# DETERMINING A GENETIC DISTANCE

The classical method for determining the genetic distance between the loci of two allele pairs known to reside on the same homologous chromosome pair of an organism involves observing the phenotypes of the offspring of one of two particular breedings. During the course of Thomas Hunt Morgan's work on fruit flies, he happened to carry out both breedings and was rewarded not only with the first clear evidence of crossing over but also with the first unambiguous genetic-distance data. Morgan's experiments and data are used here to illustrate the procedure.

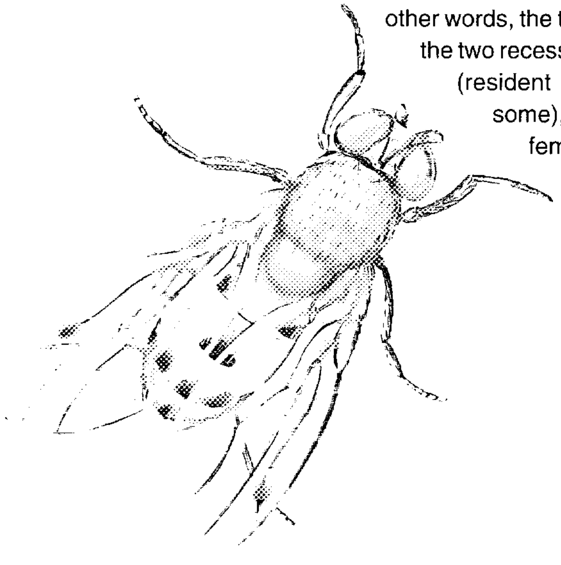
The allele pairs in question reside on one of the homologous autosome pairs of *Drosophila melanogaster.* One allele pair affects eye color: a dominant allele *A* that specifies red eye color and a recessive allele *a* that specifies purple eye color. The other allele pair affects wing length: a dominant allele *B* that specifies wild-type wings and a recessive allele *b* that specifies vestigial (very short) wings.

The participants in the first breeding are a female fruit fly that is heterozygous for both traits (and therefore has red eyes and normal wings) and a male fruit fly that is homozygous for both recessive trait variants (and therefore has purple eyes and vestigial wings). Furthermore, the female is known to be a product of the breeding *AABB* × *aabb.* Therefore the distribution of the alleles *A, a, B,* and *b* on the homologous autosome pair of the female is known: Both dominant alleles (*A* and *B*) reside on one member of the homologous autosome pair, and both recessive alleles (*a* and *b*) reside on the other member. Such an allele distribution is denoted by writing the genotype of the female as *AB/ab.* The distribution of the alleles *a, a, b,* and *b* on the homologous autosome pair of the male is also known (because the male is homozygous for both traits) and is denoted in a similar fashion as *ab/ab.* Thus the first breeding can be symbolized by

$$AB/ab \text{ female} \times ab/ab \text{ male}. \qquad (1)$$

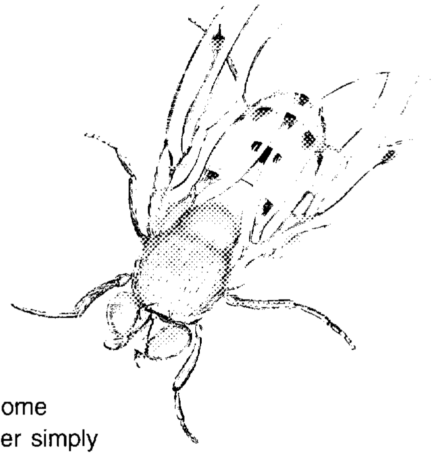
Meioses in the heterozygous female that involve no crossovers between the two loci yield two types of eggs: those possessing the chromosome with the allele combination *AB* and those possessing the chromosome with the allele combination *ab.* In other words, the two dominant alleles and the two recessive alleles remain linked (resident on the same chromosome), just as they are in the female herself. But those meioses in the female that involve a single crossover between the two loci (or any odd number of crossovers) yield in addition two other types of eggs: those possessing a chromosome with the allele combination *Ab* and those possessing a chromosome with the allele combination *aB.* In other words, a single crossover between the two loci establishes linkage between one dominant and one recessive allele. On the other hand, meioses in the doubly homozygous male, whether or not they invove crossovers between the two loci, yield sperms possessing only the allele combination *ab.* Thus the offspring of breeding 1 possess four genotypes, each corresponding to one of the four possible phenotypes:

$$AB/ab \text{ female} \times ab/ab \text{ male} \longrightarrow AB/ab + ab/ab + Ab/ab + aB/ab.$$



Morgan examined more than 2800 progeny of breeding 1 and found that 47.2 percent had red eyes and normal wings (*AB/ab*), 42.1 percent had purple eyes and vestigial wings (*ab/ab*), 5.3 percent had red eyes and vestigial wings (*Ab/ab*), and 5.4 percent had purple eyes and normal wings (*aB/ab*). All the offspring exhibiting the last two phenotypes (the combinations of one recessive trait variant and one dominant trait variant) result only from crossovers during meioses in the female parent. Thus the data indicate that the probability of new allele linkages being formed by crossing over is 0.107 = 0.053 + 0.054. That value for the so-called recombination fraction corresponds to a genetic distance of about 12 centimorgans. (The relationship between recombination fraction and genetic distance is presented in "Classical Linkage Mapping" in "Mapping the Genome.")

The participants in the other breeding that provides unambiguous recombination-fraction data are, like the participants in breeding 1, a doubly heterozygous female and a doubly homozygous-recessive male. How-



ever, the second female is known to be a product of the breeding *Ab/Ab* × *aB/aB* (rather than the breeding *AB/AB* × *ab/ab*). Therefore the distribution of alleles on her homologous autosome pair is *Ab/aB* (rather than *AB/ab*). (The difference in allele distributions of the two doubly heterozygous females is often referred to as a difference in linkage phase.) The second breeding is thus symbolized by

*Ab/aB* female × *ab/ab* male. (2)

Breeding 2 yields offspring that exhibit the same genotypes and phenotypes as breeding 1:

*Ab/aB* female × *ab/ab* male
*Ab/ab* + *aB/ab* + *AB/ab* + *ab/ab*.

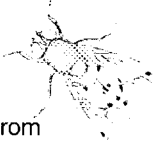
Morgan examined more than 2300 progeny of breeding 2 and found that 41.3 percent had red eyes and vestigial wings (*Ab/ab*), 45.7 percent had purple eyes and normal wings (*aB/ab*), 6.7 percent had red eyes and normal wings (*AB/ab*), and 6.3 percent had purple eyes and vestigial wings (*ab/ab*). Again, all the offspring exhibiting the last two phenotypes result only from crossovers during meioses in the female parent. Thus the data indicate that the recombination fraction for the two allele pairs is 0.130, which corresponds to a genetic distance of about 15 centimorgans.

Note that the two data sets yield different values for the same genetic distance. However, the difference between the values is within the statistical uncertainties associated with measurements of probabilistic events. Note also that the same genetic distance could in principle be determined by carrying out the reciprocal of breeding 1 or breeding 2 (that is, a breeding between a doubly heterozygous male and a doubly homozygous-recessive female). Then, the crossovers detected are those that occur during meioses in the male parent rather than in the female parent. However, for some unknown reason crossing over simply does not occur in male fruit flies. But fruit flies are exceptional in that respect, and genetic distances for other species can be determined by carrying out either breeding 1, say, or its reciprocal.

Breedings 1 and 2 are those that provide unambiguous recombination-fraction data. As an example of the ambiguities that can arise, consider the fruit-fly breeding

*AB/ab* female × *AB/ab* male. (3)

Assume first that crossing over between the two loci does not occur during meioses in the female parent. Then the offspring of breeding 3 exhibit two phenotypes: red eyes and normal wings (*AB/AB* and *AB/ab*) and purple eyes and vestigial wings (*ab/ab*). Now assume that crossing over does occur during meioses in the female parent. Then among the offspring of breeding 3 are some that exhibit the two other possible phenotypes: red eyes and vestigial wings (*Ab/ab*) and purple eyes and normal wings (*aB/ab*). All offspring that exhibit those two phenotypes result only from crossing over. However, crossing over also leads to offspring that exhibit one of the phenotypes produced in the absence of crossing over, namely, red eyes and normal wings (*Ab/AB* and *aB/AB*). In other words, whereas the offspring produced by breeding 1 or 2 can unambiguously be sorted by phenotype into two categories—those that are the result of crossovers and those that are not—the offspring resulting from breeding 3 cannot be so sorted because meioses that do and do not involve crossovers result in the doubly dominant phenotype.



The reader can accept on faith or verify personally that breedings 1 and 2 are the only breedings that provide unambiguous recombination-fraction and hence genetic-distance data. Note, in addition, that obtaining even ambiguous data requires that one parent be doubly heterozygous.

Determining a genetic distance is thus relatively easy when the breeding of the organism in question can be manipulated at will. But determining the genetic distance between the loci of two human allele pairs is much more difficult, since the breeding of humans cannot be manipulated, the genotypes and allele distributions of human parents are not always known, and human breedings generally produce so few offspring that the statistical uncertainty in the measured recombination fraction is large.

cell displays a characteristic pattern of dark and light bands when stained with an appropriate dye (see "Chromosomes: The Sites of Hereditary Information"). Because the banding pattern characteristic of a pair of homologous metaphase chromosomes varies along the length of the chromosomes, it can also be used to identify different regions of the chromosomes. The advent of chromosome banding led to recognition of the occasional occurrence of aberrant chromosomes. (The incidence of aberrant chromosomes, like the incidence of gene mutations, can be increased by exposure to x rays or other mutagenic agents.) Several types of chromosome aberrations, or rearrangements, were noted, including translocations (the exchange of chromosome regions between nonhomologous chromosomes) and inversions (the reversal of the orientation of a chromosome region).

Obviously a chromosome rearrangement can lead to changes in the complement of genes present on a chromosome or to changes in their relative locations. The gene (or genes) affected by a chromosome rearrangement (as determined from genetic data) can then be assigned a locus within the rearranged chromosome region. Although the locus so obtained is inexact, it is better than the alternative of knowing nothing at all about the locus. Knowledge of the whereabouts on a chromosome of a gene then serves to "anchor" a genetic-linkage map including that gene to the chromosome. (Recall that a linkage analysis provides only distances between genes on a chromosome; additional information is required to locate the genes relative to the chromosome itself.)

Chromosome rearrangements and gene mutations are but two examples of naturally rare phenomena that, once noted, are exploited to gain basic information about genes. Another example is the exceptional behavior of the cells that compose the salivary glands of *Drosophila* (and other insects of the order Diptera). In 1933 the American zoologist Theophilus Shickel Painter (1889–1969) and independently two German geneticists discovered that the chromosomes in those cells were microscopically visible during interphase. (Interphase chromosomes are usually not microscopically visible because they have not yet condensed in preparation for mitosis.) For some unknown reason the salivary cells of *Drosophila* undergo not a single round but many successive rounds of chromosome duplication during the S phase of interphase (see "The Eukaryotic Cell Cycle"). The numerous (on the order of a thousand) copies of each chromosome remain closely associated along their lengths, forming a fiber sufficiently thick to be microscopically visible. Because such "polytene" chromosomes are not condensed, sites of chromosome rearrangements can be pinpointed with much greater resolution.

**The Rise of Molecular Genetics**. By 1940 many genes were known to exist, and a goodly number of the known genes had been assigned to particular regions of particular chromosomes. But the gene remained an abstract concept. No one knew what genes do or even of what they are made. A speculation about what genes do had appeared as early as 1903, when the French geneticist Lucien Claude Cuénot (1866–1951) proposed that inherited coat-color differences in mice were due to the action of different genes. And in 1909 the English physician Archibald Edward Garrod (1857–1936) established that the human disease alkaptonuria was inherited as a recessive trait variant and proposed that the unmistakable symptom of the disease (urine that blackens after being excreted) was due to accumulation in the urine of a metabolic product that normally is degraded with the help of a certain enzyme. (An enzyme is a protein that catalyzes a biochemical reaction.) But Cuénot's and Garrod's proposals were regarded as mere speculation for many years. Then, in 1941, the American geneticist George Wells Beadle (1903–1989) and the American biochemist Edward Lawrie Tatum (1909–1975) clearly demonstrated the connection between the genes an organism possesses and the biochemicals it is able to synthesize.

Beadle and Tatum's work focused on the bread mold *Neurospora crassa*. Because wild-type spores of *N. crassa* can be cultured in the laboratory on a minimal growth medium (one containing only sucrose, inorganic salts, and the vitamin biotin), they reasoned that the mold must possess enzymes that help convert those simple molecules into all the other necessities of life. By exposing *N. crassa* to ultraviolet light, Beadle and Tatum produced a very few mutant spores that could not be cultured on a minimal growth medium but could be cultured on a growth medium containing a single additional nutrient (vitamin $B_6$, for example). They concluded that the x rays had caused a mutation in a gene that somehow directs the synthesis of an enzyme involved in the synthesis of the nutrient. Evidence in support of such a conclusion accumulated, and in 1948 the American geneticist and biochemist Norman Harold Horowitz (1915–) propounded the famous one gene–one enzyme hypothesis. Molecular genetics was born. Horowitz's hypothesis has since been modified to state that one gene directs the synthesis of one protein, or, more precisely, one polypeptide chain, since some proteins contain more than one polypeptide chain.

Beadle and Tatum's work on *N. crassa* demonstrated the value of studying such a simple organism. Attention soon turned to even simpler organisms—bacteria. The bacterium *Escherichia coli*, a tenant of the vertebrate gut, gained particular favor. As a result of studies begun soon after World War II by François Jacob (1920–), Joshua Lederberg (1925–), Jacques Lucien Monod (1910–1976), and Elie Leo Wollman (1917–), more is known about the genes of *E. coli*, including their regulation, than of any other living organism. Attention also focused on viruses, the simplest of all organisms, and in particular on the viruses that infect bacteria, known as bacteriophages or simply phages. (Viruses are composed of a nucleic acid core encased in a protein coat. They are not living organisms in the sense that they lack the machinery for biosynthesis. They can, however, reproduce—by usurping the biosynthetic machinery of the cells they infect—and pass their characteristics from generation to generation through the medium of genes just as cellular organisms do.) In the United States the so-called Phage Group, led by Max Delbruck (1906–1981), Alfred Day Hershey (1908–), and Salvador Edward Luria (1912–1991), aroused interest in the interaction between phages and bacteria as a model system for studying the fundamental mechanisms of heredity. Work by the Phage Group included developing quantitative methods for studying the life cycles of phages and later the discovery that phages can transfer bacterial genes from one bacterial strain to another, a process called transduction. (Transduction was to become a progenitor of recombinant-DNA technology.) The promiscuous exchange of genetic material between different strains of bacteria and between bacteria and their viruses facilitated the mapping of genes and the identification of their functions.

What genes are made of was the other big question about genes in the 1940s. In 1925 Wilson, reversing his previous stance, espoused protein as the genetic material. The idea of a proteinaceous genetic material was subsequently widely accepted for more than two decades, primarily because the nonproteinaceous component of chromosomes, DNA (deoxyribonucleic acid), was thought by chemists to have a structure that rendered it incapable of carrying any kind of message. However, in 1944 the American bacteriologists Oswald Theodore Avery (1877–1955) and his colleagues presented strong evidence that the genetic material was DNA. Their evidence was the ability of DNA extracted from dead members of a pathogenic strain of *Streptococcus pneumoniae* to impart the inherited characteristic of pathogenicity to live members of a nonpathogenic strain of the same bacterium. (We now know that the mechanism involved in the transformation from nonpathogenicity to pathogenicity is DNA recombination, of which crossing over is a specific example.) And in 1952 Hershey and another member of the Phage Group, the American geneticist Martha Chase (1927–), showed that DNA is the component of a phage that enters a bacterium and thus presumably directs the synthesis of new phages within the infected bacterium. Nevertheless, despite the accumulating evidence, DNA was not widely accepted as the genetic material.

Then in 1953 James Dewey Watson (1928–) and Francis Harry Compton Crick (1916–) proposed a structure for DNA that accounted for its ability to self-replicate and to direct the synthesis of proteins. The structure they proposed is of course the famous double helix, which, like two-ply embroidery floss, is composed of two strands coiled helically about a common axis. Each strand is a polymer of deoxyribonucleotides, and each deoxyribonucleotide contains a phosphate group, the residue of the sugar deoxyribose, and the residue of one of four nitrogenous organic bases (adenine, cytosine, guanine, and thymine). The deoxyribonucleotides are linked together in a manner such that alternating phosphate groups and sugar residues form a backbone off which the bases project. Hereditary information is encoded in the order, or sequence, of bases along the strands. The two strands are coiled about the helix axis in a manner such that the backbones form the boundaries of a space within which the bases are contained. Each base on one strand is linked by hydrogen bonds to a base on the other strand; the members of each "base pair" lie in a plane that is essentially perpendicular to the axis of the helix. Of the ten theoretically possible base pairs, only two so-called complementary pairs are found in DNA: the pair adenine and thymine and the pair cytosine and guanine. Thus the order of the bases on one strand is precisely related to the order of the bases on the other strand, and the two strands are said to be complementary. Further details are presented in "DNA: Its Structure and Components."

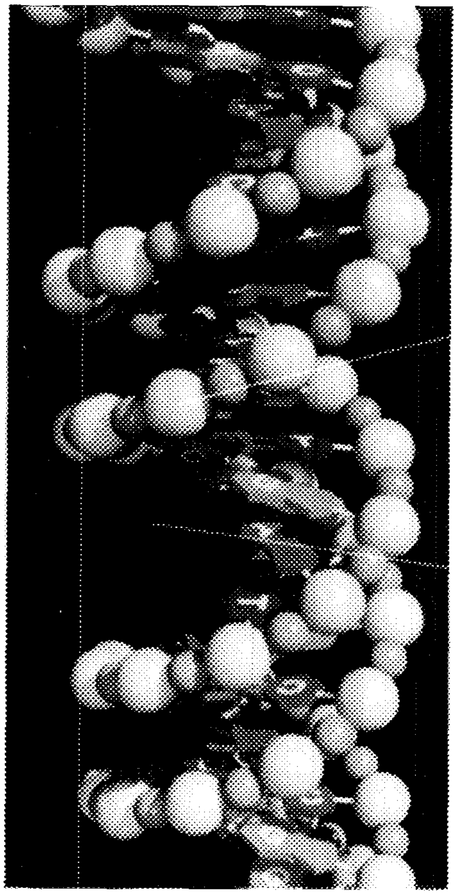
Watson and Crick arrived at their structure for DNA with the help of x-ray diffraction data for DNA fibers obtained by Maurice Hugh Frederick Wilkins (1916–) and Rosalind Franklin (1920–1957) and of the observation in 1950 by Erwin Chargaff (1905–) that the number of molecules of adenine in any of various DNA samples equals the number of molecules of thymine and that the number of molecules of cytosine equals the number of molecules of guanine. In addition, following the example of the American chemist Linus Carl Pauling (1901–), who in 1951 had worked out the details of a helical polypeptide structure (the so-called $\alpha$ helix), they made liberal use of ball-and-stick models.

Molecules of DNA are exceptional among biological macromolecules in two respects. First, they are very long relative to their width. If the diameter of the double helix could be increased to that of a strand of angel-hair pasta, the length of the DNA molecule in a typical human chromosome would be about 12 kilometers. Second, although single-helical configurations are not uncommon in biological macromolecules, the double-helical configuration of DNA is unique. One might wonder why DNA is double-stranded. After all, normally only one of the strands directs protein synthesis, the two strands are replicated separately, and some viruses manage quite nicely with only single-stranded DNA. The evolutionary advantage of double-stranded DNA is thought to lie in the fact that, if one strand is damaged, the other strand can provide the information required to repair the damaged strand.

The base-pairing feature of DNA immediately suggested that each strand of DNA could serve as the template for directing the synthesis of a complementary strand. The result would be two identical double-stranded DNA molecules, each containing one new and one old strand. The suggestion that DNA replication is "semiconservative" was proved correct (for the DNA of *E. coli* and a higher plant) several years after the double-helical DNA structure was proposed. The details of DNA replication, however, are very complex, involving a number of enzymes. One enzyme first uncoils a portion of the DNA molecule, and another separates the two strands. Then an enzyme called a DNA polymerase, using one of the separated DNA strands as a template, catalyzes the polymerization of free deoxyribonucleoside triphosphates into a strand that is complementary to the template. Some features of the process are detailed in "DNA Replication."

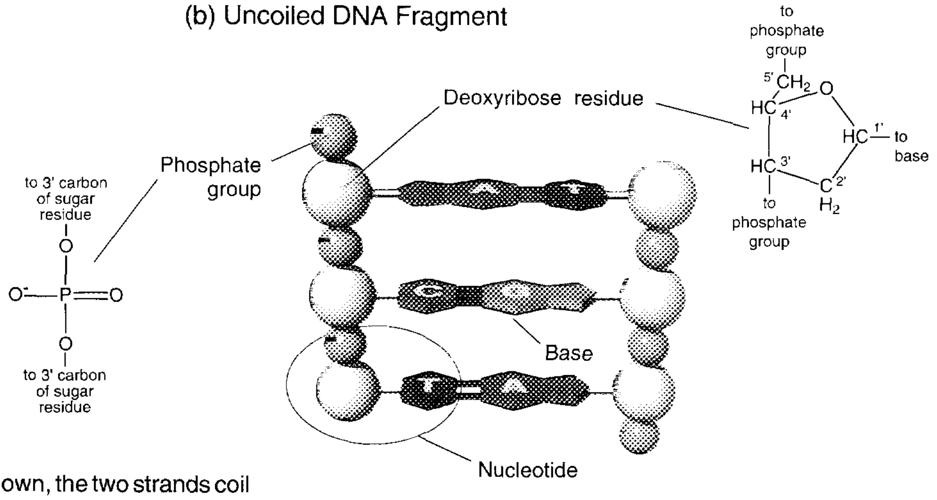
Now that genes were known to direct the synthesis of proteins and to be made of DNA, the next problem was to determine the relationship between DNA and proteins. The first clue about the relationship came in 1949 when Pauling presented evidence that the hemoglobin present in humans suffering from sickle-cell anemia differed *structurally* from the hemoglobin in humans not suffering from that inherited disease. (Hemoglobin is composed of two copies each of two polypeptides, the so-called $\alpha$ and $\beta$ chains. The $\alpha$ chain contains 141 amino acids, and the $\beta$ chain contains 150 amino acids.) What features of a protein affect its structure? By the 1940s biochemists were beginning to realize that the structure of a protein is determined not so much by which amino acids it contains but more by the sequence of the amino acids along the

# DNA: its structure and components



(a) Computer-generated Image of DNA
(by Mel Prueitt)

(b) Uncoiled DNA Fragment



The usual configuration of DNA is shown in (a). Two chains, or strands, of repeated chemical units are coiled together into a double helix. Each strand has a "backbone" of alternating deoxyribose residues (larger spheres) and phosphate groups (smaller spheres). Free deoxyribose, $C_5O_4H_{10}$, is one of a class of organic compounds known as sugars; the phosphate group, $(PO_4)^{-3}$, is a component of many other biochemicals.

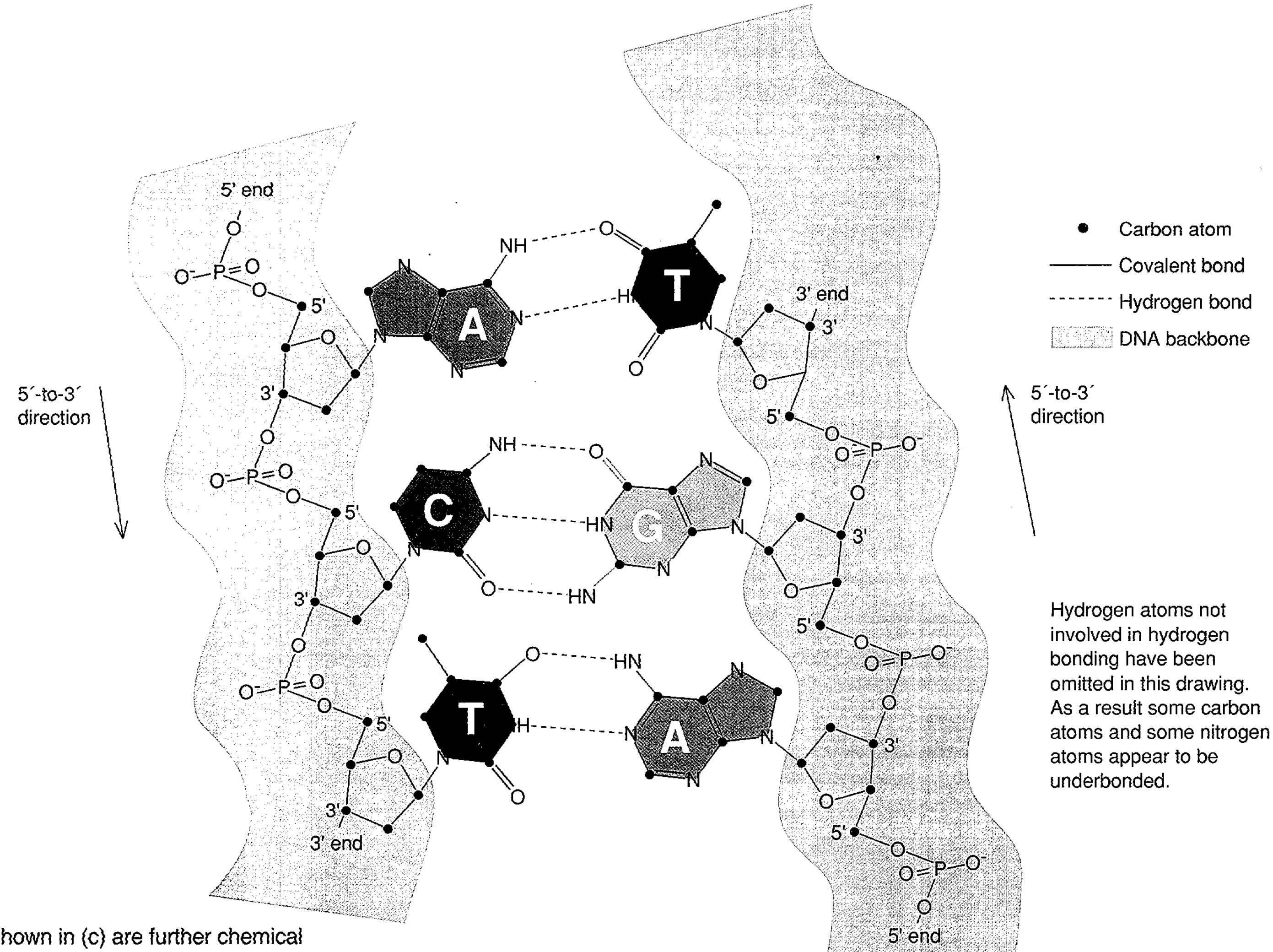
Attached to each sugar residue is one of four essentially planar nitrogenous organic bases: adenine (A), cytosine (C), guanine (G), or thymine (T). The plane of each base is essentially perpendicular to the helix axis. Encoded in the order of the bases along a strand is the hereditary information that distinguishes, say, a robin from a human and one robin from another.

As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are complementary bases), and C is always linked to G (C and G are complementary bases). Thus the order of the bases along one strand is dictated by and can be inferred from the order of the bases along the other strand. (The two strands are said to be complementary.) The pairing of A only with T and of C only with G is the feature of DNA that allows it to serve as a template not only for its own replication but also for the synthesis of proteins (see "DNA Replication" and "Protein Synthesis"). Note that the members of a base pair are essentially coplanar.

All available evidence indicates that each eukaryotic chromosome contains a single long molecule of DNA, only a small portion of which is shown here. Furthermore, the ends of each DNA molecule, called telomeres, have a special base sequence and a somewhat different structure.

Shown in (b) is an uncoiled fragment of (a) containing three complementary base pairs. From the chemist's viewpoint, each strand of DNA is a polymer made up of four repeated units called deoxyribonucleotides, or simply nucleotides. The four nucleotides are regarded as the monomers of DNA (rather than the sugar residue, the phosphate group, and the four base residues) because the *nucleotides are the units added as a strand of DNA is being synthesized* (see "DNA Replication").

A particular nucleotide is commonly designated by the symbol for the base it contains. Thus T is a symbol not only for the base thymine (more precisely, the thymine residue) but also for the indicated nucleotide. Also shown are chemical and structural details of the backbone components. Note that four carbon atoms of the sugar residue and its one oxygen atom form a pentagon in a plane parallel to the helix axis, and that the fifth carbon atom of the sugar residue projects out of that plane.



Hydrogen atoms not involved in hydrogen bonding have been omitted in this drawing. As a result some carbon atoms and some nitrogen atoms appear to be underbonded.

Shown in (c) are further chemical and structural details of the DNA segment shown in (b). The planes of the three base pairs have been rotated into the plane of the sugar residues. Details of particular note include the following.

Linking any two neighboring sugar residues is an -O–P–O- "bridge" between the 3′ carbon atom of one of the sugars and the 5′ carbon atom of the other sugar. (The designations 3′ (three prime) and 5′ (five prime) arise from a standard system for numbering atoms in organic molecules.) When a DNA molecule is broken into fragments, as it must be before it can be studied, the breaks usually occur at one of the four covalent bonds in each bridge.

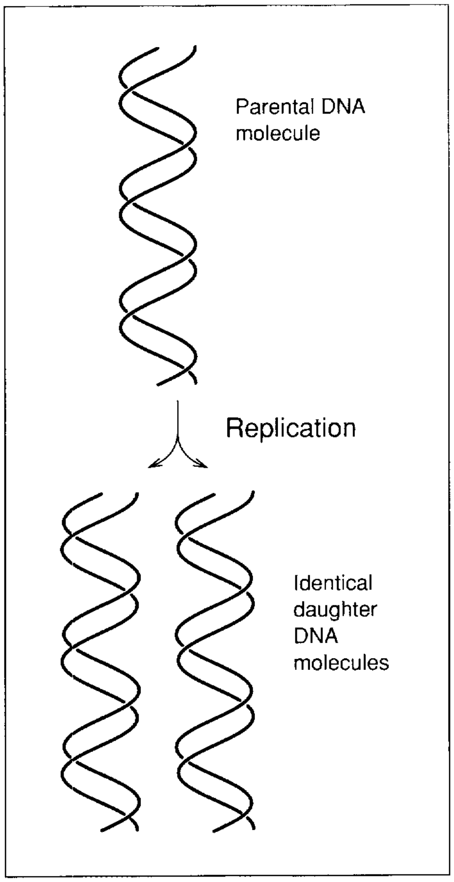
Because deoxyribose has an asymmetric structure, the ends of each strand of a DNA fragment are different. At one end the terminal carbon atom in the backbone is the 5′ carbon atom of the terminal sugar (the carbon atom that lies outside the planar portion of the sugar), whereas at the other end the terminal carbon atom is the 3′ carbon atom of the terminal sugar (a carbon atom that lies within the planar portion of the sugar).

The two complementary strands of DNA are antiparallel. In other words, arrows drawn from, say, the 5′ end to the 3′ end of each strand have opposite directions. Most of the enzymes that move along a backbone in the course of catalyzing chemical reactions move in the 5′-to-3′ direction. The composition of a DNA fragment is represented symbolically in a variety of ways. However, all of the representations focus on the order, or sequence, of the nucleotides (and hence the bases) along the strands of the fragment. For example, the most complete representation for the fragment shown above is

5′-ACT-3′
3′-TGA-5′.

The most abbreviated representation, ACT (or, equivalently, AGT), gives the sequence of only one strand (since the sequence of the complementary strand can be inferred from the given sequence) and follows the convention that the left-to-right direction corresponds to the 5′-to 3′ direction.

# DNA REPLICATION

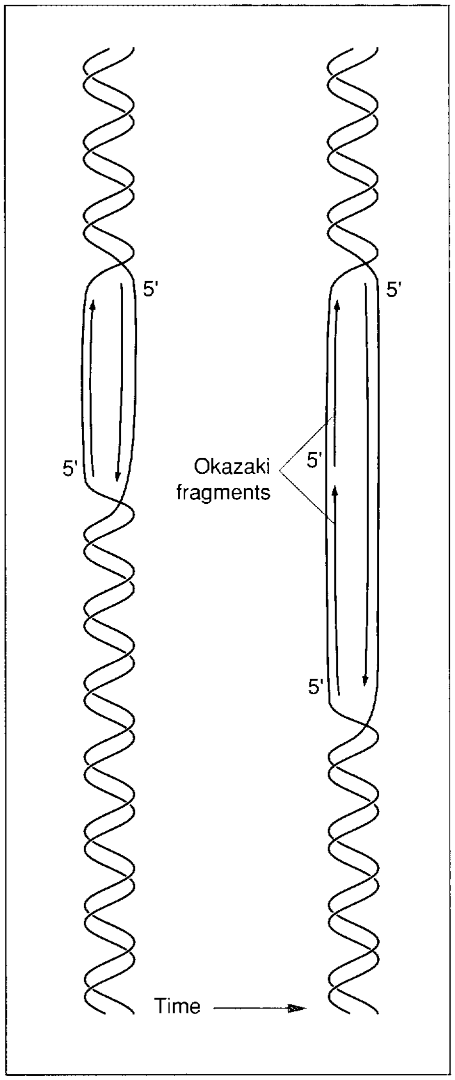


An overall description of DNA replication is quite simple. Each strand of a parent DNA molecule serves as the template for synthesis of a complementary strand. The result is two daughter DNA molecules, each composed of one parental strand and one newly synthesized strand and each a duplicate of the parent molecule. But this overall simplicity, illustrated above, is misleading, since DNA replication involves the intricate and coordinated interplay of more than twenty enzymes. The most important general feature of DNA replication is its extremely high accuracy. A "proofreading" capability of DNA polymerase, the enzyme that catalyzes the basic chemical reaction involved in replication, guarantees that only about one per billion of the bases in a newly synthesized strand differs from the complement of the corresponding base in the template strand.

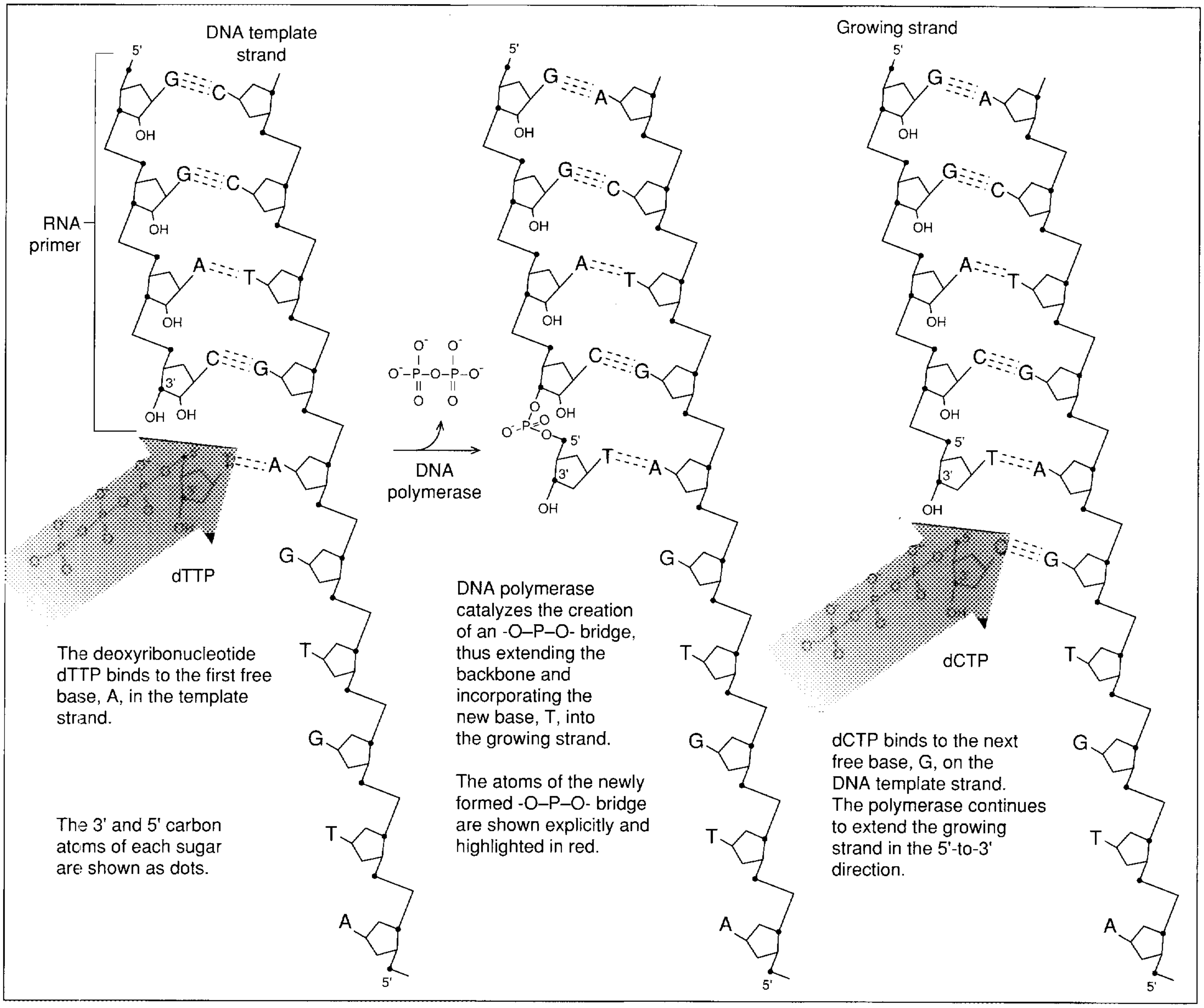
A more detailed description of DNA replication should note first that replication of a chromosomal DNA molecule does not begin at one end of the molecule and proceed uninterruptedly to the other end. Instead, scattered along the molecule are numerous occurrences of a particular base sequence, and each occurrence of that sequence serves as an "origin of replication" for a portion of the molecule. Thus different portions of a DNA molecule are replicated separately. Baker's yeast, *Saccharomyces cerevisiae*, is one of the few eukaryotes for which the base sequence of its origins of replication is now known. Knowledge of the base sequence of an organism's origins of replication is necessary in the creation of artificial chromosomes of the organism, synthetic entities that are treated by the organism's cellular machinery just as its own chromosomes are treated. The cloning vectors known as YACs are an example of artificial chromosomes.

Replication of the portion of a DNA molecule flanked by two origins of replication begins with the action of enzymes that move along the parental DNA, progressively uncoiling and denaturing (separating into single strands) the double helix. Uncoiling and denaturation expose the bases in each parental strand and thereby enable the bases to direct the order in which deoxyribonucleotides are added by DNA polymerase to the strand being synthesized.

Because, as shown in the figure at right, DNA polymerase elongates a growing chain of deoxyribonucleotides only in the 5′-to-3′ direction (arrows), one of the new DNA strands can be synthesized continuously but the other strand must be synthesized in short pieces called Okazaki fragments. (The Okazaki fragments shown here are much shorter than they are in reality.) The discontinuous synthesis of one of the new strands is the source of additional complexities in replicating the very ends, the telomeres, of a DNA molecule.



As shown in the figure on the next page, the participants in the chemical reaction by which each portion of a DNA strand is synthesized include a "primer," the enzyme DNA polymerase, a DNA template (a parental strand), and a supply of free deoxyribonucleoside triphosphates (dNTPs). The usual primer is a very short strand of RNA, generally containing between four and twelve ribonucleotides. (RNA is a single-stranded nucleic acid; its structure is very similar to that of a strand of DNA. Because the sugar residue in RNA is derived from ribose rather than deoxyribose, the repeated units in RNA are



The deoxyribonucleotide dTTP binds to the first free base, A, in the template strand.

The 3' and 5' carbon atoms of each sugar are shown as dots.

DNA polymerase catalyzes the creation of an -O–P–O- bridge, thus extending the backbone and incorporating the new base, T, into the growing strand.

The atoms of the newly formed -O–P–O- bridge are shown explicitly and highlighted in red.

dCTP binds to the next free base, G, on the DNA template strand. The polymerase continues to extend the growing strand in the 5'-to-3' direction.

called ribonucleotides rather than deoxyribonucleotides.) A primer is required because DNA polymerase catalyzes the addition of a deoxyribonucleotide to an existing chain of nucleotides (either ribonucleotides or deoxyribonucleotides) but not the de novo synthesis of a chain of deoxyribonucleotides. The action of each parental strand as a template is based on hydrogen bonding between complementary bases. In particular, a base in a parental strand hydrogen bonds to the dNTP containing the complementary base. As a result, the dNTP is fixed in a position such that the DNA polymerase can exert its catalytic action on the triphosphate group of the dNTP and the 3′ hydroxyl group of the 3′-terminal sugar of the primer. The result is the addition of a deoxyribonucleotide to the primer and the release of a pyrophosphate group, $(P_2O_7)^{-4}$. The next deoxyribonucleotide in the template strand fixes its complementary dNTP into position, the DNA polymerase moves further along the chain being elongated, and addition of another deoxyribonucleotide is effected by action of the polymerase on the triphosphate group of the dNTP and the hydroxyl group of the sugar of the deoxyribonucleotide just previously added. Successive repetitions of the process and eventual replacement of the RNA primer with DNA lead to formation of double-stranded DNA identical to the parental DNA.

polypeptide chain. Then in 1957 Vernon Martin Ingram (1924–) demonstrated that the sixth amino acid in the $\beta$ chain of normal hemoglobin is glutamic acid, whereas the sixth amino acid in the $\beta$ chain of sickle hemoglobin is valine. Otherwise, the amino-acid sequences of both $\beta$ chains are identical. Ingram's work suggested that the function of DNA was to determine the order in which amino acids are assembled into proteins.

DNA itself could not, however, be the template for the synthesis of proteins, since DNA is sequestered in the nucleus of a eukaryotic cell, whereas proteins were known to be synthesized in the cytoplasm outside the nucleus. Perhaps an intermediary substance was involved, one that receives hereditary information from DNA in the nucleus and then moves to the cytoplasm, where it serves as the template for protein synthesis. A likely candidate for such an intermediary was the other known nucleic acid, namely ribonucleic acid, or RNA, which is found primarily in the cytoplasm. Like DNA, RNA is a polymer of four different nucleotides, but the nucleotides are ribonucleotides containing the sugar ribose, which differs from deoxyribose in possessing a hydroxyl group on its $2'$ carbon atom. Another difference is that the base thymine is absent from RNA, being replaced by the base uracil (U), which lacks the extra-ring methyl group of thymine but, like thymine, hydrogen bonds with adenine. The final difference between DNA and RNA is that RNA is usually single-stranded.

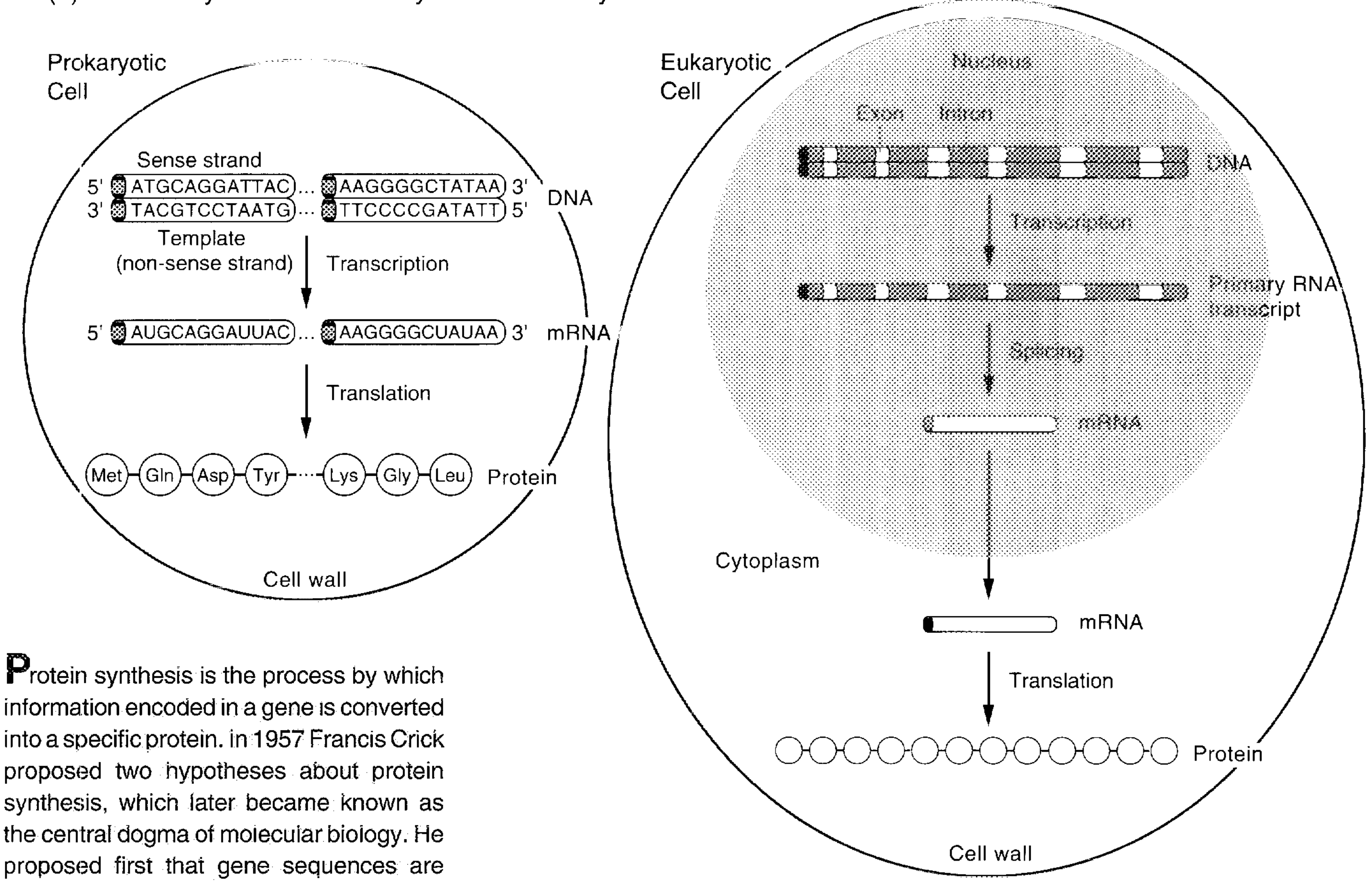
That RNA is the intermediary between DNA and proteins soon became the working hypothesis of biochemists, and the details of protein synthesis were worked out in the fifties and sixties. Briefly, a segment of DNA (a gene) serves as the template for the synthesis, in the nucleus, of so-called messenger RNA (mRNA), a process called transcription and similar to DNA replication. The mRNA then enters the cytoplasm, where it serves as the template for the ordered assembly of amino acids into a protein, a process called translation. Details of transcription and translation are illustrated "Protein Synthesis."

The last general problem about the relation between DNA and proteins was to crack the code relating the sequence of deoxyribonucleotides that constitutes a gene to the sequence of amino acids that constitutes a protein. Experiments performed in 1961 by Crick and the British molecular biologist Sydney Brenner (1927–) suggested that the code was a triplet code, or, in other words, that a sequence of three adjacent deoxyribonucleotides (a codon) specifies each amino acid. The genetic code was completely cracked by 1966, thanks primarily to the independent efforts of two groups, one led by Marshall Warren Nirenberg (1929–) and the other by Har Gobind Khorana (1922–). As shown in "The Genetic Code," eighteen of the twenty amino acids are specified by two or more codons. The redundancy of the code implies that gene mutations involving single-base substitutions do not necessarily result in a change in an amino acid.

Now that what seemed the major questions about the material and mechanisms of heredity had been answered, was anything fascinating left to learn? Or would

# PROTEIN SYNTHESIS

(a) Protein Synthesis in Prokaryotic and Eukaryotic Cells



Protein synthesis is the process by which information encoded in a gene is converted into a specific protein. In 1957 Francis Crick proposed two hypotheses about protein synthesis, which later became known as the central dogma of molecular biology. He proposed first that gene sequences are "collinear" with protein sequences. In other words, the linear arrangement of subunits (deoxyribonucleotides) composing a gene corresponds to the linear arrangement of subunits (amino acids) composing a protein. Second, Crick proposed that a segment of RNA (a ribonucleotide sequence) acts as an intermediate translator between the deoxyribonucleotide sequence and the amino-acid sequence, or, in other words, that genetic information flows from DNA to RNA to protein. Crick had no experimental evidence to support his hypotheses. But very shortly Charles Yanofsky and Seymour Benzer working independently, provided the first evidence in support of the collinearity hypothesis. Their experiments showed that mutations in the genes of *E. coli* and of the T4 bacteriophage produced parallel changes in amino-acid sequences. And as details of protein synthesis were worked out, the role of RNA as an intermediary was also established.

Shown in (a) is an overview of protein synthesis in a prokaryotic cell. In the first stage, called transcription, a DNA segment, a gene, serves as a template for the synthesis of a single-stranded RNA segment called a messenger RNA (mRNA). The base sequence of the mRNA is complementary to the base sequence of one strand of the gene (the template, or "non-sense," strand) and is therefore identical to the base sequence of the other strand of the gene (the "sense" strand). The one exception to the identity is that the base U (uracil) replaces the base T. (Recall that in RNA uracil, rather than thymine, is the base complementary to adenine.)

In the second stage of protein synthesis, called translation, the mRNA serves as the template for the stringing together of amino acids into a protein. The protein is assembled according to the genetic code. That is, the succession of codons (triplets of adjacent ribonucleotides) that compose the mRNA dictates the succession of amino acids that compose the protein. (A listing of codons and corresponding amino acids is presented in "The Genetic Code.") Although transcription and translation are depicted here as if they occurred at different times, translation of a prokaryotic mRNA often begins before its synthesis by transcription is complete.

Also shown in (a) is an overview of protein synthesis in a eukaryotic cell. Unlike prokaryotic genes, most eukaryotic genes are composed of stretches of protein-coding sequences (exons) interrupted by longer stretches of noncoding sequences (introns). Both the exons and introns within a eukaryotic gene are transcribed. The resulting primary transcript is then spliced; that is, each intron is removed and the adjacent exons are linked together.

The shortened RNA is now an mRNA, an RNA that contains only protein-coding sequences. The mRNA leaves the nucleus and in the cytoplasm is translated into a protein according to the genetic code. Thus transcription and translation are of necessity temporally separated in eukaryotic cells.

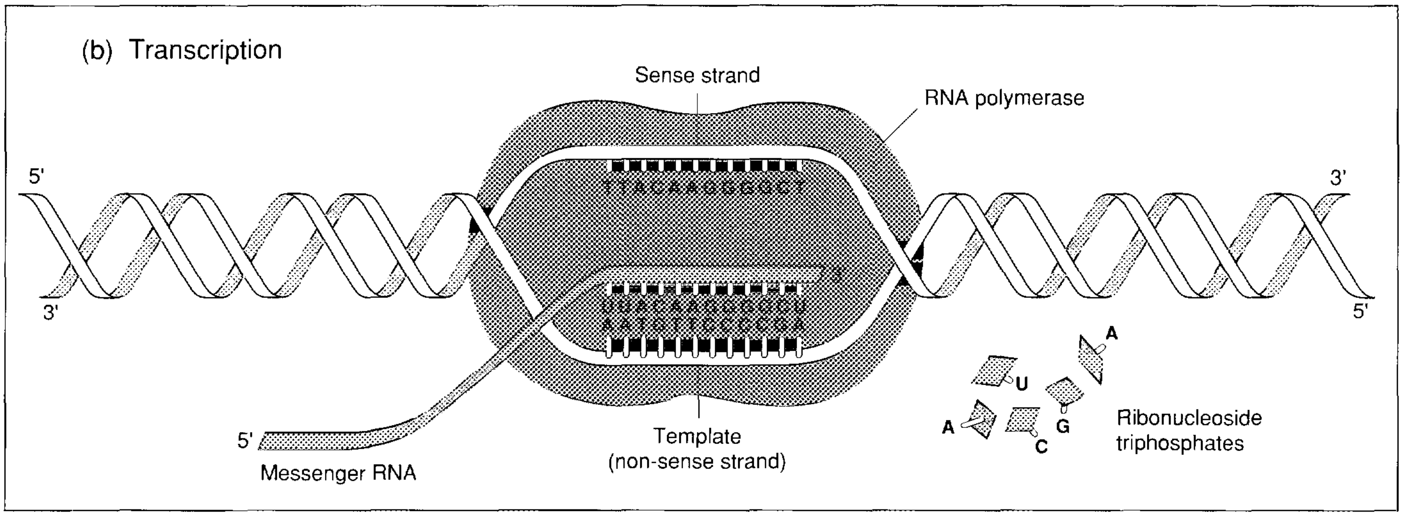
The overviews in (a) illustrate that, as Crick had postulated, genetic information flows from DNA to RNA to protein within both prokaryotic and eukaryotic cells. One important exception to the central dogma is the class of viruses known as retroviruses, of which the AIDS virus is an example. Retroviruses store genetic information in RNA and then convert the information to DNA—a reversal of the usual information flow that is known as reverse transcription.

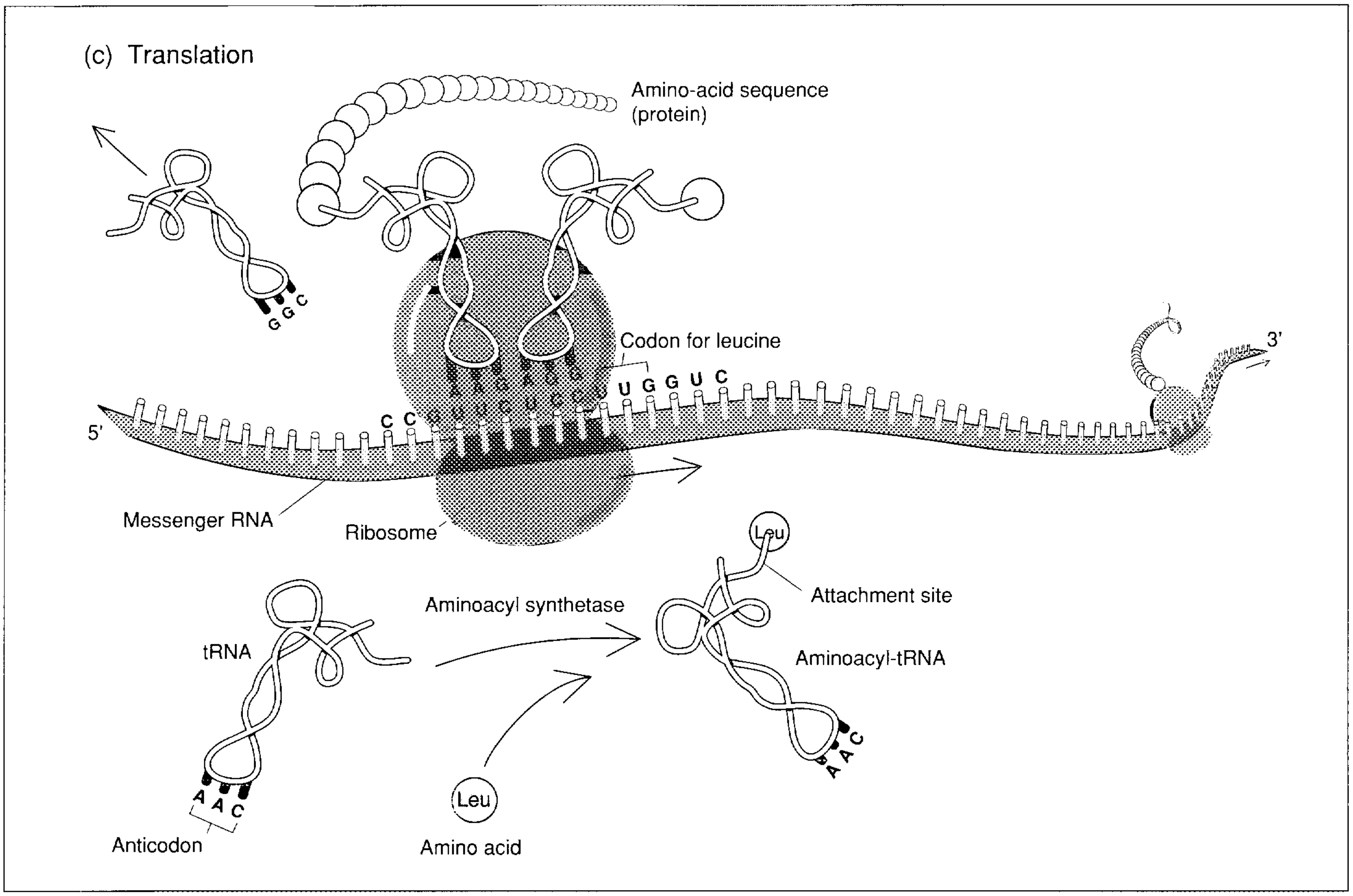
Details of transcription and translation are shown in (b) and (c) respectively. Transcription begins when an enzyme, an RNA polymerase, binds to a particular segment of a gene called the promoter. The double helix then uncoils and separates into two strands, exposing a small number of bases. The RNA polymerase facilitates hydrogen bonding between an exposed base in the template strand and its complementary base in a free ribonucleoside triphosphate (NTP) and then between the next exposed base in the template strand and its complementary base in another free NTP. While the two NTPs are held in proximity by the hydrogen bonds, the RNA polymerase catalyzes the formation of an -O–P–O- bridge between them, thus forming a chain of two covalently linked ribonucleotides. (See "DNA Replication" for details about formation of -O–P–O- bridges.) A third NTP is hydrogen-bonded to the third exposed base in the template strand and is covalently linked to the second ribonucleotide in the chain. The RNA polymerase moves along the template in the 3′-to-5′ direction, continuing to unwind and separate the double helix and to elongate the RNA chain in the 5′-to-3′ direction by catalyzing the addition of successive ribonucleotides. At the same time, the distorted DNA in the wake of the polymerase rewinds. After the gene is fully transcribed, the polymerase separates from the double helix. If the gene transcribed is a eukaryotic gene, the newly minted RNA is spliced and the resulting mRNA enters the cytoplasm through pores in the nuclear membrane.

As shown in (c), translation occurs with the help of transfer RNA molecules (tRNAs) and ribosomes. Each tRNA is a tiny, cloverleaf-shaped molecule that serves as an adapter: At one end it contains a triplet of ribonucleotides (an anticodon) that binds with a complementary codon on the mRNA strand, and at the other end it has an attachment site for a single amino acid. Many varieties of tRNAs exist. An important difference between one tRNA and another is the presence of a different anticodon on the central cloverleaf stem. The number of different anticodons found in the various tRNAs is less than the number of codons in the genetic code. That is so because the base pairing between the third base of the mRNA codon and the first base of the tRNA anticodon can depart from the usual Watson-Crick rules. For example, G can pair with U in addition to C.

Ribosomes are very large molecules composed of ribosomal RNA (rRNA) and approximately fifty different proteins. As a ribosome travels along an mRNA it catalyzes the reactions that lead to synthesis of the protein encoded in the mRNA. Thousands of ribosomes exist within each cell.

Before a tRNA molecule participates in translation, it must be converted to an aminoacyl-tRNA (become attached to the amino acid corresponding to its anticodon). Each of the twenty amino acids found in proteins can be attached to at least one type of tRNA, and most can be attached to several. The binding between tRNA and amino acid is cata-





(c) Translation

lyzed by one of a group of enzymes. Those exquisitely specific enzymes, called aminoacyl synthetases, are in fact the agents by which the genetic information in mRNA is decoded.

Translation begins when an aminoacyl-tRNA containing the amino acid methionine and a ribosome bind to an initiation sequence near the 5′ end of the mRNA. The initiation sequence consists of the START codon AUG, to which the aminoacyl-tRNA binds through complementary base pairing. A second aminoacyl-tRNA, which contains an anticodon complementary to the second mRNA codon, binds to the mRNA. Then the amino acid on the first aminoacyl-tRNA is joined by a peptide bond to the amino acid on the second aminoacyl-tRNA, thus creating a chain of two amino acids dangling off the end of the second aminoacyl-tRNA. The process continues as the ribosome moves along the mRNA (in the 5′-to-3′ direction) and as peptide bonds are formed between successive amino acids. When the ribosome reaches a STOP codon within the mRNA, the ribosome detaches from the mRNA, and the completed protein is released into the cytoplasm.

The process of translation is fast: A single ribosome can translate up to fifty ribonucleotides per second. Furthermore, at any one time numerous ribosomes may be traveling along a single mRNA, each producing a molecule of the same protein. Thus a protein needed for diverse tasks within the cell can be quickly and efficiently produced.

Note: Published only recently (in June 1992) was strong evidence that the formation of peptide bonds between amino acids during translation is catalyzed not by some protein enzyme within a ribosome but instead by an RNA component of the ribosome. That news is exciting but not completely unexpected, since the ability of RNA to function as a catalyst in other situations had been demonstrated in the early 1980s. In particular, the primary transcript of a ribosomal-RNA gene of the protozoan *Tetrahymena thermophila* had been shown to effect its own splicing and the catalytic action of an RNA-protein complex that processes the primary transcripts of certain transfer-RNA genes had been ascribed to the RNA component of the complex rather than the protein component.

# THE GENETIC CODE

What triplet of ribonucleotides directs the addition of, say, the amino acid alanine to a protein that is being synthesized? Of lysine? Of any one of the twenty amino acids found in proteins? That was the problem to be faced after advancement of the ideas that a gene is a string of deoxyribonucleotide triplets, that the string of deoxyribonucleotide triplets is transcribed into a string of ribonucleotide triplets, and that the string of ribonucleotide triplets is translated into a string of amino acids–a protein. The results of research on the problem is condensed in the genetic code, a listing of the sixty-four possible ribonucleotide triplets and the amino acid (or translation command) corresponding to each. Fortunately for those who worked on the problem, the genetic code is organism-independent. That is, the same genetic code is used by virtually all organisms.

Researchers began to crack the genetic code in the early 1960s. Marshall Nirenberg and his collaborators added a synthetic RNA, consisting entirely of repetitions of a single ribonucleotide, say U, to a bacterial extract that contained everything necessary for protein synthesis except RNA. The result was a string of the amino acid phenylalanine. They concluded that the ribonucleotide triplet UUU codes for phenylalanine. Other ribonucleotide triplets were decoded by performing similar experiments with synthetic RNAs containing only A's, C's, or G's or various combinations of ribonucleotides. By 1966 research teams led by Har Gobind Khorana and Marshall Nirenberg had cracked the entire genetic code.

(a) RNA Codons for the Twenty Amino Acids

| | Second base | | | | |
|---|---|---|---|---|---|
| First base | U | C | A | G | Third base |
| U | Phe | Ser | Tyr | Cys | U |
| | Phe | Ser | Tyr | Cys | C |
| | Leu | Ser | STOP | STOP | A |
| | Leu | Ser | STOP | Trp | G |
| C | Leu | Pro | His | Arg | U |
| | Leu | Pro | His | Arg | C |
| | Leu | Pro | Gln | Arg | A |
| | Leu | Pro | Gln | Arg | G |
| A | Ile | Thr | Asn | Ser | U |
| | Ile | Thr | Asn | Ser | C |
| | Ile | Thr | Lys | Arg | A |
| | Met (start) | Thr | Lys | Arg | G |
| G | Val | Ala | Asp | Gly | U |
| | Val | Ala | Asp | Gly | C |
| | Val | Ala | Glu | Gly | A |
| | Val | Ala | Glu | Gly | G |

Amino-acid abbreviations

| | | |
|---|---|---|
| Ala | = | Alanine |
| Arg | = | Arginine |
| Asp | = | Aspartic acid |
| Asn | = | Asparagine |
| Cys | = | Cysteine |
| Glu | = | Glutamic acid |
| Gln | = | Glutamine |
| Gly | = | Glycine |
| His | = | Histidine |
| Ile | = | Isoleucine |
| Leu | = | Leucine |
| Lys | = | Lysine |
| Met | = | Methionine |
| Phe | = | Phenylalanine |
| Pro | = | Proline |
| Ser | = | Serine |
| Thr | = | Threonine |
| Trp | = | Tryptophan |
| Tyr | = | Tyrosine |
| Val | = | Valine |

Shown in (a) is the usual representation of the genetic code. The letters U, C, A, and G are symbols for the ribonucleotides containing the bases uracil, cytosine, adenine, and guanine, respectively. The symbols in the body of the table are three-letter abbreviations for the amino acids. To find the amino acid specified by a particular codon (say the codon CAG), locate the first nucleotide (C) along the left side of the table and the second nucleotide (A) along the top of the table. Their intersection pinpoints one of four amino acids. Of those four the one aligned with the third nucleotide (G) is the amino acid in question. Thus the amino acid glutamine (Gln) is specified by the three-nucleotide sequence CAG.

Shown in (b) is another version of the genetic code, one expressed in terms of DNA codons instead of RNA codons. Each single-stranded deoxyribonucleotide triplet listed in (b) is the sequence of the so-called sense strand of a DNA codon—the strand that does not serve as a template for synthesis of RNA. Note that most of the amino acids are specified by at least two codons. For example, phenylalanine is specified by two codons: TTT and TTC. Arginine is specified by a total of six codons: CGT, CGC, CGA, CGG, AGA, and AGG. In general, the more an amino acid is used in protein synthesis the likelier it is to be specified by more than one codon. Note also the start codon (ATG) and the three stop codons (TAA, TGA, and TAG) that are used to signal the beginning and end of protein synthesis. The substantive difference between the two versions of the genetic code is that in (b) the deoxyribonucleotide T replaces the ribonucleotide U.

(b) DNA Codons for the Twenty Amino Acids

| Ala | Arg | Asp | Asn | Cys | Glu | Gln | Gly | His | Ileu | Leu | Lys | Met (START) | Phe | Pro | Ser | Thr | Trp | Tyr | Val | STOP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GCA | AGA | GAT | AAT | TGT | GAA | CAA | GGA | CAT | ATA | TTA | AAA | ATG | TTT | CCA | AGT | ACA | TGG | TAT | GTA | TAA |
| GCG | AGG | GAC | AAC | TGC | GAG | CAG | GGG | CAC | ATT | TTG | AAG | | TTC | CCG | AGC | ACG | | TAC | GTG | TAG |
| GCT | CGA | | | | | | GGT | | ATC | CTA | | | | CCT | TCA | ACT | | | GTT | TGA |
| GCC | CGG | | | | | | GGC | | | CTG | | | | CCC | TCG | ACC | | | GTC | |
| | CGT | | | | | | | | | CTT | | | | | TCT | | | | | |
| | CGC | | | | | | | | | CTC | | | | | TCC | | | | | |

molecular genetics degenerate into clearing up details here and details there? Some thought so, and bemoaned the passing of a golden age. But in reality another era, and one just as golden, was opening, thanks to development of techniques for manipulating and analyzing DNA.

## The Techniques of Molecular Genetics

The late 1960s mark the beginning of the recombinant-DNA revolution. During the ensuing years it became possible to make billions of identical copies of segments of DNA by cloning (duplicating) each segment individually as a recombinant DNA molecule in the bacterium *Escherichia coli.* The significance of that breakthrough was enhanced by other new developments, including the ability to separate fragments of DNA that differ in length by only a few nucleotide pairs, to determine the nucleotide sequences of cloned segments of DNA, to create specific mutations in cloned genes, and to introduce cloned eukaryotic genes into experimental organisms.

Those startling developments arose from advances during the previous decade in nucleic-acid biochemistry and in bacterial and phage genetics. Basic features of the replication, repair, and recombination of DNA and of the synthesis of proteins had been elucidated, and identification and isolation of the enzymes that catalyze the chemical reactions involved had allowed those processes to be reproduced in vitro. The action of phages as carriers of genetic material between different strains of *E. coli* had been utilized to isolate individual *E. coli* genes. The rates of transcription of *E. coli* genes had been determined (by measuring the amounts of RNA transcribed from the different genes) and had been found to be regulated, that is, to vary from gene to gene and in response to external stimuli. The observed regulation of gene expression in *E. coli* had been traced to the interaction of certain proteins with regulatory sequences in its genome. By 1968 about a hundred genes had been ordered on the genetic maps of phages, and about fifteen hundred genes had been ordered on the genetic map of *E. coli.*

On the other hand, essentially nothing was known about the structure of eukaryotic genes, their regulation, or their organization in chromosomal DNA molecules. Even the major difference between prokaryotic and eukaryotic genes—the presence of introns in the latter—had not yet been discovered. Most frustrating was the lack of a methodology for studying eukaryotic genomes analogous to the phage-bacteria system for studying the organization, rearrangement, and functions of phage and bacterial genomes.

But in 1968 techniques began to be developed that exploit the cellular machinery and the biosynthetic products of bacteria to replicate, manipulate, and analyze eukaryotic genes and to manufacture eukaryotic proteins. Improvements during the past twenty years in recombinant-DNA techniques have produced an explosion of knowledge about eukaryotic genes and about the organization and rearrangements of DNA in eukaryotic genomes, including the human genome.

This section briefly describes some of the techniques that are employed in the study of DNA and points out some of the facts about DNA the techniques have helped to reveal. The chronological approach will be more or less abandoned, and none of the contributions will be attributed to their originators.

A description of the preparation of a sample of DNA is appropriate as a preliminary to this section. The usual preparation procedure involves treating a large number of cells (typically about 5 million) of the organism in question with a detergent, which dissolves cellular membranes and dissociates the proteinaceous component of the chromosomes from the DNA. Then the membrane components and the proteins are removed with an organic solvent such as a chloroform-phenol mixture, and the DNA is precipitated with ethanol as a highly viscous liquid. The mass of the DNA in such a sample is small, about 30 micrograms in the case of human DNA and correspondingly smaller in the case of DNA extracted from organisms with smaller genomes.

It is worth noting that no DNA sample prepared in the above manner contains intact DNA molecules. The mechanical aspects of sample preparation (such as stirring and pipetting) invariably break some of the covalent bonds of the DNA backbones. That accidental fragmentation is usually of little consequence, however, because most of the techniques employed to study DNA at the molecular level are applicable only to stretches of DNA shorter than the intact molecules found in chromosomes. In fact, deliberate fragmentation, by either mechanical or biochemical means, is the first step in many of the techniques to be described below.

The length of a DNA molecule or fragment is expressed in terms of the number of base pairs it contains. (Because the structure of DNA is regular, number of base pairs is directly proportional to physical length.) The average length of the intact DNA molecules within human chromosomes, for example, is about 130 million base pairs, which corresponds to a physical length of about 4.5 centimeters. The lengths of the known human genes are much shorter, ranging from less than a hundred base pairs for the transfer-RNA genes to over a million base pairs for the Duchenne muscular-dystrophy gene and the cystic-fibrosis gene.

We turn now to the means for manipulating and analyzing DNA.

**Fractionation by Copy Number and Repetitive DNA.** The mid 1960s brought to light a surprising feature of eukaryotic DNAs: their content of multiple identical or nearly identical copies of various sequences. The various repeated sequences are collectively called repetitive DNA, and, depending on the species, repetitive DNA is estimated to constitute between 3 and 80 percent of the total. (Between 25 and 35 percent of the human genome, and of other mammalian genomes, is repetitive DNA.) In contrast, the DNAs of viruses and prokaryotes contain no or very little repetitive DNA. The phenomenology of repetitive DNA is complex and not yet fully explored. A few of the repeated sequences are genes, but most have no known function. The multiple copies of some repeated sequences are situated one after the other; the known lengths of the repeated units in such tandem repeats range from two base pairs to several thousand base pairs. Some tandem repeats occur at only one location within a genome; others, called interspersed tandem repeats, occur at many locations. Like the multiple copies of an interspersed tandem repeat, the multiple copies of other repeated sequences are scattered here and there within a genome; the known lengths of such interspersed repeats range from about a hundred base pairs to seven thousand base pairs. And finally the copy numbers of the various repeated sequences range from less than ten to over a million. Two of the many repeated sequences found in the human genome are the GT sequence, an interspersed tandem repeat that consists of between fifteen and thirty tandem repetitions of the sequence 5′-GT and has a copy number on the order of a hundred thousand, and the *Alu* sequence, an interspersed repeat that is about three hundred base pairs in length and has a copy number close to 2 million.

The existence of repetitive DNA became known from comparison of the renaturation kinetics of prokaryotic and eukaryotic DNAs. Recall that the natural configuration of DNA is double-stranded. However, DNA can be separated into single strands (denatured) by, say, heating an aqueous solution of the DNA to about 100°C. When the temperature of a thermally denatured sample of DNA is lowered, random encounters among the single-stranded fragments lead to renaturation, or the re-establishment of hydrogen bonds between complementary fragments. The kinetics of the renaturation can be monitored by, for example, measuring the time dependence of the absorption of ultraviolet light by the sample, since single- and double-stranded DNA have different capacities to absorb ultraviolet light.

Consider the renaturation of two samples of denatured DNA, one prepared by breaking the genome of *E. coli* into equal-length fragments and the other prepared by breaking, into fragments of the same length as the *E. coli* fragments, a hypothetical DNA molecule of the same total length as the *E. coli* genome but composed of multiple repetitions of a single sequence. Each single-stranded *E. coli* fragment is complementary to only one of the many single-stranded fragments in the first sample, whereas each single-stranded hypothetical fragment is complementary to one-half of the equally numerous single-stranded fragments in the second sample. Obviously, then, the hypothetical sample renatures more rapidly, at least initially, than the *E. coli* sample, and therefore the graphs of fraction renatured versus time for the two samples are different. This example illustrates why renaturation-kinetics data are the source of information about the presence of repetitive DNA.

Other types of information can be extracted from renaturation-kinetics data. Consider the renaturation of the *E. coli* genome and the genome of the virus known as T4, each broken into fragments of the same length. Both genomes contain essentially no repetitive DNA, but the sample of *E. coli* DNA contains a greater number of fragments because the *E. coli* genome (which contains about 5,000,000 base pairs of DNA) is larger than the T4 genome (which contains about 170,000 base pairs of DNA). Therefore the *E. coli* genome renatures less rapidly than the T4 genome. In other words, renaturation kinetics provides information about the relative sizes of genomes. Furthermore, because the rate at which hydrogen bonds are established between fragments of single-stranded DNA that have similar but not identical base sequences depends on the degree of similarity of the base sequences of the fragments, the kinetics of the joint renaturation of samples of DNA from different species provides an estimate of the overall similarity of the base sequences of the DNAs.

Today renaturation is most often used to fractionate fragments of DNA by copy number, that is, to separate a DNA sample into components containing highly repetitive DNA, less highly repetitive DNA, and single-copy DNA. Such a separation narrows the search for genes, most of which occur only once within a genome and hence are contained in the single-copy fraction.

**Fragmenting DNA with Restriction Enzymes.** Until 1970 DNA molecules were of necessity fragmented by mechanical means, such as forcing a sample through a syringe. Mechanical fragmentation has disadvantages: Identical pieces of DNA are not fragmented at the same points, and the lengths of the resulting fragments vary widely. Then came discovery of restriction enzymes (or, more precisely, type II restriction endonucleases), biochemicals capable of "cutting" double-stranded DNA not only in a reproducible manner but also into less widely varying lengths. In particular, a restriction enzyme recognizes and binds to an enzyme-specific, very short sequence within a DNA segment and catalyzes the breaking of two particular oxygen-phosphorus-oxygen (-O–P–O-) bridges, one in each backbone of the segment. The locations along a stretch of DNA of the sequence recognized by a restriction enzyme are called restriction sites.

The -O–P–O- bridges broken by a restriction enzyme usually lie within the recognition sequence of the enzyme. For example, the restriction enzyme *Eco*RI recognizes and binds to the sequence

$$\begin{array}{c} 5'\text{-GAATTC-}3' \\ 3'\text{-CTTAAG-}5' \end{array}$$

and, if allowed to interact with a sample of DNA for a sufficiently long time (to completely "digest" the DNA), cuts the DNA within every occurrence of that sequence. Note that the sequence recognized by *Eco*RI, like the sequences recognized by many other restriction enzymes, is palindromic; in other words, the 5′-to-3′ sequence of one strand is identical to the 5′-to-3′ sequence of the other strand.

The average length of the restriction fragments produced by *Eco*RI, a "6-base cutter" (a restriction enzyme that recognizes a 6-base-pair sequence), can be estimated to be about 4000 base pairs, since DNA is approximately a random sequence of four base pairs and any given sequence of six base pairs occurs on average every $4^6 = 4096$

base pairs within such a sequence. (Note, however, that the observed average length of the fragments produced by an $N$-base cutter sometimes differs considerably from the estimate of $4^N$.) Fragments with a shorter average length can be obtained by complete digestion with, say, a 4-base cutter, and fragments with a longer average length can be obtained by complete digestion with a restriction enzyme that recognizes a sequence longer than 6 base pairs or by partial digestion with a 6-base cutter, which leaves some of the restriction sites uncut.

A majority of the many restriction enzymes available today, including *Eco*RI, cut DNA in a fashion such that the resulting fragments terminate in a very short section of single-stranded DNA. For example *Eco*RI cuts the DNA segment

5′- . . . GAATTC . . . -3′
3′- . . . CTTAAG . . . -5′

into the fragments

5′- . . . G-3′
3′- . . . CTTAA-5′

and

5′-AATTC . . . -3′
3′-G . . . -5′

Note that the single-stranded ends of the two *Eco*RI restriction fragments are complementary. The utility of such "sticky" ends in the creation of recombinant DNA molecules will be described below.

A brief natural history of restriction enzymes is presented in "Restriction Enzymes," as well as a listing of a few of the many available.

**Fractionating DNA Fragments by Length: Gel Electrophoresis**. Because DNA fragments are negatively charged, they are subject to an electrical force when placed in an electric field. In particular, DNA fragments placed in a gel (a porous, semisolid material) move through the gel in a direction opposite to the direction of an applied electric field. Furthermore, the rate at which a fragment travels is approximately inversely proportional to the logarithm of its length. Therefore gel electrophoresis is a means for separating DNA fragments by length. Details of the technique are described in "Gel Electrophoresis."

But what is the point of separating fragments of DNA by length? After all, the lengths of the fragments obtained either by breaking a DNA molecule mechanically or by cutting it with a restriction enzyme bear no relation to the functioning of the molecule within a cell. Nevertheless, gel electrophoresis, particularly of restriction fragments, is of great utility in the study of DNA. For example, consider the genome of the phage known as $\lambda$ (lambda), a double-stranded DNA molecule about 50,000 base pairs in length. When many copies of the $\lambda$ genome are completely digested with *Eco*RI and the resulting restriction fragments are subjected to gel electrophoresis, groups of

# RESTRICTION ENZYMES

Like the immune systems of vertebrate eukaryotes, the restriction enzymes of bacteria combat foreign substances. In particular, restriction enzymes render the DNA of, say, an invading bacteriophage harmless by catalyzing its fragmentation, or, more precisely, by catalyzing the breaking of certain -O–P–O- bridges in the backbones of each DNA strand. The evolution of restriction enzymes helped many species of bacteria to survive; their discovery by humans helped precipitate the recombinant-DNA revolution.

Three types of restriction enzymes are known, but the term "restriction enzyme" refers here and elsewhere in this issue to type II restriction endonucleases, the only type commonly used in the study of DNA. (A nuclease is an enzyme that catalyzes the breaking of -O–P–O- bridges in a string of deoxyribonucleotides or ribonucleotides; an endonuclease catalyzes the breaking of internal rather than terminal -O–P–O- bridges.) Many restriction enzymes have been isolated; more than seventy are available commercially. Each somehow recognizes and binds to its own restriction sites, short stretches of double-stranded DNA with a specific base sequence. Having bound to one of its restriction sites, the enzyme catalyzes the breaking of one particular -O-P–O- bridge in each DNA strand.

The accompanying table lists a few of the more commonly used restriction enzymes and the organism in which each is found. The first three letters of the name of a restriction enzyme are an abbreviation for the species of the source organism and are therefore customarily italicized. The next letter(s) of the name designates the strain of the source organism, and the terminal Roman numeral denotes the order of its discovery in the source organism.

| Restriction Enzyme | Source Organism | Base Sequence of Restriction Site |
|---|---|---|
| *Bam*HI | *Bacillus amyloliquefaciens* | 5'-GGATCC-3'<br>3'-CCTAGG-5' |
| *Eco*RI | *Escherichia coli* | 5'-GAATTC-3'<br>3'-CTTAAG-5' |
| *Hae*III | *Haemophilus aegyptius* | 5'-GGCC-3'<br>3'-CCGG-5' |
| *Hind*II | *Haemophilus influenzae* | 5'-GT(C or T)(A or G)AC-3'<br>3'-CA(G or A)(T or C)TG-5' |
| *Mbo*I | *Moraxella bovis* | 5'-GATC-3'<br>3'-CTAG-5' |
| *Not*I | *Nocardia otitidis* | 5'-GCGGCCGC<br>3'-CGCCGGCG |
| *Taq*I | *Thermus aquaticus* | 5'-TCGA<br>3'-AGCT |

Also listed in the table are the base sequences of the restriction sites of the enzymes. The red line separates the ends of the resulting fragments. The restriction sites of many of the known restriction enzymes and of all the restriction enzymes listed in the table have palindromic base sequences. That is, the 5′-to-3′ base sequence of one strand is the same as the 5′-to-3′ base sequence of its complementary strand. Both the bridges broken by a restriction enzyme that recognizes a palindromic sequence lie within or at the ends of the sequence.

Note that most of the restriction enzymes in the table make "staggered" cuts; that is, they produce fragments with protruding single-stranded ends. Those "cohesive," or "sticky," ends are very useful. Suppose that a sample of human DNA and a sample of phage DNA are both fragmented with the same restriction enzyme, one that makes staggered cuts. When the resulting fragments are mixed, they will tend to hydrogen bond with each other because of the complementarity of their sticky ends. In particular, some human DNA fragments will hydrogen bond to some phage DNA fragments. And that bonding is the first step in the creation of a recombinant DNA molecule.

A final point about restriction enzymes is the problem of how the DNA of a bacterium avoids being chopped up by the friendly fire of the restriction enzyme(s) it produces. Evolution has solved that problem also. A bacterium that produces a type II restriction endonuclease produces in addition another enzyme that catalyzes the modification of restriction sites in its own DNA in a manner such that they cannot serve as binding sites for the restriction enzyme.

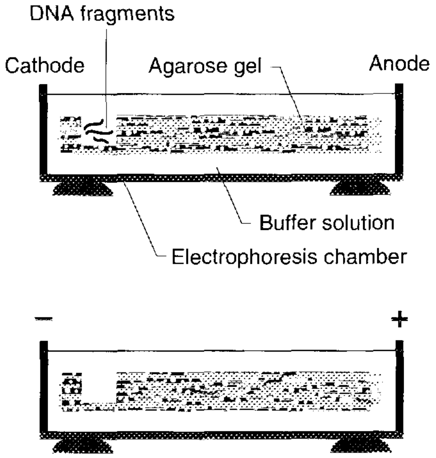
# GEL ELECTROPHORESIS

Historically gel electrophoresis was first applied to separating proteins essentially according to mass, but the technique was adapted to separating fragments of DNA (or RNA) essentially according to fragment length. The technique works on DNA because the phosphate groups of a DNA fragment are negatively charged, and therefore, under the influence of an electric field, the fragment migrates through a gel (a porous, semisolid medium) in a direction opposite to that of the field. Furthermore, the rate at which the fragment migrates through the gel is approximately inversely proportional to the logarithm of its length.

Gel electrophoresis of DNA is carried out with two types of electric field. Conventional gel electrophoresis employs a field that is temporally constant in both direction and magnitude. In contrast, pulsed-field gel electrophoresis employs a field that is created by pulses of current and therefore varies periodically from zero to some set value. More important, the direction of the electric field also varies because different pulses flow through pairs of electrodes at different locations. (Note, however, that the time-averaged direction of the electric field is along the length of the gel.) The advantage of such a pulsed field is that it prevents long DNA fragments, fragments longer than about 50,000 base pairs, from jackknifing within the structural framework of the gel and thus allows the long fragments to migrate through the gel in a length-dependent manner, just as shorter fragments migrate in a constant electric field.

The gel employed is usually a solidified aqueous solution of agarose, a purified form of agar. By varying the concentration of agarose in the gel, conventional gel electrophoresis can be applied to samples containing DNA fragments with average lengths between a few hundred base pairs and tens of thousands of base pairs. (Another gel used for conventional electrophoresis is polyacrylamide, which is particularly suited to separating fragments with lengths less than about a thousand base pairs and is therefore the gel of choice for sequencing.) Conventional gel electrophoresis in an agarose gel is illustrated in (a); details of the technique are as follows.

(a) Conventional Gel Electrophoresis



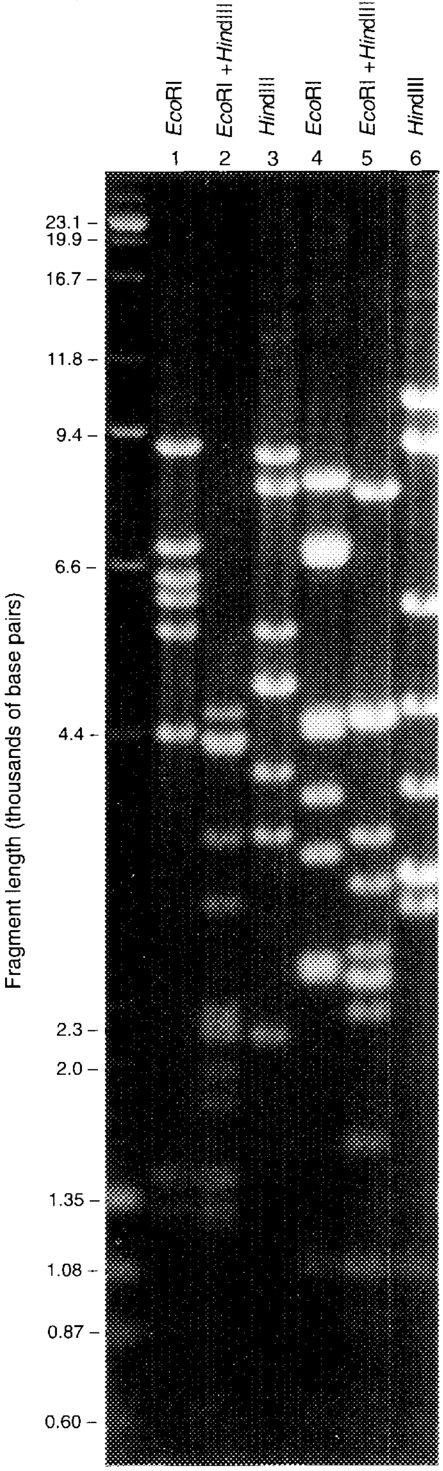
Agarose is dissolved in a hot buffer solution, and the gel solution is allowed to solidify into a thin slab in a casting tray in which the teeth of a comb-like device are suspended. After the gel has solidified, the comb is removed. The "wells" formed by the teeth of the comb are the receptacles into which the samples of DNA are loaded. The thickness of the gel is about 5 millimeters; its length and width are much greater and vary with the purpose of the electrophoresis. Before being loaded with the DNA sample(s), the gel is immersed in a conducting buffer solution in an electrophoresis chamber.

Before a DNA sample is loaded into a well, it is mixed with a dense solution of sucrose or glycerol to prevent the DNA from escaping into the buffer solution. Into one well is loaded a gel-calibration sample, a sample containing fragments of known lengths. As shown in (a), the flow of electricity through the gel causes the fragments to migrate toward the positive electrode. The shorter fragments move more easily through the gel and therefore travel farther.

The positions of the fragments after electrophoresis can be detected by soaking the gel in a solution of ethidium bromide, which binds strongly to DNA and emits visible light when illuminated with ultraviolet light. In a photograph of the ultraviolet-illuminated gel, the fragments appear as light bands. The ethidium-bromide visualization technique makes the positions of all the fragments in the gel visible. An alternative visualization technique detects only certain fragments (see "Hybridization Techniques").

The above description of gel electrophoresis might suggest that the sample of DNA contains but one copy of each fragment. In reality the sample must contain many copies of each fragment, and each band seen in the image of the length-separated fragments contains many fragments, all of which have the same length but not necessarily the same sequence.

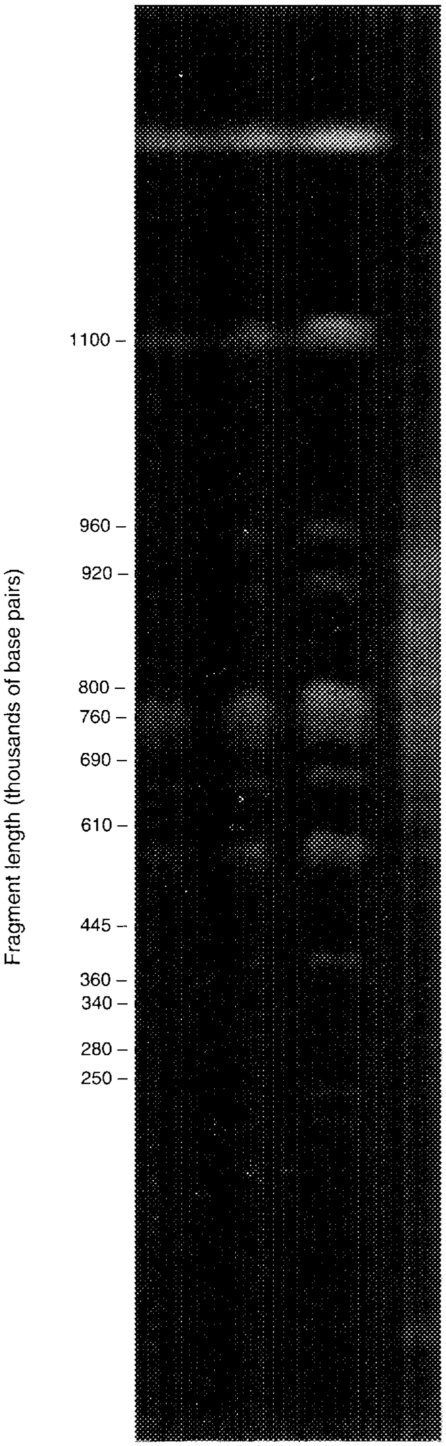
(b) Conventional Gel Electrophoresis of Fragmented Human DNA Segments



Shown in (b) are the results of conventional gel electrophoresis of six different samples of human DNA. Samples 1, 2, and 3 consisted of the restriction fragments produced by cutting the same cloned segment of human DNA with *Eco*RI alone (a 6-base cutter), with both *Eco*RI and *Hin*dIII (another 6-base cutter), and with *Hin*dIII alone, respectively. Samples 4, 5, and 6 consisted of the restriction fragments produced by cutting a different cloned segment of human DNA again with *Eco*RI alone, with both *Eco*RI and *Hin*dIII, and with *Hin*dIII alone, respectively. The leftmost lane of the gel contains fragments of the lengths indicated. Note that all the restriction fragments are well resolved.

Shown in (c) are the results of pulsed-field gel electrophoresis of three identical samples, each containing all sixteen of the intact DNA molecules that compose the genome of the yeast *Saccharomyces cerevisiae.* The four longest chromosomal DNA molecules are not resolved; all four are located in the topmost band. The remaining twelve chromosomal DNA molecules, however, are well resolved. The indicated lengths of the resolved DNA molecules were determined from the positions, in the rightmost lane of the gel, of the fragments in a calibration sample. Even longer fragments, fragments with lengths up to about 5 million base pairs, can be separated by increasing the duration of the pulses.

(c) Pulsed-field Electrophoresis of Intact DNA Molecules of *Saccharomyces cerevisiae*



DNA fragments are found in the gel at locations corresponding to lengths of 3400, 4900, 5300, 6000, 7900, and 22,000 base pairs. That set of six *Eco*RI restriction-fragment lengths is unique to the $\lambda$ genome and hence can be used as an identifying characteristic of the genome, a characteristic called its *Eco*RI restriction-fragment fingerprint. Only viral genomes can be fingerprinted with a 6-base cutter such as *Eco*RI. Complete digestion of the much larger bacterial and eukaryotic genomes with a 6-base cutter yields so many restriction fragments that gel electrophoresis produces an essentially continuous smear of fragments rather than a relatively small number of well-separated fragments. However, a short segment of a large genome can be fingerprinted with a 6-base cutter, provided many copies of the segment are available.

Note that the *Eco*RI restriction-fragment fingerprint of the $\lambda$ genome provides no information about the order of the restriction fragments along the $\lambda$ genome. More information is needed to order the fragments and thereby construct an *Eco*RI restriction-site map of the $\lambda$ genome, a map showing the distances between its *Eco*RI restriction sites. One way to get the additional information is to carry out two digestions, one of which is complete and the other only partial. The complete digestion produces fragments such that the length of each is equal to the distance between some two adjacent restriction sites; the partial digestion produces some fragments such that the length of each is equal to the distance spanned by three or more adjacent restriction sites. Together the length data obtained from the two digestions provide sufficient information to order the fragments and construct the restriction-site map.

The restriction-fragment fingerprints of cloned segments of a large genome have found application in the efforts to "map" the segments, that is, to arrange the segments in the order in which they appear along the genome. The principle behind this application is as follows. Suppose that the restriction-fragment fingerprints of two segments of a genome include a number of restriction-fragment lengths in common. Calculations based on the distribution of restriction sites along the genome and on the number of restriction-fragment lengths in common lead to a value for the probability that the two fragments overlap and therefore contain pieces of DNA that are contiguous along a chromosomal DNA molecule. (See "Physical Mapping—A One-Dimensional Jigsaw Puzzle" in "Mapping the Genome.")

This discussion of gel electrophoresis concludes by noting that the electric field used to carry out the procedure is usually a constant electric field. However, in such a field long DNA fragments (fragments longer than about 50,000 base pairs) tend to become trapped at arbitrary locations in the gel and thus do not migrate through the gel in a length-dependent manner. But fragments that long or longer are of interest, and separating them by length is sometimes desirable. For example, making a *Not*I restriction-site map of a human chromosome involves gel electrophoresis of restriction fragments that are on average 1,000,000 base pairs long. (*Not*I is an 8-base cutter; the estimated average length of the fragments it produces, namely $4^8 = 65,536$ base pairs, differs considerably from the observed average length because the recognition sequence of that restriction enzyme includes several occurrences of the dinucleotide sequence 5′-CG, which happens to be rare in mammalian genomes. *Not*I is one of a group of "infrequent cutters," all of which contain at least one occurrence of the sequence 5′-CG and produce fragments with average lengths ranging from 100,000 base pairs to 1 million base pairs.) Length separation of long fragments can be accomplished by using an electric field that varies intermittently in direction but has a time-averaged direction along the length of the gel. Such a "pulsed" field allows long DNA fragments to wind their way through the molecular framework of the gel. As shown in "Gel Electrophoresis," pulsed-field electrophoresis can separate even the very long DNA molecules extracted intact from yeast chromosomes. (Note that pulsed-field gel electrophoresis of long fragments requires preparation of the DNA sample by special methods because the accidental fragmentation involved in the method described at the beginning of this section cannot be tolerated when DNA molecules are to be studied either intact or as the long, reproducibly cut fragments produced by a restriction enzyme such as *Not*I.)

**Amplifying DNA.** Most of the techniques currently used to analyze a segment of DNA require the availability of many copies of the segment. Two methods for "amplifying" a DNA segment are now at hand: molecular cloning, which was developed in the 1970s, and the polymerase chain reaction (PCR), which was developed less than a decade ago.

*Amplification by Molecular Cloning.* Molecular cloning involves replication of a foreign DNA segment by a host organism, usually the bacterium *E. coli*. However, a segment of DNA that has entered an *E. coli* cell will not be replicated by the cell unless the segment has first been combined with a cloning "vector," a DNA molecule that the cell does replicate. The combination of the segment to be cloned, the "insert," and the vector is called a recombinant DNA molecule.

The phenomenon of transduction, discovered in 1952, had shown that DNA from the genome of one strain of *E. coli* is sometimes incorporated into the genome of a phage without affecting the ability of the phage to be replicated in another strain of *E. coli*. In other words, the phage genome was known to act as a vector, a DNA molecule that carries foreign DNA into a host cell, where it is then replicated. Nevertheless, the earliest cloning vectors were plasmids, small DNA molecules found in and replicated by bacteria. (Plasmids, like the genomes of bacteria, are circular DNA molecules. They are, however, much smaller than bacterial genomes. Some plasmids are replicated only when their hosts replicate and occur as single copies. The replication of other plasmids is not coordinated with host-cell replication; such plasmids occur as multiple copies.) The plasmid first used was one of a number that had been studied intensively because they contain genes that confer on the bacteria in which they reside the ability to survive in the presence of antibiotics. Today two vectors in addition to phage genomes and plasmids are also widely used: cosmids, which are replicated in *E. coli*, and yeast artificial chromosomes (YACs), which are replicated in the single-celled eukaryotic organism *Saccharomyces cerevisiae* (baker's yeast). Both cosmids and YACs are synthetic rather than naturally occurring DNA molecules.

The first step in molecular cloning is to make the recombinant DNA molecules in vitro. The following is a description of the procedure employed when the vector is a plasmid that contains a single restriction site for *Eco*RI embedded within a gene for resistance to ampicillin. Digestion of a population of such plasmids with *Eco*RI produces "linearized" plasmids with sticky ends. Inserts with identical sticky ends are formed by digesting the DNA to be cloned also with *Eco*RI. When the linearized plasmids and the inserts are mixed together, along with an enzyme called a DNA ligase, the sticky ends of some inserts hydrogen bond to the sticky ends of the linearized plasmids. The backbones of such hydrogen-bonding products are then covalently linked by the DNA ligase into recombinant DNA molecules (here recombinant plasmids). Note that the ligation mixture also contains some nonrecombinant plasmids because some linearized plasmids simply recyclize.

A more detailed description of the making of recombinant DNA molecules with plasmids and other vectors is presented in the article "DNA Libraries." Here we point out only that different vectors are used to clone inserts of different lengths. Plasmids carry inserts that are usually about 4000 base pairs long, $\lambda$ phages carry inserts that are usually four to five times longer, and YACs carry inserts that are usually more than one hundred times longer. (The great lengths of the inserts carried by YACs implies that YAC cloning, like pulsed-field gel electrophoresis, requires a special method of DNA preparation.)

The next step in molecular cloning with plasmids is to expose a population of *E. coli* cells to the ligation mixture in the hope that one recombinant plasmid will enter each of a reasonable fraction of the cells. Entry of a plasmid into an *E. coli* cell is said to transform the cell, provided the plasmid is replicated by the cell. The mechanism by which a plasmid (or a YAC) enters a host cell is not completely understood, but several empirical methods have been found that increase the efficiency of transformation (number of cells transformed per unit mass of recombinant DNA molecules). In contrast, the mechanism by which a phage enters (infects) a host cell is fairly well understood and is inherently more efficient.

After the *E. coli* cells have been exposed to the ligation mixture, the solution containing the exposed cells is diluted, a small amount of the diluted solution is transferred to each of a number of culture dishes containing a solid growth medium, and the cells are allowed to divide. (Dilution of the exposed cells assures that only a relatively small number of cells is transferred to each culture dish.) The aggregate, or colony, of cells produced by successive divisions of a single cell is called a clone of the single cell. Each member of a clone that arises from a transformed cell contains at least one copy of the plasmid and, if the transforming plasmid was a recombinant plasmid, at least one copy of the insert.

Because the goal of molecular cloning is not only to obtain many copies of the insert within a recombinant DNA molecule but also to do so in as short a time as possible, one criterion for a host cell is a short generation time. The generation times of both *E. coli* and yeast are suitably short. For example, the generation time of *E. coli* is about 20 minutes. Thus a single *E. coli* cell can, under suitable conditions, multiply into more than a billion cells in about 10 hours.

The final step in plasmid cloning is to identify the clones arising from cells transformed by recombinant plasmids. Recall that the *Eco*RI restriction site of the plasmid used in this example lies within its ampicillin-resistance gene. Assume that each host cell itself contained a plasmid carrying a gene for resistance to ampicillin. Then only those clones that arose from cells transformed by a recombinant plasmid possess an *inoperative* ampicillin-resistance gene (because the insert interrupts the gene). Using that fact to identify the clones of interest involves transferring a portion of each clone from the culture dish to some other vessel in a manner that preserves the positions of the clones. Ampicillin is then added to the other vessel, and the positions of the clones that die are noted. The clones at the corresponding positions on the culture dish are the clones desired. Other ingenious tricks have been devised to identify the desired clones.

The sample of DNA to be cloned usually consists of many different fragments, all from the same source. Examples are the large sets of fragments obtained by cutting, say, the mouse genome or the human X chromosome with a restriction enzyme. Then each recombinant DNA molecule contains a different fragment of the source DNA, and each host cell entered by a recombinant DNA molecule gives rise to a clone of a different fragment. A collection of such clones is called a DNA library—a mouse-genome DNA library, say, or a human-X-chromosome DNA library. The article "DNA Libraries" describes molecular cloning more fully and discusses the problems it presents.

*Amplification by PCR.* Unlike cloning, the polymerase chain reaction is carried out entirely in vitro and, more important, is capable of amplifying a specific one of the many fragments that may be present in a DNA sample. The selectivity of the reaction implies that it is also a means for detecting the presence of the fragment being amplified. Details of the reaction are presented in "The Polymerase Chain Reaction and Sequence-tagged Sites" in "Mapping the Genome."

**Sequencing DNA.** The ultimate in detailed information about a fragment of DNA is its base sequence. The process of obtaining that information is called sequencing. Two sequencing methods were developed in 1977, both based on essentially the same principle but each realizing the goal in a different way. Let $b_1b_2b_3 \ldots b_N$ be the base sequence of the fragment to be sequenced. Consider the set of subfragments $\{b_1, b_1b_2, b_1b_2b_3, \ldots, b_1b_2b_3 \ldots b_N\}$. Assume that such a set of subfragments can be generated and, equally important, can be separated into four subsets: the subset A consisting of those subfragments that end in the base A; the subset C consisting of those subfragments that end in C; the subset G consisting of those subfragments that ends in the base G; and the subset T consisting of those subfragments that end in the base T. Note that together the four subsets compose the set $\{b_1, b_1b_2, b_1b_2b_3, \ldots, b_1b_2b_3 \ldots b_N\}$. The subsets A, C, G, and T are subjected to electrophoresis, each in a different "lane" of a gel (a different strip of gel parallel to the direction of the applied electric field). After electrophoresis each subfragment is located in one of the four lanes according to its length. Suppose that the shortest subfragment, $b_1$, appears in the A lane of the gel; that the next longer subfragment, $b_1b_2$, appears in the T lane; that the next longer subfragment, $b_1b_2b_3$, appears in the G lane; . . . ; and that the longest subfragment, $b_1b_2b_3 \ldots b_N$, appears in the T lane. Then the base sequence of the fragment is ATG . . . T.

Obviously the above description of the principle of the two sequencing methods has avoided the question of how the four subsets of subfragments are generated. The procedures for doing so are described in "DNA Sequencing" in "Mapping the Genome."

Although sequencing is still a tedious and expensive process, the information so obtained is crucial to identification of the DNA mutations that cause inherited disorders and to a broad understanding of the functioning and evolution of genes and genomes. Much effort is being devoted to increasing the speed and decreasing the cost of current sequencing methods and to searching for new methods.

**Hybridization: Detecting the Presence of Specific DNA Sequences.** The two single-stranded DNA fragments produced by denaturation of a (double-stranded) DNA fragment will, under appropriate conditions, renature (form a double-stranded fragment by hydrogen bonding) because the single-stranded fragments are complementary along the entirety of their lengths. (Recall that two single-stranded fragments are complementary if and only if the 5′-to-3′ base sequence of one is the complement of the 3′-to-5′ base sequence of the other.) Similarly, hydrogen bonding between an RNA fragment and a complementary single-stranded DNA fragment will form a double-stranded DNA-RNA fragment, a phenomenon called hybridization. (Hybridization between the RNA transcript of an *E. coli* gene and the template strand of the gene was the technique used in the 1960s to measure the rates of transcription of various *E. coli* genes.) The term "hybridization" now also includes the hydrogen bonding that occurs between any two single-stranded nucleic-acid fragments that are complementary along only some portion (usually a relatively short portion) of their lengths.

Hybridization is widely used to detect the presence of a particular DNA segment in a sample of DNA. If the sample consists of a set of cloned DNA fragments, each cloned fragment is denatured and then allowed to interact with a solution containing many copies of a radioactively labeled "probe," a relatively short stretch of single-stranded DNA whose sequence is identical to or complementary to some unique portion of the segment of interest. Under the right conditions the probe hybridizes only to the cloned fragment (or fragments) that contains the segment of interest, and the radioactivity of the probe identifies the fragment to which the probe has hybridized. For example, suppose that the sample is a complete set of cloned human DNA fragments and the segment of interest is the interspersed tandem repeat $(5'\text{-GT})_n$. Examples of a probe for that segment are the single-stranded fragments with the sequences $(5'\text{-AC})_7$ and $(5'\text{-GT})_7$. Because the segment $(5'\text{-GT})_n$ appears at numerous locations in the human genome, such a probe hybridizes to numerous cloned fragments but only to those containing the interspersed tandem repeat (or a portion thereof). If the sample to be interrogated with a probe is instead a solution containing many different DNA fragments, the fragments must first be separated and immobilized, usually by gel electrophoresis. If the probe is sufficiently short, hybridization can be carried out directly on the gel. Usually, however, the length-separated fragments are first transferred from the gel to a nitrocellulose filter. The procedure, called Southern (or gel-transfer) hybridization, is illustrated in "Hybridization Techniques."

In-situ hybridization is a variation of hybridization in which the sample to be interrogated with a probe consists of the intact DNA molecules within metaphase chromosomes. The metaphase chromosomes are spread out on a microscope slide and partially denatured. The probe copies are labeled with a fluorescent molecule and allowed to interact with the denatured chromosomes. The presence of bound probe is detected by observing the chromosomes with a fluorescence microscope. An example of the fluorescence signal obtained by using the technique is shown in "Hybridization Techniques." In-situ hybridization provides information about which chromosome contains the segment of interest and its approximate location on the chromosome.

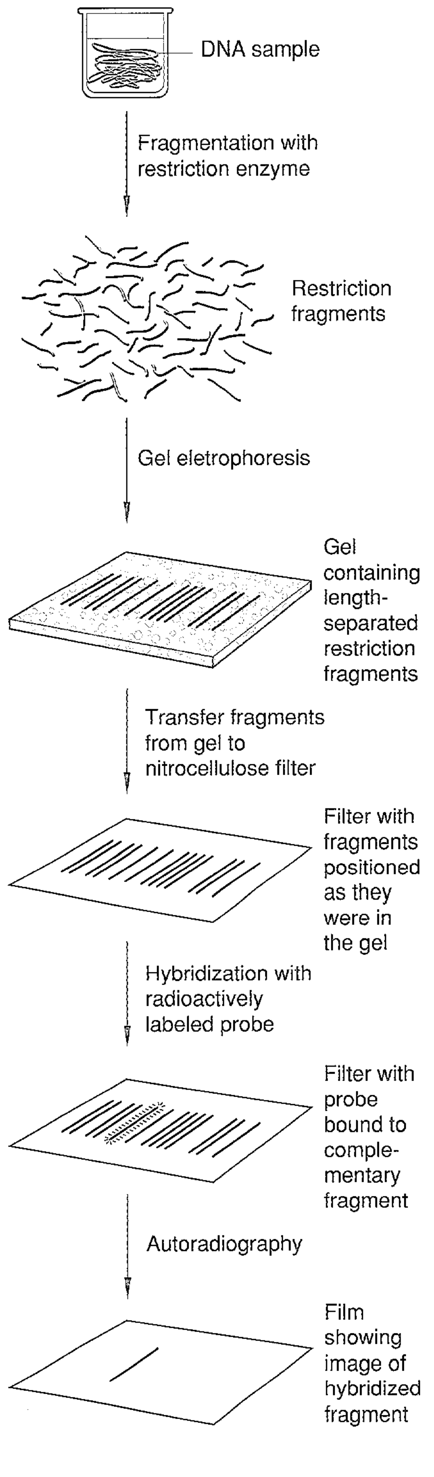
This section on the techniques of molecular genetics concludes with an application that not only requires the use of almost all the techniques described but also is of particular significance to the efforts to arrange cloned fragments of human DNA in the same order as they appear in the intact DNA molecules of human chromosomes. The application involves the use of long cloned fragments of human DNA to obtain an upper limit on the length of the segment of DNA that separates the chromosomal locations of any two short cloned fragments of human DNA (such as those provided by plasmid, phage, or cosmid cloning). The long fragments, which are produced by cutting human genomic DNA with an infrequent cutter, are subjected to pulsed-field gel electrophoresis and then to Southern hybridization. Two different probes are used separately in the hybridization; each is unique to one of the two short cloned fragments. If both probes hybridize to the same long fragment, then both short fragments lie within the long fragment. In other words, the chromosomal locations of the short fragments are separated by a length of DNA no longer than the length of the long fragment to which both probes hybridized.

# HYBRIDIZATION TECHNIQUES

Southern hybridization is a technique for identifying, among a sample of many different DNA fragments, the fragment(s) containing a particular nucleotide sequence. As depicted in (a), the sample has typically been fragmented with a restriction enzyme. The restriction fragments are subjected to gel electrophoresis to separate them by length and immobilize them. The length-separated fragments are then transferred to a filter paper made of nitrocellulose, a procedure called blotting. (Note that blotting preserves the locations of the fragments.) The filter is washed first with a solution that denatures the fragments and then with a solution containing many copies of a radioactively labeled, single-stranded "probe" whose sequence is identical to or complementary to some unique portion of the sequence of interest. The probe hybridizes (hydrogen bonds) to only the denatured fragments containing the complement of its sequence and hence the sequence of interest. The unbound probe is washed away, and the filter is dried and placed in contact with x-ray film. The radioactivity of the bound probe exposes the film and creates an image, an autoradiogram, of the fragment(s) to which the probe has bound. Southern hybridization is particularly useful for detecting variations among different members of a species in the lengths of the restriction fragments originating from a particular region of the organism's genome (see "Modern Linkage Mapping with Polymorphic DNA Markers" in "Mapping the Genome").

The number of fragments "picked out" by a probe depends on the number of times the sequence of interest occurs in the sample DNA. If the sequence occurs only once (if a probe for, say, a single-copy gene is being used), the probe picks out one or at most two fragments (provided the probe is shorter than any of the fragments in the sample). On the other hand, if the sequence of interest occurs more than once (if a probe for a multiple-copy gene or a repeated sequence is being used), the probe picks out a larger number of fragments. Furthermore, the hybridization conditions (temperature and salinity of the probe solution) can be adjusted so that either exact complementarity or a lesser degree of complementarity is required for binding of the probe.

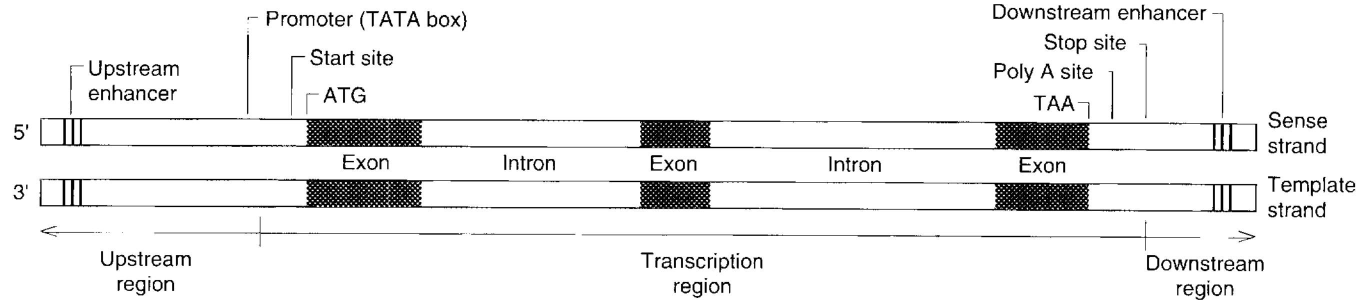
(a) Southern Hybridization


DNA sample
Fragmentation with restriction enzyme
Restriction fragments
Gel eletrophoresis
Gel containing length-separated restriction fragments
Transfer fragments from gel to nitrocellulose filter
Filter with fragments positioned as they were in the gel
Hybridization with radioactively labeled probe
Filter with probe bound to complementary fragment
Autoradiography
Film showing image of hybridized fragment

In-situ hybridization is a variation of hybridization in which the sample consists of the complement of chromosomes within a cell arrested at metaphase. The metaphase chromosomes are spread out and partially denatured on a microscope slide, the probe is labeled with a fluorescent dye, and the bound probe is imaged with a fluorescence microscope. Shown in (b) is the fluorescence signal resulting from in-situ hybridization of a probe for the human telomere to human metaphase chromosomes. (A telomere is a special sequence at each end of a eukaryotic DNA molecule that protects the molecule from enzymatic degradation and prevents shortening of the molecule as it is replicated. The sequence of the human telomere was discovered by Robert K. Moyzis and his colleagues, who also provided evidence that all vertebrates share the same telomeric sequence. Note that, as expected, the probe has bound only to the terminal regions of each chromosome. (Micrograph courtesy of Julie Meyne.)

(b) Results of In-Situ Hybridization of Human-Telomere Probe to Human Chromosomes

# THE ANATOMY OF A EUKARYOTIC PROTEIN GENE



Each eukaryotic gene is placed in one of three classes according to which of the three eukaryotic RNA polymerases is involved in its transcription. The genes for RNAs are transcribed by RNA polymerases I and III. The genes for proteins, the class first brought to mind by the word "gene" and the class focused on here, are transcribed by RNA polymerase II (*pol* II).

Shown above are the components of a prototypic protein gene. By convention the sense strand of the gene, the strand with the sequence of DNA bases corresponding to the sequence of RNA bases in the primary RNA transcript, is depicted with its 5′-to-3′ direction coincident with the left-to-right direction. (Often only the sense strand of a gene is displayed.) The left-to-right direction thus coincides with the direction in which the template strand is transcribed. The terms "upstream" and "downstream" describe the location of one feature of a gene relative to that of another. Their meanings in that context are based on regarding transcription as a directional process analogous to the flow of water in a stream.

The start site is the location of the first deoxyribonucleotide in the template strand that happens to be transcribed. It defines the beginning of the transcription region of the gene. Note that the start site lies upstream of the DNA codon (ATG) corresponding to the RNA codon (AUG) that signals the start of translation of the transcribed RNA. The transcription region ends at some nonspecific deoxyribonucleotide between 500 and 2000 base pairs downstream of the poly A site. Within the poly A site are sequences that, when transcribed, signal the location at which the primary RNA transcript is cleaved and equipped with a "tail" composed of a succession of ribonucleotides containing the base A. (The poly A tail is thought to aid the transport of messenger RNA from the nucleus of a cell to the cytoplasm.) Note that the poly A site lies downstream of the DNA codon (here TAA) corresponding to one of the RNA codons (UAA) that signals the end of translation of the transcribed RNA.

Within the transcription region are exons and introns. Exons tend to be about 300 base pairs long; each is a succession of codons uninterrupted by stop codons. Introns, on the other hand, are not uninterruped successions of codons, and the RNA segments transcribed from introns are spliced out of the primary RNA transcript before translation. A few protein genes contain no introns (the human α–interferon gene is an example), most contain at least one, and some contain a large number (the human thyroglobulin gene contains about forty). Generally the amount of DNA composing the introns of a protein gene is far greater than the amount composing its exons.

Close upstream of the start site is a promoter sequence, where *pol* II binds and initiates transcription. A common promoter sequence in eukaryotic genes is the so-called TATA box, which has the consensus sequence 5′-TATAAA and is located at a variable short distance (about 30 base pairs) upstream of the start site.

The region upstream of the promoter and, less frequently, the downstream region or the transcription region itself contain sequences that control the rate of initiation of transcription. Although expression of a protein gene is regulated at a number of stages in the pathway from gene to protein, control of replication initiation is the dominant regulatory mechanism. (Primary among the other regulatory mechanisms is control of splicing.) The regulated expression of a gene (the when, where, and degree of expression) is the key to phenotypic differences between the various cells of a multicellular organism and also between organisms that possess similar genotypes.

Initiation of transcription is controlled mainly by DNA sequences (*cis* elements) and by certain proteins, many but not all of which are sequence-specific DNA-binding proteins (*trans*-acting transcription factors). Thus both temporal and cellular specificities of transcription control are governed by the availability of the different *trans*-acting transcription factors. Interactions of transcription factors with *cis* elements and with each other lead to formation of complex protein assemblies that control the ability of *pol* II to initiate transcription. Most of the complexes enhance transcription initiation, but some act as repressors. Enhancers and repressors can be located as far as 10,000 base pairs away from the transcription region.

Class I and class III genes differ from protein genes not only in their anatomies but also in the promoters, *cis* elements, and *trans*-acting factors involved in their transcription.

## Genes and Genomes: What the Future Holds

The techniques described in the preceding section, and others not mentioned, have greatly increased our knowledge of the molecular anatomies of genes. Previously, a gene for a protein was defined narrowly as a segment of DNA that is transcribed into a messenger RNA, which in turn is translated into the protein. The definition considered more appropriate today includes not only the protein-coding segment of the gene (its transcription region) but also its sometimes far-flung regulatory regions (see "The Anatomy of a Eukaryotic Protein Gene"). The regulatory regions contain DNA sequences that help determine whether and at what rate the gene is expressed (or, equivalently, the protein is synthesized). Some of the genes of a multicellular organism, its "housekeeping" genes, are expressed at more or less the same level in essentially all of its cells, regardless of type. Others are expressed only in certain types of cells or only at certain times. Gene regulation is, in fact, the key not only to appropriate functioning of the organism but also to its development from a single cell. In addition, gene regulation may also be responsible for the striking phenotypic differences between higher apes and humans despite the negligible differences between the structures of their proteins. "The Anatomy of a Eukaryotic Protein Gene" presents also a few details about the mechanisms of gene regulation.

Despite the accumulating knowledge, it is safe to say that what is known about genes, particularly human genes, is far less than what remains to be learned. The total number of human genes can now be only crudely estimated, remarkably few have been localized to particular regions of particular chromosomes, and even fewer have been sequenced or studied in sufficient detail to understand their regulation. Other outstanding questions include the mechanisms by which the expression of genes is coordinated and the effects of gene mutations on morphology, physiology, and pathology.

The techniques of molecular genetics are also providing information about genomes as a whole, opening the way to comparative studies of genome anatomy, organization, and evolution. For example, the available evidence indicates remarkable similarities between the mouse genome and the human genome, despite the 60 million years that have elapsed since rodents and primates diverged from a common ancestor. The similarities lie not only in the base sequences of genes but also in their linkages. Perhaps the conserved linked genes represent units of some higher, as yet unknown operational feature. The same may be true also of repetitive DNA, about which we now know so little. In time, when those and other genomes have been sequenced in their entireties, the observed similarities and differences will be a rich source of answers and new questions about the operation and evolution of genomes. ■

## Further Reading

James A. Peters, editor. 1964. *Classic Papers in Genetics.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

J. Herbert Taylor, editor. 1965. *Selected Papers on Molecular Genetics.* New York: Academic Press.

John Cairns, Gunther S. Stent, and James D. Watson, editors. 1966. *Phage and the Origins of Molecular Biology.* Cold Spring Harbor, New York: Cold Spring Harbor Laboratory of Quantitative Biology.

John C. Kendrew. 1968. *The Thread of Life: An Introduction to Molecular Biology.* Cambridge, Massachusetts: Harvard University Press.

René J. Dubos. 1976. *The Professor, the Institute, and DNA.* New York: The Rockefeller University Press.

Franklin H. Porugal and Jack S. Cohen. 1977. *A Century of DNA: A History of the Discovery of the Structure and Function of the Genetic Substance.* Cambridge, Massachusetts: The MIT Press.

Horace Freeland Judson. 1979. *The Eighth Day of Creation.* New York: Simon and Schuster.

James D. Watson. 1980. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA.* New York: W. W. Norton and Co.

James D. Watson and John Tooze. 1981. *The DNA Story: A Documentary History of Gene Cloning.* San Francisco: W. H. Freeman and Company.

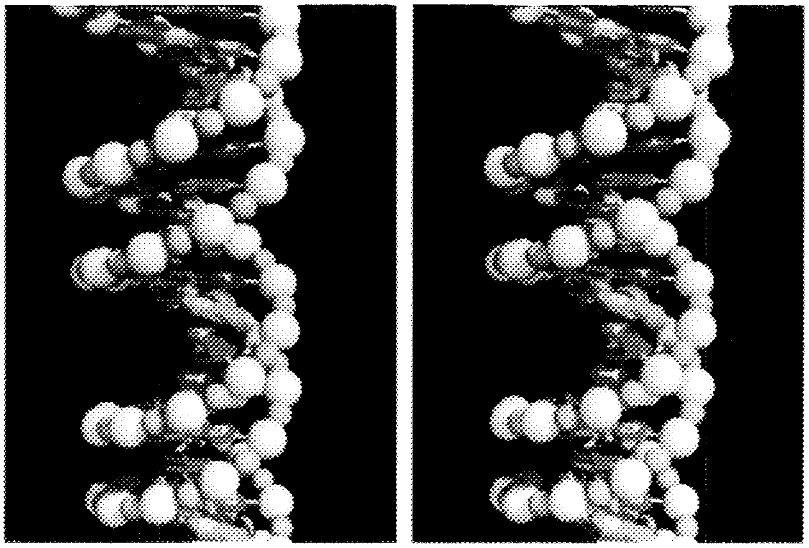
James D. Watson, Nancy H. Hopkins, Jeffrey W. Roberts, Joan Argetsinger Steitz, and Alan M. Weiner. 1987. *Molecular Biology of the Gene.* Menlo Park, California: The Benjamin/Cummings Publishing Company, Inc.

David A. Micklos and Greg A. Freyer. 1990. *DNA Science: A First Course in Recombinant DNA Technology.* New York: Cold Spring Harbor Laboratory Press.

James Darnell, Harvey Lodish, and David Baltimore. 1990. *Molecular Cell Biology*, second edition. New York: W. H. Freeman and Company.

Maxine Singer and Paul Berg. 1991. *Genes & Genomes: A Changing Perspective.* Mill Valley, California: University Science Books.

**Robert P. Wagner** is a consultant to the Laboratory's Life Sciences Division and Professor Emeritus of Zoology at the University of Texas, Austin, the institution from which he received his Ph.D. His work at the Laboratory focuses on the activities of the Center for Human Genome Studies. He has taught undergraduate and graduate genetics for over thiry-five years and has authored or co-authored six books and many research and review articles on various aspects of genetics. His numerous honors and awards include fellowships from the National Research Council and the Guggenheim Foundation and election as a fellow of the American Association for the Advancement of Science and as president of the Genetics Society of America.



To create a stereoscopic image of DNA from the two images on this page, focus your eyes on a distant object above the page and then move the images up into your line of sight, holding the page 12 to 18 inches away and being careful to keep your eyes focused at infinity. If your eyes have not shifted, you should be aware of three images. Concentrate on the middle one, which is the desired stereoscopic image. You may have to practice a few times and should be sure the page and your head are vertical.