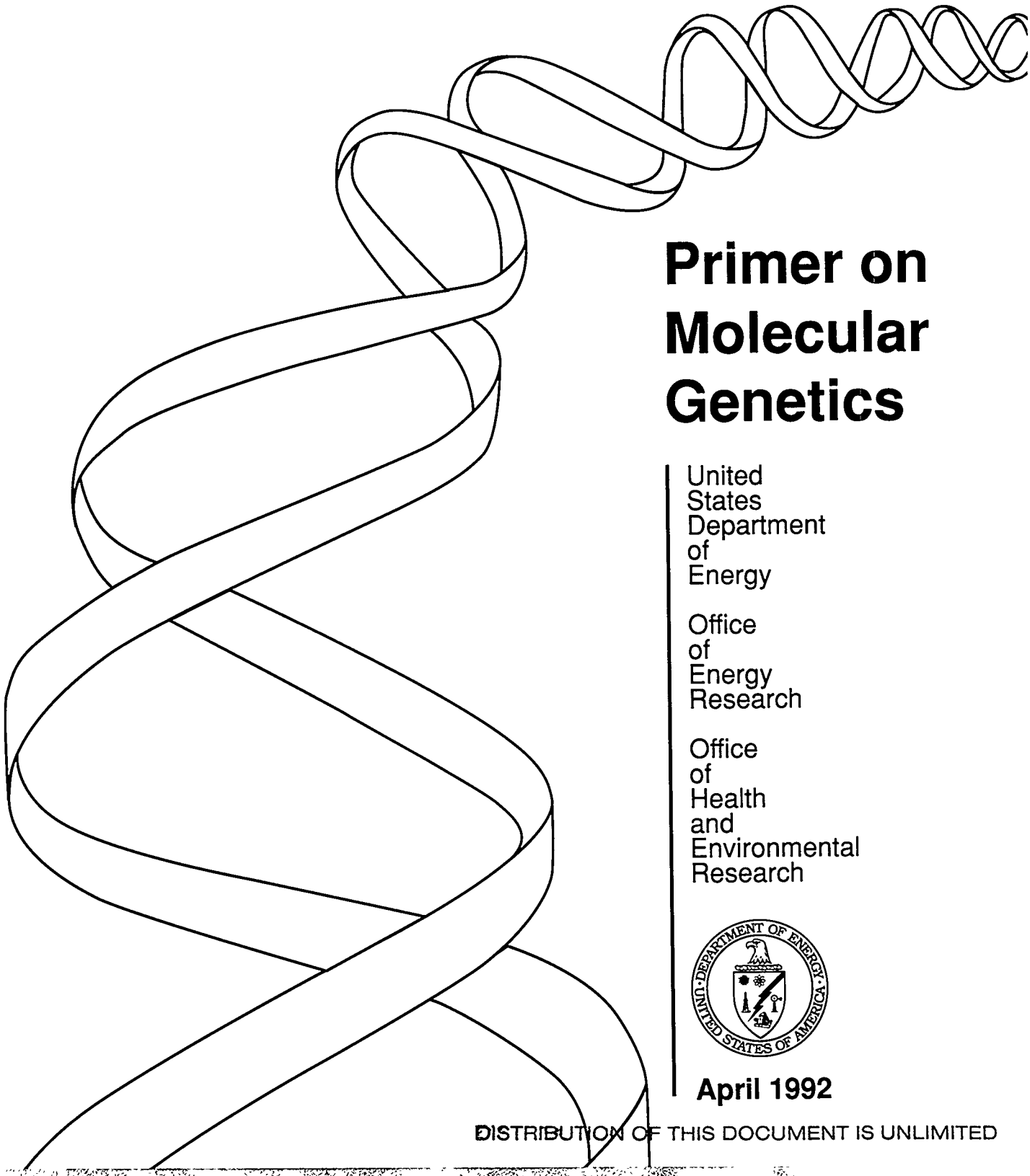
2

DOE

Human Genome Program



Primer on Molecular Genetics

United States Department of Energy

Office of Energy Research

Office of Health and Environmental Research

April 1992

The "Primer on Molecular Genetics" is taken from the April 1992 draft of the DOE *Human Genome 1991–92 Program Report*, which is expected to be published in May 1992. The primer is intended to be an introduction to basic principles of molecular genetics pertaining to the genome project. The material contained herein is not final and may be incomplete.

For additional copies of the primer or the completed report when published, contact:

Human Genome Management Information System
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6050

Voice: 615/574-7582, FTS 624-7582
Fax: 615/574-9888, FTS 624-9888
Internet: "yustln@ornl.gov"
BITNET: "yustln@ornlstc.bitnet"

ORNL/M--2026

DE92 010680

DOE

Human Genome Program

Primer on Molecular Genetics

Date Published: April 1992

U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research
Washington, D.C. 20585

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Contents

Primer on Molecular Genetics

Revised and expanded by Denise Casey (HGMIS) from the primer contributed by Charles Cantor and Sylvia Spengler (Lawrence Berkeley Laboratory) and published in the *Human Genome 1989–90 Program Report.*

Introduction

The complete set of instructions for making an organism is called its genome. It contains the master blueprint for all cellular structures and activities for the lifetime of the cell or organism. Found in every nucleus of a person's 10 trillion cells, the human genome consists of tightly coiled threads of deoxyribonucleic acid (DNA) and associated protein molecules, organized into structures called chromosomes (Fig. 1).

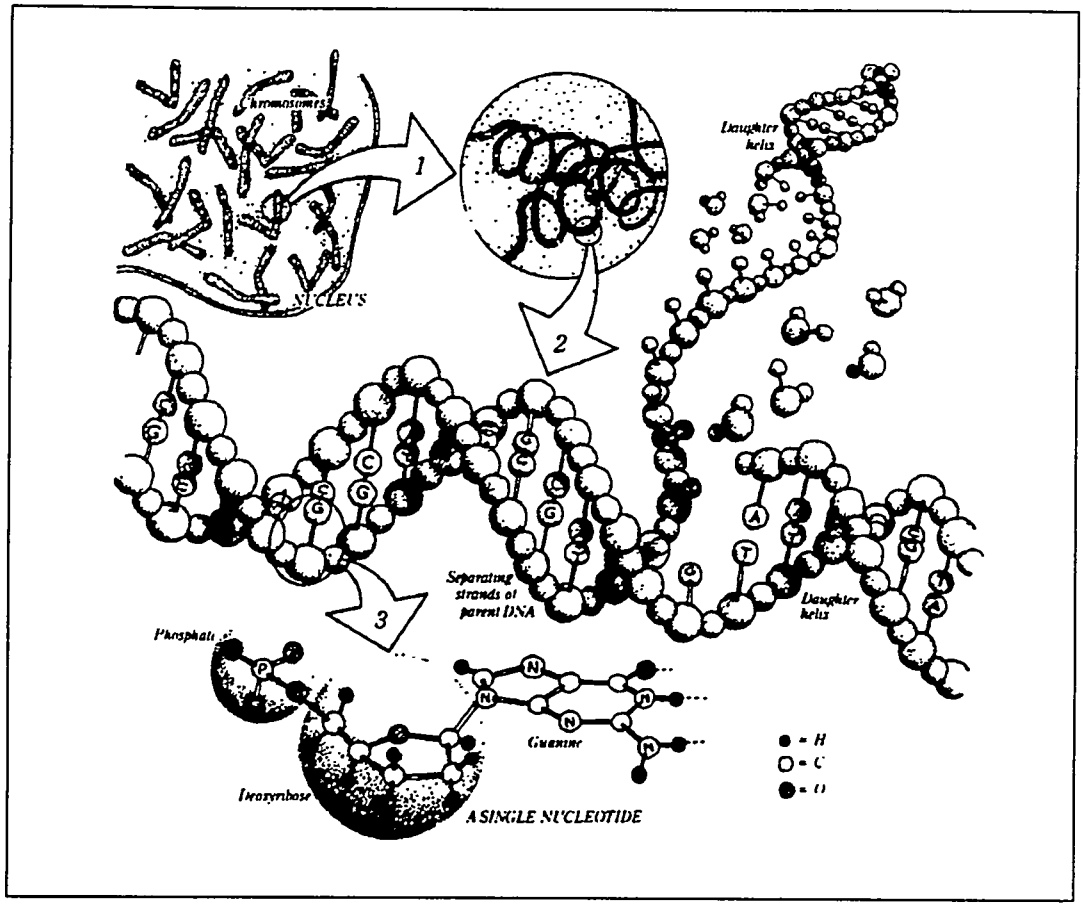


Fig. 1. The Human Genome at Four Levels of Detail. *Apart from reproductive cells (gametes) and mature red blood cells, every cell in the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA (1, 2). Each strand of DNA consists of repeating nucleotide units composed of a phosphate group, a sugar (deoxyribose), and a base (guanine, cytosine, thymine, or adenine) (3). Ordinarily, DNA takes the form of a highly regular double-stranded helix, the strands of which are linked by hydrogen bonds between guanine and cytosine and between thymine and adenine. Each such linkage is a base pair (bp); some 3 billion bp constitute the human genome. The specificity of these base-pair linkages underlies the mechanism of DNA replication illustrated here. Each strand of the double helix serves as a template for the synthesis of a new strand; the nucleotide sequence (i.e., linear order of bases) of each strand is strictly determined. Each new double helix is a twin, an exact replica, of its parent. (Figure and caption text provided by the LBL Human Genome Center.)*

Primer on Molecular Genetics

If unwound and tied together, the strands of DNA would stretch more than 5 feet but would be only 50 trillionths of an inch wide. For each organism, the components of these slender threads encode all the information necessary for building and maintaining life, from simple bacteria to remarkably complex human beings. Understanding how DNA performs this function requires some knowledge of its structure and organization.

DNA

In humans, as in other higher organisms, a DNA molecule consists of two strands that wrap around each other to resemble a twisted ladder whose sides, made of sugar and phosphate molecules, are connected by "rungs" of nitrogen-containing chemicals called bases. Each strand is a linear arrangement of repeating similar units called nucleotides, which are each composed of one sugar, one phosphate, and a nitrogenous base (Fig. 2). Four different bases are present in DNA—adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases arranged along the sugar-phosphate backbone is called the DNA sequence; the sequence specifies the exact genetic instructions required to create a particular organism with its own unique traits.

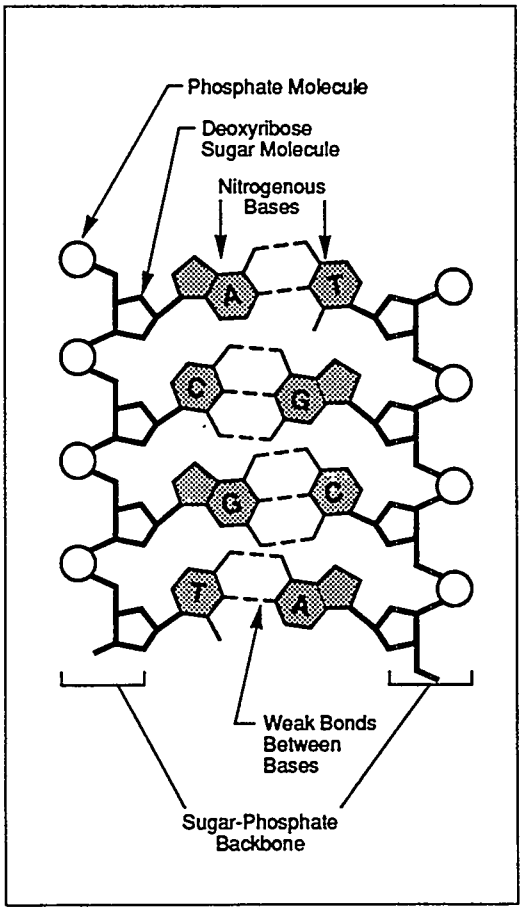


Fig. 2. DNA Structure. *The four nitrogenous bases of DNA are arranged along the sugar-phosphate backbone in a particular order (the DNA sequence), encoding all genetic instructions for an organism. Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G). The two DNA strands are held together by weak bonds between the bases. A gene is a segment of a DNA molecule (ranging from fewer than 1 thousand bases to several million), located in a particular position on a specific chromosome, whose base sequence contains the information necessary for protein synthesis.*

The two DNA strands are held together by weak bonds between the bases on each strand, forming base pairs (bp). Genome size is usually stated as the total number of base pairs; the human genome contains roughly 3 billion bp (Fig. 3).

Each time a cell divides into two daughter cells, its full genome is duplicated; for humans and other complex organisms, this duplication occurs in the nucleus. During cell division the DNA molecule unwinds and the weak bonds between the base pairs break, allowing the strands to separate. Each strand directs the synthesis of a complementary new strand, with free nucleotides matching up with their complementary bases on each of the separated strands. Strict base-pairing rules are adhered to—adenine will pair only with thymine (an A-T pair) and cytosine with guanine (a C-G pair). Each daughter cell receives one old and one new DNA strand (Figs. 1 and 4). The cell's adherence to these base-pairing rules ensures that the new strand is an exact copy of the old one. This minimizes the incidence of errors (mutations) that may greatly affect the resulting organism or its offspring.

Genes

Each DNA molecule contains many genes—the basic physical and functional units of heredity. A gene is a specific sequence of nucleotide bases, whose sequences carry the information required for constructing proteins, which provide the structural components of cells and tissues as well as enzymes for essential biochemical reactions. The human genome is estimated to comprise at least 100,000 genes.

Human genes vary widely in length, often extending over thousands of bases, but only about 10% of the genome is known to include the protein-coding sequences (exons) of genes. Interspersed within many genes are intron sequences, which have no coding function. The balance of the genome is thought to consist of other noncoding regions (such as control sequences and intergenic regions), whose functions are obscure. All living organisms are composed largely of proteins; humans can synthesize at least 100,000 different kinds. Proteins are large, complex molecules made up of long chains of subunits called amino acids. Twenty different kinds of amino acids are usually found in proteins. Within the gene, each specific sequence of three DNA bases (codons) directs the cell's protein-synthesizing machinery to add specific amino acids. For example, the base sequence ATG codes for the amino acid methionine. Since 3 bases code for 1 amino acid, the protein coded by an average-sized gene (3000 bp) will contain 1000 amino acids. The genetic code is thus a series of codons that specify which amino acids are required to make up specific proteins.

The protein-coding instructions from the genes are transmitted indirectly through messenger ribonucleic acid (mRNA), a transient intermediary molecule similar to a single strand of DNA. For the information within a gene to be expressed, a complementary RNA strand is produced (a process called transcription) from the DNA template in the nucleus. This

Comparative Sequence Sizes	Bases
• Largest known continuous DNA sequence (yeast chromosome 3)	350 Thousand
• *Escherichia coli* (bacterium) genome	4.6 Million
• Largest yeast chromosome now mapped	5.8 Million
• Entire yeast genome	15 Million
• Smallest human chromosome (Y)	50 Million
• Largest human chromosome (1)	250 Million
• Entire human genome	3 Billion

Fig. 3. Comparison of Known Sequence with Chromosome and Genome Sizes. *A comparison of the size of the largest DNA fragment for which the sequence has been determined, with approximate chromosome and genome sizes of model organisms and humans. A major focus of the Human Genome Project is the development of sequencing schemes that are faster and more economical.*

mRNA is moved from the nucleus to the cellular cytoplasm, where it serves as the template for protein synthesis. The cell's protein-synthesizing machinery then translates the codons into a string of amino acids that will constitute the protein molecule for which it codes (Fig. 5). In the laboratory, the mRNA molecule can be isolated and used as a template to synthesize a complementary DNA (cDNA) strand, which can then be used to locate the corresponding genes on a chromosome map. The utility of this strategy is described in the section on physical mapping, p. 11.

Chromosomes

The 3 billion bp in the human genome are organized into 24 distinct, physically separate microscopic units called chromosomes. All genes are arranged linearly along the chromosomes. The nucleus of most human cells contains 2 sets of chromosomes, 1 set given by each parent. Each set has 23 single chromosomes—22 autosomes and an X or Y sex chromosome. (A normal female will have a pair of X chromosomes; a male will have an X

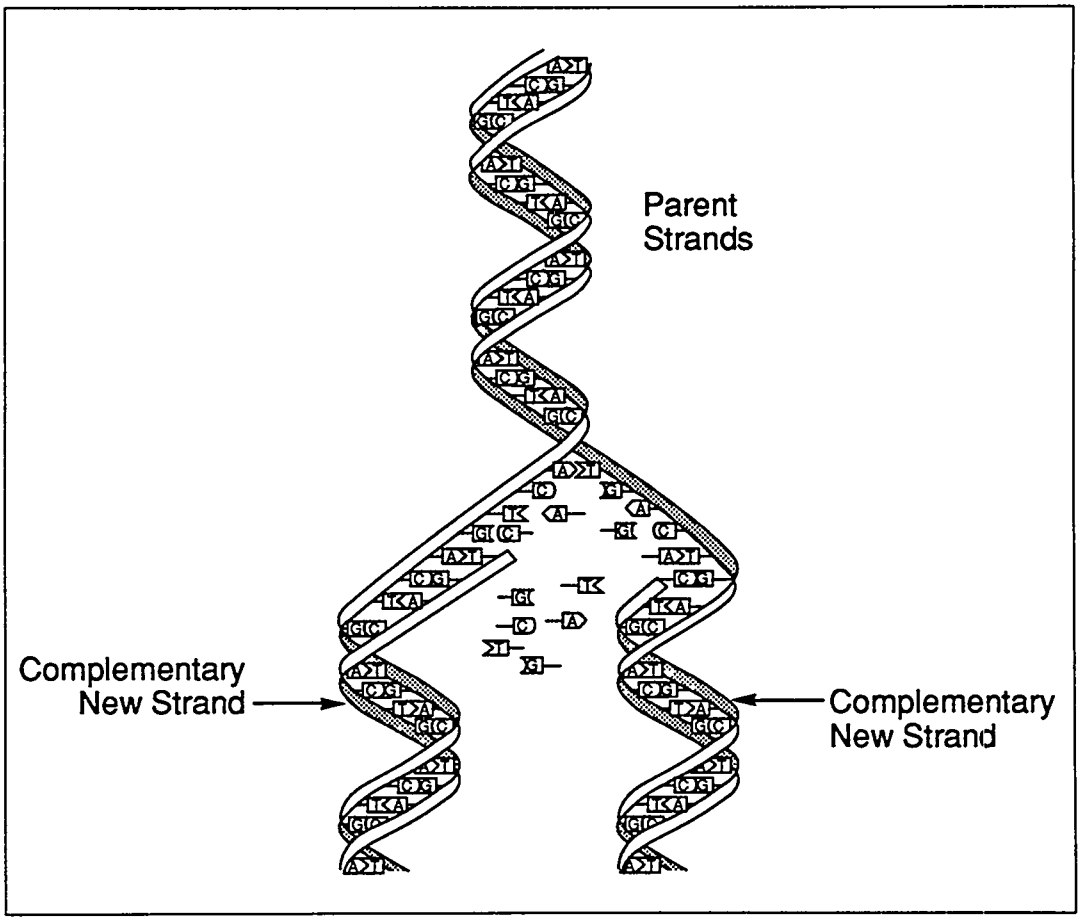


Fig. 4. DNA Replication. *During replication the DNA molecule unwinds, with each single strand becoming a template for synthesis of a new, complementary strand. Each daughter molecule, consisting of one old and one new DNA strand, is an exact copy of the parent molecule. [Source: adapted from* Mapping Our Genes—The Genome Projects: How Big, How Fast? *U.S. Congress, Office of Technology Assessment, OTA-BA-373 (Washington, D.C.: U.S. Government Printing Office, 1988).]*

and Y pair.) Chromosomes contain roughly equal parts of protein and DNA; chromosomal DNA contains an average of 150 million bases. DNA molecules are among the largest molecules now known.

Chromosomes can be seen under a light microscope and, when stained with certain dyes, reveal a pattern of light and dark bands reflecting regional variations in the amounts of A and T vs G and C. Differences in size and banding pattern allow the 24 chromosomes to be distinguished from each other, an analysis called a karyotype. A few types of major chromosomal abnormalities, including missing or extra copies of a chromosome or gross breaks and rejoinings (translocations), can be detected by microscopic examination; Down's syndrome, in which an individual's cells contain a third copy of chromosome 21, is diagnosed by karyotype analysis (Fig. 6). Most changes in DNA, however, are too subtle to be detected by this technique and require molecular analysis. These subtle DNA abnormalities (mutations) are responsible for many inherited diseases such as cystic fibrosis and sickle cell anemia or may predispose an individual to cancer, major psychiatric illnesses, and other complex diseases.

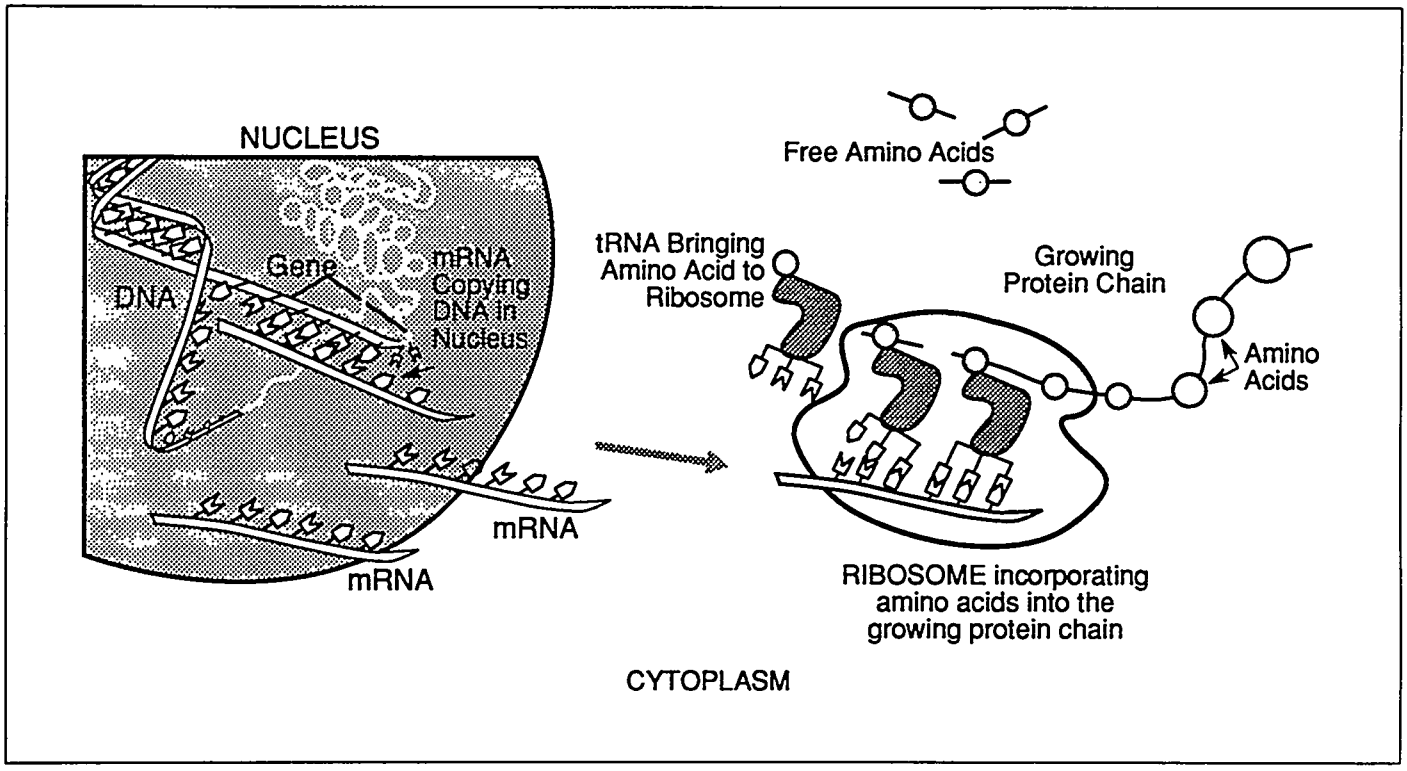


Fig. 5. Gene Expression. *When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA in a process similar to DNA replication. The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where triplets of bases (codons) forming the genetic code specify the particular amino acids that make up an individual protein. This process, called translation, is accomplished by ribosomes (cellular components composed of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein. (Source: see Fig. 4.)*

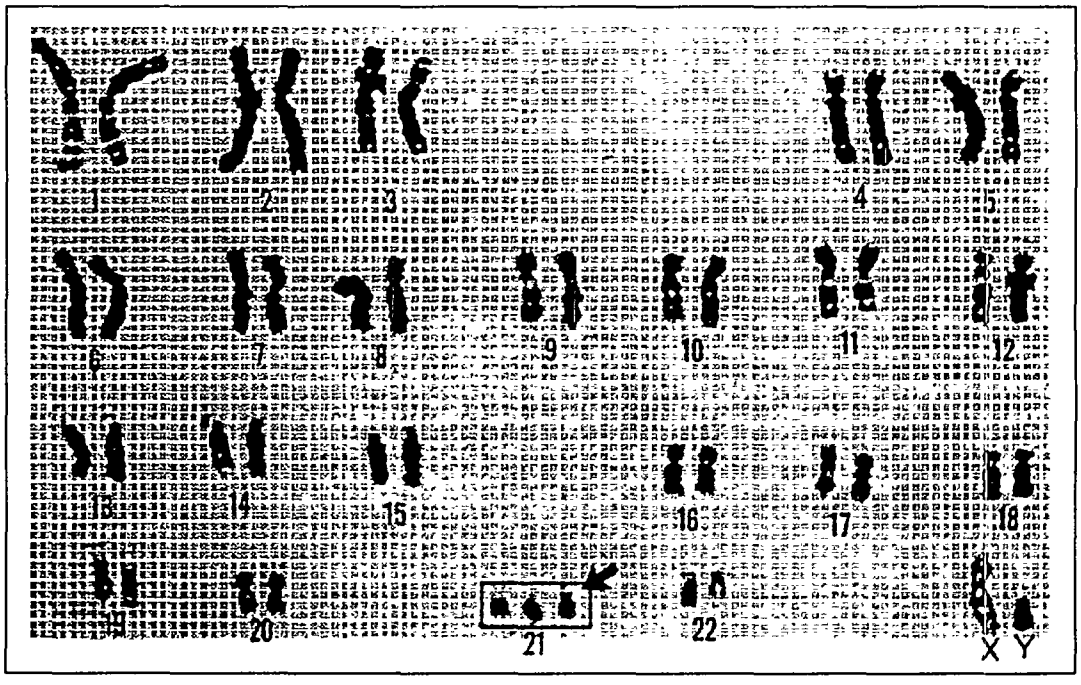


Fig. 6. Karyotype. Microscopic examination of chromosome size and banding patterns allows medical laboratories to identify and arrange each of the 24 different chromosomes (22 pairs of autosomes and one pair of sex chromosomes) into a karyotype, which then serves as a tool in the diagnosis of genetic diseases. The extra copy of chromosome 21 in this karyotype identifies this individual as having Down's syndrome.

Mapping and Sequencing the Human Genome

A primary goal of the Human Genome Project is to make a series of descriptive diagrams—maps—of each human chromosome at increasingly finer resolutions. Mapping involves (1) dividing the chromosomes into smaller fragments that can be propagated and characterized and (2) ordering (mapping) them to correspond to their respective locations on the chromosomes. After mapping is completed, the next step is to determine the base sequence of each of the ordered DNA fragments. The ultimate goal of genome research is to find all the genes in the DNA sequence and to develop tools for using this information in the study of human biology and medicine. Improving the instrumentation and techniques required for mapping and sequencing—a major focus of the genome project—will increase efficiency and cost-effectiveness. Goals include automating methods and optimizing techniques to extract the maximum useful information from maps and sequences.

A genome map describes the order of genes or other markers and the spacing between them on each chromosome. Human genome maps are constructed on several different scales or levels of resolution. At the coarsest resolution are genetic linkage maps, which depict the relative chromosomal locations of DNA markers (genes and other identifiable DNA sequences) by their patterns of inheritance. Physical maps describe the chemical characteristics of the DNA molecule itself.

Geneticists have already charted the approximate positions of over 2300 genes, and a start has been made in establishing high-resolution maps of the genome (Fig. 7). More-precise maps are needed to organize systematic sequencing efforts and plan new research directions.

Mapping Strategies

Genetic Linkage Maps

A genetic linkage map shows the relative locations of specific DNA markers along the chromosome. Any inherited physical or molecular characteristic that differs among individuals and is easily detectable in the laboratory is a potential genetic marker. Markers can be expressed DNA regions (genes) or DNA segments that have no known coding function but whose inheritance pattern can be followed. DNA sequence differences are especially useful markers because they are plentiful and easy to characterize precisely.

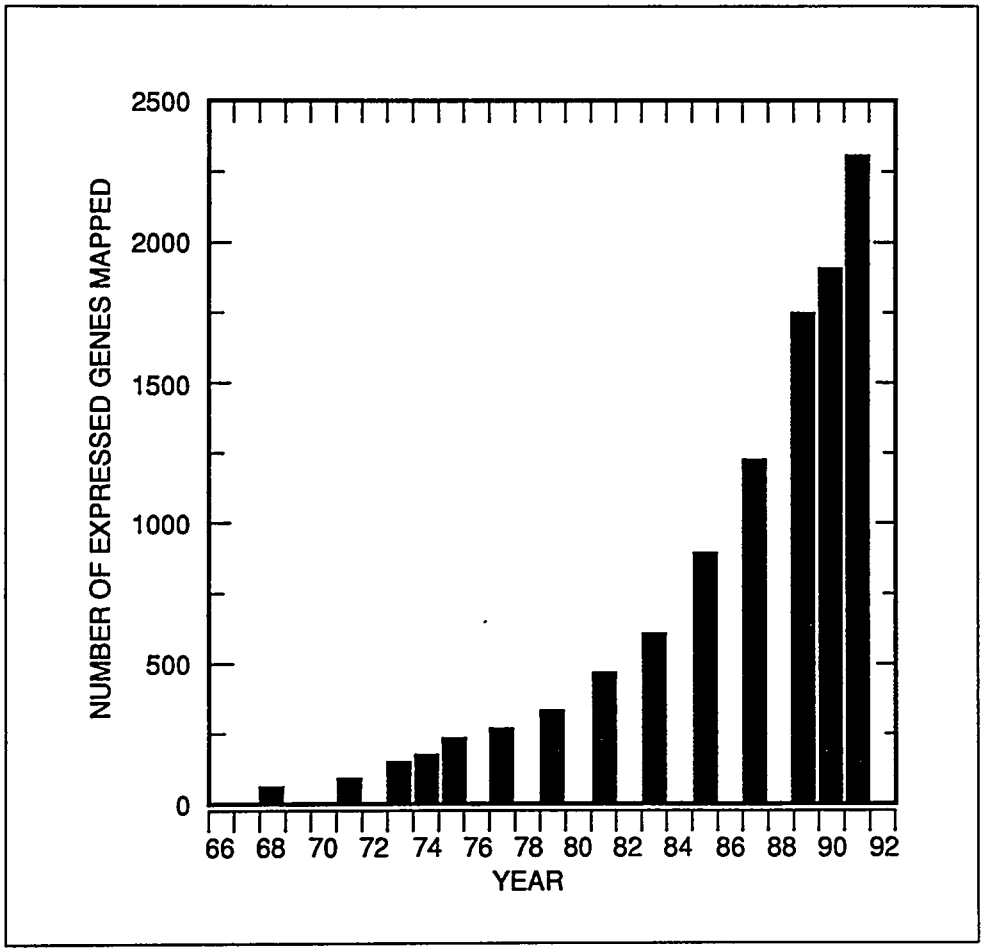


Fig. 7. Assignment of Genes to Specific Chromosomes. *The number of genes assigned (mapped) to specific chromosomes has greatly increased since the first autosomal (i.e., not on the X or Y chromosome) marker was mapped in 1968. Most of these genes have been mapped to specific bands on chromosomes. The acceleration of chromosome assignments is due to (1) a combination of improved and new techniques in chromosome sorting and band analysis, (2) data from family studies, and (3) the introduction of recombinant DNA technology. [Source: adapted from Victor A. McKusick, "Current Trends in Mapping Human Genes," The FASEB Journal 5(1), 12 (1991).]*

Primer on Molecular Genetics

Markers must be polymorphic to be useful in mapping; that is, alternative forms must exist among individuals so that they are detectable among different members in family studies. Polymorphisms are variations in DNA sequence that occur on average once every 300 to 500 bp. Variations within exon sequences can lead to observable changes, such as differences in eye color, blood type, and disease susceptibility. Most variations occur within introns and have little or no effect on an organism's appearance or function, yet they are detectable at the DNA level and can be used as markers. Examples of these types of markers include (1) restriction fragment length polymorphisms (RFLPs), which reflect sequence variations in DNA sites that can be cleaved by DNA restriction enzymes (see box, p. 13), and (2) variable number of tandem repeat sequences, which are short repeated sequences that vary in the number of repeated units and, therefore, in length (a characteristic easily measured). The human genetic linkage map is constructed by observing how frequently two markers are inherited together.

Two markers located near each other on the same chromosome will tend to be passed together from parent to child. During the normal production of sperm and egg cells, DNA strands occasionally break and rejoin in different places on the same chromosome or on the other copy of the same chromosome (i.e., the homologous chromosome). This process (called meiotic recombination) can result in the separation of two markers originally on the same chromosome (Fig. 8). The closer the markers are to each other—the more "tightly linked"—the less likely a recombination event will fall between and separate them. Recombination frequency thus provides an estimate of the distance between two markers.

On the genetic map, distances between markers are measured in terms of centimorgans (cM), named after the American geneticist Thomas Hunt Morgan. Two markers are said to be 1 cM apart if they are separated by recombination 1% of the time. A genetic distance of 1 cM is roughly equal to a physical distance of 1 million bp (1 Mb). The current resolution of most human genetic map regions is about 10 Mb.

The value of the genetic map is that an inherited disease can be located on the map by following the inheritance of a DNA marker present in affected individuals (but absent in unaffected individuals), even though the molecular basis of the disease may not yet be understood nor the responsible gene identified. Genetic maps have been used to find the exact chromosomal location of several important disease genes, including cystic fibrosis, sickle cell disease, Tay-Sachs disease, fragile-X syndrome, and myotonic dystrophy.

HUMAN GENOME PROJECT GOALS

	Resolution
• Complete a detailed human genetic map	2 Mb
• Complete a physical map	0.1 Mb
• Acquire the genome as clones	5 kb
• Determine the complete sequence	1 bp
• Find all the genes	

With the data generated by the project, investigators will determine the functions of the genes and develop tools for biological and medical applications.

One short-term goal of the genome project is to develop a high-resolution genetic map (2 to 5 cM); recent consensus maps of some chromosomes have averaged 7 to 10 cM between genetic markers. Genetic mapping resolution has been increased through the application of recombinant DNA technology, including in vitro radiation-induced chromosome fragmentation and cell fusions (joining human cells with those of other species to form hybrid cells) to create panels of cells with specific and varied human

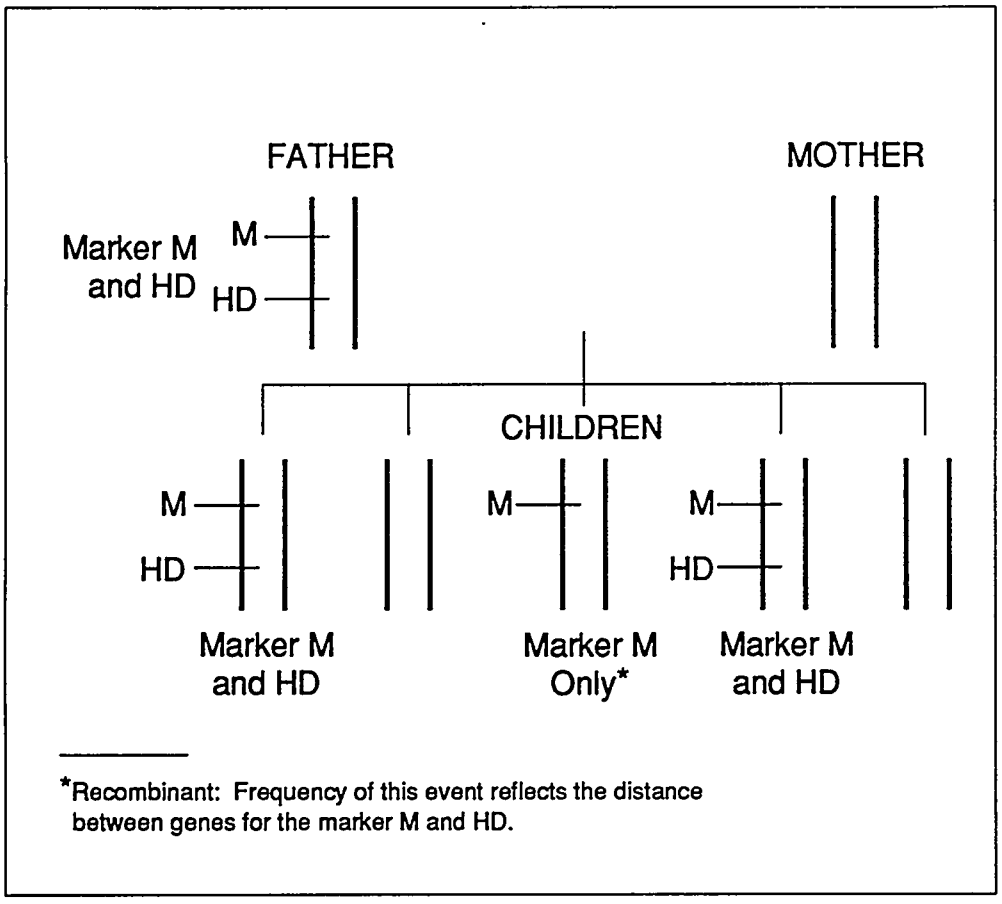


Fig. 8. Constructing a Genetic Linkage Map. *Genetic linkage maps of each chromosome are made by determining how frequently two markers are passed together from parent to child. Because genetic material is sometimes exchanged during the production of sperm and egg cells, groups of traits (or markers) originally together on one chromosome may not be inherited together. Closely linked markers are less likely to be separated by spontaneous chromosome rearrangements. In this diagram, the vertical lines represent chromo-some 4 pairs for each individual in a family. The father has two traits that can be detected in any child who inherits them: a short known DNA sequence used as a genetic marker (M) and Huntington's disease (HD). The fact that one child received only a single trait (M) from that particular chromosome indicates that the father's genetic material recombined during the process of sperm production. The frequency of this event helps determine the distance between the two DNA sequences on a genetic map .*

chromosomal components. Assessing the frequency of marker sites remaining together after radiation-induced DNA fragmentation can establish the order and distance between the markers. Because only a single copy of a chromosome is required for analysis, even nonpolymorphic markers are useful in radiation hybrid mapping. [In meiotic mapping (described above), two copies of a chromosome must be distinguished from each other by polymorphic markers.]

Physical Maps

Different types of physical maps vary in their degree of resolution. The lowest-resolution physical map is the chromosomal (sometimes called cytogenetic) map, which is based on the distinctive banding patterns observed by light microscopy of stained chromosomes. A complementary DNA (cDNA) map shows the locations of expressed DNA regions (exons) on the chromosomal map. The more detailed cosmid contig map depicts the order of overlapping DNA fragments spanning the genome. A macrorestriction map describes the order and distance between enzyme cutting (cleavage) sites. The highest-resolution physical map is the complete elucidation of the DNA base-pair sequence of each chromosome in the human genome. Physical maps are described in greater detail below.

Low-Resolution Physical Mapping

Chromosomal map. In a chromosomal map, genes or other identifiable DNA fragments are assigned to their respective chromosomes, with distances measured in base pairs. These markers can be physically associated with particular bands (identified by cytogenetic staining) primarily by in situ hybridization, a technique that involves tagging the DNA marker with an observable label (e.g., one that fluoresces or is radioactive). The location of the labeled probe can be detected after it binds to its complementary DNA strand in an intact chromosome.

As with genetic linkage mapping, chromosomal mapping can be used to locate genetic markers defined by traits observable only in whole organisms. Because chromosomal maps are based on estimates of physical distance, they are considered to be physical maps. The number of base pairs within a band can only be estimated.

Until recently, even the best chromosomal maps could be used to locate a DNA fragment only to a region of about 10 Mb, the size of a typical band seen on a chromosome. Improvements in fluorescence in situ hybridization (FISH) methods allow orientation of DNA sequences that lie as close as 2 to 5 Mb. Modifications to in situ hybridization methods, using chromosomes at a stage in cell division (interphase) when they are less compact, increase map resolution to around 100,000 bp. Further banding refinement might allow chromosomal bands to be associated with specific amplified DNA fragments, an improvement that could be useful in analyzing observable physical traits associated with chromosomal abnormalities.

cDNA map. A cDNA map shows the positions of expressed DNA regions (exons) relative to particular chromosomal regions or bands. (Expressed DNA regions are those transcribed into mRNA.) cDNA is synthesized in the laboratory using the mRNA molecule as a template; base-pairing rules are followed (i.e., an A on the mRNA molecule will pair with a T on the new DNA strand). This cDNA can then be mapped to genomic regions.

Because they represent expressed genomic regions, cDNAs are thought to identify the parts of the genome with the most biological and medical significance. A cDNA map can provide the chromosomal location for genes whose functions are currently unknown. For disease-gene hunters, the map can also suggest a set of candidate genes to test when the approximate location of a disease gene has been mapped by genetic linkage techniques.

High-Resolution Physical Mapping

The two approaches to high-resolution physical mapping are termed "top-down" (producing a macrorestriction map) and "bottom-up" (resulting in a contig map). With either strategy (described on pp. 14–15) the maps represent ordered sets of DNA fragments that are generated by cutting genomic DNA with restriction enzymes (see Restriction Enzymes box at right, p. 13). The fragments are then amplified by cloning or by polymerase chain reaction (PCR) methods (see DNA Amplification, pp. 18–20). Electrophoretic techniques are used to separate the fragments according to size into different bands, which can be visualized by direct DNA staining or by hybridization with DNA probes of interest. The use of purified chromosomes separated either by flow sorting from human cell lines or in hybrid cell lines allows a single chromosome to be mapped (see Separating Chromosomes box at right).

A number of strategies can be used to reconstruct the original order of the DNA fragments in the genome. Many approaches make use of the ability of single strands of DNA and/or RNA to hybridize—to form double-stranded segments by hydrogen bonding between complementary bases. The extent of sequence homology between the two strands can be inferred from the length of the double-stranded segment. Fingerprinting uses restriction map data to determine which fragments have a specific sequence (fingerprint) in common and therefore overlap. Another approach uses linking clones as probes for hybridization to chromosomal DNA cut with the same restriction enzyme.

Restriction Enzymes: Microscopic Scalpels

Isolated from various bacteria, restriction enzymes recognize short DNA sequences and cut the DNA molecules at those specific sites. (A natural biological function of these enzymes is to protect bacteria by attacking viral and other foreign DNA.) Some restriction enzymes (rare-cutters) cut the DNA very infrequently, generating a small number of very large fragments (several thousand to a million bp). Most enzymes cut DNA more frequently, thus generating a large number of small fragments (less than a hundred to more than a thousand bp).

On average, restriction enzymes with

- 4-base recognition sites will yield pieces 256 bases long,
- 6-base recognition sites will yield pieces 4000 bases long, and
- 8-base recognition sites will yield pieces 64,000 bases long.

Since hundreds of different restriction enzymes have been characterized, DNA can be cut into many different small fragments.

Separating Chromosomes

Flow sorting

Pioneered at LANL, flow sorting employs flow cytometry to separate, according to size, chromosomes isolated from cells during cell division when they are condensed and stable. As the chromosomes flow singly past a laser beam, they are differentiated by analyzing the amount of DNA present, and individual chromosomes are directed to specific collection tubes.

Somatic cell hybridization

In somatic cell hybridization, human cells and rodent tumor cells are fused (hybridized); over time, after the chromosomes mix, human chromosomes are preferentially lost from the hybrid cell until only one or a few remain. Those individual hybrid cells are then propagated and maintained as cell lines containing specific human chromosomes. Improvements to this technique have generated a number of hybrid cell lines, each with a specific single human chromosome.

Primer on Molecular Genetics

Macrorestriction maps: Top-down mapping. In top-down mapping, a single chromosome is cut (with rare-cutter restriction enzymes) into large pieces, which are ordered and subdivided; the smaller pieces are then mapped further. The resulting macrorestriction maps depict the order of and distance between sites at which rare-cutter enzymes cleave (Fig. 9a). This approach yields maps with more continuity and fewer gaps between fragments than contig maps (see below), but map resolution is lower and may not be useful in finding particular genes; in addition, this strategy generally does not produce long stretches of mapped sites. Currently, this approach allows DNA pieces to be located in regions measuring about 100,000 bp to 1 Mb.

The development of pulsed-field gel (PFG) electrophoretic methods has improved the mapping and cloning of large DNA molecules. While conventional gel electrophoretic methods separate pieces less than 40 kb (1 kb = 1000 bases) in size, PFG separates molecules up to 10 Mb, allowing the application of both conventional and new mapping methods to larger genomic regions.

Contig maps: Bottom-up mapping. The bottom-up approach involves cutting the chromosome into small pieces, each of which is cloned and ordered. The ordered fragments form contiguous DNA blocks (contigs). Currently, the resulting "library" of clones

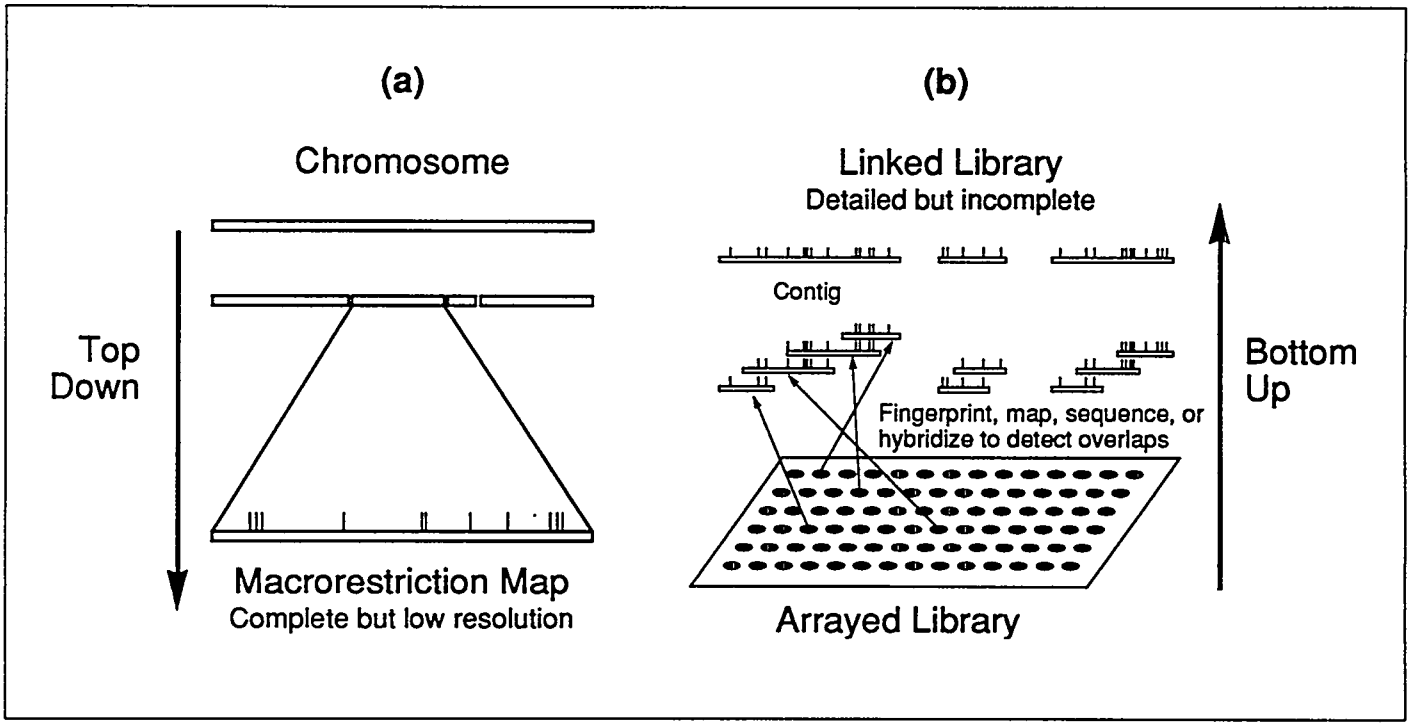


*Fig. 9. **Physical Mapping Strategies.** Top-down physical mapping **(a)** produces maps with few gaps, but map resolution may not allow location of specific genes. Bottom-up strategies **(b)** generate extremely detailed maps of small areas but leave many gaps. A combination of both approaches is being used. [Source: Adapted from P. R. Billings et al., "New Techniques for Physical Mapping of the Human Genome," The FASEB Journal 5(1), 29 (1991).]*

varies in size from 10,000 bp to 1 Mb (Fig. 9b, p. 14). An advantage of this approach is the accessibility of these stable clones to other researchers. Contig construction can be verified by FISH, which localizes cosmids to specific regions within chromosomal bands.

Contig maps thus consist of a linked library of small overlapping clones representing a complete chromosomal segment. While useful for finding genes localized to a small area (under 2 Mb), contig maps are difficult to extend over large stretches of a chromosome because all regions are not clonable. DNA probe techniques can be used to fill in the gaps, but they are time consuming. Figure 10 is a diagram relating the different types of maps.

Technological improvements now make possible the cloning of large DNA pieces, using artificially constructed chromosome vectors that carry human DNA fragments as large as 1 Mb. These vectors are maintained in yeast cells as artificial chromosomes (YACs). (For more explanation, see DNA Amplification, pp. 18–20.) Before YACs were developed, the largest cloning vectors (cosmids) carried inserts of only 20 to 40 kb. YAC methodology drastically reduces the number of clones to be ordered; many YACs span entire human genes. A more detailed map of a large YAC insert can be produced by subcloning, a process in which fragments of the original insert are cloned into smaller-insert vectors. Because some YAC regions are unstable, large-capacity bacterial vectors (i.e., those that can accommodate large inserts) are also being developed.

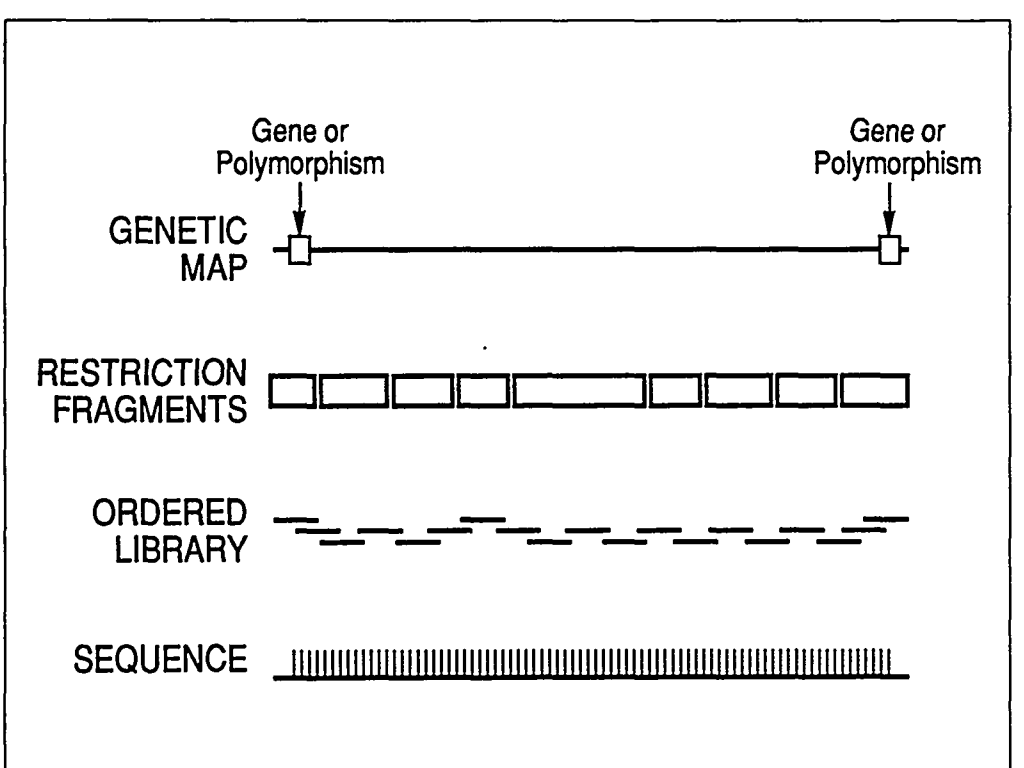


Fig. 10. Types of Genome Maps. *At the coarsest resolution, the genetic map measures recombination frequency between linked markers (genes or polymorphisms). At the next resolution level, restriction fragments of 1 to 2 Mb can be separated and mapped. Ordered libraries of cosmids and YACs have insert sizes from 40 to 400 kb. The base sequence is the ultimate physical map. Chromosomal mapping (not shown) locates genetic sites in relation to bands on chromosomes (estimated resolution of 5 Mb); new in situ hybridization techniques can place loci 100 kb apart. This direct strategy links the other four mapping approaches. [Source: see Fig. 9.]*

Sequencing Technologies

The ultimate physical map of the human genome is the complete DNA sequence—the determination of all base pairs on each chromosome. The completed map will provide biologists with a Rosetta stone for studying human biology and enable medical researchers to begin to unravel the mechanisms of inherited diseases. Much effort continues to be spent locating genes; if the full sequence were known, emphasis could shift to determining gene function. The Human Genome Project is creating research tools for 21st-century biology, when the goal will be to understand the sequence and functions of the genes residing therein.

Achieving the goals of the Human Genome Project will require substantial improvements in the rate, efficiency, and reliability of standard sequencing procedures. While technological advances are leading to the automation of standard DNA purification, separation, and detection steps, efforts are also focusing on the development of entirely new sequencing methods that may eliminate some of these steps. Sequencing procedures currently involve first subcloning DNA fragments from a cosmid or bacteriophage library into special sequencing vectors that carry shorter pieces of the original cosmid fragments (Fig. 11). The next step is to make the subcloned fragments into sets of nested fragments differing in length by one nucleotide, so that the specific base at the end of each successive fragment is detectable after the fragments have been separated by gel electrophoresis. Current sequencing technologies are discussed on p. 21.

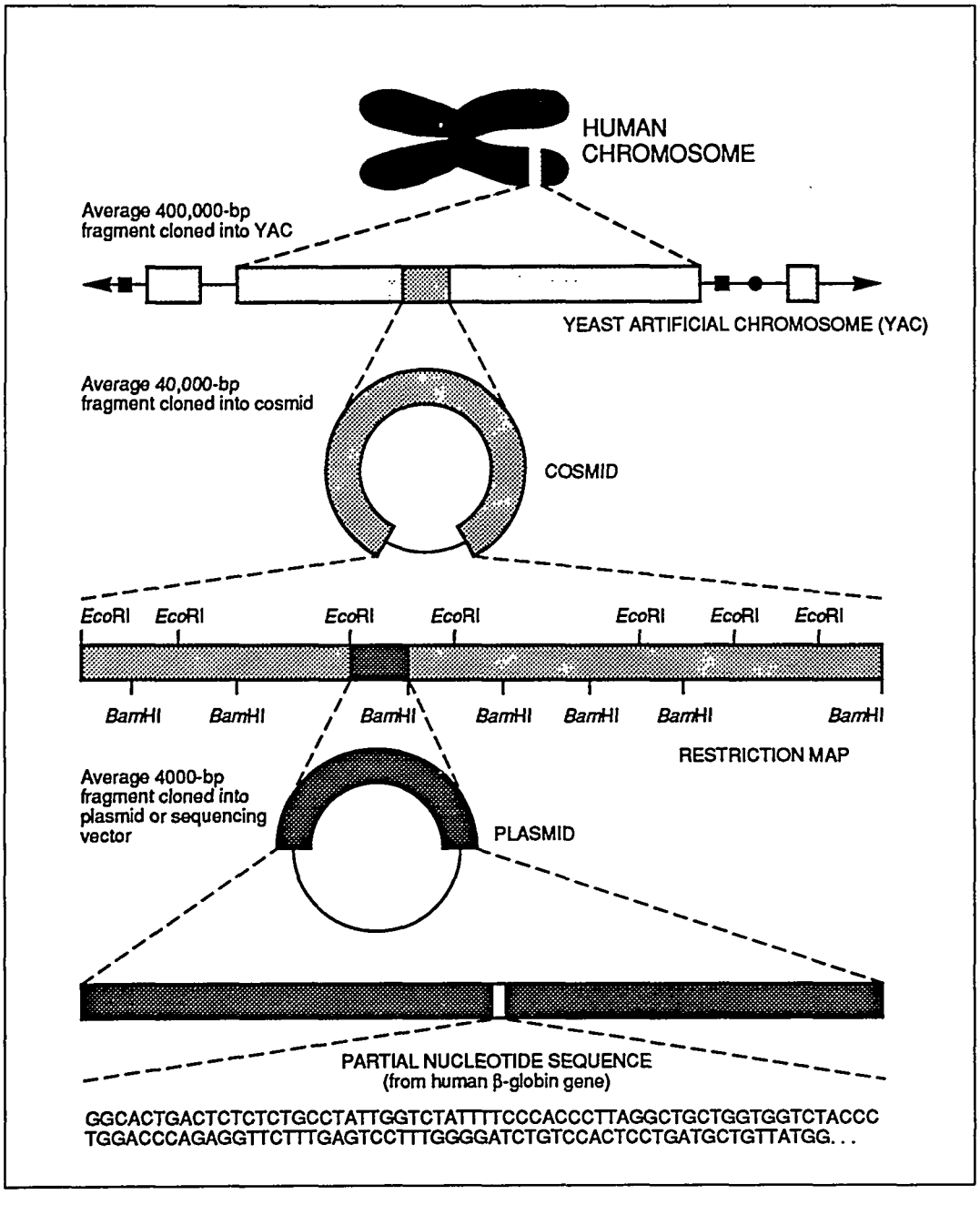


Fig. 11. Constructing Clones for Sequencing. *Cloned DNA molecules must be made progressively smaller and the fragments subcloned into new vectors to obtain fragments small enough for use with current sequencing technology. Sequencing results are compiled to provide longer stretches of sequence across a chromosome. (Source: adapted from David A. Micklos and Greg A. Freyer,* DNA Science, A First Course in Recombinant DNA Technology, *Burlington, N.C.: Carolina Biological Supply Company, 1990.)*

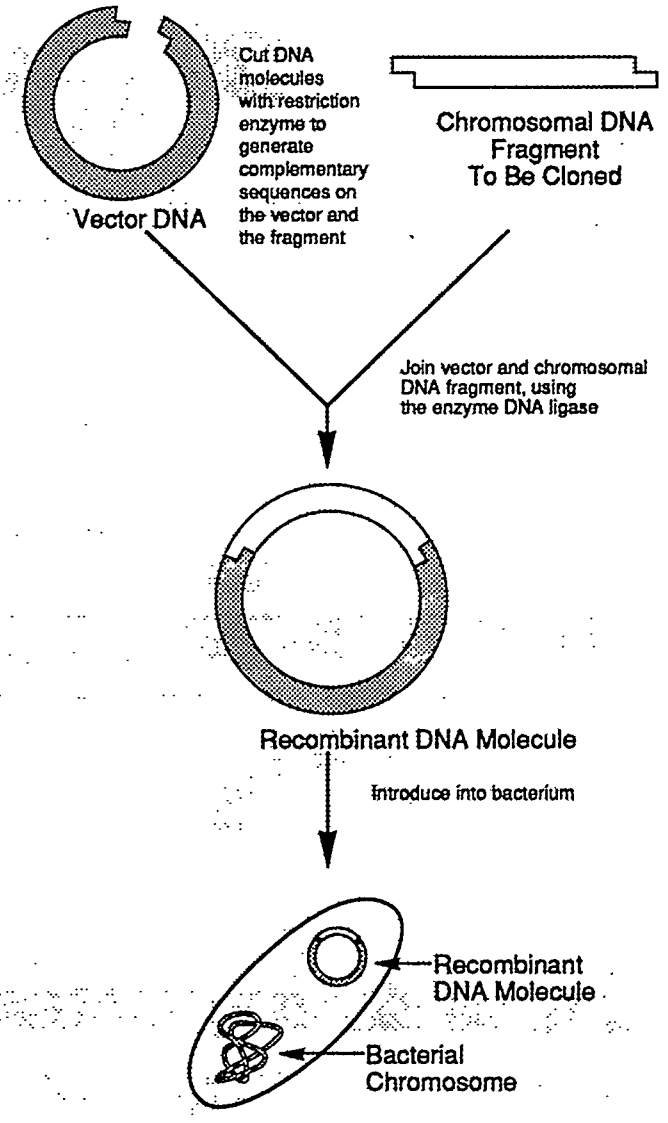
DNA Amplification: Cloning and PCR

Cloning (in vivo DNA amplification)

Cloning involves the use of recombinant DNA technology to propagate DNA fragments inside a foreign host. The fragments are usually isolated from chromosomes using restriction enzymes and then united with a carrier (a vector). Following introduction into suitable host cells, the DNA fragments can then be reproduced along with the host cell DNA. Vectors are DNA molecules originating from viruses, bacteria, and yeast cells. They accommodate various sizes of foreign DNA fragments ranging from 12,000 bp for bacterial vectors (plasmids and cosmids) to 1 Mb for yeast vectors [yeast artificial chromosomes (YACs)]. Bacteria are most often the hosts for these inserts, but yeast and mammalian cells are also used **(a)**.

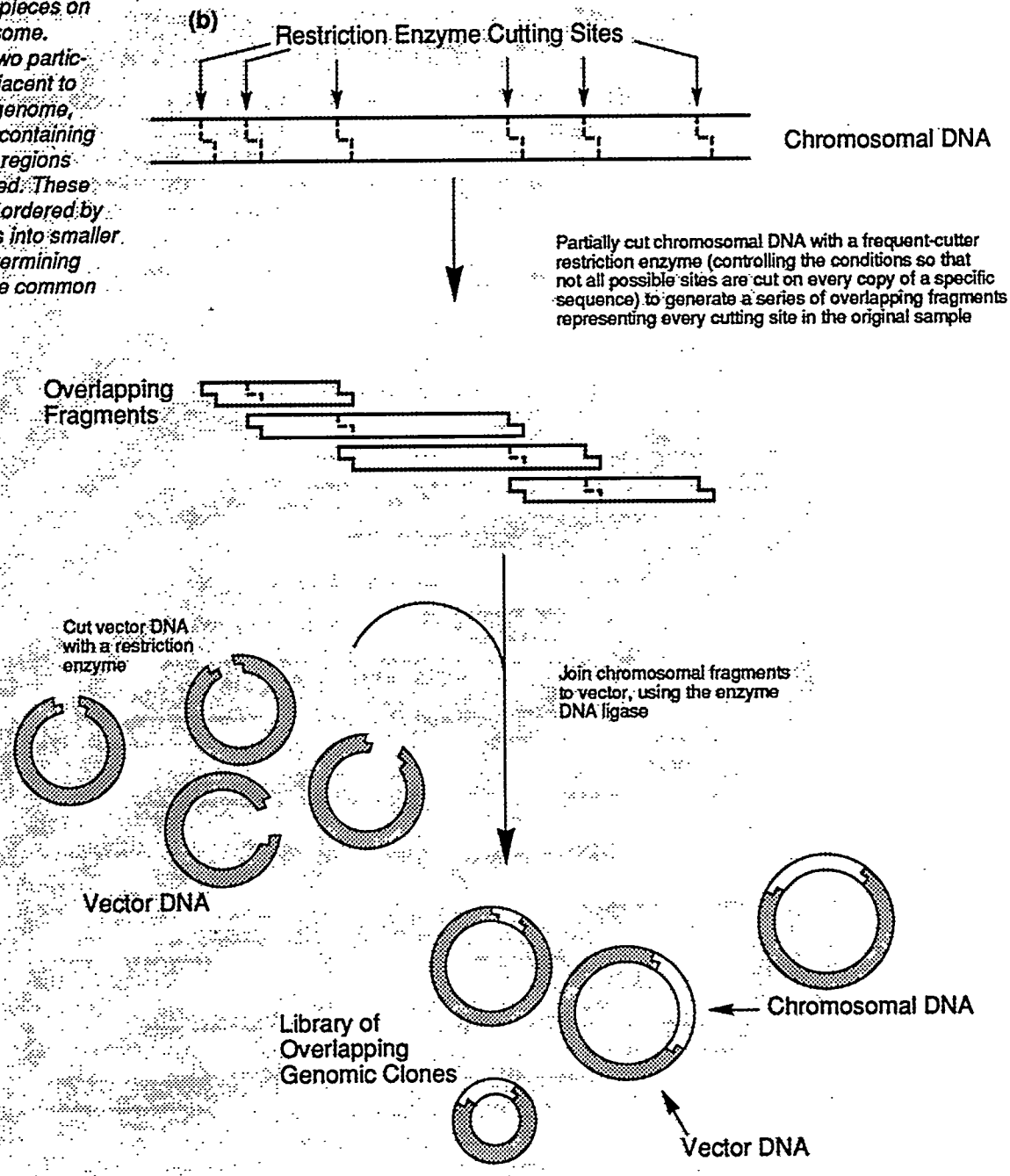
Cloning procedures provide unlimited material for experimental study. A random (unordered) set of cloned DNA fragments is called a library. Genomic libraries are sets of overlapping fragments encompassing an entire genome **(b)**. Also available are chromosome-specific libraries, which consist of fragments derived from source DNA enriched for a particular chromosome. (See Separating Chromosomes box, p. 13.)

(a)



(a) ***Cloning DNA in Plasmids.*** *By fragmenting DNA of any origin (human, animal, or plant) and inserting it in the DNA of rapidly reproducing foreign cells, billions of copies of a single gene or DNA segment can be produced in a very short time. DNA to be cloned is inserted into a plasmid (a small, self-replicating circular molecule of DNA) that is separate from chromosomal DNA. When the recombinant plasmid is introduced into bacteria, the newly inserted segment will be replicated along with the rest of the plasmid.*

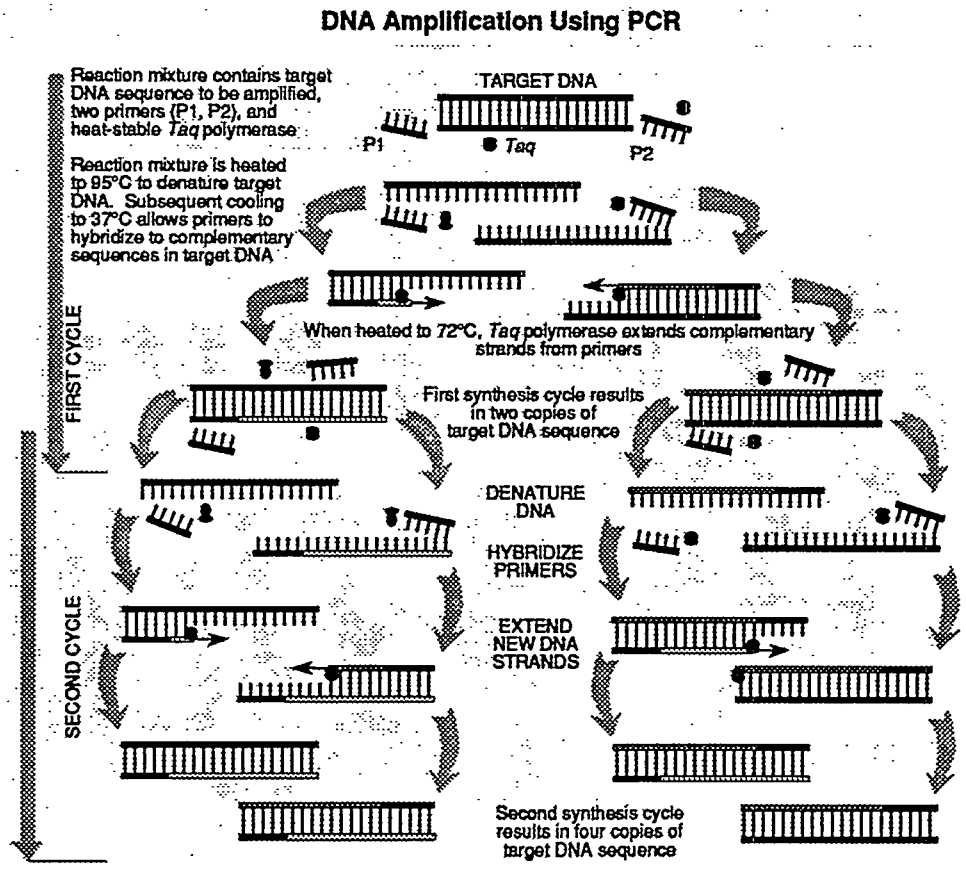
(b) ***Constructing an Overlapping Clone Library.*** *A collection of clones of chromosomal DNA, called a library, has no obvious order indicating the original positions of the cloned pieces on the uncut chromosome. To establish that two particular clones are adjacent to each other in the genome, libraries of clones containing partly overlapping regions must be constructed. These clone libraries are ordered by dividing the inserts into smaller fragments and determining which clones share common DNA sequences.*



Polymerase chain reaction (PCR) (in vitro DNA amplification)

Described as being to genes what Gutenberg's printing press was to the written word, PCR can amplify a desired DNA sequence of any origin (virus, bacteria, plant, or human) hundreds of millions of times in a matter of hours, a task that would have required several days with recombinant technology. PCR is especially valuable because the reaction is highly specific, easily automated, and capable of amplifying minute amounts of sample. For these reasons, PCR has also had a major impact on clinical medicine, genetic disease diagnostics, forensic science, and evolutionary biology.

PCR is a process based on a specialized polymerase enzyme, which can synthesize a complementary strand to a given DNA strand in a mixture containing the 4 DNA bases and 2 DNA fragments (primers, each about 20 bases long) flanking the target sequence. The mixture is heated to separate the strands of double-stranded DNA containing the target sequence and then cooled to allow (1) the primers to find and bind to their complementary sequences on the separated strands and (2) the polymerase to extend the primers into new complementary strands. Repeated heating and cooling cycles multiply the target DNA exponentially, since each new double strand separates to become two templates for further synthesis. In about 1 hour, 20 PCR cycles can amplify the target by a millionfold.



Source: *DNA Science*, see Fig. 11.

Current Sequencing Technologies

The two basic sequencing approaches, Maxam-Gilbert and Sanger, differ primarily in the way the nested DNA fragments are produced. Both methods work because gel electrophoresis produces very high resolution separations of DNA molecules; even fragments that differ in size by only a single nucleotide can be resolved. Almost all steps in these sequencing methods are now automated. Maxam-Gilbert sequencing (also called the chemical degradation method) uses chemicals to cleave DNA at specific bases, resulting in fragments of different lengths. A refinement to the Maxam-Gilbert method known as multiplex sequencing enables investigators to analyze about 40 clones on a single DNA sequencing gel. Sanger sequencing (also called the chain termination or dideoxy method) involves using an enzymatic procedure to synthesize DNA chains of varying length in four different reactions, stopping the DNA replication at positions occupied by one of the four bases, and then determining the resulting fragment lengths (Fig. 12).

These first-generation gel-based sequencing technologies are now being used to sequence small regions of interest in the human genome. Although investigators could use existing technology to sequence whole chromosomes, time and cost considerations make large-scale sequencing projects of this nature impractical. The smallest human chromosome (Y) contains 50 Mb; the largest (chromosome 1) has 250 Mb. The largest continuous DNA sequence obtained thus far, however, is approximately 350,000 bp, and the best available equipment can sequence only 50,000 to 100,000 bases per year at an approximate cost of $1 to $2 per base. At that rate, an unacceptable 30,000 work-years and at least $3 billion would be required for sequencing alone.

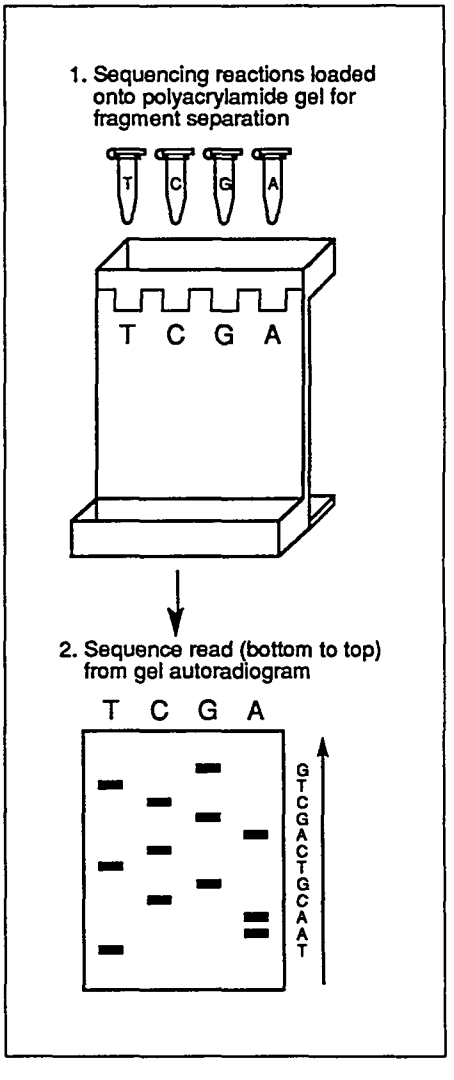


Fig. 12. DNA Sequencing. *Dideoxy sequencing (also called chain-termination or Sanger method) uses an enzymatic procedure to synthesize DNA chains of varying lengths, stopping DNA replication at one of the four bases, and then determining the resulting fragment lengths. Each sequencing reaction tube (T, C, G, and A) in the diagram contains*

- *a DNA template, a primer sequence, and a DNA polymerase to initiate synthesis of a new strand of DNA at the point where the primer is hybridized to the template;*
- *the four deoxynucleotide triphosphates (dATP, dTTP, dCTP, and dGTP) to extend the DNA strand;*
- *one labeled deoxynucleotide triphosphate (using a radioactive element or dye); and*
- *one* didoxynucleotide triphosphate, *which terminates the growing chain wherever it is incorporated. Tube A has* didATP, *tube C has* didCTP, *etc.*

For example, in the A reaction tube the ratio of the dATP to didATP *is adjusted so that each tube will have a collection of DNA fragments with a* didATP *incorporated for each adenine position on the template DNA fragments. The fragments of varying length are then separated by electrophoresis (1) and the positions of the nucleotides analyzed to determine sequence. The fragments are separated on the basis of size, with the shorter fragments moving faster and appearing at the bottom of the gel. Sequence is read from bottom to top (2). (Source: see Fig. 11.)*

Sequencing Technologies Under Development

A major focus of the Human Genome Project is the development of automated sequencing technology that can accurately sequence 100,000 or more bases per day at a cost of less than $.50 per base. Specific goals include the development of sequencing and detection schemes that are faster and more sensitive, accurate, and economical. Many novel sequencing technologies are now being explored, and the most promising ones will eventually be optimized for widespread use.

Second-generation (interim) sequencing technologies will enable speed and accuracy to increase by an order of magnitude (i.e., 10 times greater) while lowering the cost per base. Some important disease genes will be sequenced with such technologies as (1) high-voltage capillary and ultrathin electrophoresis to increase fragment separation rate and (2) use of resonance ionization spectroscopy to detect stable isotope labels.

Third-generation gel-less sequencing technologies, which aim to increase efficiency by several orders of magnitude, are expected to be used for sequencing most of the human genome. These developing technologies include (1) enhanced fluorescence detection of individual labeled bases in flow cytometry, (2) direct reading of the base sequence on a DNA strand with the use of scanning tunneling or atomic force microscopies, (3) enhanced mass spectrometric analysis of DNA sequence, and (4) sequencing by hybridization to short panels of nucleotides of known sequence. Pilot large-scale sequencing projects will provide opportunities to improve current technologies and will reveal challenges investigators may encounter in larger-scale efforts.

Partial Sequencing To Facilitate Mapping, Gene Identification

Correlating mapping data from different laboratories has been a problem because of differences in generating, isolating, and mapping DNA fragments. A common reference system designed to meet these challenges uses partially sequenced unique regions (200 to 500 bp) to identify clones, contigs, and long stretches of sequence. Called sequence tagged sites (STSs), these short sequences have become standard markers for physical mapping.

Because coding sequences of genes represent most of the potentially useful information content of the genome (but are only a fraction of the total DNA), some investigators have begun partial sequencing of cDNAs instead of random genomic DNA. (cDNAs are derived from mRNA sequences, which are the transcription products of expressed genes.) In addition to providing unique markers, these partial sequences [termed expressed sequence tags (ESTs)] also identify expressed genes. This strategy can thus provide a means of rapidly identifying most human genes. Other applications of the EST approach include determining locations of genes along chromosomes and identifying coding regions in genomic sequences.

End Games: Completing Maps and Sequences; Finding Specific Genes

Starting maps and sequences is relatively simple; finishing them will require new strategies or a combination of existing methods. After a sequence is determined using the methods described above, the task remains to fill in the many large gaps left by current mapping methods. One approach is single-chromosome microdissection, in which a piece is physically cut from a chromosomal region of particular interest, broken up into smaller pieces, and amplified by PCR or cloning (see DNA Amplification, pp. 18–20). These fragments can then be mapped and sequenced by the methods previously described.

Chromosome walking, one strategy for filling in gaps, involves hybridizing a primer of known sequence to a clone from an unordered genomic library and synthesizing a short complementary strand (called "walking" along a chromosome). The complementary strand is then sequenced and its end used as the next primer for further walking; in this way the adjacent, previously unknown, region is identified and sequenced. The chromosome is thus systematically sequenced from one end to the other. Because primers must be synthesized chemically, a disadvantage of this technique is the large number of different primers needed to walk a long distance. Chromosome walking is also used to locate specific genes by sequencing the chromosomal segments between markers that flank the gene of interest (Fig. 13).

The current human genetic map has about 1000 markers, or 1 marker spaced every 3 million bp; an estimated 100 genes lie between each pair of markers. Higher-resolution genetic maps have been made in regions of particular interest. New genes can be located by combining genetic and physical map information for a region. The genetic map basically describes gene order. Rough information about gene location is sometimes available also, but these data must be used with caution because recombination is not equally likely at all places on the chromosome. Thus the genetic map, compared to the physical map, stretches in some places and compresses in others, as though it were drawn on a rubber band.

The degree of difficulty in finding a disease gene of interest depends largely on what information is already known about the gene and, especially, on what kind of DNA alterations cause the disease. Spotting the disease gene is very difficult when disease results from a single altered DNA base; sickle cell anemia is an example of such a case, as are probably most major human inherited diseases. When disease results from a large DNA rearrangement, this anomaly can usually be detected as alterations in the physical map of the region or even by direct microscopic examination of the chromosome. The location of these alterations pinpoints the site of the gene.

Identifying the gene responsible for a specific disease without a map is analogous to finding a needle in a haystack. Actually, finding the gene is even more difficult, because even close up, the gene still looks like just another piece of hay. However, maps give clues on where to look; the finer the map's resolution, the fewer pieces of hay to be tested.

Once the neighborhood of a gene of interest has been identified, several strategies can be used to find the gene itself. An ordered library of the gene neighborhood can be constructed if one is not already available. This library provides DNA fragments that can be

screened for additional polymorphisms, improving the genetic map of the region and further restricting the possible gene location. In addition, DNA fragments from the region can be used as probes to search for DNA sequences that are expressed (transcribed to RNA) or conserved among individuals. Most genes will have such sequences. Then individual gene candidates must be examined. For example, a gene responsible for liver disease is likely to be expressed in the liver and less likely in other tissues or organs. This type of evidence can further limit the search. Finally, a suspected gene may need to be sequenced in both healthy and affected individuals. A consistent pattern of DNA variation when these two samples are compared will show that the gene of interest has very likely been found. The ultimate proof is to correct the suspected DNA alteration in a cell and show that the cell's behavior reverts to normal.

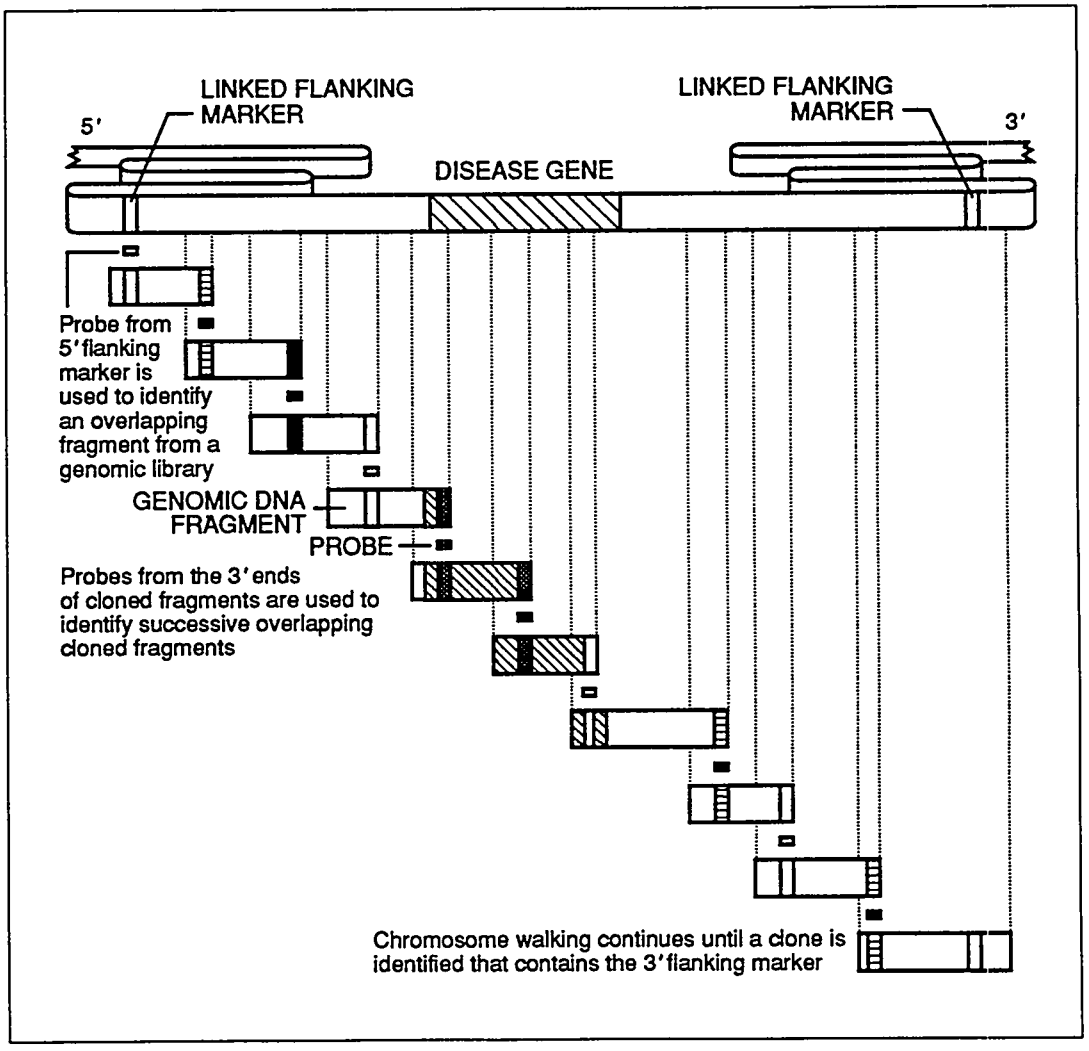


Fig. 13. Cloning a Disease Gene by Chromosome Walking. *After a marker is linked to within 1 cM of a disease gene, chromosome walking can be used to clone the disease gene itself. A probe is first constructed from a genomic fragment identified from a library as being the closest linked marker to the gene. A restriction fragment isolated from the end of the clone near the disease locus is used to reprobe the genomic library to find an overlapping clone. This process is repeated several times to walk across the chromosome and reach the flanking marker on the other side of the disease-gene locus. (Source: see Fig. 11.)*

Model Organism Research

Most mapping and sequencing technologies were developed from studies of nonhuman genomes, notably those of the bacterium *Escherichia coli,* the yeast *Saccharomyces cerevisiae,* the fruit fly *Drosophila melanogaster,* the roundworm *Caenorhabditis elegans,* and the laboratory mouse *Mus musculus.* These simpler systems provide excellent models for developing and testing the procedures needed for studying the much more complex human genome.

A large amount of genetic information has already been derived from these organisms, providing valuable data for the analysis of normal gene regulation, genetic diseases, and evolutionary processes. Physical maps have been completed for *E. coli,* and extensive overlapping clone sets are available for *S. cerevisiae* and *C. elegans.* In addition, sequencing projects have been initiated by the NIH genome program for *E. coli, S. cerevisiae,* and *C. elegans.*

Mouse genome research will provide much significant comparative information because of the many biological and genetic similarities between mouse and man. Comparisons of human and mouse DNA sequences will reveal areas that have been conserved during evolution and are therefore important. An extensive database of mouse DNA sequences will allow counterparts of particular human genes to be identified in the mouse and extensively studied. Conversely, information on genes first found to be important in the mouse will lead to associated human studies. The mouse genetic map, based on morphological markers, has already led to many insights into human biology. Mouse models are being developed to explore the effects of mutations causing human diseases, including diabetes, muscular dystrophy, and several cancers. A genetic map based on DNA markers is presently being constructed, and a physical map is planned to allow direct comparison with the human physical map.

Informatics: Data Collection and Interpretation

Collecting and Storing Data

The reference map and sequence generated by genome research will be used as a primary information source for human biology and medicine far into the future. The vast amount of data produced will first need to be collected, stored, and distributed. If compiled in books, the data would fill an estimated 200 volumes the size of a Manhattan telephone book (at 1000 pages each), and reading it would require 26 years working around the clock (Fig. 14).

Because handling this amount of data will require extensive use of computers, database development will be a major focus of the Human Genome Project. The present challenge is to improve database design, software for

HUMAN GENETIC DIVERSITY:
The Ultimate Human Genetic Database

- Any two individuals differ in about 3×10^6 bases (0.1%).
- The population is now about 5×10^9.
- A catalog of all sequence differences would require 15×10^{15} entries.
- This catalog may be needed to find the rarest or most complex disease genes.

database access and manipulation, and data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. Databases need to be designed that will accurately represent map information (linkage, STSs, physical location, disease loci) and sequences (genomic, cDNAs, proteins) and link them to each other and to bibliographic text databases of the scientific and medical literature.

Interpreting Data

New tools will also be needed for analyzing the data from genome maps and sequences. Recognizing where genes begin and end and identifying their exons, introns, and regulatory sequences may require extensive comparisons with sequences from related species such as the mouse to search for conserved similarities (homologies). Searching a database for a particular DNA sequence may uncover these homologous sequences in a known gene from a model organism, revealing insights into the function of the corresponding human gene.

Correlating sequence information with genetic linkage data and disease gene research will reveal the molecular basis for human variation. After a disease gene is identified, however, the altered protein specified by the flawed gene must still be compared with the normal version to identify the abnormality that causes disease. Once the error is pinpointed, researchers must try to determine how to correct it in the human body, a task that will require knowledge about how the protein functions and in which cells it is active.

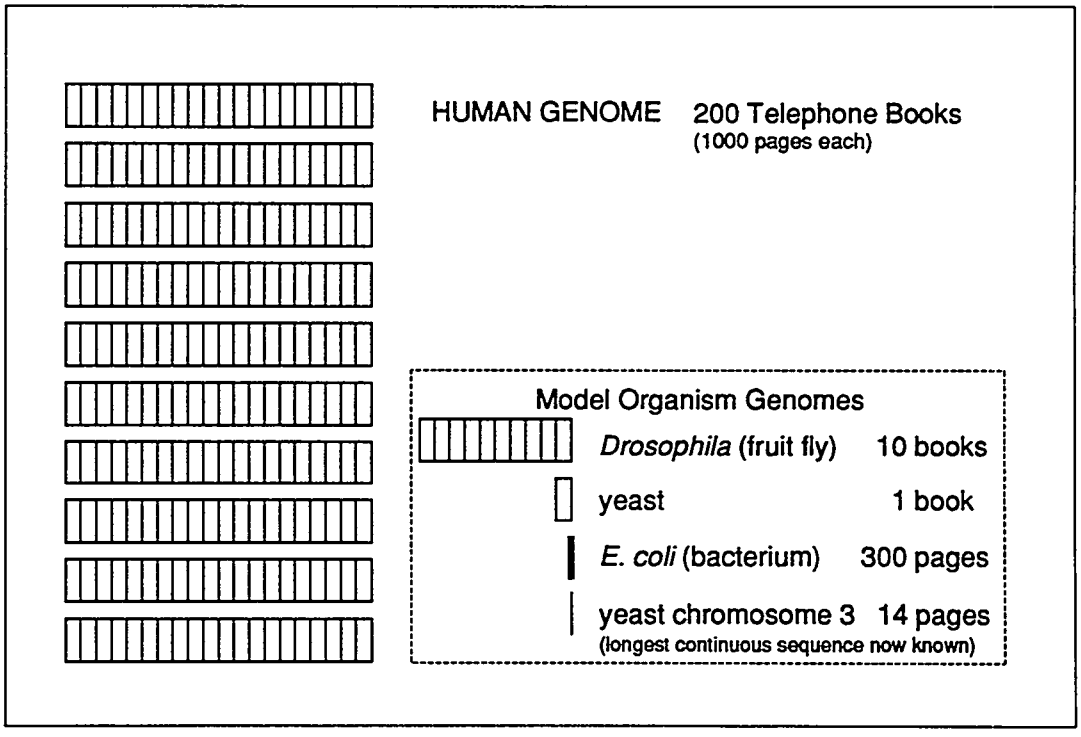


Fig. 14. Magnitude of Genome Data. *If the DNA sequence of the human genome were compiled in books, the equivalent of 200 volumes the size of a Manhattan telephone book (at 1000 pages each) would be needed to hold it all. New data-analysis tools will be needed for understanding the information from genome maps and sequences.*

Correct protein function depends on the three-dimensional (3D), or folded, structure the proteins assume in biological environments; thus, understanding protein structure will be essential in determining gene function. DNA sequences will be translated into amino acid sequences, and researchers will try to make inferences about functions either by comparing protein sequences with each other or by comparing their specific 3-D structures (Fig. 15).

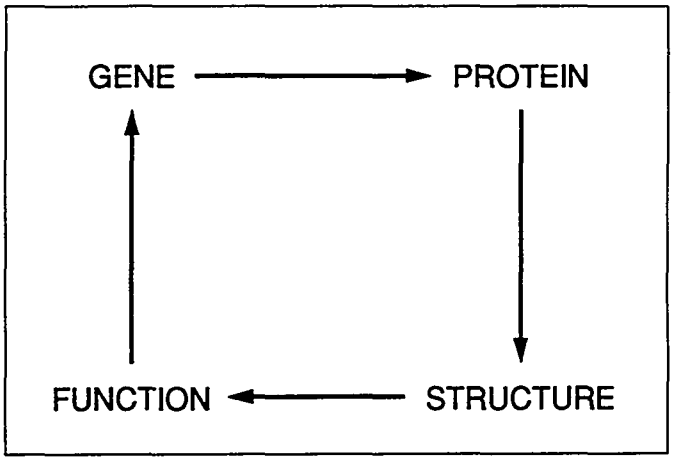


Fig. 15. Understanding Gene Function. *Understanding how genes function will require analyses of the 3-D structures of the proteins for which the genes code.*

Because the 3-D structure patterns (motifs) that protein molecules assume are much more evolutionarily conserved than amino acid sequences, this type of homology search could prove more fruitful. Particular motifs may serve similar functions in several different proteins, information that would be valuable in genome analyses. Currently, however, only a few protein motifs can be recognized at the sequence level. Continued development of analytic capabilities to facilitate grouping protein sequences into motif families will make homology searches more successful.

Mapping Databases

The Genome Database (GDB), located at Johns Hopkins University (Baltimore, Maryland), provides location, ordering, and distance information for human genetic markers, probes, and contigs linked to known human genetic disease. GDB is presently working on incorporating physical mapping data. Also at Hopkins is the Online *Mendelian Inheritance in Man* (OMIM) database, a catalog of inherited human traits and diseases.

The Human and Mouse Probes and Libraries Database (located at the American *Type Culture* Collection in Rockville, Maryland) and the GBASE mouse database (located at Jackson Laboratory, Bar Harbor, Maine) include data on RFLPs, chromosomal assignments, and probes from the laboratory mouse.

Sequence Databases

Nucleic Acids (DNA and RNA)

GenBank®, the European Molecular Biology Laboratory (EMBL) sequence database, and the DNA Database of Japan (DDBJ) house over 70 Mb of sequence from more than 2500 different organisms. Compiled from both direct submissions and journal scans, GenBank is supported at IntelliGenetics (Mountain View, California) and Los Alamos National Laboratory (LANL) through a contract from the NIH National Institute of General Medical Sciences. Although responsibility for GenBank will move to the National Center for Biotechnology Information (NCBI) of the National Library of Medicine in September 1992, LANL will continue to handle direct data submissions from authors. International collaborations with EMBL and DDBJ will also continue. NCBI is also developing GenInfo, a data archive that will eventually offer integrated access to other databases.

Proteins

The major protein sequence databases are the Protein Identification Resource [(PIR), National Biomedical Research Foundation], Swissprot, and GenPept (both distributed with GenBank). In addition to sequence information, they contain information on protein motifs and other features of protein structure.

Impact of the Human Genome Project

The atlas of the human genome will revolutionize medical practice and biological research into the 21st century and beyond. All human genes will eventually be found, and accurate diagnostics will be developed for most inherited diseases. In addition, animal models for human disease research will be more easily developed, facilitating the understanding of gene function in health and disease.

Researchers have already identified genes for a number of diseases known to be caused by a single gene, such as cystic fibrosis, Duchenne muscular dystrophy, myotonic dystrophy, neurofibromatosis, and retinoblastoma. As research progresses, investigators will also uncover the mechanisms for diseases caused by several genes or by a gene interacting with environmental factors. Genetic susceptibilities have been implicated in many major disabling and fatal diseases including heart disease, stroke, diabetes, and several kinds of cancer. The identification of these genes and their proteins will pave the way to more-effective therapies and preventive measures. Investigators determining the underlying biology of genome organization and gene regulation will also begin to understand how humans develop from single cells to adults, why this process sometimes goes awry, and what changes take place as people age.

New technologies developed for genome research will also find myriad applications in industry, as well as in projects to map (and ultimately improve) the genomes of economically important farm animals and crops.

While human genome research itself does not pose any new ethical dilemmas, the use of data arising from these studies presents challenges that need to be addressed before the data accumulate significantly. To assist in policy development, the ethics component of the Human Genome Project is funding conferences and research projects to identify and consider relevant issues, as well as activities to promote public awareness of these topics.