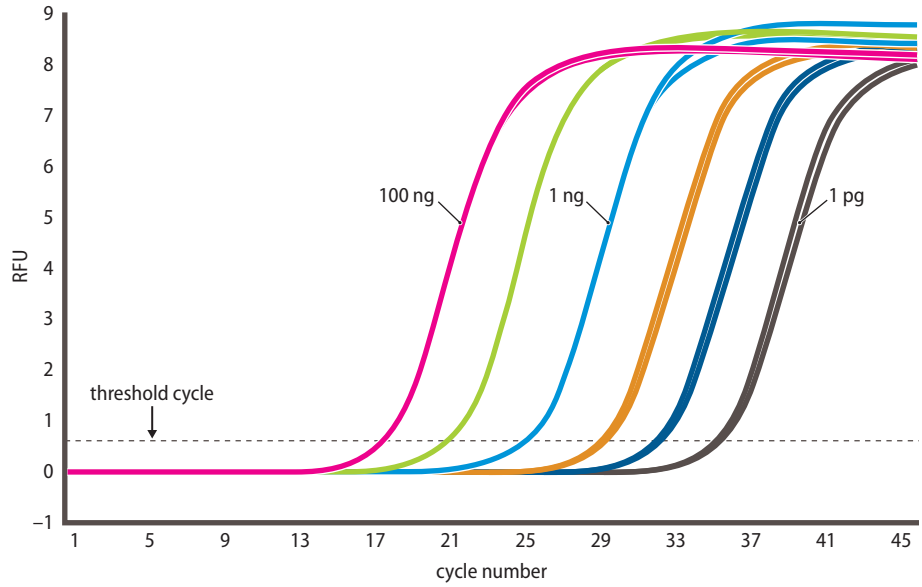


Figure 19.72 Principle of qRT-PCR. In the exponential phase of PCR amplification, the amount of product obtained by qRT-PCR is proportional to the amount of RNA in the sample. Thus, 100 ng of RNA will give a defined amount of relative fluorescence units (RFU) (~18 cycles at the threshold) before a sample with 1 ng (~25 cycles) or 1 pg (~35 cycles).
http://www.genehk.com/news/news_20100106.php

cDNA copy of the RNA of interest is first synthesized by reverse transcription. The cDNA products are then amplified by PCR, utilizing a second primer within the transcript of interest. This step is designed to yield a conveniently sized PCR product for analysis on agarose gels or by machines that monitor dye binding to DNA.

As illustrated in Figure 19.72, the amount of product obtained by qRT-PCR will be proportional to the amount of RNA present in the sample as long as every sample being compared is within the exponential phase of PCR amplification—that is, before one of the components in the reaction becomes limiting and the amount of amplification decreases. Machines that automatically quantify the amount of PCR products after each cycle in "real time" (through binding to a dye) have greatly simplified these assays, though proper controls must always be included to ensure that the amount of the PCR product is correlated with that of the input RNA.

A distinct advantage of qRT-PCR is that its extreme sensitivity allows very small amounts of RNA to be detected. A disadvantage of this approach is that only a limited segment of the RNA molecule corresponding to the region between the chosen primers will be detected. As a result, if the RNA exists in multiple forms within the sample, for example, this type of information will be missed in the analysis.

➔ The use of qRT-PCR to characterize the role of a lncRNA is described in Experimental approach 13.4.

Reporter genes can simplify the detection of specific RNAs

A common method for assaying the expression of a particular gene is to fuse its regulatory region to a **reporter gene**, which encodes a gene product that can be readily detected. Commonly, the reporter genes encode proteins, such as β-galactosidase (encoded by the *E. coli lacZ* gene), which breaks down particular galactose derivatives to give a colored product; luciferase, which catalyzes a chemical reaction that emits light; and GFP from jellyfish, which has intrinsic fluorescence. Recently, reporters encoding visualizable RNAs, such as a selected RNA denoted Spinach that fluoresces upon binding a specific compound, have also been developed. The availability of easily detected indicators of expression makes it possible to screen thousands of microbial colonies for those containing the coding and/or regulatory regions of a particular gene (as seen in Experimental approach 14.2).

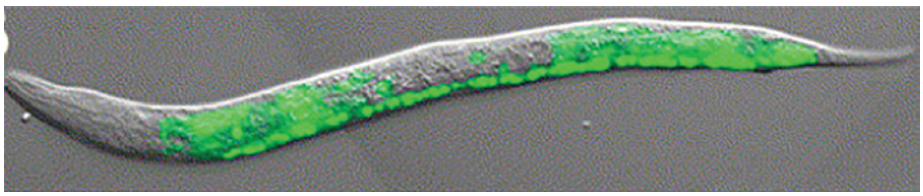


Figure 19.73 The use of reporter genes *in vivo*. Expression of a cyclin E:GFP reporter in *C. elegans*. GFP fused to the promoter of the cyclin E gene (cye-1:GFP) reports on its transcriptional activity. Shown is a larval stage L1 animal with strong expression of the *cye-1* reporter gene in a row of dividing cells (P-lineage descendants) in the ventral nerve cord (the line of cells in the middle of the animal), as well as in other tissues.

From Brodigan, T. M., Liu, J., Park, M. Kipreos, E. T., and Krause, M. (2003). Cycline E expression during development in *C. elegans*. *Developmental Biology* **254**: 102–115.

Reporter genes can be quickly and repeatedly assayed, making it possible to study the way gene expression changes with time. Reporter genes also can be used to monitor gene expression that might be limited to certain tissues within the whole organism, as illustrated in Figure 19.73. The reporter can be fused to the promoter of the gene to reflect the transcriptional regulation in operation within the organism. The reporter also can be fused directly to an open reading frame (ORF) of interest such that the reporter reflects both the transcriptional and translational regulation of the gene. In this case, the reporter protein would only be synthesized if the gene is first transcribed from the endogenous promoter and translated using the endogenous sequences. As an alternative to monitoring expression of genes in cells with reporters, specific transcripts also may be detected by using the FISH approach described for detecting specific DNA fragments in Section 19.11.

DNA microarray analysis is used to determine expression genome-wide

The availability of whole genome sequences has made it possible to develop microarrays containing nucleotide fragments for much or all of the genome, which can then be used to detect specific RNA sequences on a genome-wide scale. As we encountered for aCGH in Section 19.11, the microarray contains many different DNA fragments that are immobilized on the surface of a specialized slide or chip, as depicted in Figure 19.74. The DNA molecules can be either synthetic oligonucleotides or larger fragments that have been generated by PCR. Microarrays can be constructed, for example, so that each spot contains a DNA fragment corresponding to a portion of every known ORF. These microarrays can then be used to monitor transcription in a population of cells by isolating the pool of RNA, generating fluorescently labeled cDNA copies of this population of RNA, and then hybridizing the mixture of cDNAs to the microarray. A labeled cDNA will hybridize to a spot in the microarray that contains a complementary DNA fragment. The relative intensity of each spot on the microarray reflects the transcript level for the corresponding gene.

Expression of the same set of genes under different conditions (say, in the wild-type organism, as compared with a mutant) can be compared by labeling each cDNA pool with a different fluorescent dye. Both samples are then used to probe the same microarray, and the relative intensities of each dye on each spot can be determined. If red and green dyes are used, genes that are expressed at equal levels under both conditions will yield a yellow spot on the array, whereas differences in the ratio of expression will give rise to a spot that is either green or red in color. This approach can be used to assess relative gene expression under two different conditions, such as in the presence and absence of a stress such as heat shock.

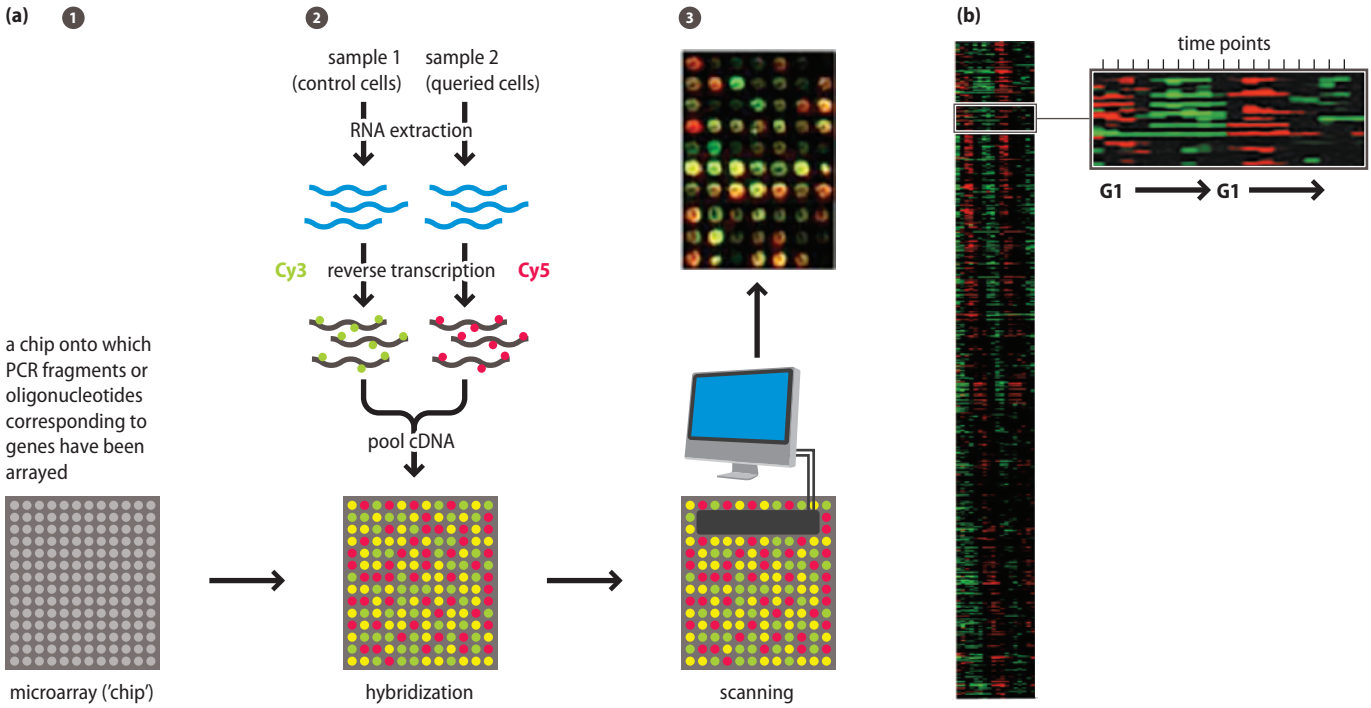


Figure 19.74 Microarray analysis. (a) Steps involved in microarray analysis. A chip is prepared that contains sequences representative of a select group of genes or an entire genome (step 1). RNA is isolated from two sets of cells: the control cells and the queried cells (for example, cells grown under a special growth condition or a mutant strain versus a wild-type control). The RNA is reverse-transcribed to generate cDNA, and each cDNA pool is labeled with a different fluorophore (in this case, Cy3 which emits green light, and Cy5 which emits red light). The labeled RNAs are pooled and hybridized to the microarray (step 2). Each spot on the chip is excited with a laser (at the Cy3 and Cy5 excitation wavelengths), and the emission is captured by a camera (step 3). If a particular transcript is expressed more robustly in the control than the non-control cells, the spot will appear green, while higher expression in the queried culture will yield a red spot. Equal expression gives a yellow signal. (b) Genome-wide microarray analysis of cell cycle-dependent gene expression in *S. cerevisiae*. Cells were treated such that they progressed synchronously through the cell cycle, and the expression of every gene was examined at fixed time intervals. Each row represents the expression pattern of a single gene, and each column represents a time point. At the left-most time point, the cells are found in G1, and the cells proceed through two full cell cycles as the time course progresses. Red indicates genes with increased transcription relative to the control (RNA from asynchronous cells), while green indicates genes with decreased transcription. The data are displayed in the form of clusters, meaning that genes that have a similar expression pattern are placed next to each other in the presentation. In the right panel, which is an enlargement of a small section from the left panel, the expression pattern in these rows is similar.

Adapted from Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O, Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell-cycle regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridisation. *Molecular Biology of the Cell* **9**: 3273–3297.

Deep sequencing is used to determine genome-wide expression patterns

A final approach of increasing importance is the simple direct sequencing of cDNAs generated from an RNA population of interest, known as RNA-Seq. As sequencing methods have improved, it is now straightforward to sequence many millions of cDNAs. If enough sequences are read, the data from such experiments can be used to evaluate the amount of a species in a population, in particular when directly compared with a related population. Deep sequencing can effectively be applied to different subpopulations of RNAs, such as RNAs below a certain size or RNAs that are processed in a specific way, to learn more about the transcriptome. For example, to identify unprocessed RNAs with a 5′ triphosphate, the total RNA population can be treated with a terminator exonuclease enzyme that will degrade all processed RNAs with a 5′ cap or monophosphate before the remainder of the population is subjected to deep sequencing.

One caveat to deep sequencing for examining genome-wide RNA levels, which also applies to the microarray approaches we have just described, is that the need to generate a cDNA copy of the RNA can bias the population that is recovered for sequencing. For example, an RNA will be missed if reverse transcriptase cannot make a full copy of the transcript due to the presence of inhibitory secondary structures or nucleotide modifications. In some experiments, this limitation of reverse transcriptase falling off is actually exploited to gain insights into the positions of the structures and modifications. Another permutation takes advantage of the observation that other nucleotide modifications are read through by promiscuous reverse transcriptases that commonly insert a different complementary nucleotide than the one specified by the nucleotide in the primary RNA species. For these modifications, the mutational readout of a sequencing experiment can be used to identify the sites of modification on the RNA species of interest.

19.13 DETECTION OF SPECIFIC PROTEINS

It is often desirable to follow the fate of a specific protein. Proteins can be followed in a cell lysate or in their native location in the cell. In either case, there are two main approaches for detecting proteins—the use of antibodies directed either against the protein or against a tag added to the protein by cloning. In this section, we discuss the detection of proteins in cell lysates and also consider an approach for evaluating the relative levels of specific proteins in two different samples. We discuss the detection of proteins inside cells in Section 19.16.

Western blots rely on antibodies to detect proteins

A powerful method for detecting specific proteins takes advantage of the immune system's ability to generate antibodies that bind to proteins with very high affinity and specificity. Antibodies can be obtained by exposing the appropriate animal (often mouse, rabbit, goat, or chicken) to either a purified protein or a short peptide corresponding to a region of that protein. Generally, two classes of antibodies are used—monoclonal antibodies, which are obtained from a clonal population of immune cells that express only one type of antibody and thus react with only one epitope, or one antigenic determinant on the protein; and polyclonal antibodies, which are a collection of different antibodies that often react with different epitopes of the same protein.

The western blot, or **immunoblot**, detects a specific protein immobilized on a membrane using antibodies directed against the protein (or a tag fused to the protein, as will be discussed next). In this method, illustrated in Figure 19.75, proteins are separated by SDS-PAGE and transferred from the gel to a hydrophobic or positively charged membrane. The membrane is then incubated with specific antibodies in the presence of a protein-rich blocking agent such as a solution containing BSA, which reduces non-specific binding of the antibody. After the membrane is washed, the bound antibody can be detected in a variety of ways, most of which rely on "secondary" antibodies that bind to the "primary" antibody. These "secondary" antibodies themselves can be directly labeled with radioactivity, coupled to a fluorescent molecule, or linked to an enzyme such as horseradish peroxidase, which produces a colored or fluorescent band at the location of the protein of interest when incubated with an appropriate substrate.

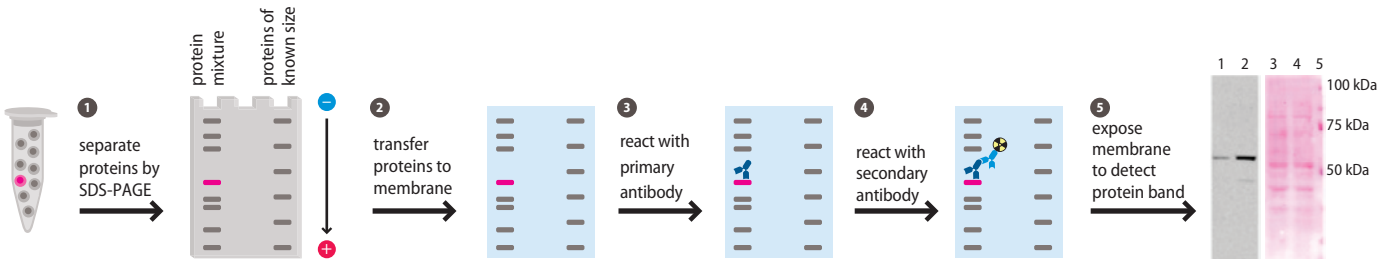


Figure 19.75 Western blot analysis. A protein mixture containing a protein of interest (pink) is loaded on a polyacrylamide gel and separated by electrophoresis, as described in Figure 19.52 (step 1). Proteins of known size (often labeled with colored dyes) that serve as molecular weight markers can also be loaded. The proteins are then transferred to the filter in a special chamber, using an electric field (step 2). The filter can then be stained with dye, such as Ponceau S Red, which reveals the proteins on the filter and shows the positions of the molecular weight markers (see pink-stained filter on the far right). The filter is then exposed to the primary antibody that recognizes the protein of interest (step 3). This reaction takes place in a buffer that minimizes non-specific binding of the antibody to irrelevant proteins (known as cross-reactivity). The filter with the bound primary antibody is washed several times to remove unbound antibody and then exposed to a secondary antibody that is conjugated to a detectable molecule, such as a radioactive isotope or a fluorophore (step 4). The filter is again washed to remove unbound secondary antibody and then processed to detect the secondary antibody using autoradiography (to detect radioactivity) or a phosphorimager (to detect fluorescence) (step 5). In the example shown, two samples containing a mixture of proteins were run on an SDS-PAGE gel and transferred to a filter. The filter was then stained with Ponceau S Red (on the right), revealing the proteins in both lanes that were transferred to the filter (lanes 3 and 4: note the presence of many protein bands). The Ponceau S Red staining also shows the locations of the molecular weight markers (lane 5). The same filter was then treated with antibodies against the protein of interest, followed by secondary antibodies that recognize the primary antibody. The filter was then exposed to reveal the location of the band corresponding to the protein of interest (lanes 1 and 2).

The accuracy of this method depends upon the specificity of the primary antibody—some antibodies (especially polyclonal ones that may recognize several epitopes) may react with other proteins, in addition to the intended target. It is thus important to include the appropriate controls (such as a lysate missing the protein of interest) to confirm the specificity of the primary antibody.

The addition of specific tags can be used to detect proteins

It can be difficult or time-consuming to generate antibodies to specific proteins. An alternative approach is to introduce a protein tag that is fused to either the N- or C-terminus of the protein of interest, and then to use an antibody that binds specifically to that tag. A summary of commonly used tags is given in Figure 19.76. Some of the tags comprise only a few amino acids, while other tags correspond to an entire small protein. Very specific antibodies have been developed for most of the tags, allowing for sensitive detection of proteins carrying these tags. As described in Section 19.8, several of the tags also bind to specific matrices, which can be used to purify the tagged proteins.

The addition of specific tags can also be used to reduce the levels of specific proteins or remove proteins from specific compartments

Sometimes it is useful to be able to remove a specific protein from a complex or compartment of a cell in a regulated manner. Special tags have also been developed for these purposes. Commonly, a protein sequence that promotes rapid degradation, termed "degron," is fused to the protein of interest. One version of this approach takes advantage of a plant degron that responds to auxin. Since mammalian cells normally do not encounter auxin, if this degron is fused to a protein of interest, auxin treatment can lead to specific degradation of the target protein. Other regulated tags can be utilized to remove proteins from a specific intracellular

Commonly used protein tags

tag name	size	source and comments
Myc	10 amino acids (EQKLISEEDL)	derived from the human c-Myc proteins
hemagglutinin epitope (HA)	9 amino acids (YPYDVPDYA)	derived from the influenza hemagglutinin glycoprotein
FLAG	8 amino acids (DYKDDDDK)	a synthetic peptide, can be cleaved by enteropeptidase
hexa histidine-tag (6 × His)	6 (or more) amino acids (HHHHHH)	a synthetic peptide that binds to a resin containing nickel (Ni^{2+})
vesicular stomatitis virus glycoprotein (VSV-G)	11 amino acids (YTDIEMNRLGK)	derived from a viral glycoprotein needed for the budding of vesicular stomatitis virus
V5	14 amino acids (GKPIPNPLLGLDST)	derived from a small epitope present on the P and V proteins of the paramyxovirus of simian virus 5 (SV5)
calmodulin-binding peptide (CBP)	26 amino acids	derived from muscle myosin light-chain kinase, binds to resin containing calmodulin
glutathione-*S*-transferase (GST)	220 amino acids	naturally occurring protein that binds to resin containing glutathione
green fluorescent protein (GFP)	238 amino acids	a naturally flurorescent protein from the jellyfish, *Aequorea victoria*, commonly used in applications such as light microscopy to visualize directly the location of the tagged protein
maltose-binding protein (MBP)	371 amino acids	*E. coli* protein, binds to resin containing amylose

Figure 19.76 Commonly used protein tags. For short tags, the amino acid sequence is shown. A "synthetic tag" means that it was devised by scientists, rather than being derived from a naturally occurring protein.

compartment or complex by inducing tight binding interactions with a partner protein localized to another compartment. It can be advantageous to even introduce multiple tags to allow purification or detection by two different approaches or to allow both detection and depletion.

The relative levels of proteins in two different samples can be determined by isotope labeling followed by mass spectroscopy

As we have seen for DNA and RNA, sometimes it is desirable to detect and compare two collections of protein molecules. A recent innovation in the area of mass spectroscopy has made it possible to compare the relative amounts of many protein species in two samples (collected under different conditions, for example). This quantitative proteomic approach is referred to as **s**table **i**sotope **l**abeling with **a**mino acids in **c**ulture (SILAC) (see Figure 19.77). The basic idea behind the approach is to grow cells under two different conditions, each in the presence of an amino acid containing a different isotope. For example, one population exposed to a stress condition might be grown in the presence of arginine labeled with ^{13}C atoms, while the control unstressed population is grown in the presence of the usual ^{12}C. As a result of this differential isotopic labeling, any protein fragment containing an arginine will have a mass 6 Da heavier when grown in the presence of ^{13}C-labeled arginine, compared to the equivalent fragment from the culture grown with ^{12}C-labeled arginine.

The use of SILAC to identify ubiquitinated proteins is described in Experimental approach 14.3.

Following whole cell labeling, protein extracts are prepared and the two samples digested with proteases in preparation for analysis by mass spectrometry, which we discussed in Section 19.9. The samples are then injected simultaneously into the mass spectrometer where the ratio of peak intensities can provide information about the relative abundance of the peptides under different conditions.

Although it is not yet possible to resolve and identify all proteins in the cell with this approach, a substantial number can be followed in a single experiment,

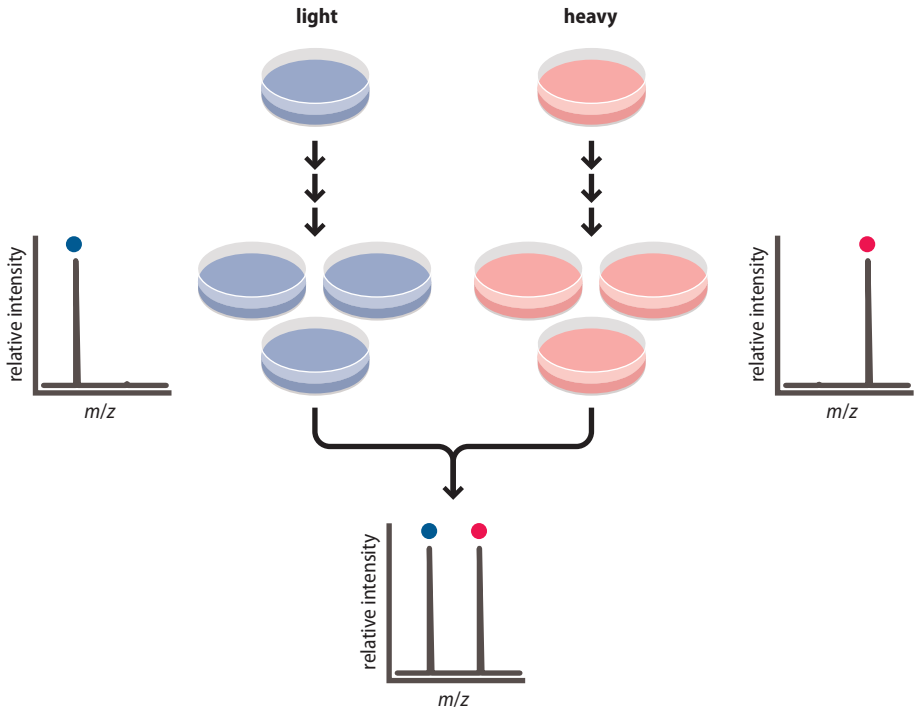


Figure 19.77 The principle of SILAC. To compare the protein composition of cells from two different sources, each cell type is labeled by its growth in medium containing an amino acid that is itself labeled with one of two different isotopes—for example, ^{12}C versus ^{13}C. In the example shown, the cells on the left were grown in "light" arginine containing ^{12}C (purple) while the cells on the right were grown in "heavy" arginine containing ^{13}C (light red). The incorporation of the ^{13}C-labeled arginine into proteins results in a mass shift of the corresponding peptides relative to the same peptides from cells grown in the presence of ^{12}C-labeled arginine; this shift can be detected by a mass spectrometer. The samples are combined, and the relative amounts of each peptide are compared. In the example shown, the peptide is present at similar levels in both samples.

yielding significant insights into global protein levels in the face of a given insult to the cell. As the technology is further developed, this method will undoubtedly become increasingly sophisticated.

19.14 DETECTION OF INTERACTIONS BETWEEN MOLECULES

The goal of the many techniques we have discussed in earlier sections of this chapter is to isolate and identify a single biological molecule of interest, such as a specific RNA or a specific protein. However, since no biological molecule acts alone, we are frequently interested in examining the interaction between molecules—especially interactions between proteins and between proteins and nucleic acids. Thus, in this section, we will consider some of the approaches that are commonly used to study binding between biological molecules. In Section 19.15, we will discuss ways to assay these interactions on a genome-wide level.

Co-immunoprecipitation and co-purification can be used to identify and study interacting proteins

Among the simplest methods used to study interactions between macromolecules are co-immunoprecipitation and co-purification. If certain molecules of interest remain associated with one protein (for instance, during many purification steps), the molecules are said to co-purify. Thus, co-purification can be exploited to identify molecules that interact with one another in the cell.

Co-immunoprecipitation is a method for isolating protein complexes using an antibody that binds to just one specific protein in the complex. An antibody against a protein of interest is added to a mixture of proteins or to a cell extract. The antibody–protein complex is then precipitated from solution by adding microscopic beads that are covered with proteins that bind tightly to **any** antibody. The beads

are relatively heavy and readily fall to the bottom of the test tube, bringing with them the protein that binds to the immobilized antibody. If any other proteins are bound to the protein of interest, they, too, will precipitate along with the rest of the complex. The binding partners that are precipitated in this way can then be identified by the method of mass spectrometry, as described in Section 19.9.

An alternative to using antibodies relies instead on generating a fusion protein containing an affinity tag (as we learned about in Sections 19.8 and 19.12), which can be used to precipitate the tagged protein, along with other macromolecules that may bind to it. For example, a fusion protein containing GST can be precipitated by beads coated with glutathione. Sometimes the association between proteins is stabilized by the addition of chemical crosslinking agents, which covalently link proteins that are within a defined distance in a complex. An interaction can be confirmed by performing the assay in reverse, in which the binding partner is precipitated or purified and the association of the original protein is examined.

Experimental approach 18.2 describes yet another assay (termed APEX) whereby the protein of interest is tagged with ascorbate peroxidase (APEX), which biotinylates nearby proteins, allowing them to be identified.

Protein binding sites in DNA and RNA can be identified by co-immunoprecipitation and co-purification

Co-immunoprecipitation or co-purification can also be used to isolate DNA or RNA fragments that bind to a protein that is precipitated by the antibody or by an affinity tag. The DNA or RNA can then be identified using the microarray approaches we learned about in Sections 19.11 and 19.12 or the sequencing methods we learned about in Section 19.9. As for co-immunoprecipitation and co-purification assays to detect interactions between proteins, appropriate controls (such as samples lacking the protein of interest) and follow-up experiments need to be carried out to rule out "non-specific" binding to DNA or RNA.

A shift in DNA or RNA mobility in gels can be used to measure binding affinity

The rate at which a given macromolecule migrates in an electrophoretic gel depends upon its size. If a molecule forms a stable complex with one or more additional macromolecules, the complex will migrate more slowly in a non-denaturing gel than the individual molecule alone. An approach, termed an **electrophoretic mobility shift assay (EMSA)**, is frequently used to measure the strength of protein binding to a DNA fragment (an example of which is given in Experimental approach 16.2) or an RNA species.

We discuss binding equilibria in more detail in Section 3.2.

In an EMSA of protein binding to DNA, a radiolabeled DNA fragment is incubated with a protein and then loaded onto a non-denaturing gel, to preserve protein–DNA interactions, as depicted in Figure 19.78. DNA with bound protein will migrate more slowly in the gel than the free DNA fragment. The relative amounts of bound and free DNA in each resulting band can be quantified, thus giving a measure of how much complex is formed at that concentration of binding partners. If this is repeated for a series of samples that each contains a different concentration of protein, a binding curve can be constructed and an equilibrium binding constant can be determined.

Precise binding sites on DNA and RNA can be identified by footprinting

The precise site on a DNA fragment or RNA molecule to which a protein binds can be identified using **footprinting**. The underlying principle of this method, illustrated in Figure 19.79, is that a protein bound to DNA or RNA can protect the region of contact from chemicals or enzymes that cleave nucleic acids.

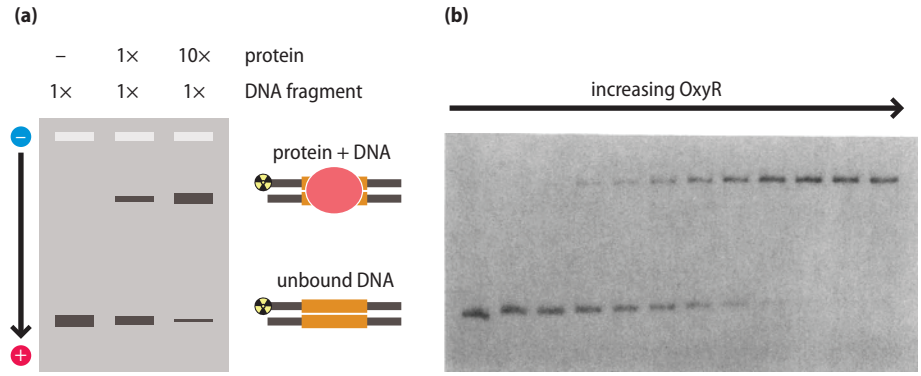


Figure 19.78 EMSA. A non-denaturing gel is used to assay binding of a protein to a DNA fragment. The first lane on the left shows the migration of the DNA fragment (dark gray band) on its own. In these experiments, the DNA is typically labeled with radioactivity or a fluorophore. The middle lane shows the migration of the DNA fragment when protein is present; the protein–DNA complex is larger than the DNA fragment alone and therefore migrates more slowly through the gel than the free DNA. The right lane shows the migration of the DNA fragment in the presence of ten times more protein than in the middle lane. This leads to a higher proportion of the DNA fragments binding to the protein, and thus more of the DNA fragments migrate slowly. (b) An example of an EMSA, showing how a radiolabeled DNA fragment bound to the OxyR protein is shifted to a more slowly migrating form as increasing amounts of the OxyR protein are added.

Adapted from Tartaglia, L. A., et al. (inc. Stortz). (1992) Multidegenerate DNA recognition by the oxyR transcriptional regulator. *Journal of Biological Chemistry* **267**: 2038–2045.

To carry out a footprinting experiment, many copies of a particular DNA or RNA molecule are labeled at one end with a radioactive atom (such as ^{32}P) or a fluorescent tag, and then treated with a chemical or endonuclease that cleaves the nucleic acid. The reaction is performed under conditions in which each DNA or RNA molecule in the solution is cleaved only once on average. This generates a set of fragments, with each fragment 1 bp longer than the next. The DNA is denatured to separate the strands, the resulting fragments are separated on a polyacrylamide gel, and the gel is autoradiographed.

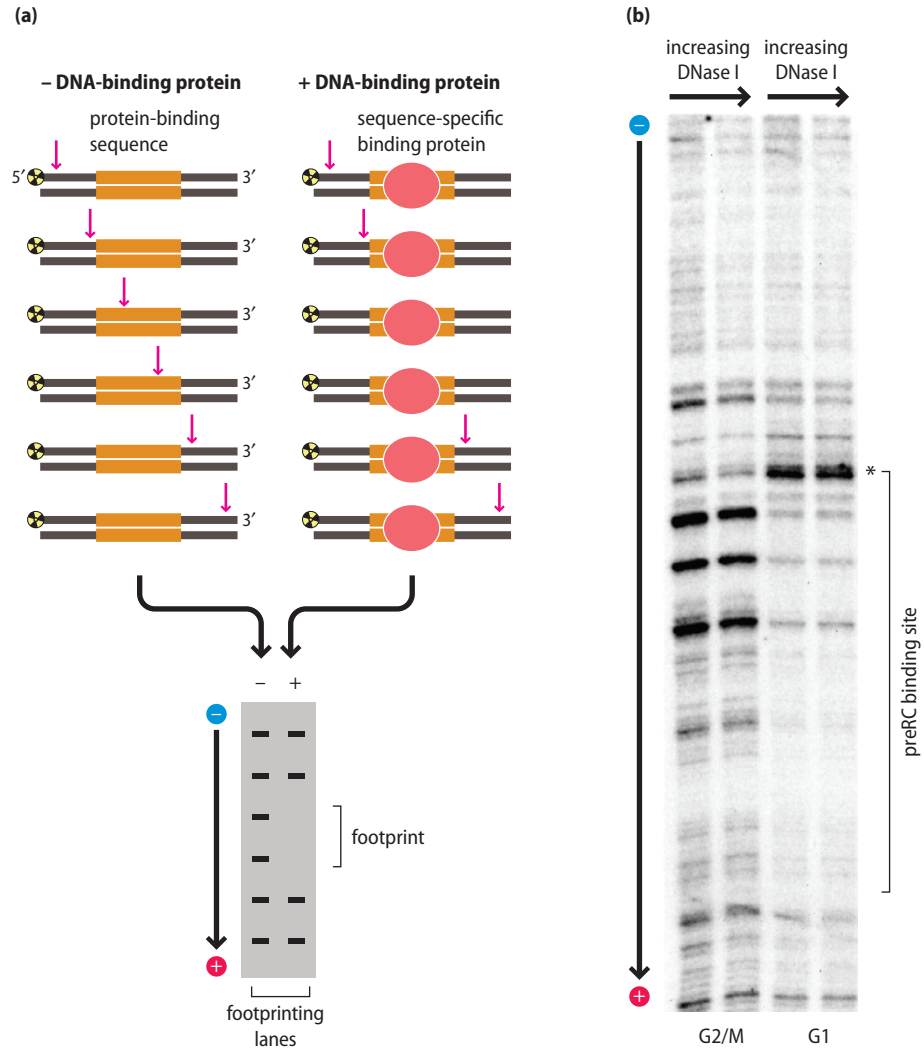


Figure 19.79 DNA footprinting assay. (a) DNA is labeled at one end and subjected to limited digestion, producing a ladder of fragments when electrophoresed on a gel, with smaller DNA fragments migrating further than larger fragments. If a protein is bound to the DNA during the digestion, it will protect a portion of the DNA from cleavage. See text for a more complete description. (b) An example of footprinting in a whole cell lysate. The region that is spanned by the black bar contains an *S. cerevisiae* origin of replication. When the cells are in G2/M, the origin is not bound by the pre-replication complex (preRC; see Chapter 6) and is thus exposed to DNase I digestion, generating the observed bands. In G1, however, the origin is bound by the preRC, protecting it from DNase I activity and generating a footprint that reveals the protected region. The dark band (indicated by an *) illustrates the paradoxical effect of protein binding sometimes enhancing endonuclease digestion at certain positions in the DNA.

Adapted from Labib et al. 2001. MCM2-7 Proteins are essential components of prereplicative complexes that accumulate cooperatively in the nucleus during G1-phase and are required to establish but not maintain, the S-phase checkpoint. *Molecular Biology of the Cell.* **12**: 3658–3667.

If no protein is bound, a ladder of bands corresponding to the collection of cleavage products will appear. If, however, a protein is bound to the DNA or RNA before it is incubated with the chemical cleavage agent or enzyme, the nucleic acid will be cleaved everywhere, except where it is protected by the protein. The resulting ladder on the gel will show fainter bands or no bands in the protected region.

Base-pairing interactions between two different RNA molecules can also be examined by variations of the footprinting technique. Again, the region where the base-pairing takes place is expected to have a different cleavage pattern with and without the partner molecule.

Two-hybrid analysis is a genetic approach for detecting interactions between proteins

A common method for studying protein interactions that is carried out *in vivo* and lends itself to large-scale screening is called the **two-hybrid assay**. This method is generally carried out in budding yeast, although it can be done in bacteria (*E. coli*) as well.

The yeast two-hybrid assay is carried out using a specially constructed strain of yeast containing a reporter gene that must be activated by a transcription factor such as GAL4, which contains distinct DNA-binding and activation domains. The reporter gene can confer growth under a particular condition or encode an enzyme such as β-galactosidase that allows visual detection of gene activation. As illustrated in Figure 19.80, GAL4 is re-engineered to consist of two separate polypeptides, one containing the DNA-binding domain and the other containing the activation domain. Thus, the reporter gene can only be activated if the fragment containing the activation domain associates with the DNA-binding domain.

How can these separate proteins be induced to form a complex? The DNA-binding domain is fused to an additional domain called the "bait," and the activation domain is fused to a domain called the "prey." If the bait and the prey bind to each other, then the DNA-binding fragment can bind to the regulatory site in the DNA and recruit the activation domain, thus turning on the reporter gene.

This system can be used to test many different combinations of bait and prey proteins, thereby serving as a convenient and rapid screen for interaction partners. Recombinant DNA techniques are used to construct a library of constructs with different proteins (or protein fragments) as bait or prey. These are then introduced into yeast cells, which can be assayed for particular combinations of bait and prey proteins that bind to one another, thus stimulating transcription of the reporter gene. Variations of the two-hybrid approach also can be used to detect protein–RNA and RNA–RNA interactions. Given the possibility of false-positives (interactions that do not normally take place in the cell), however, it is prudent to follow up all of these approaches with additional experiments that further test the detected interaction.

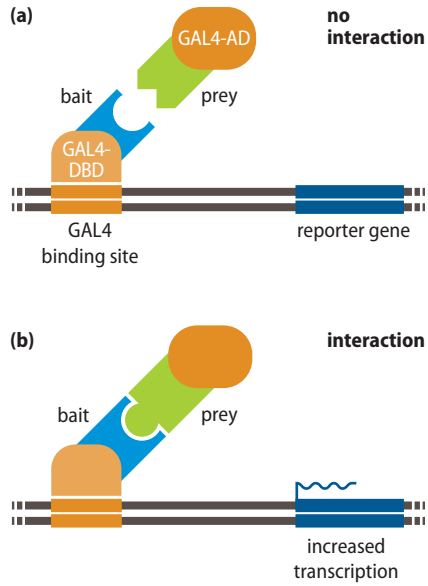


Figure 19.80 Yeast two-hybrid assay. A protein or domain of interest (the "bait," in blue) is fused to the GAL4 DNA-binding domain (orange, DBD), while a different protein fragment, or a collection of protein fragments, the "prey" (in green), is fused to the activation domain (orange, AD). (a) If the bait and prey do not interact, the reporter gene will not be expressed. (b) Transcription of the reporter gene is only activated if the "bait" and "prey" bind to each other.

Spectroscopic signals provide sensitive approaches for detecting molecular interactions

Many biological molecules are fluorescent, meaning that they absorb light at one wavelength but emit it at a different (longer) wavelength. The chemical environment can substantially affect the absorption and emission properties of a fluorescent moiety, and thus fluorescence can be used to inform scientists about the properties of a molecule. For example, tryptophan side chains have different emission spectra, depending on whether they are buried in the hydrophobic core of a

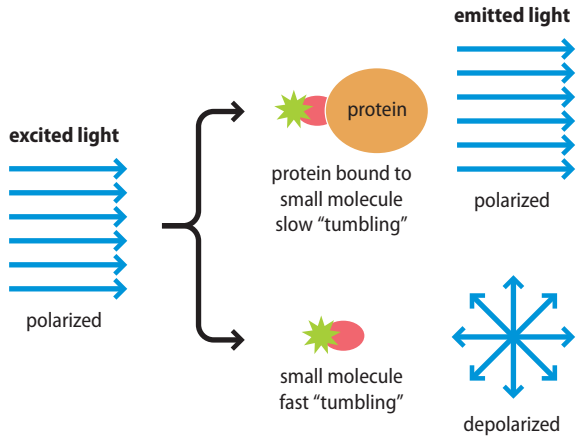


Figure 19.81 Monitoring binding interactions by fluorescence anisotropy. A fluorophore that is excited by polarized light will emit light that is polarized along the same direction. However, if the molecules are tumbling very rapidly (bottom), they will orient randomly, and thus each will emit light in random directions. The result is that the total emitted light is unpolarized. If, however, a large molecule binds to the fluorescently tagged molecule and slows down its rate of tumbling, the molecules will not have time to randomly orient after being excited and will thus emit light that is polarized along the same direction.

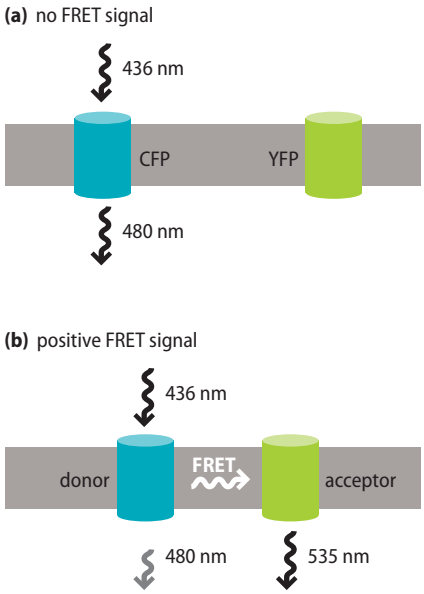


Figure 19.82 Principle of FRET. (a) The fluorescent proteins CFP (cyan fluorescent protein) and YFP (yellow fluorescent protein) have different absorption and emission spectra. Light with wavelength 436 nm will excite CFP. However, YFP has a different absorption spectrum and so does not absorb light of wavelength 436 nm and therefore does not fluoresce. (b) If the two fluorescent molecules are sufficiently close to each other, following excitation by light with wavelength 436 nm, some of the energy of the excited CFP (the FRET donor) can be transferred to the YFP molecule (the FRET acceptor). This transfer of energy excites YFP, which emits light at the YFP emission wavelength (535 nm). The donor and acceptor must be close to each other for energy transfer to occur, thus making measurements of acceptor fluorescence a sensitive measure of distance between the two molecules.

protein or lie on the surface where they are exposed to solvent. Tryptophan fluorescence can therefore be monitored to follow folding and unfolding of a purified protein in solution.

The fluorescence of a molecule can either increase or decrease when another species binds to it. Changes in fluorescence intensity can therefore be monitored to measure the binding of one molecule to another. Even if a molecule lacks intrinsic fluorescence, a fluorescent dye can be covalently attached and its fluorescence monitored. Proteins can also be engineered to have tryptophan side chains on the surface, whose fluorescence can then be monitored to follow a binding interaction. The instrument used to measure fluorescence is called a fluorimeter.

Another way in which fluorescence can be used to monitor binding interactions takes advantage of the fact that a fluorophore that is excited by polarized light will also emit polarized light, as depicted in Figure 19.81. If the molecules containing the fluorophore are tumbling rapidly in solution, after excitation, the fluorophores in solution all reorient in a random fashion and the sum of the fluorescent emissions no longer has a maximum in any particular direction. If, however, a larger molecule binds to some of the fluorescently tagged species and slows their rate of tumbling, some of the emitted light will still be polarized along the same direction because the molecules did not all have time to reorient in a random way. A method that takes advantage of this property is referred to as fluorescence anisotropy (where anisotropy means not the same in all directions). By plotting the intensity of emitted polarized light as a function of increasing concentration of the unlabeled partner, it is possible to determine the rate constant for the binding reaction (the binding constant).

Another approach that also takes advantage of fluorescently labeled molecules is **Förster resonance energy transfer** (also called **fluorescence resonance energy transfer or FRET**). This method can be used to monitor interactions between two molecules, as well as to measure the distance between them. A FRET experiment requires two different fluorophores with distinct absorption and emission spectra, as depicted in Figure 19.82. If the molecules are sufficiently close to one

another (within 5 nm), the energy emitted by the excited donor (that is, after exposure to the excitation wavelength) can be absorbed by the acceptor, causing the acceptor to fluoresce. By exciting the donor molecule and then measuring the fluorescence of the acceptor, it is possible to monitor interactions between the two molecules with a high degree of accuracy. Since the efficiency of the energy transfer depends strongly on distance (varying as $1/r^6$, where r is the distance between the donor and acceptor), the intensity of the emitted light can even be used to measure the distance between the donor and the acceptor. We shall see in Section 19.16 how FRET measurements can be used in whole cells to detect molecular interactions *in vivo*.

We learn more about binding constants in Section 3.2.

Surface plasmon resonance can be used to monitor the kinetics of binding interactions

Another sensitive technique for studying interactions between two molecules relies on surface plasmon resonance (SPR)-based technology. This method is used to measure binding interactions where one of the interacting species has been immobilized on a surface. Like the fluorescence experiments we have just described, this approach can yield quantitative information on molecular interactions. A key difference is that, rather than directly measuring the equilibrium binding constant, SPR is used to measure both the association and dissociation rates, from which the K_d can then be calculated. The ability to monitor binding kinetics in real time provides additional kinetic parameters that cannot be deduced from the equilibrium binding constant alone.

We learned about the thermodynamics of binding equilibria in Chapter 3.

As illustrated in Figure 19.83a, one of the molecular species of interest (the "bait") is immobilized on a thin gold film that lies on top of a glass surface. Buffer containing the other molecular species of interest (the "prey") is then flowed across the surface. If the prey has an affinity for the immobilized bait, more and more prey will bind to the bait until the surface is saturated. If fresh buffer is then flowed across the surface, the prey molecules will dissociate at a characteristic rate until no more molecules are bound.

We learn more about association and dissociation rates in Section 3.2.

The binding of molecules to the bait immobilized on the surface and their subsequent dissociation produce changes in the refractive index or in how the light propagates, in the immediate vicinity of the surface layer. (The physical principles underlying this phenomenon are beyond the scope of this book.) As a consequence of the refractive index change, the angle at which the incident light is refracted changes. A plot of the change in the resonance angle as a function of

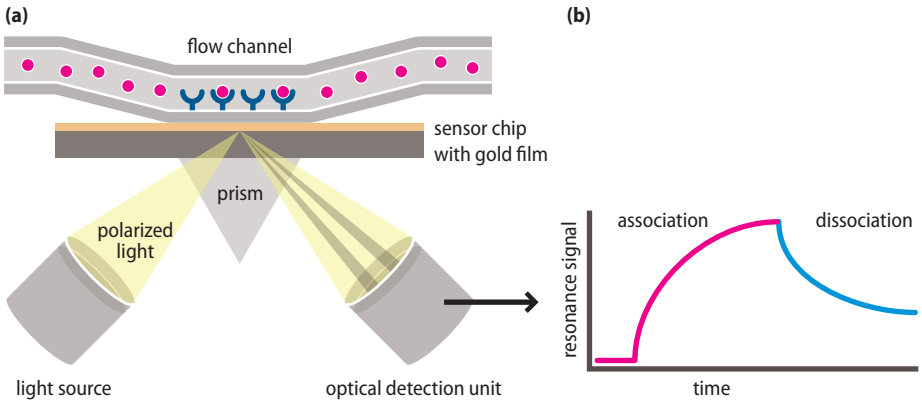


Figure 19.83 Surface plasmon resonance. (a) A glass sensor chip contains a thin layer of gold on which proteins or other molecules (shown in dark blue) can be immobilized. A solution containing a binding partner (pink spheres) flows across the upper surface. At the same time, a light is shone at an angle on the lower glass surface and the refracted light is detected. As molecules in the solution bind to the immobilized protein, the refractive index of the glass changes (due to a physical phenomenon known as surface plasmon resonance). This alters the angle at which the refracted light exits the sensor chip. (b) The changes in the refractive index are recorded by the optical detection unit. The changes can be monitored in real time. The pink curve represents the change in refractive index as more protein is bound, while the blue curve represents the change in refractive index as the binding partner is being washed away.

Cooper, M. (2002). Optical biosensors in drug discovery. *Nature Reviews Drug Discovery* **1**:515–528.

time, known as a sensorgram (an example of which is depicted in Figure 19.83b), is then used to evaluate binding.

The thermodynamic parameters of a binding reaction can be measured using isometric titration calorimetry

We learned about the thermodynamics of binding equilibria in Chapter 3.

The equilibrium binding constant gives a direct measure of the free energy of binding, as given by the equation $\Delta G = -RT \ln K$. While we have seen different methods for determining the equilibrium binding constant, from which the free energy of binding can be calculated, it is also possible to measure the free energy directly using isothermal titration calorimetry (or ITC). This method can be used to determine all three thermodynamic parameters that characterize a binding interaction: enthalpy (ΔH), entropy (ΔS), and Gibbs free energy (ΔG), which are related by the Gibbs free energy equation, $\Delta G = \Delta H - T\Delta S$.

While information about the strength of a binding reaction is given by ΔG, the relative contribution of entropy versus enthalpy provides valuable additional information. For example, a binding reaction in which an unstructured region becomes ordered upon complex formation will have a larger change in entropy than a comparable reaction in which both binding partners are already well ordered on their own. Since ITC is performed in solution, it is free of potential artifacts that may arise with methods that require one or both molecules to be labeled or immobilized. A disadvantage is that relatively large amounts of pure material are needed for each experiment.

The apparatus used for ITC is depicted in Figure 19.84a. There are two cells, one containing a reference buffer and one containing a sample solution with one of

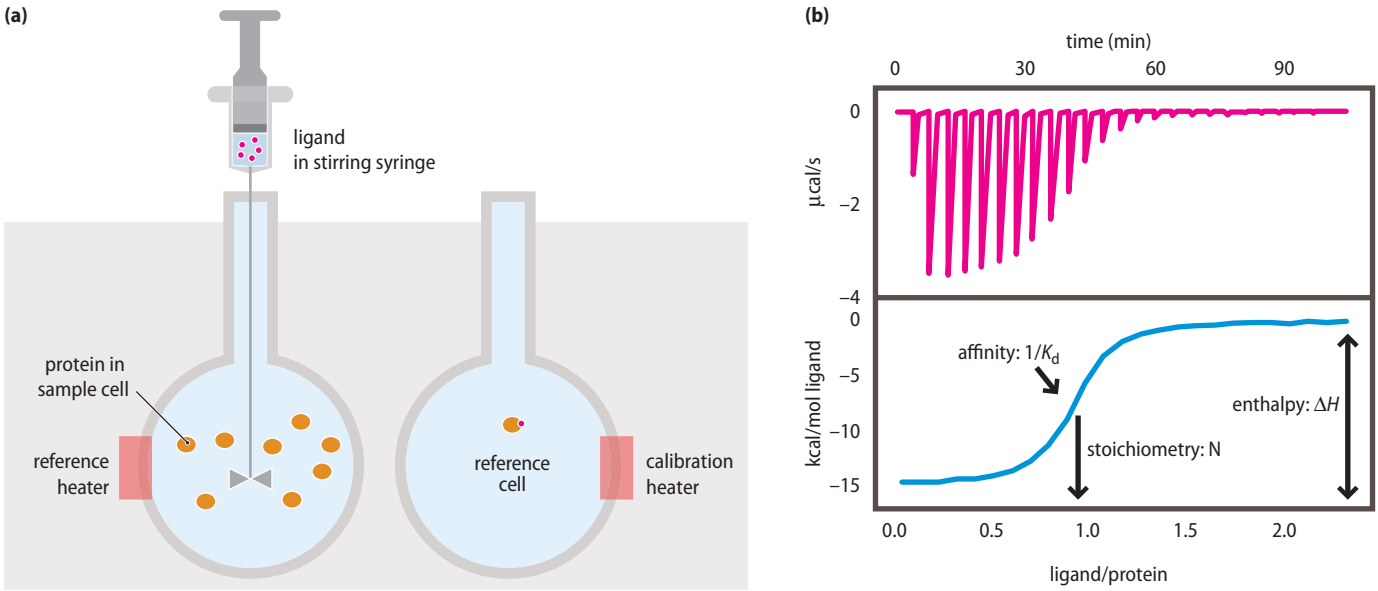


Figure 19.84 ITC. (a) The sample cell (left) and reference cell (right) are maintained at a constant temperature. A small amount of ligand is injected and binds to the protein, which causes the temperature to change slightly as heat is either absorbed or released by the binding reaction. The reference heater adjusts the sample cell temperature to return to that of the reference cell, and the resulting change in heat is recorded as in panel (b). (b) Top panel shows heat absorbed with each successive injection. As the binding sites on the proteins in the sample cell become occupied, less and less ligand is bound as its concentration exceeds that of the protein. Lower panel shows how the area for each peak (or trough) can be used to plot the heat for each injection and to determine the binding constant and enthalpy of the binding reaction.

(b) With permission from Harvey T McMahon, http://www.endocytosis.org/

the molecules of interest. A sensitive temperature detection system ensures that the sample cell remains at the same temperature as the reference cell—any temperature change in the sample cell will trigger either an increase or a decrease in the current going to the cell, depending upon whether the sample cell needs to be heated or cooled slightly, so that its temperature once again matches that of the reference cell. This configuration is then used to measure the heat that is either released or absorbed when a ligand is introduced into the sample cell.

If a small amount of the ligand is injected into the sample cell and the ligand binds to the molecule of interest, there will be a very slight change in the temperature of the sample, which will result in a temporary trough or spike in the current as the temperature is brought back to that of the reference cell (see Figure 19.84b). After repeated injections, each one adding more ligand to the sample cell, saturation will be reached as the concentration of the ligand exceeds the K_d and little of the injected ligand binds. By repeatedly injecting small amounts of the ligand, the relative change in heat for each injection can be measured and plotted as a function of the ligand to determine the equilibrium constant.

19.15 GENOME-WIDE DETECTION OF INTERACTIONS BETWEEN MOLECULES

With the escalating ease with which the sequence of an entire genome or transcriptome can be determined, as we discussed in Section 19.10, deep sequencing is increasingly being used to identify molecular interactions, on a genome-wide scale, between proteins and DNA, proteins and RNA, and between two regions of DNA or base-pairing RNAs. There are many different permutations of these approaches, most given clever acronyms, but we will only discuss a subset here to illustrate the general approaches.

Cross-linking, together with co-immunoprecipitation or co-purification, can be used to identify protein-binding across a genome

One application of co-immunoprecipitation to identify proteins that are bound to a particular region of chromosomal DNA in the cell is called **chromatin immunoprecipitation (ChIP)** and is illustrated schematically in Figure 19.85. In this method, cells are treated for a limited period with a chemical reagent that covalently cross-links proteins to DNA. Once the bound proteins are immobilized in this way, the chromosomal DNA is sheared into relatively small fragments. Antibodies against a DNA-binding protein of interest are then added to the mix to immunoprecipitate the protein, together with the cross-linked piece of DNA. A reagent is then added, which reverses the cross-links, releasing the DNA into solution and allowing the sequence to be determined.

All of the sites in a genome to which a particular protein binds can then be identified by hybridization to a microarray, sometimes referred to as "ChIP on chip" or "ChIP-chip" or directly by deep sequencing, sometimes referred to "**ChIP-Seq**." PCR can also be used to amplify just a particular sequence of interest by using appropriately designed primers, as described in Section 19.3.

Sometimes, instead of wanting to know what regions are bound by a protein, scientists want to know what regions are **free** from proteins, and thus more accessible to modification by chemical agents or transposons. In one such assay for protein-free regions, termed assay for transposase-accessible chromatin

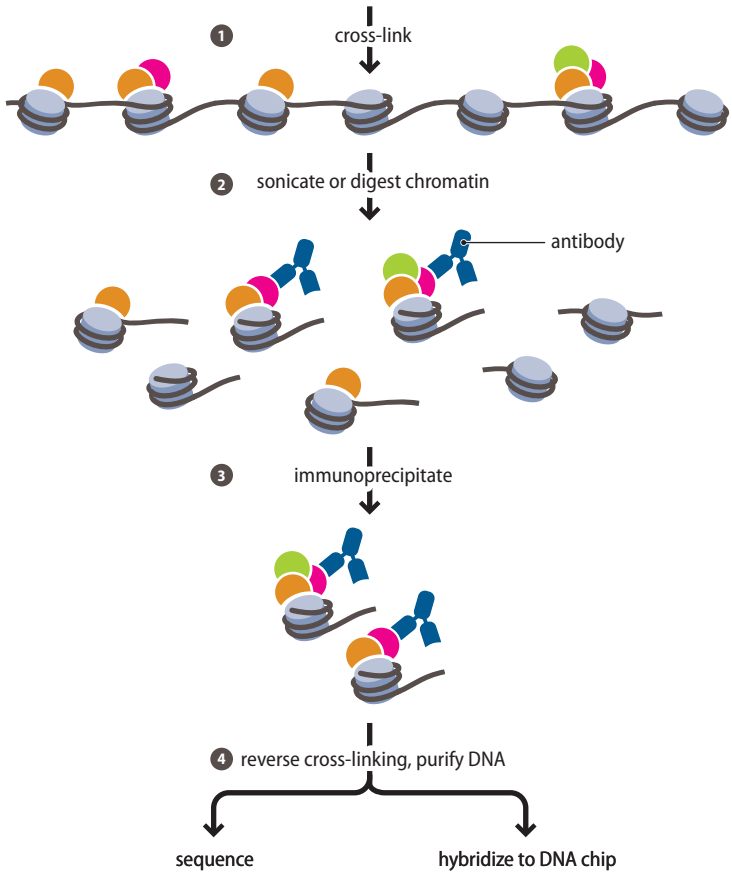


Figure 19.85 ChIP. In ChIP, chromatin is treated to chemically cross-link proteins to the DNA (step 1), after which the DNA is digested by restriction enzymes or mechanically sheared (step 2). The DNA fragments and their associated proteins are then subjected to immunoprecipitation, using an antibody (in dark blue) against the protein of interest (pink), which isolates only those DNA fragments that are bound to this protein (step 3). The cross-links are then removed, and the DNA is isolated. Linkers are then added to the DNA ends and the DNA is amplified by PCR (step 4), and the amplified DNA is analyzed by one of a variety of methods, including hybridization to a DNA chip or sequencing.

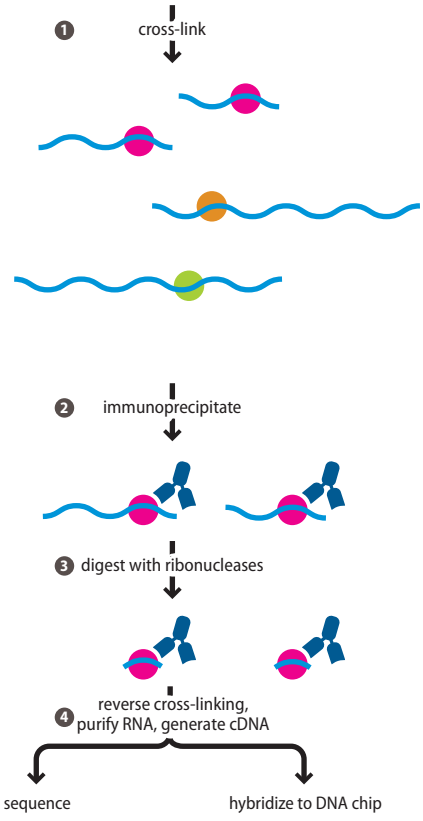


Figure 19.86 CLIP. In CLIP, cross-linking is used to stabilize protein–RNA interactions (step 1), after which the protein of interest (pink) is immunoprecipitated (step 2). RNA not protected by the protein is then digested with ribonucleases (step 3). After the cross-links are reversed, the RNA is isolated and cDNA is generated and amplified (step 4), after which the amplified DNA is analyzed by one of a variety of methods, including hybridization to a DNA chip or sequencing.

combined with sequencing (ATAC-Seq), transposition is induced in a pool of cells. The transposon will preferably insert into accessible chromatin, which is not tightly bound by other proteins. Subsequent sequencing of the chromosomal DNA from the pool of cells will reveal what sites have preferential transposon integration, thus indicating they are more likely protein-free.

Cross-linking, together with co-immunoprecipitation or co-purification, can be used to identify protein-binding across a transcriptome

An approach similar to ChIP-Seq can be used to identify RNA bound to a particular protein, as illustrated in Figure 19.86. In this method, known as cross-linking and immunoprecipitation (CLIP), proteins are cross-linked to their target RNAs in living cells using short-wave UV irradiation, thus generating a covalent bond between protein and RNA. The particular protein of interest can be immunoprecipitated, along with its bound RNA, as in the case of ChIP. After the complex is precipitated, extraneous RNA that extends beyond the interaction of interest can be digested with ribonucleases, leaving the RNA fragment that contacts the protein and is thus protected from digestion.

In a variant of this CLIP method, the ribonucleotide 4-thiouridine is added to cells and is therefore incorporated biosynthetically into the cellular RNA. RNA containing this nucleotide can more efficiently be cross-linked to proteins using lower-energy UV light. This method is referred to as photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP).

In all cases, the cross-linking can be reversed to allow isolation of the RNA and the generation of cDNA. The amplified cDNA can either be sequenced using the methods described in Section 19.10 or be hybridized to a DNA microarray, making it possible to identify the sites in the RNA to which various proteins bind.

Extensions of the co-immunoprecipitation and co-purification approaches are now being exploited to interrogate the processes of transcription and translation on a genome-wide level. For example, the mRNA regions that are bound to ribosomes, and are thus protected from nuclease digestion, can be isolated and identified in an approach termed **ribosome profiling** or Ribo-Seq. This process, which is summarized in Figure 19.87, begins with the preparation of a cell lysate that is treated with a general RNase. mRNAs that are bound by ribosomes will yield an ~30-nucleotide region that is protected from nuclease digestion. Importantly, as other RNA-associated proteins will also be present in the lysate, the purification of monosomal ribosomes, along with the protected mRNA segment, using a sucrose gradient increases the quality of the sample. The RNA bound to the ribosome is isolated, and linkers are attached to its ends to allow for the generation of cDNA. This cDNA is then used for sequencing to reveal the genes that were being translated (and the precise position where ribosomes are found). In general terms, the relative abundance of the sequences obtained per gene is proportional to the expression level of that gene.

RNA polymerase similarly can be purified and the associated RNAs identified to obtain a global picture of the RNAs actively being transcribed. Global profiling approaches for both RNA polymerases and ribosomes can provide many insights missed by studies of individual genes. For instance, ribosome profiling allows the quantification of the mRNAs that are being translated at a particular time and the identification of protein-coding genes that might have been missed by standard annotation because they are short or have unanticipated start sites. The approaches can be carried out for strains with mutations in components of the transcription or translation apparatus to identify how those components contribute to steps in the respective processes. Finally, chemical inhibitors can be added to trap all of the polymerases or ribosomes in a specific step, allowing the consequences to be examined on a global level.

Ligation allows interacting DNA regions or interacting RNAs to be identified

In addition to wanting to know where proteins bind to DNA along a chromosome, we may want to know what DNA sequences are near each other in three-dimensional space, which can be taken to indicate chromosomal regions that are interacting with each other. Such functional interactions are seen, for example, between enhancer sequences that interact with promoters, which are often located a great distance away.

A technique called chromosome conformation capture has been developed to identify DNA sequences that interact or are often in close proximity in the cell. Again, the approach first involves cross-linking to stabilize any interactions, as depicted in Figure 19.88. The DNA is then digested or sheared to give shorter fragments. The difference between this approach and the other methods described in this section is that the DNA ends are first ligated to each other before the cross-link is reversed. The resulting ligated fragment contains sequences from two chromosomal regions that may be significantly separated on the chromosome or are even located on different chromosomes.

We discuss ribosome profiling in Experimental approach 12.1.

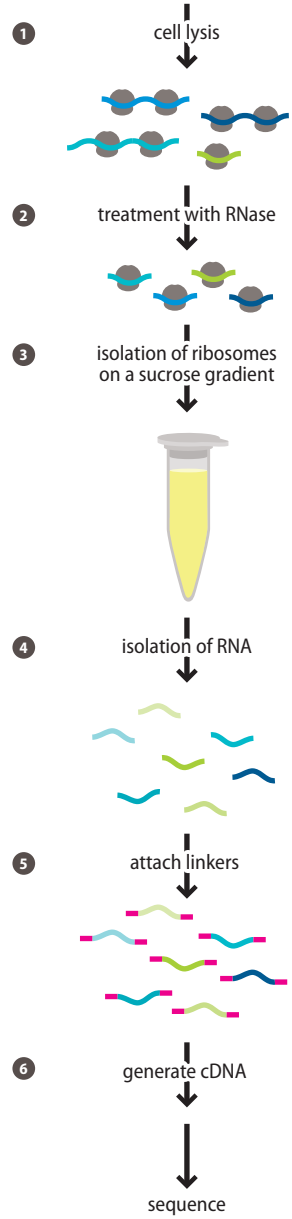


Figure 19.87 Ribosome profiling. Cells are lysed (step 1) and treated with RNase (step 2) to remove all RNA regions that are not ribosome-bound. Monosomal ribosomes are then purified by separating them on a sucrose gradient (step 3), after which the RNA that was bound to the ribosome is isolated (step 4). This RNA is then converted to cDNA by ligating it to primers (step 5) and converting it to cDNA with reverse transcriptase (step 6). These cDNA fragments are then sequenced and aligned to the transcriptome to reveal the mRNAs that were being translated in the cells. The number of sequences recovered for each gene should be proportional to the number of ribosomes that were engaged in translating the gene's mRNA.

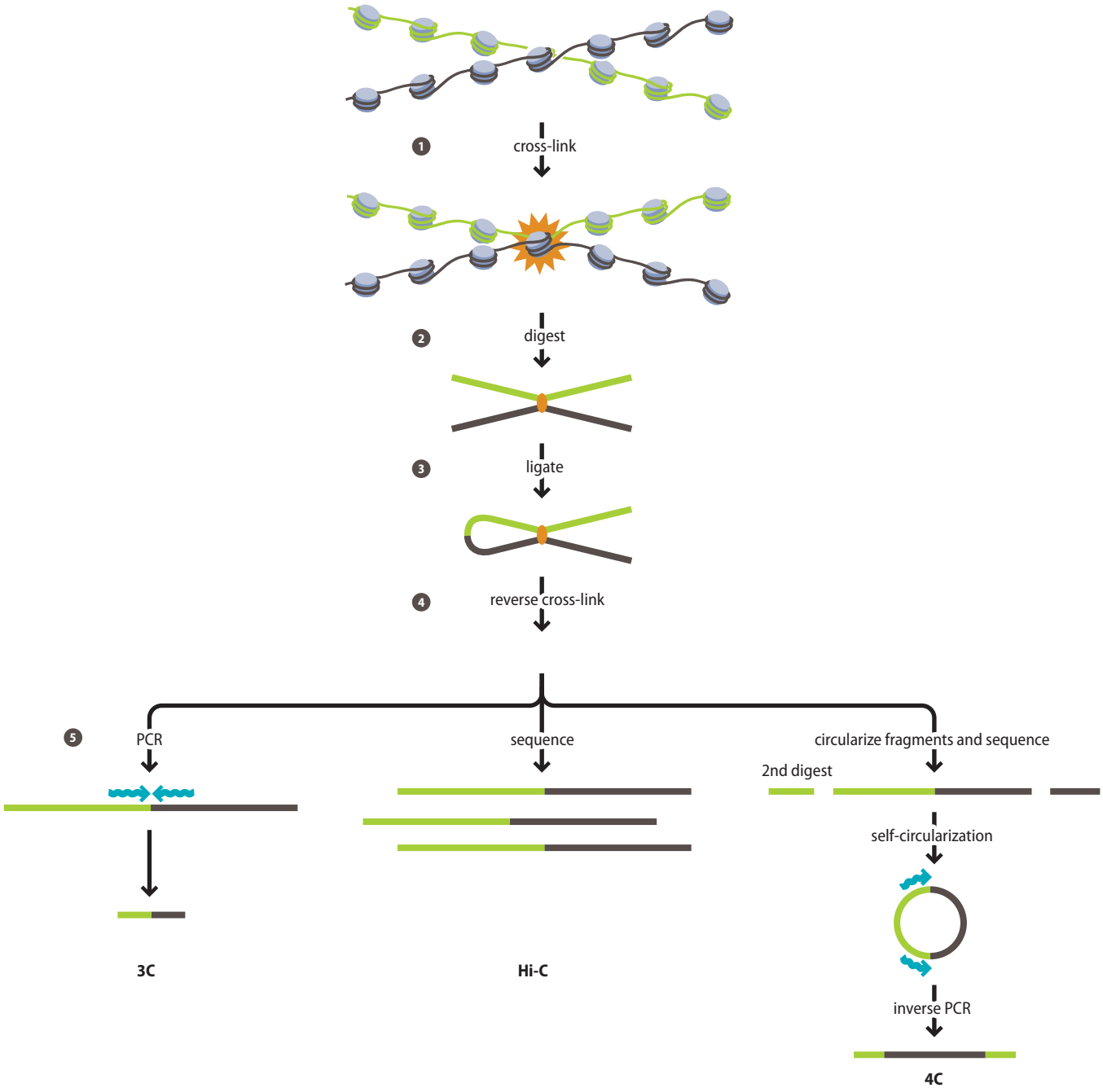


Figure 19.88 Chromosome conformation capture. In chromosome conformation capture, cross-linking is used to stabilize DNA–DNA interactions (step 1), after which DNA is digested or sheared (step 2). With the addition of ligase, nearby ends are ligated to each other (step 3). After the cross-links are reversed (step 4), the composition of the chimeric fragment can be determined (step 5) for individual regions of interest (3C) or on a genome-wide scale (Hi-C). When it is desirable to determine all interactions with one region of interest (4C), the chimeric fragments are digested with another enzyme and then circularized to simplify sequencing (step 5).

When this approach is applied to the interaction between two pieces of DNA, it is denoted 3C. However, with deep sequencing, the approach can also be carried out on a genome-wide basis. When all interactions with one region of interest are interrogated, the technique is denoted 4C; when all interactions across the genome are interrogated at once, it is denoted Hi-C.

In 4C, sequencing is simplified by digesting the chimeric fragments with a different enzyme and then circularizing the molecules. While the results from 3C, 4C, and Hi-C are standardly used to infer chromosomal domains and their conformation, we need to attach the caveat that the ligation of two fragments could be due to other factors such as their binding to the same protein complex. It is also important to remember that the results reflect the average of all the cells in the population, and thus the interactions detected may only exist in a subset of cells at the time of sampling.

Chromosomal interactions also can be tested by microscopy with the FISH approach described in Section 19.11. FISH has the advantage of examining single cells but usually is confined to examining a more limited number of interactions.

Experimental approach 4.4 describes Hi-C experiments to characterize the organization of bacterial chromosomes..

It may also be of interest to know what RNA molecules are in close proximity in a cell, often because they are base-paired. Methods related to the 3C, 4C, and Hi-C approaches, which also involve cross-linking followed by ligation, are used to identify interacting RNAs. In one method, denoted cross-linking ligation and sequencing of hybrids (CLASH), the RNAs are cross-linked to an RNA-binding protein such as Argonaute protein, which binds to microRNAs. Bound fragments are truncated with RNase and then ligated. After the cross-linking is reversed, the chimeric RNAs are isolated, converted to DNA, and sequenced. As we saw for CLIP, cross-linking can be enhanced by the incorporation of a photoactivatable ribonucleoside.

Experimental approach 13.1 describes one application of this type of cross-linking approach to identify mRNA targets of miRNAs.

A critical factor in the analysis of all of the deep sequencing approaches described in this section is the bioinformatic processing of the resulting sequences. This processing requires care in the determination of sequence ends, methods of normalization, and elimination of false-positive results, to name just a few of the critical parameters.

19.16 IMAGING CELLS AND MOLECULES

The ability to visualize cells and to localize specific cellular components has been fundamental to our understanding of many aspects of biology. In this section, we discuss a range of microscopy techniques that are used to image cells and molecules. Microscopy makes it possible to determine the subcellular localization of particular proteins or cellular structures, to follow the dynamics of cellular components over time, to examine the expression patterns of a gene of interest, and to determine whether two proteins interact *in vivo*. In addition, the analysis of single cells reveals information that is masked when we examine an entire population. For example, Figure 19.89 shows the variation in gene expression observed between different *E. coli* cells, all of which are genetically identical. These cell-to-cell differences cannot be observed when assessing the protein levels in a population by western blot analysis, for example. Cells can be visualized live or treated with chemical cross-linkers (a process known as fixation) to preserve intracellular structures. Moreover, cells can be observed whole or they can be sectioned. The method of choice is dictated by the biology question at hand.

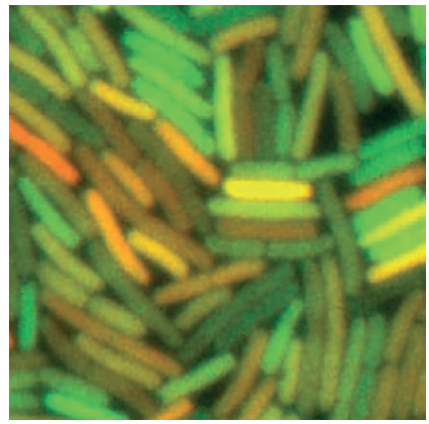


Figure 19.89 Cell-to-cell variation in gene expression. Each of these *E. coli* cells is expressing two proteins, one fused to CFP and one to YFP. The promoters of the *CFP* and *YFP* genes are the same. Although these cells are genetically identical and they are all growing in the same environment, different cells exhibit different levels of CFP (in green) and YFP (in red). This observation was made possible because the expression pattern was monitored on a cell-by-cell basis.

From Elowitz, M. B. et al. (2002). Stochastic Gene Expression in a single cell. *Science* **297**: 1183.

Many cell features can be visualized by light microscopy

Light microscopy is so-called because it uses light refraction to detect various cellular features. The features can be visualized because light passing through the sample will be absorbed differently by cell pigments or by the differing thicknesses

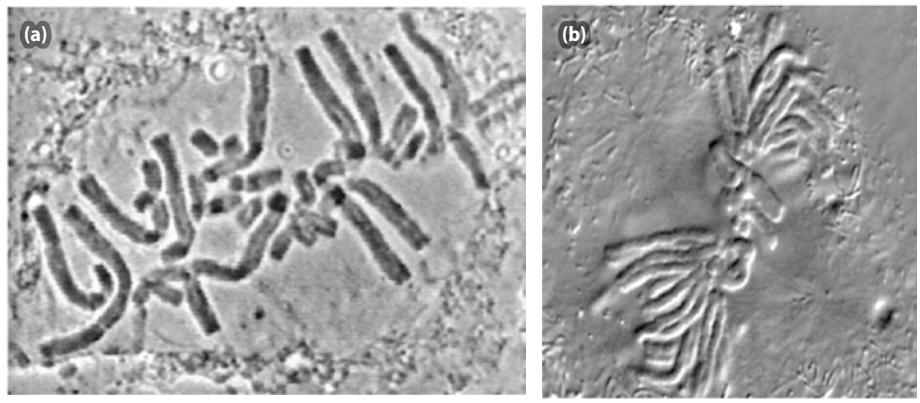


Figure 19.90 Phase contrast and DIC images of metaphase chromosomes. (a) A phase image of metaphase chromosomes from a flattened endosperm cell of the African blood lily *Haemanthus katherinae*. (b) A DIC image of metaphase chromosomes from newt lung cells. Note that the chromosomes in the phase image have light and dark regions; by contrast, DIC creates a three-dimensional image, but the chromosomes look uniform in composition.

Figure 19.79a from inoeu S. and Oldenberg, R. (1998) Microtubule dynamics in mitotic spindle displayed by polarized light microscopy. *Molecular Biology of the Cell* **9**: 1603–1607. Figure 19.79b from Conly Rieder's website, http://www.wadsworth.org/bms/SCBlinks/web_mit2/res_mit.htm.

of parts of the cell. However, most biological structures are similar in density to the cytoplasm, so various methods have been developed to increase the contrast between parts of the cell. These usually involve special lenses and filters that alter the light path within the microscope. Two common types of light microscopy are **phase contrast** and **differential interference contrast (DIC)**, images from which are shown in Figure 19.90. Phase contrast often provides a penetrating image of the cell, while the DIC method provides a three-dimensional-like image of the cell. Light microscopy can be used in combination with various stains or dyes that further increase the contrast and highlight particular cellular structures.

Intracellular structures can be seen by fluorescence microscopy in both fixed and live samples

Fluorescence microscopy makes use of the light emitted by certain dyes or proteins when they are illuminated with light of a shorter wavelength (known as the excitation wavelength, as depicted in Figure 19.91a). The characteristic excitation and emission wavelengths depend upon the chemical structure of the fluorescent molecule; scientists often use several different fluorescent molecules concomitantly, to highlight multiple cellular structures at once.

A variety of approaches have been developed to visualize specific cellular components using fluorescent molecules. These fall into two main categories: the use of fluorescent dyes or fluorophore-conjugated antibodies that are added to the cells, and the expression of fluorescent protein tags, such as **GFP** (**green fluorescent protein**), fused to the protein of interest. The former approach is usually used on fixed samples, because cells often need to be permeabilized to allow the dyes or antibodies to penetrate the cell. The fluorescent protein fusion approach can be used with either live or fixed cells.

A common example of a fluorescent dye is DAPI, which associates with double-stranded DNA (see Figure 19.91b). Other fluorescent dyes bind to different subcellular structures, such as actin filaments, membranes, or mitochondria. However, most cellular proteins or structures cannot be stained with a fluorescent chemical, and instead specific antibodies are used. Antibodies are generated in a host organism (such as mouse or rabbit) by injecting a protein or a peptide into the host. This, in turn, induces the immune system, which recognizes these proteins as foreign, to produce antibodies against certain protein domains. These antibodies, referred to as "primary antibodies," recognize and bind to the protein of interest through the antibody's variable, or Fab, domains. To detect the primary

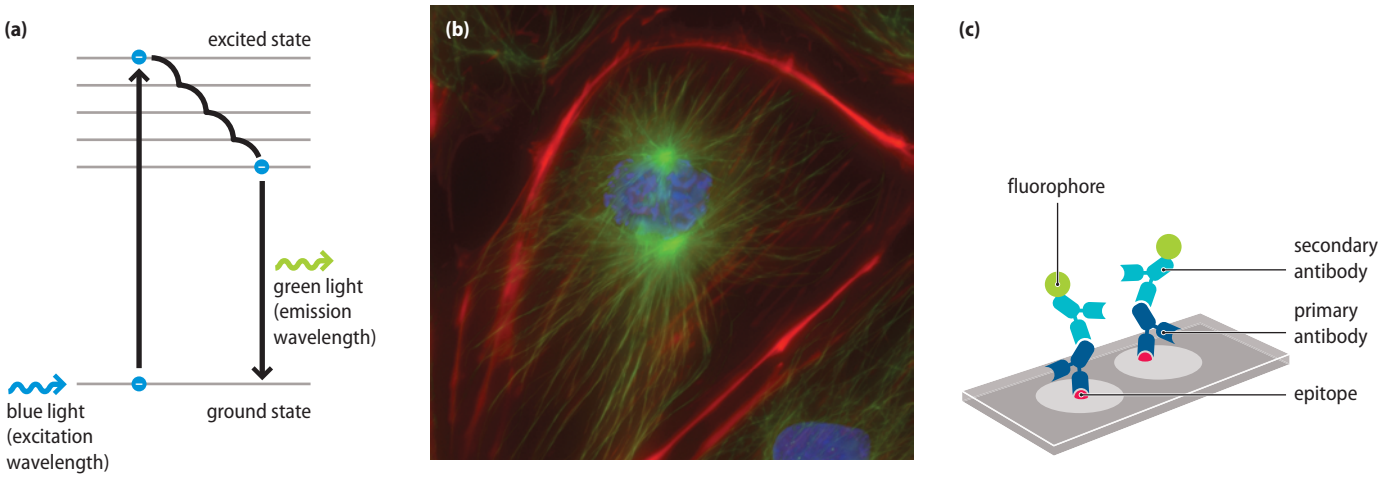


Figure 19.91 Fluorescence microscopy. (a) Each fluorophore or fluorescent protein has a specific excitation wavelength and emission wavelength. When the fluorophore is illuminated with high-energy light at the excitation wavelength, an electron moves to an excited state. On its return to the ground state, energy is released in the form of light at the emission wavelength. In fluorescence microscopy, the specimen is placed on a slide on the microscope stage and is illuminated at the excitation wavelength. The light emitted by the fluorescent molecule can be visualized through the microscope's eyepiece or captured by a specialized digital camera. (b) Multiple fluorescent dyes or fluorophores can be used at once. In the image shown, the blue signal corresponds to DNA, the green signal corresponds to microtubules, and the red signal corresponds to actin filaments. These structures can be distinguished from one another because they are labeled with fluorophores that have different excitation and emission wavelengths. (c) Indirect immunofluorescence. To detect proteins for which antibodies are available, the cells are placed on a slide, permeabilized, and treated with a primary antibody (dark blue) that recognizes the protein of interest (in red). The sample is then treated with a secondary antibody (light blue) that recognizes the primary antibody and is conjugated to a fluorophore. Thus, upon exposure to the appropriate excitation wavelength, the fluorophore associated with a secondary antibody will reveal the location of the protein of interest. This method is called indirect immunofluorescence because the fluorophore is not directly associated with the protein.

Figure 19.91b provided courtesy Wadsworth Center, New York State Department of Health.

antibody, secondary antibodies against the primary antibody's constant, or Fc, domains are made in a different host (for example, goat or donkey). The secondary antibodies can be conjugated to a fluorophore.

To detect the protein of interest, cells are fixed, placed on a microscope slide, permeabilized, and exposed first to the primary antibody. The slide is then exposed to the fluorophore-conjugated secondary antibody, as shown in Figure 19.91c. The fluorescent signal from the secondary antibody can then be visualized under the microscope to reveal the location of the target protein. This method is called **indirect immunofluorescence**. As with any experiment using antibodies, it is important to establish that the primary antibody is specific to the protein of interest and does not cross-react with other proteins.

Since a highly specific antibody against the protein of interest is not always available, as we learned in Section 19.13, scientists sometimes create fusion proteins between the protein and an epitope for which there are good antibodies. The fusion protein is then expressed in the cell so that its location can be detected by indirect immunofluorescence, using primary antibodies against the tag. While this is a powerful method, a potential pitfall is that the fusion protein may not fully mimic the endogenous protein.

Significant information about the localization of a molecule can be obtained by immunofluorescence, but it is a methodology that cannot be used on live cells because it requires cell fixation and permeabilization. Consequently, other methods are used to monitor the localization of a molecule for a period of time in live cells. In many cases, proteins are fused to the jellyfish **GFP**, the structure of which is depicted in Figure 19.92a, but other naturally fluorescent proteins and their derivatives that emit and fluoresce at different wavelengths can also be used, as shown in

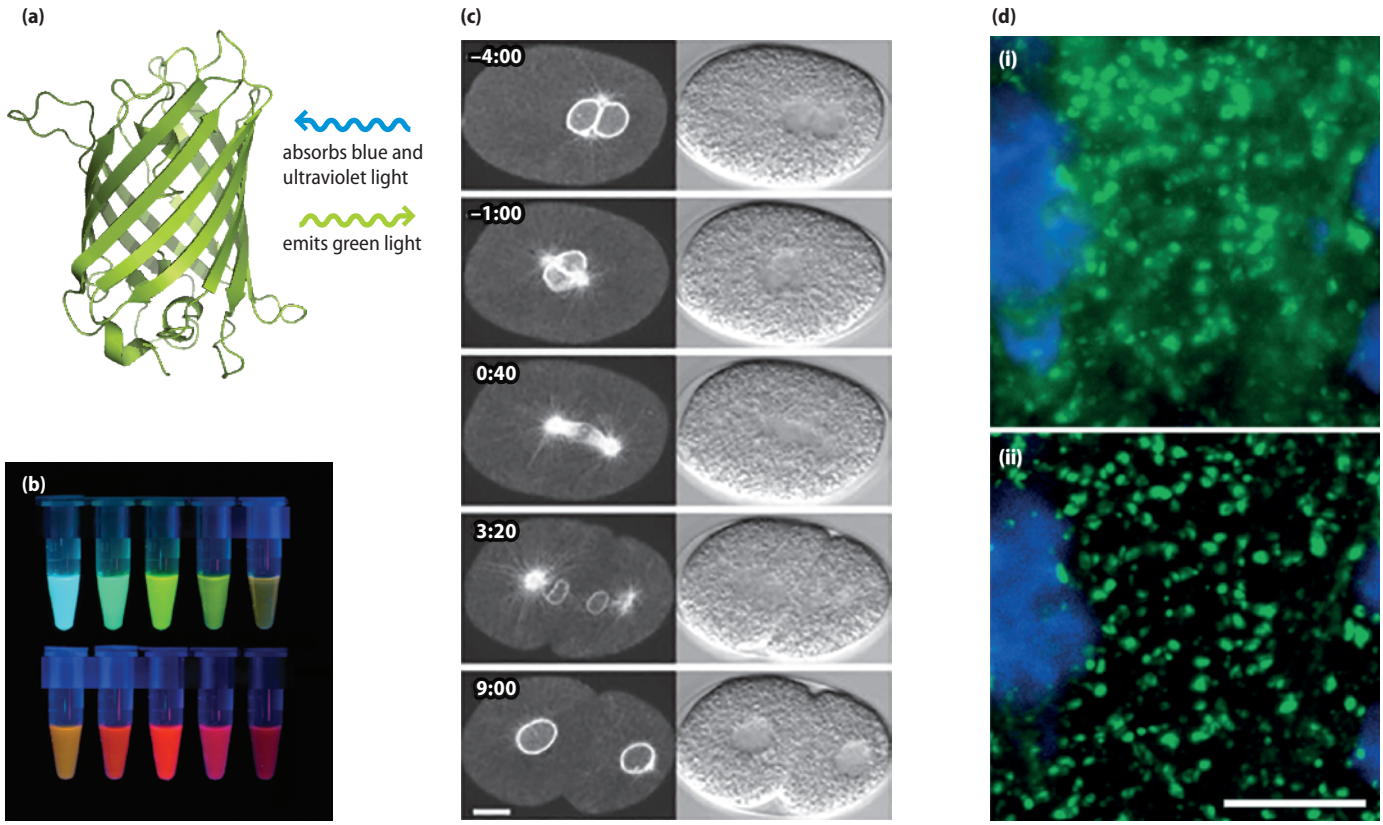


Figure 19.92 Fluorescent proteins. (a) The structure of GFP (Protein Data Bank (PDB) code 1EMA). (b) A series of fluorescent proteins, engineered from naturally occurring proteins, that emit at different wavelengths. (c) A series of images (fluorescence on the left, DIC on the right) taken of a *C. elegans* embryo expressing both GFP–lamin, which outlines the nuclear envelope (seen exclusively in time point 9:00) and GFP–tubulin (visible on its own at time point 0:40, after nuclear envelope breakdown) during cell division. These types of images are typically taken using a camera that detects light intensity, rather than color, resulting in a black-and-white image. These images can be pseudocolored, as was done in Figure 19.93. The times (minutes:seconds) are relative to anaphase onset. (d) A comparison of the images produced by conventional fluorescence microscopy (i) and confocal microscopy (ii). Notice the reduction in out-of-focus fluorescence in (ii). Both images show axon boutons found in human neurons.

Panel (b) from http://www.tsienlab.ucsd.edu/Images.htm; panel (c) reproduced with permission from Franz C, Askjaer P, Antonin W, et al. Nup155 regulates nuclear envelope and nuclear pore complex formation in nematodes and vertebrates. *The EMBO Journal*, 2005;**24**:3519–3531; panel (d) reproduced from Sweet R, Fish K, and Lewis D. Mapping synaptic pathology within cerebral cortical circuits in subjects with schizophrenia. *Frontiers in Human Neuroscience*, 2010; **4**: 44. Copyright © 2010 Sweet, Fish and Lewis.

Figure 19.92b. The localization of proteins tagged with these fluorescent proteins can be followed over time by recording a series of images (such as those shown in Figure 19.92c). In these types of experiments, researchers generate a gene fusion between their protein of interest and a fluorescent protein, such as GFP, and then express them *in vivo*. As is the case with immunofluorescence, these cells are then illuminated on the microscope stage with the appropriate excitation wavelength, and images are acquired using a sensitive digital camera.

Different types of microscopes can be used to detect fluorescence in biological specimens. In a conventional microscope, the specimen is flooded with light at the appropriate wavelength, resulting in the excitation of the fluorophores throughout the exposed area. Some of these molecules are in the microscope's plane of focus; others are either above or below this plane. Such out-of-focus fluorescence can result in a blurry image. Confocal microscopy overcomes this problem by using lasers and a special microscope configuration that minimizes the exposed area to a point while blocking out-of-focus fluorescence. To obtain an image, the laser rapidly scans the specimen, and the data from all the points are collected and assembled

into an image. Figure 19.92d illustrates the difference in the images obtained by conventional fluorescence microscopy and by scanning confocal microscopy.

Co-localization between two proteins can be detected by FRET

When several different cellular components are detected by distinct dyes, antibodies, or tags, we can determine if all of these components are found in the same location, a situation referred to as **co-localization**. It is important to note, however, that co-localization does not necessarily mean that these components physically interact—for example, that they are part of the same complex. FRET, which we described in Section 19.14, can also be applied to determine the distance between molecules within the cell. To do so, two proteins are expressed as a fusion to a FRET-compatible pair of fluorescent proteins, such as YFP and CFP. If the two proteins are close to each other (namely 10–100 angstroms (Å)), the energy generated by the excitation of the CFP will be transferred to the YFP, which will get excited and emit light that is detectable by microscopy. If the proteins are further away from each other, no such energy transfer will occur.

It is also possible to use FRET to characterize the organization of a multiprotein complex. For example, the nuclear pore complex consists of dozens of proteins. To determine how these proteins are organized relative to each other, researchers have measured the distances between pairs of proteins and created a map of interactions among the individual components of the nuclear pore, as illustrated in Figure 19.93.

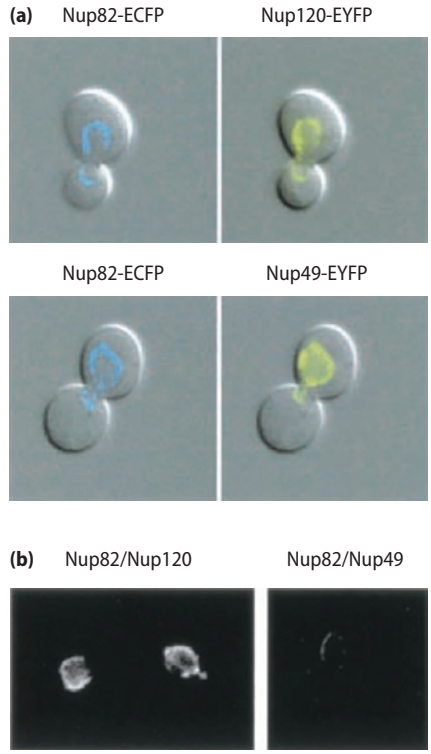


Figure 19.93 Analysis of yeast nuclear pore complex architecture by FRET. (a) Yeast cells co-expressing the nuclear pore complex protein Nup82 fused to enhanced cyan fluorescence protein (ECFP) and Nup120 fused to enhanced yellow fluorescent protein (EYFP) (top panel) or yeast cells expressing Nup82–ECFP and Nup49–EYFP (bottom panel) were excited with either the ECFP or EYFP excitation wavelengths, and the emission of either protein was recorded. The fluorescence images are superimposed on DIC images of the same cells. The images show that Nup82 co-localizes with Nup120 and Nup49 since the distributions of the proteins are similar. (b) When these cells were analyzed by FRET (by being excited with only the ECFP wavelength and the EYFP emission then being recorded), a FRET signal could be seen between Nup82 and Nup120, but not between Nup82 and Nup49, indicating that Nup120 is relatively close to Nup82 in the nuclear pore complex, but Nup49 is not.

Adapted from Damelin and Silver (2002). *In situ* analysis of spatial relationships between proteins of the nuclear pore complex. *Biophysical Journal* **83**: 3626–3636.

The movement of molecules can be followed using photobleaching or photoactivatable/photoswitchable GFP

Researchers can follow the movement of proteins and other cellular structures using live cell imaging. In the case of protein movement, the protein of interest is often tagged with a fluorescent protein, such as GFP, which can be visualized by microscopy. However, when studying an abundant protein that uniformly fills a particular cellular compartment, it is difficult to determine whether the protein moves within that compartment. Such information can be useful because lack of movement can suggest that the protein is tethered to an immobile structure. Several techniques that take advantage of properties of fluorescent proteins (or other fluorophores) can be used to examine movement within live cells.

Photobleaching occurs when a fluorophore is exposed to light of sufficiently high intensity that it destroys its ability to fluoresce. This can be a problem when taking repeated images of a cell, such as when acquiring time-lapse images, since the resulting photobleaching causes the fluorescence of a fluorescent protein to diminish over time. However, this irreversible loss of fluorescence can be exploited to assess the movement or diffusion of the protein within a cellular compartment. In this approach, called **fluorescence recovery after photobleaching (FRAP)**, a small region of interest (smaller than the total area occupied by the fluorescent protein) is illuminated repeatedly over a short period of time with high-intensity light to bleach the fluorescent protein in that region, as shown in Figure 19.94a. If the protein is mobile within the compartment, the photobleached protein will diffuse away from the exposed area and be replaced by undamaged fluorescent protein molecules, resulting in the recovery of fluorescence signal in that illuminated zone. However, if the fluorescent protein is immobile within that cellular compartment, it will neither leave the bleached area nor be replaced by unaffected protein. Consequently, the

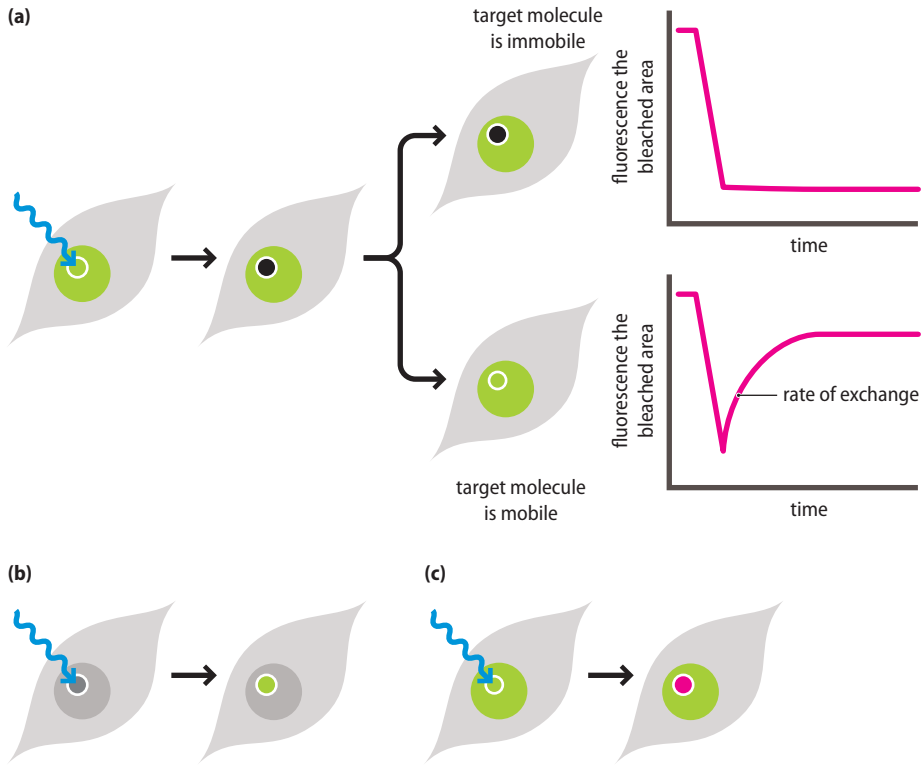


Figure 19.94 Following protein movement in live cells with FRAP and photoactivatable or photoswitchable proteins. (a) FRAP (fluorescence recovery after photobleaching). The cells in the diagram are expressing a fluorescent protein (for example, GFP) in the area indicated in green. A small area (circled in white) is rapidly bleached such that the fluorescent proteins in this area lose the ability to fluoresce (shown as a black circle). The recovery of fluorescence in this region is monitored over time. If the fluorescent proteins are immobile (that is, fixed to their location, and not free to diffuse), there will be no fluorescence recovery in the bleached area. If, however, the fluorescent proteins are mobile, the fluorescence in the bleached area will recover and the rate of recovery will be indicative of the protein's mobility. (b) PA-GFP (photoactivatable-GFP). In the example shown, a protein fused to PA-GFP localizes to a particular cellular compartment (shown in dark gray) but is not visible. Following exposure to the appropriate wavelength, the PA-GFP in the exposed region becomes fluorescent (in green). (c) Photoswitchable proteins. In the example shown, the photoswitchable GFP fills a particular cellular compartment and is visible by fluorescence microscopy (shown in green). Exposure to the appropriate wavelength will cause the protein in the exposed region to switch from emitting light in one wavelength to another (in this case, red).

bleached area will remain dark. By measuring the rate of recovery of fluorescent signal, we can estimate the rate of protein movement or diffusion in the cell.

A complementary method is **fluorescence loss in photobleaching (FLIP)** where a small region is repeatedly bleached. Unlike the FRAP method, where the photobleaching period is brief, the photobleaching in FLIP occurs over the entire course of the experiment. This method is useful, for example, when a fluorescent protein occupies two distinct cellular compartments and the investigator wants to determine if the proteins can exchange freely between the two compartments. If there is an exchange, repeated photobleaching of an area within one pool will lead to loss of fluorescence in the other pool. If, however, there is no exchange, the protein in the compartment that is not being photobleached will retain its fluorescence.

While the FRAP method can be used to determine whether a protein is mobile, it does not allow the investigator to follow the protein as it moves. To do so, one can use photoactivatable or photoswitchable fluorescence proteins. A photoactivatable protein, as its name implies, is a protein that fluoresces only after being activated by exposure to an appropriate wavelength, as depicted in Figure 19.94b. One of the commonly used photoactivatable proteins is photoactivatable GFP (PA-GFP), a derivative of GFP that normally has very low fluorescence when exposed to light at 450–550 nm. However, after exposure to the activating wavelength, PA-GFP undergoes a conformational change that causes its fluorescence to increase 100-fold when illuminated with light at a wavelength of 504 nm.

PA-GFP can facilitate the elucidation of protein dynamics by being fused to the protein of interest. Exposing a small region to the activating wavelength will cause a subpopulation of the PA-GFP-fusion proteins to fluoresce, allowing the investigator to follow their movement in a cellular compartment that may contain many additional such proteins. In addition to protein movement, PA-GFP also allows

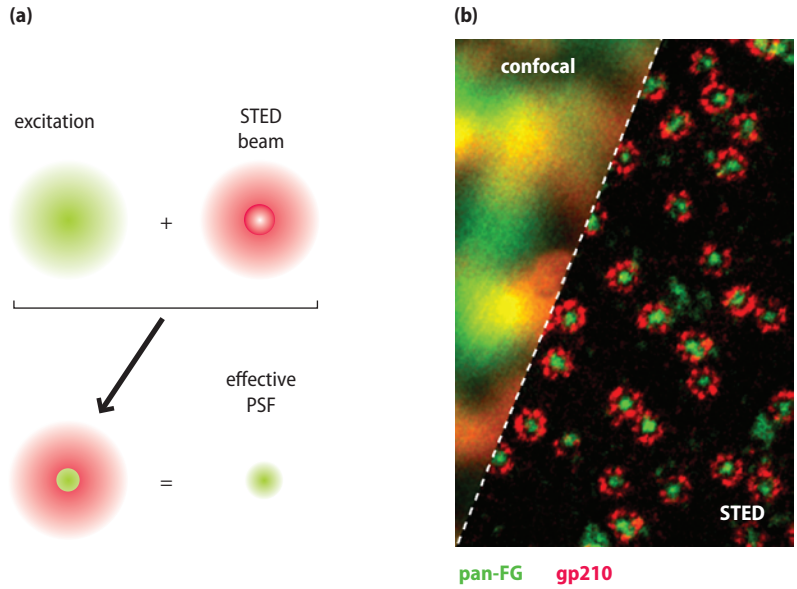


Figure 19.95 Super-resolution microscopy using STED. (a) STED limits the point spread function by using two beams. The excitation beam is shown in green, and the depletion beam is shown in red. The two beams illuminate the specimen in rapid succession, resulting in the point spread function being reduced in size. (b) Confocal fluorescence microscopy versus STED showing a side-by-side comparison of the gp210 protein subunit (red) and pan-FG (green) of the nuclear pore complex of a *Xenopus laevis* epithelial cell line. Note the much higher resolution (that is, the level of visible detail) revealed by STED.
(b) Reproduced with kind permission from Stefan W. Hell.

the measurement of protein turnover; with a GFP–fusion protein, it is not possible to observe protein turnover as new protein is being continuously synthesized. However, activation of PA-GFP distinguishes this subpopulation of fusion proteins from proteins that are synthesized after activation has taken place.

A disadvantage of PA-GFP and other photoactivatable fluorophores is that the molecule is not visible prior to activation, and thus it is sometimes difficult to identify the cellular region where the protein of interest initially resides. Photoswitchable proteins, which switch from fluorescing at one wavelength to another after being exposed to the appropriate wavelength, solve this problem (see Figure 19.94c). A commonly used, naturally occurring photoswitchable protein is called Dendra, after the soft coral *Dendronephthya* from which it was first isolated. One can use photoswitchable proteins fused to a protein of interest to identify the location of the fusion protein in the cell and, then upon photoswitching, a subpopulation that can be followed over time, as described for PA-GFP.

Super-resolution microscopy overcomes the resolution limits of light microscopy

The resolution limit of conventional fluorescence microscopy discussed earlier in this section is around 200 nm, meaning that two objects that are closer than 200 nm will appear as one. This is due to the diffraction of light creating an image that is larger than the object itself, a phenomenon known as the point spread function. Although confocal fluorescence microscopy significantly improved imaging capabilities (as Figure 19.92d depicts), the overall resolution is still limited. Super-resolution microscopy overcomes this limitation.

There are several approaches for obtaining super-resolution with biological specimens, which reach resolutions of just tens of nanometers or below (hence "super-resolution"). Some methodologies minimize the point spread function, while others use techniques that activate one fluorescent molecule at a time, allowing the location of the excited fluorescent molecule to be pinpointed. We will next discuss super-resolution methods that utilize these approaches.

The **stimulated emission depletion (STED)** method was developed to minimize the point spread function. This approach utilizes a photoactivable

fluorophore that can be excited and then turned off. The sample is illuminated rapidly by two successive beams—one that excites the fluorophore, and one doughnut-shaped beam at the so-called depletion wavelength that quenches it (shown as a STED beam in Figure 19.95a). Consequently, the molecules that are in the center of the STED beam remain excited, while molecules further away from the center return to the non-excited state, resulting in a more confined point spread function. The beams scan the specimen to generate a high-resolution image, as shown in Figure 19.95b.

A different approach for achieving super-resolution relies on the localization of a limited number of fluorescent molecules. This can be achieved by limited photoactivation of a fluorophore in photoactivated localization microscopy (PALM) or through the use of fluorophores that intrinsically switch between fluorescent and dark states (intrinsic stochastic photoblinking) in stochastic optical reconstruction microscopy (STORM).

For PALM, the specimen is imaged multiple times using a very weak beam, with only a small percentage of photoactivatable fluorophores being excited each time,

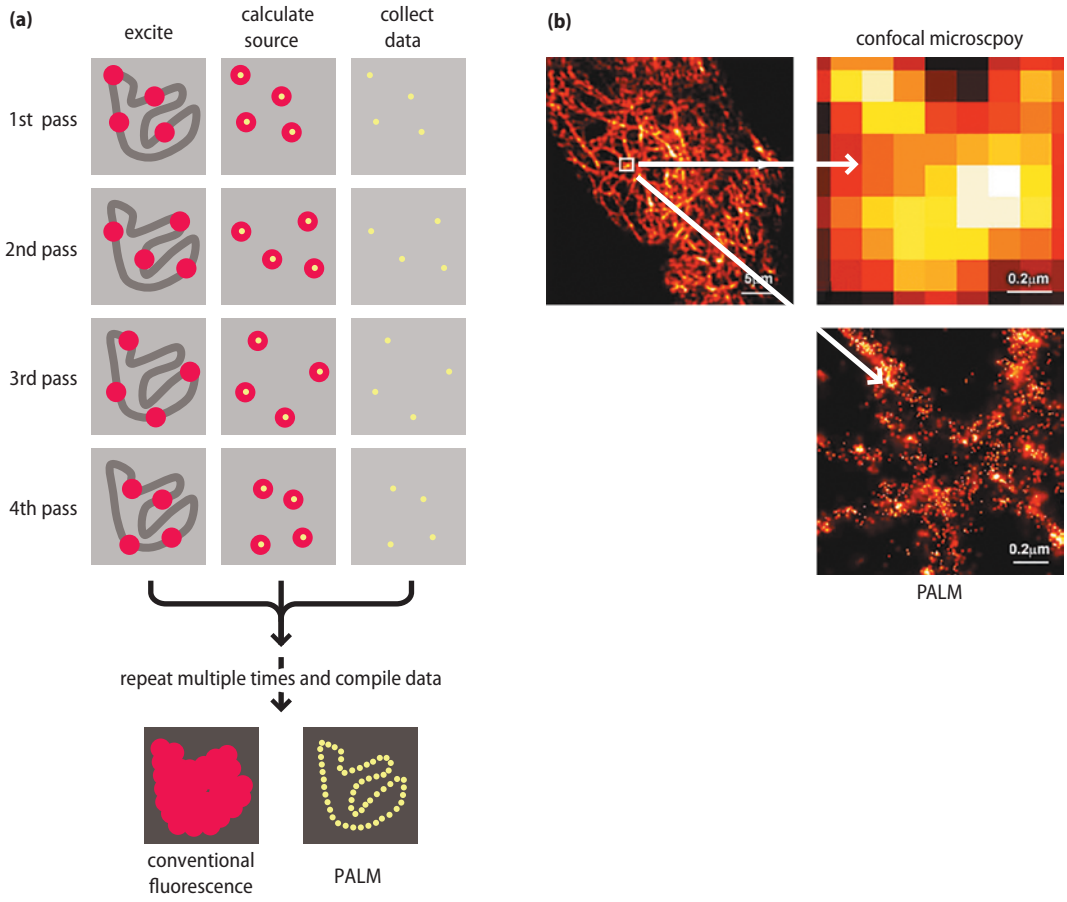


Figure 19.96 Super-resolution microscopy using PALM. (a) In the illustration shown, a cellular structure (dark gray) is expressing a photoactivatable fluorescent protein throughout its surface. The structure is imaged multiple times; only a small number of fluorescent molecules are excited each time (shown in red). The location of the fluorescent molecule within the signal is calculated and registered as a point (shown in yellow). This process is repeated multiple times, after which the raw fluorescence data and the calculated locations of the fluorescent molecules from each image are combined. While conventional or confocal fluorescence imaging results in a blurry image of the structure, the image based on the calculated locations of the fluorescing molecules gives a more precise depiction of the imaged structure. (b) Fluorescently labeled cytokeratin (a cytoplasmic intermediate filament) was imaged by confocal fluorescence microscopy (left panel). The region shown in the white box was magnified; the top panel shows the image obtained with confocal fluorescence microscopy, while the bottom panel shows the image of the same region obtained with PALM.

as depicted in Figure 19.96a. The fluorescent signal obtained from each molecule will be larger than the molecule itself, due to the point spread function. However, since each signal originates from a single molecule, it is possible to mathematically determine where within the signal the fluorescent molecule is localized. Thus, each time the specimen is imaged, the excited molecules are localized; they are then photobleached to prevent the same molecules from being resampled. To get the complete image, this process is repeated multiple times until enough data are collected. The locations of the fluorescent molecules from all the images are combined to yield a super-resolution image, as shown in Figure 19.96b.

For STORM, the molecules of interest are labeled with fluorophores that stochastically switch between fluorescent "on" and "off" states. Again, the locations of individual molecules in multiple images are combined to give the complete super-resolution image.

Electron microscopy produces images at very high resolution

The resolution of light microscopy images is limited by the wavelength of light. Although high-resolution microscopy can improve imaging resolution, a much higher resolution can be obtained using electrons, rather than light. Transmission electron microscopy (TEM) operates on the same basic principles as light microscopy, but electrons, rather than light, pass through the sample. Cells are fixed and stained with chemicals that bind different types of structures (for example, membranes), and the sample is cut into thin slices to allow the electrons to pass through. Objects whose dimensions are in the order of a few angstroms (10^{-10} m) can be detected, such that details can be visualized at near atomic levels.

In a second type of electron microscopy, **scanning electron microscopy (SEM)**, the scattering of an electron beam from a surface is monitored. The advantage of this method is the large depth of field, which allows much of the sample to be in focus at one time. This method mostly reveals the shape of the object of interest. Yeast cells imaged by both types of electron microscopy are shown in Figure 19.97. While electron microscopy can reveal structures not visible by any other imaging methodology, it is limited to fixed samples. So fluorescence microscopy remains the best approach for imaging live samples, despite its lower resolution.

Electron tomography is a method for generating a three-dimensional image using electron microscopy. The images for electron tomography are collected, while tilting the specimen by small increments around an axis. The images are then aligned to give a three-dimensional view of the object. This method is well suited for studying cellular structures, as illustrated in Figure 19.98.

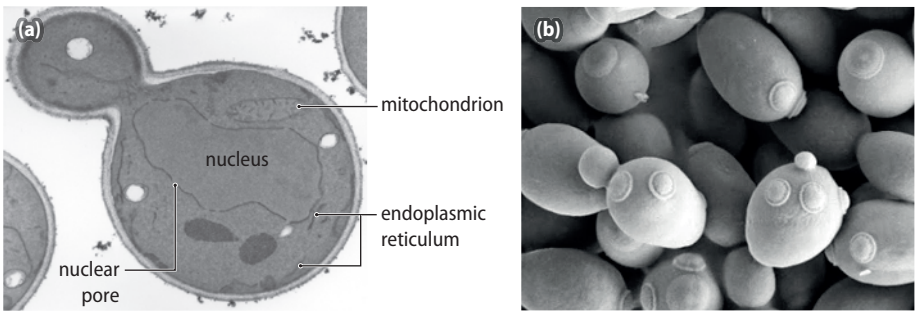


Figure 19.97 Electron microscopy. Budding yeast were examined by (a) TEM and (b) SEM. Note that TEM reveals intracellular structures, while SEM is a powerful method for examining shape.

This is free to use with attribution: https://commons.wikimedia.org/wiki/File:Saccharomyces_cerevisiae_SEM.jpg.

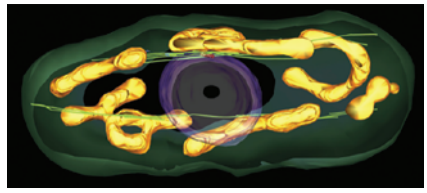


Figure 19.98 Electron tomography. An electron tomograph of a fission yeast cell, in which the three-dimensional structure of mitochondria was reconstructed (in yellow). The image also shows microtubules in light green, and the reconstruction reveals that mitochondria align with microtubules. The cell surface is shown in dark green, and the nucleus in purple.

Adapted from Hoog et al. (2007) Organisation of interphase microtubules in fission yeast analyzed by electron tomography. *Developmental Cell* **12**:349–361.

Atomic force microscopy can reveal the contour of a cell surface or molecule

Atomic force microscopy (AFM) is a method designed to characterize and analyze the surface area of a specimen at very high resolution. Figure 19.99a shows how AFM utilizes a fine tip, situated at the end of a flexible cantilever, which tracks very closely to the specimen's surface; the tip rises and falls as the cantilever moves across the specimen's surface. The surface either attracts or repels the tip due to various forces acting between the surface atoms and those of the tip, and changes in the tip's position are recorded by a laser that is reflected off the cantilever. The result is a topological representation of the specimen, which not only provides a detailed image of the specimen's shape, but also allows precise measurements of the specimen's dimensions. AFM is used to analyze complex biological molecules and structures, such as proteins bound to fragments of DNA, the topography of membranes and membrane-bound proteins, and the structure of mitotic chromosomes, as shown in Figure 19.99b.

19.17 MOLECULAR STRUCTURE DETERMINATION

Much insight into biological processes has come from information on the three-dimensional structures of biological molecules. Nearly all of the macromolecular structures depicted in this book were determined by a method known as **x-ray crystallography**, which can be used to define the structure of a molecule by

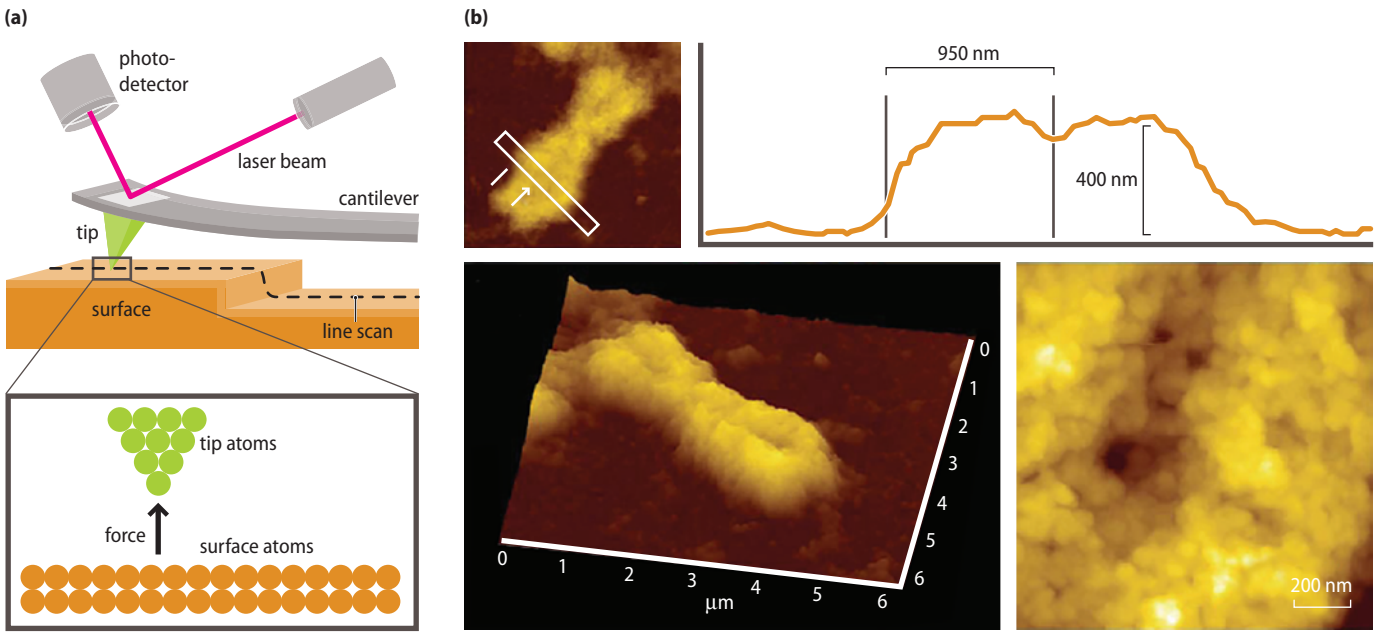


Figure 19.99 AFM. (a) General principles of AFM. The tip, which is attached to a cantilever, is moved across the surface of a sample while the laser is used to measure the relative displacement of the tip from the surface. See text for details. (b) Structure of a metaphase chromosome as determined by AFM. The chromosomes were isolated from the human cell line BALL-1. The profile (top right) is produced by calculating the average of the height in the region indicated by the box in the micrograph at left. The bottom left panel shows a three-dimensional representation of the chromosome, while the bottom right panel is a closer view of a part of the chromosome, which shows an aggregation of globular or fibrous structures.

Adapted from Ushiki and Hoshi (2008). Atomic force microscopy for imaging human metaphase chromosomes. *Chromosome research* **16**: 383–396.

examining the way in which a crystal of the molecule of interest scatters x-rays. As compared with other techniques, x-ray crystallography can produce the most accurate and detailed model of a macromolecule. Other approaches, such as nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy, can provide complementary information on protein structure and dynamics, as well as on large macromolecular assemblies. They are also used when it is not possible to obtain a crystal for analysis by x-ray crystallography.

A macromolecular structure can be determined from the diffraction of x-rays

X-ray crystallography is used to determine the three-dimensional structure of macromolecules at a level of detail that includes the position of each atom. This methodology takes advantage of the fact that the way in which molecules scatter x-rays is determined by their shape. The wavelength of x-rays used in the laboratory is around 1–2 Å, making it most appropriate for visualizing distances equivalent to the length of a covalent bond (1.2–1.5 Å). There are fundamental technical barriers to constructing an x-ray microscope that could produce a magnified image of a molecule using x-rays as the "light," so the technique of x-ray crystallography was developed instead.

Instead of scattering x-ray waves from a single molecule, x-ray crystallography uses scattering from a crystal, which consists of molecules lined up in a repeating three-dimensional array, as depicted in Figure 19.100. Purified proteins and nucleic acids can be induced to form a crystal by adjusting solution conditions—for example, by adding certain salts or organic solvents that reduce protein solubility—potentially causing the macromolecule to fall out of solution and form crystals. A trial-and-error approach is used to screen hundreds of different conditions that can potentially induce a molecule to crystallize.

When a crystal is placed in an x-ray beam, the x-ray waves are scattered by the electrons of each atom in the molecule, as shown in Figure 19.101a. We refer to this scattering as **diffraction**. Some of the scattered waves add together to produce a wave of high amplitude, whereas other waves interfere with one another and produce a wave of low amplitude. As a result, the diffraction that is recorded on an x-ray detector appears as a pattern of light and dark spots, with the intensity of each spot determined by the three-dimensional structure of the crystallized molecule. It is from many pictures like this, taken from different angles, that the information about the molecular structure is extracted.

Using sophisticated calculations, it is possible to use the intensity of each spot in the diffraction pattern to deduce the three-dimensional structure of the protein or nucleic acid molecule that has been crystallized. Additional information about the scattered waves—referred to as the phase—is needed to perform this calculation. This phase information can be obtained by a number of methods that are beyond the scope of this book. Once the amplitudes and phases of all the scattered waves are known, it is possible to mathematically reconstruct the shape of the molecule that has been crystallized.

The result of experiments to record x-ray diffraction intensities and obtain phase information is an experimental **electron density map**, which shows the outline of all the visible atoms; an example is shown in Figure 19.101b. For nearly all protein and nucleic acid structures, the hydrogen atoms cannot be visualized, and crystal structures in the PDB (Protein Data Bank) (and throughout this book) typically lack hydrogen atoms. A model of the protein is then constructed to best fit the electron density map.

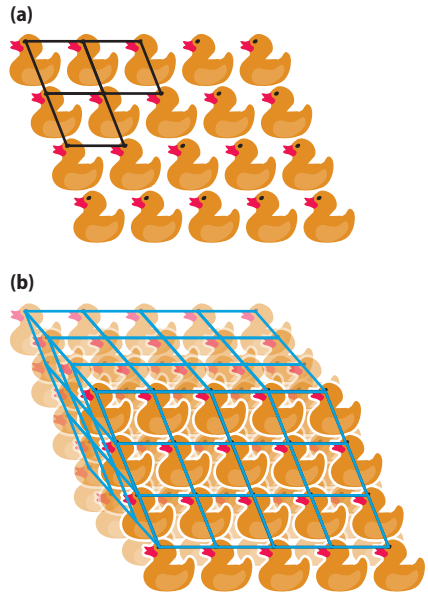


Figure 19.100 A crystal lattice. (a) A two-dimensional lattice of ducks. The parallelogram is the fundamental repeat unit of the crystal, called the unit cell (shown are three repeats of the unit). The lattice is generated by adding many identical copies of the unit cell, as indicated, along each of the two unit-cell axes. Note that each unit cell contains a complete duck, although it is made up of bits and pieces of several intact ducks. (b) A three-dimensional crystal lattice contains many repeats of the unit cell, replicated along the three unit-cell axes. (The use of ducks to illustrate crystal lattices is an established, albeit obscure, tradition in protein crystallography.)

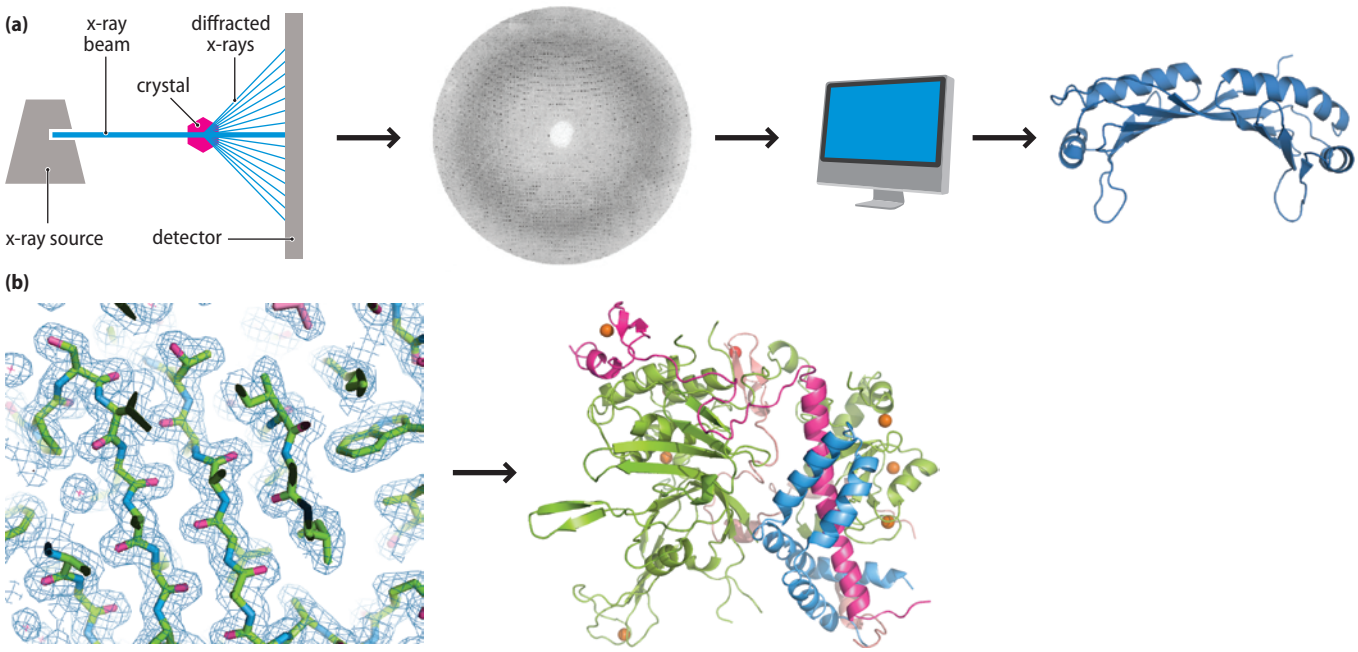


Figure 19.101 X-ray crystal structure determination. (a) An x-ray source is used to irradiate a crystal with x-rays, and the scattered x-rays are recorded on a detector. The intensity of each individual spot contains information about the structure of the molecule that has been crystallized. These intensities, together with additional information about the phase of the scattered wave, can be used in computer programs to calculate an electron density map, which is then used to construct a model of the macromolecules. The protein structure shown is a ribbon depiction of the TATA-binding protein (TBP). (b) An electron density map (blue lattice) of Ubp8, a subunit of the yeast SAGA deubiquitinating module. The molecular structure that was constructed to fit the electron density map is shown in green (carbon), blue (nitrogen), and red (oxygen). The ribbon model on the right shows four proteins in the complex: Ubp8 (green), Sgf11 (pink), Sus1 (blue), and Sgf73 (salmon); orange spheres are bound zinc atoms. (PDB codes 1TBP, 3MHH.)

Figure 19.101b courtesy of Cynthia Wolberger.

You can find the Protein Data Bank (PDB) at http://www.wwpdb.org.

The accuracy of the model will depend on the resolution of the data, which is a function of how far from the center of the image diffraction spots can be recorded. A structure determined at 1.8 Å resolution shows the positions of atoms with high accuracy, with all side chains clearly defined, whereas the side chains in a structure determined at 3.5 Å resolution are not well defined. Experimental approach 2.1 describes the application of x-ray crystallography to determine the structure of the DNA double helix. It is important to note that portions of a macromolecule that do not adopt a fixed conformation—perhaps a flexible loop or a small mobile domain—do not appear in the electron density map, irrespective of resolution. These portions of the molecule or complex are thus not present in the resulting structure.

The model for a protein (or nucleic acid) and its bound ligands is stored in a data file as a list of atoms, along with their corresponding positions in space, which are given by *x*, *y*, and *z* coordinates. A single macromolecular structure may have several thousand atoms. The coordinates for protein and nucleic acid structures, including those depicted throughout this book, can be obtained from the PDB website and displayed using a personal computer.

Cryo-electron microscopy can be used to image large macromolecular assemblies

Although x-ray crystallography can provide very detailed and accurate structures, the need to first obtain a crystal of the molecule or complex of interest is a particular barrier to structural studies of very large macromolecular complexes that are

typically difficult or impossible to crystallize. Recent advances in the technique of cryo-electron microscopy have made it possible to determine structures of large complexes, in many cases to a resolution that rivals that of x-ray crystallography. Just as a light microscope works by focusing light scattered by an object to form a magnified image, an electron microscope focuses electrons that have been scattered by a sample, forming a magnified image. The wavelength of electrons is far shorter than even that of x-rays, so electron beams can be used to image molecules at atomic resolution. Since samples in an electron microscope must be imaged in a vacuum, the sample must be preserved either by coating it with a metallic "stain" or freezing the sample in a very thin layer of ice. The highest resolution structures are determined from samples preserved in ice—hence the term **cryo-electron microscopy**, commonly referred to as cryoEM. In contrast with x-ray crystallography, cryoEM is most amenable to determining structures of large complexes, typically larger than 100–200 kDa.

To determine a structure by cryoEM, a homogeneous preparation of the complex of interest is laid down on a special grid and flash-frozen. Electron micrographs are then recorded, yielding images of the particles that are randomly oriented on the grid, as illustrated in Figure 19.102a. The image of each individual particle represents a view down a particular axis—a projection of the structure along that axis. By combining thousands of projections of the randomly oriented macromolecule, it is possible to reconstruct an image of the particle in three dimensions (see Figure 19.102b). Recent technological advances in the detectors used to record electron micrographs, along with the development of powerful new computer algorithms, have made it possible in some cases to obtain an electron density map from which an atomic resolution model can be constructed (see Figure 19.103b).

NMR spectroscopy gives structural information on macromolecules in solution

Whereas x-ray crystallography and cryoEM provide snapshots of macromolecules, they cannot be used to study molecules free in solution. **Nuclear magnetic resonance spectroscopy**, commonly referred to as **NMR**, can be used to determine structures of macromolecules free in solution. NMR is also a powerful tool for studying structural fluctuations—known as protein dynamics—and

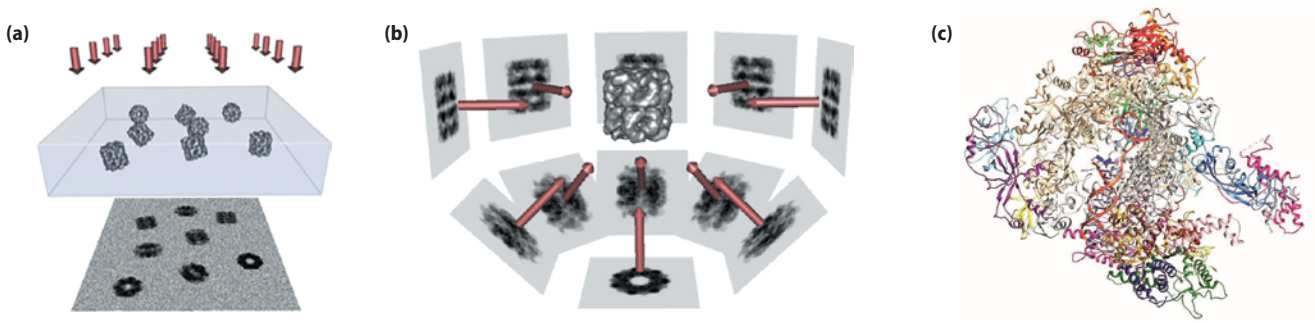


Figure 19.102 Cryo-electron microscopy: single molecule reconstruction. (a) The arrows symbolize the electron beam impinging on the specimen, which contains many randomly oriented molecules embedded in a thin layer of ice. The resulting images show the different projections of the particles. (b) This image illustrates how the different projections correspond to different views of the particles whose structure is being determined. (c) Structure of RNA polymerase III bound to a DNA template determined by cryoEM. The surface shows the electron density map that was used to construct a model (PDB code 5FJ8).
With permission from Greg Pintilie, Massachusetts Institute of Technology, Cambridge, MA, USA.

the interactions between macromolecules in solution. NMR spectroscopy can therefore provide information on proteins that complements the static picture of a protein structure provided by x-ray crystallography or cryoEM. NMR is best suited to studying relatively small proteins, although recent advances have made it possible to extend the technique to the study of proteins containing hundreds of amino acids.

NMR takes advantage of interactions between atomic nuclei or between nuclei and their environments. Nuclei with an odd number of nuclear particles—protons plus neutrons—have a property known as spin that causes them to behave like small magnets. Hydrogen atoms, ^{1}H, have spin, as do the low-abundance ^{13}C and ^{15}N isotopes of carbon and nitrogen. In absence of magnetic fields, nuclei with spin have no preferred orientation. When placed in a magnetic field, however, these spins behave like small gyroscopes, with their axes precessing around the direction of the magnetic field. The frequency with which each spin precesses is determined by the type of atom, as well as its chemical environment. Spins tend to align with the magnetic field, adding up to a quantity called magnetization that can be manipulated and detected.

If a radiofrequency electromagnetic pulse is applied in a direction perpendicular to the magnetic field, the magnetization due to the spinning nuclei can be realigned, so it becomes perpendicular to the magnetic field. Once the pulse is removed, the magnetization changes in a way that induces a current that can be detected, digitized, and analyzed as a spectrum featuring various NMR signals. The frequencies at which the different signals appear are referred to as chemical shifts and reflect the chemical environments of the nuclei they represent. The nuclei also couple with each other through space or through chemical bonds, which produces further effects on the NMR spectrum that can be used to characterize the molecule.

By manipulating the magnetic field and subjecting the sample to different types of radiofrequency pulses, it is possible to record a large variety of NMR spectra, which report on interatomic distances, bond orientations, dihedral angles, or fluctuations of these parameters. This information can be combined to determine the three-dimensional structure of the protein, as well as to provide information on molecular motions and conformational changes. An example of the spectral differences that reflect conformational differences between a phosphorylated and an unphosphorylated protein is shown in Figure 19.103.

In a simple, yet powerful, application, the position and shape of NMR signals can be used to study how one protein interacts with other proteins or with small molecules. Since chemical shifts depend upon the environment of the nuclei, the way in which chemical shifts of atoms that are in contact with binding partners change can be used to map the binding sites. It is possible to monitor the chemical shift changes of just one of the binding partners by expressing the protein in a medium enriched in ^{13}C and ^{15}N. This selective labeling makes it possible to monitor the chemical shift perturbations in the isotopically labeled protein as the unlabeled binding partner is added to the NMR sample. In some cases, forming a complex with a binding partner may induce a flexible region of a protein to form a defined structure. These newly structured residues add new features to the NMR spectrum that can be used to identify the residues and determine the structure they adopt.

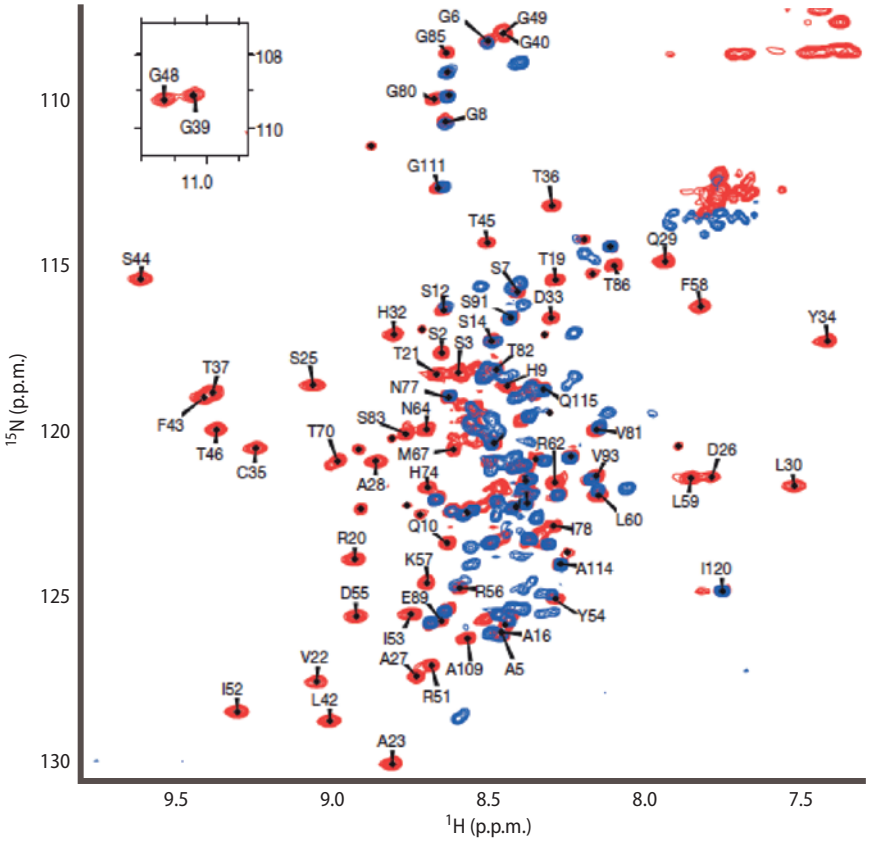


Figure 19.103 Using NMR to monitor conformational changes in a protein. Shown are NMR spectra called heteronuclear single quantum correlation (HSQC) spectra, which monitor the correlation between amide nitrogens (the ^{15}N isotope) and hydrogens (^{1}H). The red and blue peaks show the HSQC spectrum of a protein when it is phosphorylated (red) or unphosphorylated (blue). The significant difference in the red and blue spectra indicates that part of the protein significantly changes conformation when it is phosphorylated.
Bah et al., *Nature*, 2015; **519**: 106–109.

SUMMARY

Advances in our understanding of molecular biology and genomics rely on ongoing experimentation, drawing on the kinds of tools and techniques we have introduced in this chapter. It is not only our understanding of molecular biology as a discipline that continues to move forward, but also the techniques used to probe genome function. As existing techniques become enhanced and new methodologies become available, new and exciting approaches can be adopted to probe areas of interest, illuminating our understanding in new—and often unexpected—ways.

The principles of genome function presented throughout this book will, doubtlessly, continue to be refined as investigators around the world continue their research; so too will the tools and techniques used to explore these principles. It is this dynamic interplay between what we know and the experimental approaches used to further our understanding that makes molecular biology so fascinating, as we continue to uncover new questions that remain to be explored.

AN OVERVIEW OF MOLECULAR BIOLOGY METHODS

- Molecular biology methods range from genome-wide analysis to the dissection of interactions between isolated single molecules.
- A key aspect of molecular biology is the study of processes both *in vivo* and *in vitro*.
- Most molecular components and processes are studied in a relatively small range of model organisms.
- Studies of the molecular components of a wide variety of genetically tractable organisms have proved Jacques Monod's prediction: "What is true for *E. coli* is true for the elephant."

CELL CULTURE

- Both bacterial and eukaryotic cells can be grown in culture.

- Individual molecules can be purified from cultures of these relatively homogeneous populations of cells.
- ES cells and iPS cells can be grown in culture and manipulated to give rise to differentiated cells.

GENE AND GENOME MANIPULATION

- The ability to manipulate genes and genomes is central to molecular biology.
- DNA cloning, whereby specific DNA regions are isolated and propagated, is often carried out in bacteria.
- A library of clones may represent total genomic DNA, expressed genes only, or just a subset of expressed genes.
- The targeted introduction of mutations through DNA engineering (mutagenesis) allows for the generation of variant DNAs, RNAs, and proteins for study.
- Transposons and chemical mutagenesis are commonly used to introduce random mutations into the genome of an organism.
- Expression of desired genes can be silenced by siRNAs and their derivatives.
- Recombination is commonly used to introduce specific genes or mutations into the genome of an organism.
- Increasing use is being made of derivatives of CRISPR–Cas systems for genome manipulation and analysis.

THE ISOLATION AND CHARACTERIZATION OF BIOLOGICAL MOLECULES

- The isolation and characterization of molecules *in vitro* is a key method in molecular biology.
- The molecules present in a sample can be identified and quantified in a variety of ways.
- Many detection techniques are spectroscopic in nature and exploit the characteristic ways in which molecules interact with electromagnetic radiation.
- Chromatography techniques exploit the different physical characteristics of molecules to separate them from one another.
- Various approaches exist for determining the sequence (primary structure) of a biological molecule.
- Specific DNA sequences and RNA molecules can be detected using hybridization-based methods.
- Proteins can be detected by virtue of their specific interaction with antibodies.
- Genome sequence information and DNA manipulation allow for the addition of tags to specific proteins and RNA, facilitating their isolation without prior knowledge of their function.

GENOME SEQUENCING

- Advances in genome sequencing have enabled us to obtain genome sequence information with increased speed and accuracy, and at lower cost.

CHARACTERIZATION OF INTERACTIONS BETWEEN MOLECULES

- The identification of components that interact with each other has enhanced our understanding of protein, DNA, and RNA function.
- A range of biophysical techniques allows equilibrium and rate constants to be determined for the interactions between molecules.
- Advances in sequencing have facilitated the genome-wide analysis of interactions between proteins and DNA and RNA, as well as interactions between different regions of DNA and base-paired RNAs.

CELLULAR IMAGING AND MOLECULAR STRUCTURE DETERMINATION

- Advances in imaging methods, including the use of reporter proteins such as GFP, have provided important new tools for studying cells in both live and fixed samples.
- Imaging can reveal information about both molecular localization and movement within the cellular environment.
- The three-dimensional structure of molecules is most widely determined by x-ray crystallography.
- Insights into the structures of large molecular complexes are increasingly being obtained by cryo-electron microscopy.
- Valuable information about the structures of molecules in solution can be obtained by NMR spectroscopy.

ONLINE RESOURCES FOR GENOMICS AND MODEL ORGANISMS

- National Institutes for Health
  https://www.genome.gov/
  https://www.ncbi.nlm.nih.gov/Genome
- US Department of Energy Office of Science
  https://genomicscience.energy.gov/
- Major genomics centers
  J. Craig Venter Institute: https://www.jcvi.org/
  The Sanger Institute: https://www.sanger.ac.uk/
  Broad Institute of MIT and Harvard: https://www.broadinstitute.org/
  The Genome Center at Washington University: https://www.genome.wustl.edu/
  Baylor College of Medicine, Human Genome Sequencing Center: https://www.hgsc.bcm.edu/
- Microbes
  https://www.microbes.info/resources/General_Microbiology/Databases/Genetic/index.htm
- *E. coli*
  https://ecocyc.org/
- *S. cerevisiae* (budding yeast)
  https://www.pombase.org/
- *S. pombe* (fission yeast)
  https://www.genedb.org/genedb/pombe/
- *D. discoideum*
  https://dictybase.org/
- *C. elegans* (worm)
  https://wormbase.org/
- *Drosophila* (fruit fly)
  https://flybase.org/
- *Arabidopsis*
  https://www.arabidopsis.org/
- *Xenopus* (frog)
  https://www.xenbase.org/
- *D. rerio* (zebrafish)
  https://zfin.org/
- Mouse
  https://www.informatics.jax.org/
- Human
  https://www.genome.ucsc.edu/
- The Human Genome Project
  https://www.genome.gov/human-genome-project/

QUESTIONS

19.1 MODEL ORGANISMS

1. List four general properties of model organisms that have facilitated their study.

2. Different model organisms are chosen to study particular biological processes because they have distinct properties that facilitate the study. For each of the following model organisms, name one or more biological process described in the text that was characterized in that organism and explain the unique characteristics that made the organism a good choice for the study.

a. *E. coli*
b. *B. subtilis*
c. *T. thermophila*
d. *C. elegans*
e. *M. musculus*

19.2 CULTURED CELLS AND VIRUSES

1. What is an advantage in using cell culture for the study of proteins and nucleic acids?

2. Describe one advantage and one disadvantage of using eukaryotic cell lines for study.

19.3 AMPLIFICATION OF DNA AND RNA SEQUENCES

1. At the end of four cycles of PCR, you have 48 double-stranded DNA molecules containing the target sequence. How many double-stranded DNA molecules containing this sequence did you have before starting the PCR reaction?

a. 1
b. 3
c. 6

d. 12

2. Which of the following sequences would you find during first-strand cDNA synthesis, and what do the strands correspond to?
a. 5′ AGTCGATGCTAGT 3′
b. 5′ AGTCGATGCTAGT 3′
3′ TCAGCTACGATCA 5′
c. 5′ AGTCGATGCTAGT 3′
3′ UCAGCUACGAUCA 5′

19.4 DNA CLONING

1. DNA cloning has been absolutely essential in the study of genes and their function.
a. What is DNA cloning?
b. Why is DNA cloning essential to the study of genes?
c. What is a DNA vector?
d. In addition to an origin of replication, what are some other useful features of DNA vectors?

2. How does Gibson cloning differ from cloning using restriction enzymes?

3. Site-directed mutagenesis is a technique that has significantly enhanced the ability to study gene function.
a. What is site-directed mutagenesis?
b. Why is this technique so significant in the study of gene function?

4. How is a DNA library generated and what are some of the uses of these libraries?

Challenge question

5. Which of the following molecules will ligate to this sticky end:
5′ C-T-G-C-A
3′ G
a. 5′ C-T-G-C-A
3′ G
b. 5′ G
3′ C-T-G-C-A
c. 3′ G
5′ A-C-G-T-C
d. 5′ A-C-G-T-C
3′ G

19.5 UNDIRECTED GENOME MANIPULATION

1. Compare and contrast the use of chemicals and transposons in mutating genomes.

19.6 DIRECTED GENOME MANIPULATION

1. Generating a knockout mouse.
a. What is a knockout mouse?
b. The first step in the process is to create a targeting vector containing a portion of the gene of interest, the *Neo*^{R} gene, and the *HSV-tk* gene. What is the purpose of each of these components?
c. The original ES cells for recombination are taken from a gray mouse; the cells that successfully underwent homologous recombination are injected into a white mouse blastocyst, and the blastocyst is implanted in a black mouse surrogate parent. Why are the three different colored mice necessary? Explain what possible colored pups could arise, which is desirable, and why.
d. In the final step, the chimeric mice are mated with a white mouse, and the offspring are analyzed. Why is this step necessary? What are the possible outcomes of this cross?

2. Describe four uses of CRISPR–Cas systems in genome manipulation.

19.7 DETECTION OF BIOLOGICAL MOLECULES

1. Which three amino acids absorb UV light and thus allow proteins to be assayed spectroscopically?
a. Glycine, tyrosine, proline
b. Proline, tryptophan, lysine
c. Tyrosine, proline, phenyalanine
d. Tryptophan, tyrosine, phenylalanine

2. Describe how you would determine whether your DNA or RNA sample is contaminated with proteins.

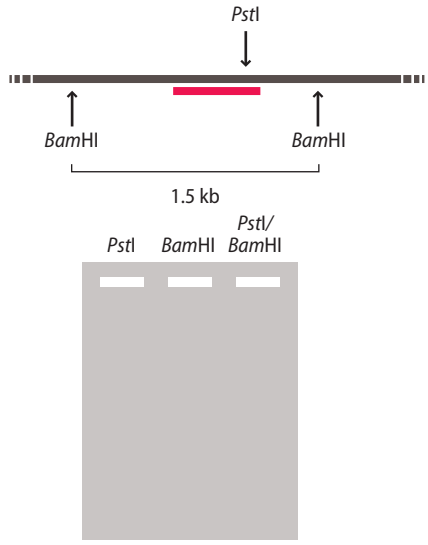
19.8 SEPARATION AND ISOLATION OF BIOLOGICAL MOLECULES

1. To be able to study the particular characteristics of a biological process, it is first necessary to isolate and purify the corresponding molecule or molecules away from other cellular components. Briefly describe the method and the general process by which one would isolate each of the following classes of cellular components:
a. Cellular organelles
b. Proteins
c. Nucleic acids

Challenge questions

2. Proteins can have domains of distinct functionality within a single polypeptide. However, in some instances, these domains can be on distinct polypeptides in a multi-subunit protein complex.
a. Describe how you would use gel electrophoresis to distinguish between a protein that has domain A and domain B on a single subunit, and a protein that has domain A and domain B on two subunits. What results would you expect?
b. After performing the gel electrophoresis experiment described in (a), you note that the bands for the two samples look identical and realize you made a mistake in preparing the gel and did not add a reducing agent in preparing the sample. How might this omission affect the results of your experiment?

3. This diagram represents a piece of DNA cloned into the unique *Bam*HI site in a 4.0 kb plasmid which does not have any other *PstI* restriction sites.



The red line indicates where a radioactively labeled probe hybridizes. Three samples of the genomic DNA are cut, one with *Pst*I, one with *Bam*HI, and one with both enzymes. The digested samples are run on

an agarose gel, blotted, and probed. Draw the results you expect to see from a Southern blot relative to a marker lane with bands at 1.0, 2.0, and 5.0 kb and explain your answer.

19.9 IDENTIFYING THE COMPOSITION OF BIOLOGICAL MOLECULES

1. You are preparing an "A" reaction for radioactive Sanger sequencing. You have added the buffer, the DNA template, the DNA polymerase, the labeled primer, and the four dNTPs to the tube. Which of the following molecules must you also add and why?

2. Histone acetyltransferases were initially discovered in an elegant experiment that detected activity of the protein in a polyacrylamide gel infused with histones (Experimental approach 4.2, Figure 1). Starting with the gel shown in that figure, design an experiment to determine the sequence of the gene encoding the protein.

19.10 OBTAINING AND ANALYZING SEQUENCES ON A GENOMIC SCALE

1. Compare and contrast classic Sanger sequencing methods to the next-generation sequencing methods that are currently available.

2. Why does obtaining a complete genomic sequence require oversequencing?

19.11 DETECTION OF SPECIFIC DNA SEQUENCES

1. Each of the following techniques can be used to examine whole genomes for insertions, deletions, and translocations. Briefly describe each method in terms of how it is unique from the other methods.

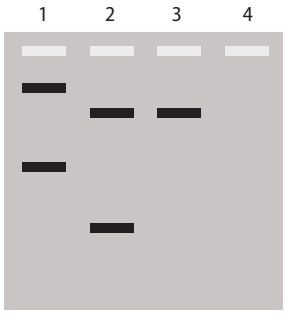
a. Karyotyping
b. FISH
c. aCGH

2. When would colony hybridization be useful?

a. To screen a genomic library
b. To test for protein production
c. To check for antibiotic resistance in a plasmid
d. To investigate RNA:RNA interactions

Challenge question

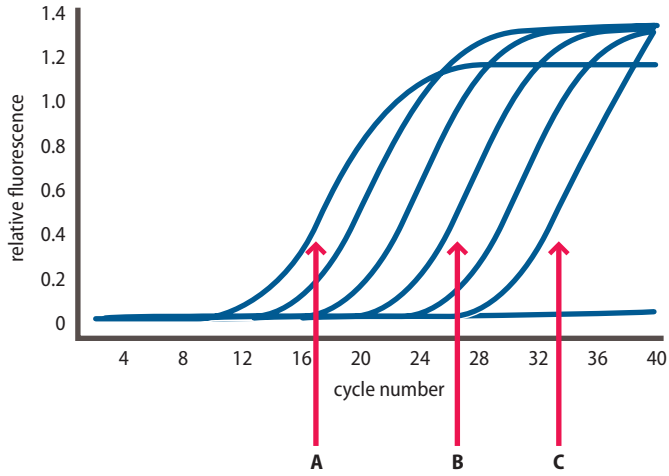
3. You have transformed your organism with a gene for GFP, which should integrate randomly into the genome. You want to work further with the most stable transformants. You extract DNA from several different transformants and perform Southern analysis with a probe for the *GFP* gene. With which of the transformed strains (1–4) would you work further and why?



19.12 DETECTION OF SPECIFIC RNA MOLECULES

1. Quantitative PCR is a way to detect relative amounts of starting material in a PCR reaction. After each cycle, the amount of PCR product is detected using fluorescence. Typical results are plotted in the figure below, with each line on the graph resulting from a single PCR reaction.

Which curve represents the sample with the most abundant starting PCR template? Explain your answer.



Challenge question

2. Analysis of the sequence for a gene of interest suggests that the four exons contained within the gene may be alternatively spliced to produce up to three alternative transcripts consisting of: (1) all four exons; (2) exons 1, 2, and 4; and (3) exons 1, 3, and 4. Researchers performed a northern blot analysis to determine if the three possible transcripts were actually produced.

a. Why did the researchers choose a northern blot?
b. To which exon should the researchers create a probe? Why?
c. Based upon the exon sizes indicated in the figure below, draw a figure that shows the expected result of the northern blot if only (1) and (3) transcripts are produced.

19.13 DETECTION OF SPECIFIC PROTEINS

1. How can one detect a specific protein if there is no antibody available for that protein?

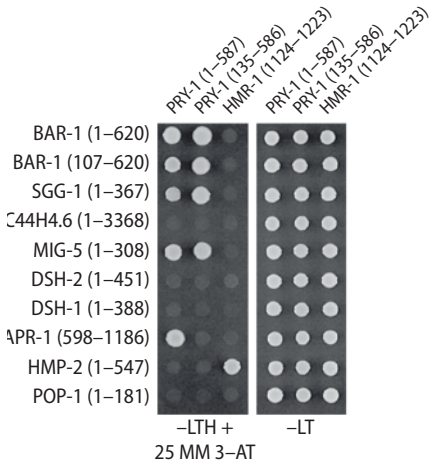
19.14 DETECTION OF INTERACTIONS BETWEEN MOLECULES

Challenge questions

1. Describe two approaches for identifying the sequences bound by a specific DNA binding protein.

2. The data in the image reproduced in the figure below show the results of a yeast two-hybrid experiment, with the parts of the proteins tested indicated with numbers. From these data, which protein binds only to **part** of one of its interactors? What does this suggest regarding how the

proteins bind to one another? Explain your answer in terms of how the yeast two-hybrid experiment works and with reference to the data shown.



19.15 GENOME-WIDE DETECTION OF INTERACTIONS BETWEEN MOLECULES

1. What is ChIP used for?
- a. To test protein:protein interactions
- b. To test DNA:RNA interactions
- c. To test DNA:protein interactions
- d. To test DNA:DNA interactions

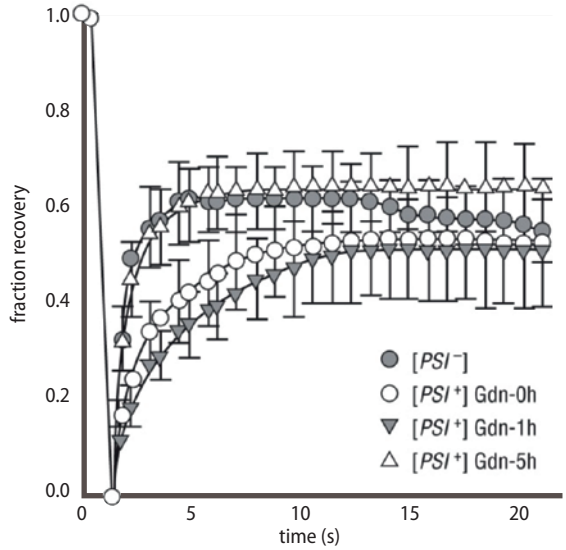
2. What can be learned from Hi-C experiments?

19.16 IMAGING CELLS AND MOLECULES

1. Figure 19.89 is an image that depicts one advantage to the use of microscopy in the study of biological processes.
- a. What does the figure represent?
- b. Why is this figure significant?
- c. Describe two additional advantages to the use of microscopy in the study of biological processes.

Challenge question

2. The graph in the figure below shows data from a FRAP experiment. Which of the proteins has the slowest movement? Explain your answer in terms of how FRAP works and with reference to the data shown.



19.17 MOLECULAR STRUCTURE DETERMINATION

1. Provide one advantage and one disadvantage of each of the following techniques for determining the three-dimensional structure of a molecule.
- a. X-ray crystallography
- b. Cryo-electron microscopy
- c. NMR