GENETIC ENGINEERING, PROCEDURES

1. Introduction

Genetics is the science dealing with the information content of an organism, especially the hereditary information passed on from one generation to another. The related discipline of molecular biology deals with the informational macromolecules that replicate and express genetic information in living systems. The term genetic engineering implies the deliberate manipulation in the laboratory of the hereditary information content (genotype) of a cell in order to alter the observable properties (phenotype) of an organism. In some sense, therefore, genetic engineering is as old as agriculture. Selective breeding in prehistoric times led to the introduction of maize and wheat as well as to the propagation of fermentative microbes. This "Mendelian engineering" is a major industry today as agricultural crops and animals are bred for yield, product composition and disease resistance.

The more contemporary meaning of genetic engineering implies a use of the techniques of molecular biology, especially recombinant DNA techniques, rather than breeding in the formation of new genotypes. Recombinant DNA molecules

are composed of two parts. First, a *vector* whose function is to provide the biochemical functions necessary for replication of the recombinant DNA molecule. Second, the *passenger* DNA that is joined to the vector and is replicated passively under control of the vector. Recombinant DNA technology allows the construction *in vitro* of DNA molecules that are not found in Nature and their subsequent introduction into organisms, resulting in new genotypes and phenotypes of the recipient. Keep in mind that, particularly in plant and animal science, the two techniques are often complementary: a gene may be introduced into an organism by recombinant DNA technology and the line with the desired properties further manipulated by breeding.

2. Analysis of DNA Information

Molecular biology is an information-based science. In this context, we can define information as the negative logarithm of the probability of a system occupying a particular state, given the total number of states available to it. In a DNA sequence there are four possible states at each position, corresponding to the four nucleic acid bases, Adenine, Guanine, Thymine, and Cytosine. A DNA of chain length n therefore has 4^n possible arrangements available to it. Although the cost in free energy terms of maintaining it is relatively small, this is an enormous amount of information. Even a small linear virus whose genome is 5000 nucleotides long would have 4^{5000} ($\sim 10^{3010}$) potential sequences; by comparison, the number of elementary particles in the universe is estimated at 10^{80}. Since there are so many potential sequences of a DNA molecule, and since DNA molecules of the same base composition can have similar biochemical properties, but very different sequences, standard biochemical techniques cannot address the most biologically important property of DNA, its information content. Genetic engineering techniques allow the analysis and manipulation of genetic information based on its nucleotide sequence.

2.1. Sequence-Dependent Cleavage of DNA by Restriction Enzymes. Bacteria in Nature are constantly exposed to exogenous DNA, primarily from bacteriophage (viruses) in the environment. Probably as a defense system, many bacteria contain a two-part DNA restriction and modification system. Restriction enzymes are of several types; the most useful for cloning, the Type II restriction enzymes, recognize specific sequences, usually 4–8 base pairs (bp) in length, and cut DNA molecules within these sequences. In Nature, restriction enzymes serve as a sort of immune mechanism: invading viruses are inactivated by restriction enzyme digestion of their DNA. The recognition sequences of restriction enzymes are short enough to be present in the host's genomic DNA many times; eg, a sequence of 6 bp would be present once every $4^6 = 4096$ bp by random chance. In the case of a "standard" bacterium, *Escherichia coli*, whose genome is 4.7×10^6 bp, chromosomal DNA would be cleaved at >1000 sites by a resident restriction enzyme, rendering it nonfunctional. In order to prevent this inactivation, bacteria encoding restriction enzymes also synthesize modification enzymes that modify (usually by methylation) DNA at the sequence recognized by the restriction enzyme, rendering the DNA refractory to digestion and thereby serving to distinguish host from

Table 1. **Sequence Specificities of Restriction Endonucleases**

Bacterial source	Enzyme	Sequence specificity[a]
E. coli / *R*	*EcoR*I	G↓A–A–T–T–C
Bacillus amyloliquefaciens H	*BamH*I	G↓G–A–T–C–C
B. globigii	*Bgl*II	A↓G–A–T–C–T
Xanthomonas malvacearum	*Xma*I	C↓C–C–G–G–G
Providencia stuarti	*Pst*I	C–T–G–C–A↓G

[a]The left end of each sequence is the 5′-end and the right end is 3′. Only one strand is shown for convenience, although the enzymes break duplex DNA. The arrow shows the position of the bonds broken.

invading DNA. A given restriction enzyme is useful in the analysis of DNAs prepared from hosts that do not contain an interfering modification enzyme; since a variety of bacteria are used as sources for restriction enzymes, virtually all DNA from laboratory microorganisms or eukaryotic cells can be cut *in vitro* for recombinant DNA experiments. Table 1 lists the recognition sequences and corresponding sites of methylation for several restriction enzymes used in molecular cloning and DNA mapping experiments (1). All of these enzymes cleave DNA prepared from laboratory strains of *E. coli*, the most common bacterial host for genetic engineering experiments.

Restriction sites provide physical markers on a DNA molecule. The fragments resulting from restriction digestion are commonly separated by electrophoresis in gel supports of polyacrylamide or agarose and visualized by staining with the dye ethidium bromide, which fluoresces strongly when inserted into DNA. The length in bp of a DNA fragment is estimated by comparison of its mobility with that of a fragment of known length. In general, the mobility of a DNA fragment during electrophoresis is inversely proportional to its chain length; this relationship holds over a range of DNA size that depends on the concentration of the gel and the conditions of electrophoresis. It is usually fairly simple to derive the restriction map of a small DNA, eg, a virus or plasmid.

When mapping longer DNA species (on the order of a whole chromosome), other techniques are used. A few enzymes are known whose recognition sequences are 8 bp long; these sequences occur relatively more rarely in a chromosome than do 6 bp sequences. In some cases, naturally occurring DNAs contain fewer sites for these “rare cutters” than would be predicted by random chance, allowing the determination of relatively simple restriction maps of whole bacterial or eukaryotic chromosomes. Separation of the very large pieces of DNA resulting from restriction digestion in these experiments requires special electrophoresis conditions to resolve the fragments (2).

2.2. Location of Specific Sequences to DNA Restriction Fragments. A second technique that is universally applied to DNAs large and small, is that of Southern blotting (Fig. 1). In these experiments, DNA fragments separated by gel electrophoresis are denatured *in situ* and transferred by capillary action to a nitrocellulose or nylon membrane, thereby making a "contact print" of the DNA in the gel. The single-stranded DNA fragments are bound irreversibly to the filter that is then immersed in a solution containing a single-stranded nucleic acid probe. The probe forms double helical base-paired hybrid regions with filter-bound DNA of complementary sequence. Solution conditions of hybridization can be set up to distinguish exact from inexact matches. The hybridized probe is detected by directly exposing the filter to X-ray film (if the probe is radioactive) or by enzymatic staining (if the probe is labeled by chemical modification). In either case the "contact print" of the gel shows the mobilities of the DNA fragments complementary to the probe. It is important to note that the specificity of Watson-Crick base pairing in DNA allows single fragments to be detected in the midst of a large excess of noncomplementary DNA. This specificity is the basis for, among other techniques, genetic fingerprinting of individual human DNAs and the use of species-specific gene probes for detection of bacterial species.

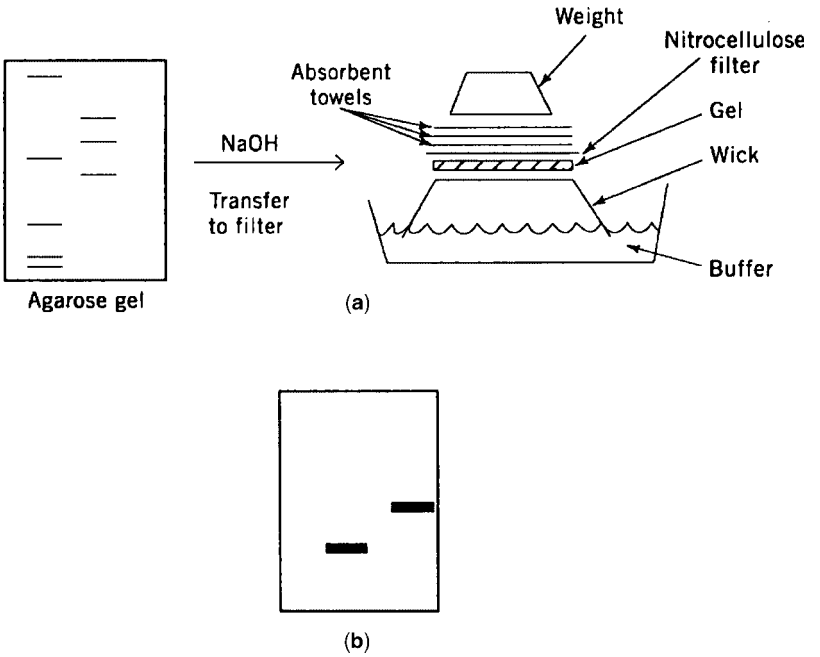

Fig. 1. Southern blot analysis of DNA showing (**a**) step 1, an agarose gel containing separated restriction fragments of DNA, denoted by (—), which is immersed in NaOH to denature the double-stranded structure of DNA, and then transferred by capillary flow to a nitrocellulose filter. In step 2, the bound DNA is allowed to hybridize to a labeled nucleic acid probe, and the unbound probe is washed off. In step 3, the filter is placed into contact with x-ray film resulting in (**b**) bands of exposure on the film which are detected after development and correspond to regions where the restriction fragment is complementary to the probe.

2.3. Restriction Sites as Genetic Markers. Classical genetic analysis uses observable phenotypes to infer genetic information. In order to derive a genetic map of a genotype, geneticists look for mutations which are linked, ie, carried on a chromosome. If two genes are carried on different chromosomes they will cosegregate, ie, appear together in the progeny of a genetic mating, half the time. [Consider the probability (0.50) of getting two heads or two tails when two coins are flipped.] If two genes are located near each other on the same chromosome, the frequency of cosegregation will be greater, somewhere between 50 and 100%.

In the early 1980s workers recognized that the presence of a restriction enzyme site is a chromosomal marker that can be assayed by Southern blotting of genomic DNA. Thus, if the pattern of a Southern blot is different for the DNA of different individuals in the population, and if one or more patterns cosegregate with a mutant gene causing a disease, the restriction pattern is a surrogate diagnostic marker for the disease. This phenomenon, termed restriction fragment linked polymorphism (RFLP), can be used to predict an inheritance pattern in the absence of any other information about the disease other than its pattern of heredity. The probes used to detect RFLPs can be used as the starting point to isolate clones of the gene encoding the disease itself (3). Similar logic forms the basis of genetic fingerprinting of individuals for forensic analysis (4).

3. Gene Isolation by Recombinant DNA Techniques

Workers in the early 1970s recognized that restriction enzymes provided tools not only for DNA mapping but also for construction of new DNA species not found in nature. A collection of recombinant DNA species consisting of many passenger sequences joined to identical vector molecules is called a **library**. Individual recombinant DNAs are isolated from single clones of the library for detailed analysis and manipulation.

3.1. Plasmid DNAs. Plasmids are nucleic acid molecules capable of intracellular extrachromosomal replication. Usually plasmids are circular DNA species, but linear and RNA plasmids are known. In Nature, plasmids can assume a variety of lifestyles. Plasmids can recombine into the host chromosome, they can be packaged into virus particles, they can replicate at high or low copy number relative to the host chromosome, and their information can affect the host phenotype. While no single plasmid is usually capable of all these behaviors, the properties of various plasmids have been used to construct vectors for a variety of purposes.

Ultimately, a plasmid is defined by its mode of DNA replication. DNA replication is initiated at a single, characteristic sequence, termed the origin. The origin sequence determines the copy number of the plasmid relative to the host chromosome and the host enzymes that are involved in plasmid replication. Two different plasmids that contain the same origin sequence are termed *incompatible*; this term does not refer to the active exclusion of one plasmid by another from the cell but rather to a stochastic process by which the two plasmids are partitioned differentially into progeny cells. A cell that contains two plasmids

of the same incompatibility group will segregate two clonal populations, each of which has one of the two plasmids in it.

Plasmids can be introduced into cells by several methods. The most common method is transformation, where the recipient cells are made competent to receive DNA by washing with a solution of Ca^{2+} or other inorganic ions. Then the naked DNA is added directly; a fraction of the cells take up the DNA and replicate it. These cells are then selected by growth in media containing an antibiotic. In many cases, discharge of a high voltage capacitor across a solution of cells renders them permeable to DNA; a phenomenon, termed electroporation, which can increase the efficiency of transformation substantially. Some but not all plasmids also transfer by conjugation, a sexual process where the DNA is donated from one cell to another after physical contact.

Most plasmids are topologically closed circles of DNA. They can be separated from the bulk of the chromosomal DNA by virtue of their resistance to alkaline solution: The double stranded structure of DNA is denatured at high pH but because the two strands of the plasmid are topologically joined they are more readily renatured. This property is exploited in rapid procedures for the isolation of plasmid DNA from recombinant microorganisms (available in Refs. (5,6)).

3.2. Plasmid Vectors for Facile Introduction of Passenger DNA and Selection of Recombinants. The map of a commonly used plasmid vector, pUC19 (7), is shown in Figure 2. Three parts of the vector are key to its utility. The origin sequence, *ori*, allows the replication of plasmid DNA in high copy number relative to the chromosome. A gene, *amp*, encoding the enzyme beta-lactamase, which hydrolyzes penicillin compounds, allows growth of plasmid-containing cells in media containing ampicillin. The third region of the plasmid allows the introduction of passenger DNA. The polylinker sequence is a chemically synthesized sequence of DNA that contains recognition sequences for a variety of restriction enzymes. A large number of these sites are unique, occurring only once in the vector. Associated with the polylinker sequence is a portion of the gene encoding *E. coli* beta-galactosidase. In the appropriate genetic background, beta-galactosidase enzymatic activity can be detected by the use of a chromogenic substrate (X-gal), which is hydrolyzed to yield a blue compound. Bacterial colonies containing the intact polylinker sequence express beta-galactosidase activity and stain blue. When the polylinker is disrupted by insertion of a passenger DNA sequence, beta-galactosidase is no longer made and the colonies do not stain to a blue color. Clones containing recombinant plasmids therefore can be identified visually in a background of nonrecombinants.

3.3. Construction of a Recombinant Plasmid by Joining Vector and Passenger DNA. The unique restriction sites in the plasmid vector DNA provide sites in the molecule for insertion of restriction-digested DNA fragments. Figure 3 shows an example: vector and passenger DNAs are digested separately with the enzyme EcoRII and the digested DNAs are mixed in the presence of an energy donor, ATP, and the enzyme DNA ligase. A fraction of the complementary ends left by EcoRI in the vector and passenger DNAs form Watson-Crick base pairs and the DNA chains are then joined covalently by DNA ligase. The recombinant DNA formed in this example is then introduced into recipient *E. coli* cells. Recipients containing the recombinant DNA are recognized by their ability to grow in the presence of ampicillin (this is diagnostic of the vector) and their

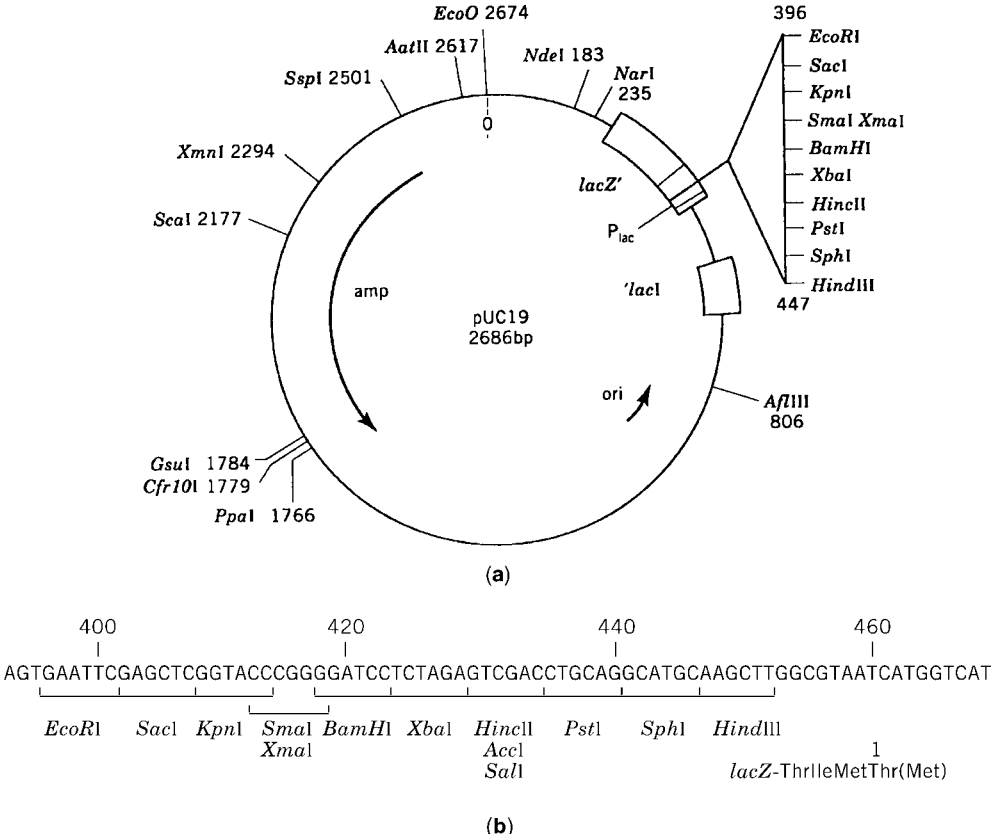

Fig. 2. (**a**) Map of pUC19, a commonly used plasmid vector where the numbers correspond to the positions of the various restriction enzyme cuts; and (**b**) nucleic acid composition of pUC19 from position 393 (5′-end) through position 469 (3′-end) (5,7).

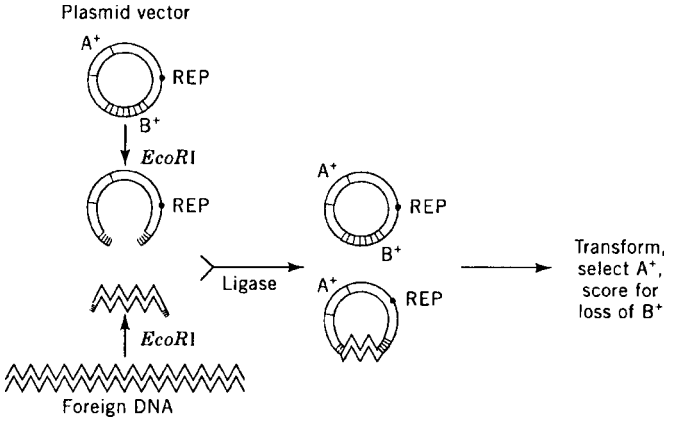

Fig. 3. Construction of a recombinant DNA by joining vector and passenger fragments where (⌊⌋⌊⌋⌊⌋) represent sticky ends. Both A and B genes represent selectable traits, so that introduction of foreign DNA in B gene leads to the loss of an identifiable function. REP represents replication and maintenance genes. (Courtesy of CRC Press.)

lack of beta-galactosidase enzymatic activity, which indicates the insertion of foreign DNA into the polylinker, sequence B in Figure 3.

3.4. Identification of the Desired Passenger Sequence in Plasmid Cloning Experiments. The objective of recombinant DNA construction is to obtain a clone of a single DNA sequence. If more than a single restriction fragment is ligated into a vector, the result is a library of clones, all of which have the same vector sequence, but with different passengers. Libraries are often described by the source of the passenger DNA. Genomic DNA libraries contain the total chromosomal DNA inserted into a vector. Copy DNA (cDNA) libraries contain passenger DNA derived by copying messenger RNA into DNA using the enzyme RNA-dependent DNA polymerase (reverse transcriptase).

The number of independent sequences in a DNA population is defined as its complexity. In a cloning experiment, the complexity of a library reflects the complexity of the passenger DNA population. In order to find a gene in a complex library, it is necessary to screen a large number of clones. The relationship between the complexity of the passenger DNA population, the size of the inserted passenger DNA fragments, and the number of clones that must be screened to have a defined probability of finding the correct sequence is given by Poisson statistics (8):

$$N = \ln(1 - P)/\ln(1 - I/G)$$

where N is the number of clones, P is the probability of finding the desired sequence, I is the size of the inserted passenger DNA, and G is the genome complexity in bp. For cloning a mammalian genomic sequence (complexity $= 3 \times 10^9$ bp) using fragments of 1×10^4 bp of DNA, a library of 1.38×10^6 independent clones is required to find a sequence with a probability of 99%. The number of clones required to identify a gene could be reduced by the insertion of larger DNAs or by the use of a less complex initial population of passenger DNA. For example, a mammalian organism synthesizes $<10^5$ mRNAs, representing only a few percent of its genomic DNA information. Thus, fewer clones would need to be screened to identify a particular coding sequence in a cDNA library.

The complexity attainable in construction of a plasmid library is limited by the efficiency of introducing the DNA into recipient bacteria. Libraries of sufficient complexity to clone mammalian genes are not normally feasible in plasmid vectors. Usually, these complex DNAs are cloned in bacteriophage or cosmid vectors (see below), which can accept larger fragments of passenger DNA. Plasmid libraries have been used to clone microbial genomes (complexity $\sim 4 \times 10^7$ bp) since correspondingly smaller library sizes are required.

Given a library of sufficient complexity, it is then necessary to find the clone of interest against the background of recombinant clones containing other passenger sequences. Three strategies are generally employed. The simplest method is to use genetic complementation of a mutant in the host. Thus, eg, a plasmid genomic library from *Streptomyces coelicolor* was transformed into a mutant *E. coli* host deficient in the metabolism of galactose (gal^-). The transformants were selected for the ability to metabolize galactose; the cloned genes were shown to direct galactose metabolism in *S. coelicolor* (9). Alternatively, genes encoding antibiotic resistance have been identified by direct phenotypic

selection. A wide variety of genes from bacteria and yeast have been detected in this fashion. In cloning a gene for which no *E. coli* phenotype can be selected, it is necessary to screen for the desired sequence. The most common screening method uses a radioactively labeled probe to hybridize to DNA from the recombinant bacteria. The agar plate containing colonies of recombinant bacteria is blotted with a sheet of nitrocellulose or nylon filter paper, thereby transferring some of the bacteria in the colony to the filter. These transferred colonies are lysed with alkali *in situ*, thereby also making the DNA single stranded. This DNA "contact print" is hybridized to the probe in the same way as restriction fragments are hybridized in Southern blotting (see above). Colonies containing DNA complementary to the probe are identified by autoradiography. Since only a portion of the original colony is transferred to the filter paper, the pattern on the autoradiogram will identify the appropriate colonies on the master plate.

The limiting factor in identifying a clone of interest is the availability of a probe. Probes may be obtained by using a homologous sequence from a previously identified clone. Thus, eg, a sequence encoding mouse beta-globin can be used to identify beta-globin genes from the human. Similarly, sequences complementary to *E. coli* ribosomal RNA will identify yeast rRNA sequences. In most screenings using heterologous probes it is necessary to hybridize the probe under conditions of lesser stringency to obviate the effects of DNA sequence variations across species. A potentially general method of identifying a probe is, first, to purify a protein of interest by chromatography or electrophoresis. Then a partial amino acid sequence of the protein is determined chemically. The amino acid sequence is used to predict likely short DNA sequences that direct the synthesis of the protein sequence; because the genetic code uses redundant codons to direct the synthesis of some amino acids, the predicted probe is unlikely to be unique. The least redundant sequence of 25–30 nucleotides is synthesized chemically as a mixture. The mixed probe is used to screen the library and the identified clones further screened, either with another probe reverse-translated from the known amino acid sequence or by directly sequencing the clones. While not all recombinant clones will encode the protein of interest, reiterative screening allows identification of the correct DNA recombinant.

If an antibody to the protein of interest is available, it is sometimes possible to use vector sequences, eg, the betagalactosidase promoter sequence, to direct the transcription of the passenger DNA into messenger RNA and the translation of that mRNA into protein that can be recognized by the antibody. While this method is somewhat less reliable than the use of nucleic acid probes, specialized vectors are available for this purpose.

3.5. Vectors for Cloning Larger Fragments of DNA. Plasmid DNAs used in molecular cloning have a practical limit in the amount of DNA that can be inserted into them. Fragments longer than this limit often accumulate deletion variants whose replication is favored over the original molecular species; this leads to loss of the original clone. In addition, the introduction of plasmids into recipient cells by transformation is relatively inefficient. When complex libraries are needed, eg, to isolate a mammalian gene, other cloning strategies are needed. These strategies are based on the replication of the bacteriophage lambda.

Lambda infects *E. coli* in either of two modes: it can lyse the cell to release more virus particles in a short time or it can insert itself into the bacterial chromosome and be replicated passively with the host, a phenomenon termed lysogeny. The DNA encoding lysogenic functions, ~40% of the bacteriophage chromosome, can be replaced by foreign DNA without interfering with lytic growth of the phage. Lambda-derived vectors use this phenomenon for cloning purposes. An example is shown in Figure 4: The vector DNA is prepared from phage grown lytically and the central "stuffer" fragment is removed and discarded, leaving the two "arms" of the vector. The vector arms are ligated to fragments of the DNA. Conditions of ligation are such that long tandem molecules of

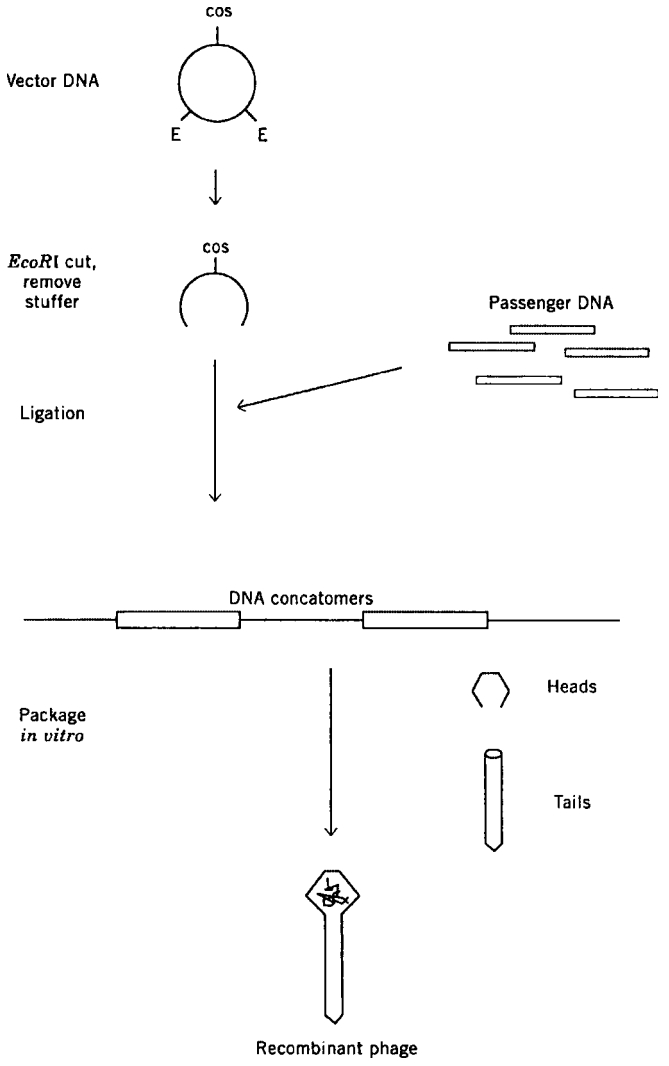

Fig. 4. Construction of recombinant phage in vectors derived from bacteriophage lambda where E represents the enzyme *EcoR*I. Other terms are defined in text.

DNA are formed, containing several vector-insert combinations linked together. The DNA is packaged *in vitro* into virus particles and the recombinant particles are mixed with host bacteria. Infection of the bacteria by recombinant packaged phage is essentially 100% efficient; this provides an advantage over the use of plasmid vectors.

Selection for phage clones containing recombinant DNA is provided by the packaging reaction. DNA packaging involves cutting of the catenated DNA at specific sequences termed *cos* (cohesive). If two *cos* sequences are separated by ~50 kb of DNA, the cut DNA will fill the phage head and a viable phage will result. If the sequences are closer together, as in the case of two arms being joined without an insert, no viable phage will result. Similarly, if two passenger sequences are inserted between arms, the resulting DNA will be too large to be packaged in the phage head.

3.6. Isolation of DNA for Phage Cloning. Because lambda-derived cloning vectors accept only a narrow size range of DNA inserts, a library constructed from completely restriction-digested DNA is unlikely to be representative of the total passenger DNA population. Because a restriction enzyme recognizing a 6-bp sequence will cleave DNA on the average of once in 4×10^3 bp and, to a first approximation, restriction sites are randomly located, many of the restriction sites will be located significantly more or less than 20 kb apart. The fragments resulting from digestion at these sites would not be packaged, and therefore would not be found in the library. In order to construct representative libraries the passenger DNA population is partially digested with a frequently cutting (4-bp recognition sequence) restriction enzyme under conditions where the average size of the products is close to 20 kb. The passenger DNA fragments are then separated by agarose gel electrophoresis or by sucrose density gradient centrifugation to eliminate those smaller and larger fragments in the digestion products. One and only one passenger DNA molecule from this preparation will yield a viable phage after the two steps of ligation to vector arms and *in vitro* packaging. If size-fractionation of the passenger DNA population were not done before ligation, two or more small fragments might ligate to each other and be cloned together, a linkage that would not reflect their true relationship.

The vector arms in phage cloning experiments are prepared in such a way to prevent their self-annealing. First, the *cos* sequences of the phage DNA are ligated together; this ensures that each recombinant phage DNA molecule will encode all of the phage functions essential for growth. Then the central "stuffer" fragment is removed by restriction digestion, separated from the arms by density gradient centrifugation, and discarded. Finally, the arms are treated with alkaline phosphatase to remove the 5′-phosphate from the "sticky ends". This insures against the self-ligation of the vector arms.

3.7. Screening of Recombinant Phage by DNA Hybridization or Antibody Recognition. The result of a phage infection of bacteria growing in a nutrient agar plate is a plaque or hole in the lawn of host bacteria. The plaque contains a clonal population of phage that result from initial infection of a single cell by a single phage. Infection typically results in a burst of 100–200 progeny phage that then infect neighboring bacteria to continue the infection process. As the bacteria reach saturation and stop growing the process stops, leaving the plaque of lysed bacteria and infectious phage particles. In screening

a recombinant phage library, the phage are plated at high density so that plaques are nearly contiguous. Then the plate is blotted with nitrocellulose or nylon filter paper. Dipping the filter into alkaline solution lyses the phage particles and denatures the DNA, making it ready for hybridization with a DNA probe. The phage from the hybridizing region of the plate are diluted to a lower density, plated, and rescreened. After two or three screenings, a clonal population of recombinant phage are present. The passenger DNA can be analyzed by Southern blotting, restriction mapping and sequencing.

If a DNA probe is not available, the recombinant phage can be screened by antibody methods. This has been used primarily in screening inserts derived by copying mRNA into cDNA using the enzyme reverse transcriptase. The cDNA population is then cloned into specialized phage that facilitate transcription and translation of the inserted cDNA. Antibody screening of the clones is done as described above. It should be noted that most eukaryotic genes contain noncoding introns inserted in the midst of the coding sequence; therefore genomic DNA libraries usually cannot be screened by antibody technology.

3.8. Cosmid Vectors. While the amount of information required for growth of a phage is large, that required for replication and selection of a plasmid is much less. Addition of cos sequences to a plasmid (hence the name cosmid) allows recombinant molecules of the appropriate size to be packaged into infectious particles. Since infection is much more efficient than transformation, use of recombinant particles overcomes a major disadvantage of plasmid vectors. In addition, the cosmid can accept >40 kb of passenger DNA without loss of infectivity, which means that smaller library sizes are required to represent a complex passenger DNA population. Once the recombinant cosmid DNA is inserted into a bacterial host cell it replicates like a plasmid. In some cases, this allows the use of phenotypic selection to identify genes. Although artifacts can arise because of the instability of recombinant cosmids, their use has been valuable in the preparation of ordered libraries representing entire bacterial or eukaryotic genomes (10).

3.9. Artificial Chromosomes for Insertion of Passenger DNA Molecules $>10^5$ bp. The sizes of the passenger DNA molecules inserted into cosmid and phage vectors are limited by the requirement for packaging of the recombinant DNA into the phage head. In addition, recombinant plasmids in *E. coli* are often unstable if they are >50–100 kb. Olson and co-workers (11) have developed a system for constructing libraries from complex DNAs that overcomes many of these limitations. Reasoning that individual eukaryotic chromosomes are extremely large DNA molecules ($>10^6$–10^7bp in most organisms), they constructed a vector that provides the sequences necessary for faithful replication and segregation of an individual yeast chromosome to the recombinant. Passenger DNA inserted into these vectors is replicated by yeast as an extra, linear chromosome in the nucleus. In addition to a cloning site and sequences allowing the preparation of the vector from *E. coli*, these YAC vectors (Fig. 5) contain sequences required for chromosomal propagation: *CEN*, the sequence that provides the information for segregation of the artificial chromosome through mitosis and meiosis; *ARS*, an autonomous replication sequence, ie, an origin of DNA replication; *TEL*, sequences directing the formation of telomeres, the ends of the chromosome. In YAC cloning experiments, large fragments of chromosomal DNA

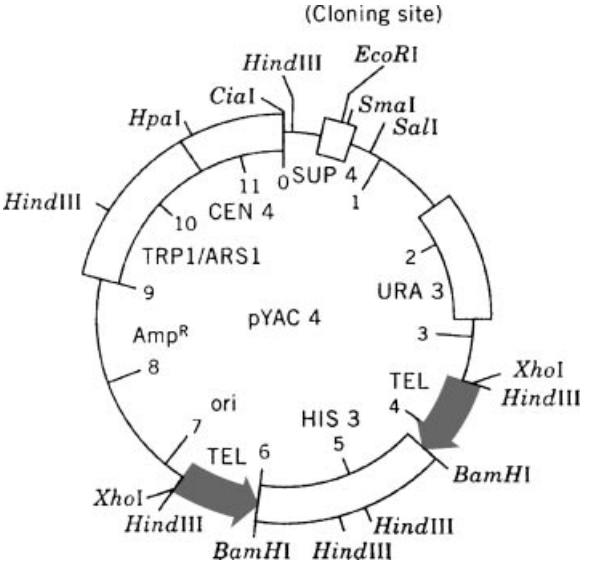

Fig. 5. Restriction map of the yeast artificial chromosome (YAC) vector used for cloning very large fragments of eukaryotic DNA. Terms defined in text (10).

are prepared by partial digestion with a restriction enzyme and size fractionation. Then the passenger DNA is ligated into the vector and the ligation mixture is transformed into yeast. The YACs are replicated as linear molecules, just as are the resident chromosomes of yeast. Screening for desired sequences is done by standard protocols, as described above.

Similarly, bacterial artificial chromosome (BAC) vectors can be used to propagate large DNAs in *E. coli*. Because of the ease of manipulation, BAC libraries are used extensively in constructing whole-genome libraries for genomic sequencing.

An artificial chromosome library can represent all the sequences of the human genome (3×10^9 bp) with a 99% probability of finding an individual sequence in as few as 50,000 clones. This library size is small enough to be carried as individual cultures in microtiter dishes and screened by automated means. Mapping and sequencing of complex genomes, as in the Human Genome Project, has used artificial chromosome libraries as the first component of the program.

3.10. Clones Linked Together to Form Larger Maps. Many eukaryotic genes are larger than the carrying capacity of a single phage or cosmid vector. Genomic DNA from eukaryotes contains noncoding sequences termed introns within the coding sequence; therefore, the DNA required to encode a protein can be much longer than that predicted from the size of the corresponding mRNA. For example, the gene for the human blood clotting protein Factor VIII is over 2×10^6 bp long, but the corresponding mRNA is about one-tenth the size.

As a result of these considerations it is usually necessary to identify clones in the library that are linked to the first clone obtained by screening. The library is rescreened using a restriction fragment from one end of the passenger DNA

segment in the first clone as a probe. The process, termed chromosome walking, may be reiterated several times until the contiguous inserts, termed contigs, cover the entire chromosomal region of interest.

Chromosome walking, while in principle capable of determining large maps, is tedious, especially when assembling phage or cosmids into contigs. In addition, the presence of repeated sequences in the probe and/or genome of interest can confound the analysis. The use of YAC or BAC clones can lead to more efficiency. More specialized "jumping" techniques can also be applied, especially in cloning genes from nearby linked restriction fragment polymorphisms (12).

4. Analysis of DNA Sequences

After a desired clone is obtained and mapped with restriction enzymes, further analysis usually depends on the determination of its nucleotide sequence. The nucleotide sequence of a new gene often provides clues to its function and the structure of the gene product. Additionally, the DNA sequence of a gene provides a guidepost for further manipulation of the sequence, eg, leading to the production of a recombinant protein in bacteria.

The sequence of a gene predicts the sequence of the protein it encodes. The relationship between nucleotide sequence of a DNA or its mRNA and the amino acid sequence of the protein it encodes is given by the genetic code. Several strategies are available for identifying protein-coding regions. The frequency at which synonomous codons are used varies in different organisms. Sequences are searched using a database of codon frequency in the organism of interest to identify the most likely coding regions. In addition, a DNA sequence can be translated in all three reading frames. A database of known protein coding sequences is then searched using the predicted amino acid sequences as a query. Statistically significant homologies can provide a clue to the structure and function of the protein(s) encoded by the cloned DNA.

4.1. Determination of DNA Sequence Information. Almost all DNA sequence is determined by enzymatic methods originated by Sanger and coworkers (13), that exploit the properties of the enzyme DNA polymerase. While a chemical method for DNA sequencing exists, its use has been supplanted for the most part in the initial determination of a sequence. Chemical sequencing (Maxam-Gilbert sequencing, Ref. 14) is more often used for mapping functional sites on DNA fragments of known sequence.

DNA polymerase enzymes all synthesize DNA by adding deoxy-nucleotides to the free 3′-OH group of an RNA or DNA primer sequence. The identity of the inserted nucleotide is determined by its ability to base pair with the template nucleic acid. The dependence of synthesis on a primer oligonucleotide means that synthesis of DNA proceeds only in a 5′ to 3′ direction; if only one primer is available, all newly synthesized DNA sequences begin at the same point.

DNA polymerases normally use 3′-deoxynucleotide triphosphates as substrates for polymerization. Given an adequate concentration of substrate, DNA polymerase will synthesize a long strand of new DNA complementary to the substrate. The use of this reaction for sequencing DNA depends on the inclusion of a single 2′,3′-dideoxynucleoside triphosphate in each of four polymerization

reactions. The dideoxynucleotides are incorporated normally in the chain in response to a complementary residue in the template; since no 3′-OH is available for further extension, polymerization is terminated. Thus, for a sequencing reaction initiated at a primer with a reaction mixture containing ddATP, at each T in the template, the enzyme will incorporate either a deoxy-A or a dideoxy-A. In the first instance, polymerization will continue; in the second, the chain will terminate. The result at the completion of the reaction is a series of nested fragments with each containing an identical 5′-end and with different lengths. In a similar fashion, inclusion in other reaction mixtures of either ddGTP, ddCTP, and ddTTP will produce nested fragments whose termini correspond to C, G, and A residues, respectively, in the template sequence (Fig. 6).

The nested oligonucleotides of a sequencing experiment are separated on the basis of chain length by electrophoresis in polyacrylamide gels. The gels separate the fragments on the basis of size, roughly proportional to the logarithm of their chain length. This means that shorter fragments are more widely separated than are long fragments. Practically, this means that sequences longer than 350–400 nt are difficult to determine from a single experiment, although longer gels can resolve fragments up to 600 nt long. The primer or newly synthesized chain is usually labeled isotopically, and the fragments are detected by autoradiography. In practice, reading the gels is the limiting step in gel-based DNA sequence analysis.

Automated sequencing systems use fluorescent dyes rather than radioactivity to detect the synthesized fragments that are size-separated by capillary electrophoresis. Twenty-four to 96 sequences can be determined in parallel and the data output is routinely 600 or more nucleotides per determination. Output from the instrument is stored as a computer file (15). The ease and efficiency of automated sequence determination makes it the method of choice for virtually all sequencing projects. Sequences are usually determined from both template strands to insure against mistakes, and it is helpful if each sequence is determined multiple times. It is also necessary to sequence across all the restriction sites used for subcloning portions of the fragment, although this is not necessary for large-scale sequencing projects like the Human Genome Project (see below).

5. Computer Analysis of DNA Sequence Information

The amount of information from a single DNA sequencing project can be staggering. Therefore, it is almost always necessary to analyze the data by computer methods. A number of commercial and purpose-built systems are available for analysis of DNA sequence information, operating on a variety of platforms, ranging from personal computers through workstations and supercomputers, depending on the intensity of the task. The field is not fully developed and research is ongoing in algorithm development, database manipulation and network applications, among others.

5.1. Assembly and Analysis of the Results from a Sequencing Project. The size of a gene almost always exceeds the data available from a single DNA sequencing experiment. It is necessary to identify contiguous regions of sequence information (contigs) and assemble these into completed projects.

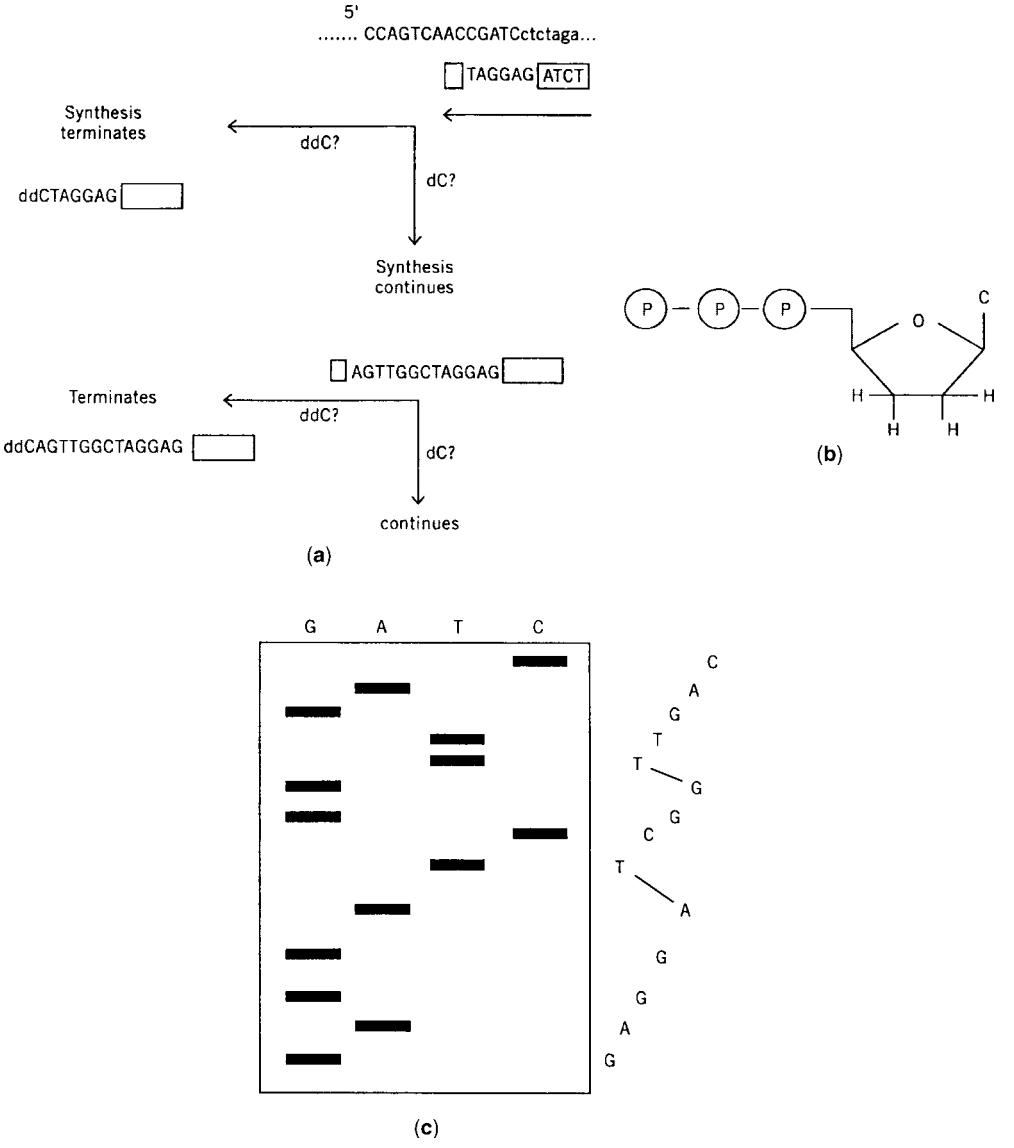

Fig. 6. DNA sequence analysis. (**a**) Simplified methodology for dideoxy sequencing. A primer, 5′-TCTA, hybridized to the template, is used to initiate synthesis by DNA polymerase. (**b**) Structure of 2′,3′-dideoxy CTP. When no 3′-OH functionality is available to support addition of another nucleotide to the growing chain, synthesis terminates once this residue is incorporated into the synthetic reaction. (**c**) Representation of a DNA sequencing gel and the sequence, read from bottom to the top of the gel, gives sequence information in the conventional 5′ to 3′ direction.

This can be done by relatively simple string matching, followed by highlighting points where two sets of overlapping information disagree. Resolution of discrepancies is then a matter of the investigator's judgment or doing more sequencing experiments across this stretch of DNA.

Automated sequence analysis allows the application of "shotgun" sequencing strategies to very complex sequences. In shotgun sequencing, sequences from very many clones are determined at random and assembled into contigs by computer analysis. Although many experts in the field were skeptical about the ability of shotgun strategies to determine complex genomes, leading them to prefer detailed mapping before sequence determination, even the human genome (3×10^9 bp) was analyzed in this fashion. Shotgun sequencing has led to the determination of hundreds of genome sequences (Refs. 16,17 and http://www.ncbi.nlm.nih.gov/Genomes/index.html). To determine protein-coding regions in DNA, the string of nucleotides in a DNA file is translated into an amino acid sequence using the genetic code. Because amino acid residues are specified by a genetic code of three nucleotides, any nucleic acid sequence can be read in three separate reading frames. The three predicted sequences are then used as arguments to search the database of previously determined genes and gene products. The databases, either GENBANK (TM), maintained by the National Library of Medicine (U.S.A.) or EMBL, maintained at the European Molecular Biology Laboratory, Heidelberg, share information and a search of one is now equivalent to the other. Databases are distributed either by electronic mail, on physical media (CD-ROM or tape) from the repositories, or by commercial software vendors as part of an analysis package. Virtually all journals in the field require deposition of the sequence file into a database before accepting a manuscript for publication.

Searching the database for a string of nucleotides or amino acids is extremely simple, but normally unenlightening. Usually, DNA and protein sequences vary among organisms so that, eg, an enzyme from the mouse and one from bacteria may have the same catalytic mechanism and function but different amino acid sequences. Search programs use statistically based scoring tables to evaluate the probability that two protein or DNA sequences being closely related. Sequences that share a large fraction of chemically related or identical amino acids are predicted to define biochemically and evolutionarily related proteins (18,19).

Other database searches may be used to predict active sites or secondary structures in proteins, likely points of mRNA initiation (promoters), translation initiation and restriction sites that could be useful in further manipulation. When more than two homologous sequences are aligned, a statistically most probable consensus can be generated as a guide to functional residues; however, finding the best multiple alignment of several sequences (eg, in defining the information necessary for promoter function) from first principles is computationally difficult and has not been implemented rigorously.

The large number of genomic sequences available makes it possible to analyze sequences for functional motifs based on evolutionary patterns and it is expected that this approach will be even more useful in the future.

6. Uses of Sequence Information

DNA sequence information is the starting point for other applications, including the expression of a gene product, the search for related sequences in biological samples, *in vitro* mutagenesis of the sequence, and structure–function studies of gene expression.

6.1. Specific Amplification of Related Sequences by the Polymerase Chain Reaction. If the sequence of a gene is known, primers that are unique to the gene can be synthesized. An oligonucleotide longer than 15–18 residues is likely to be unique even in a complex genome. Such an oligonucleotide, if hybridized to a single-stranded DNA, can be used to prime DNA polymerase so that the DNA is replicated. If two primers are made, each complementary to one strand of a gene, then the strand of each DNA located between them can be specifically replicated by DNA polymerase (see Fig. 7). The newly replicated DNA strands can be separated by heating and will then serve as template for another round of primed synthesis, leading to another doubling in the concentration of the original amplified sequence. Since the concentration of the DNA of interest doubles with each cycle, at least a 50,000-fold increase in its concentration is achievable within a few hours. This polymerase chain reaction (PCR) (20) is used in a variety of experimental manipulations and diagnostic procedures. Sequences less than a few hundred base pairs long are most efficiently amplified by PCR but even this relatively limited information can be fruitful.

PCR amplification of a DNA sequence is facilitated by the use of a heat-stable DNA polymerase [*Taq* polymerase(TM)] derived from the thermostable bacterium *Thermus aquaticus*. The thermostable polymerase allows the repeated steps of strand separation, primer annealing, and DNA synthesis to be carried out in a single reaction mixture whose temperature is cycled automatically. Each cycle consists of a high temperature step to denature the template strands, a lower temperature annealing of the primer and template, and a higher temperature synthesis step. All components of the reaction are present in the same tube.

It should be emphasized that the PCR requires specific primers for synthesis; therefore, the sequence flanking the sequence to be amplified must be known. PCR reactions are also very susceptible to contamination by other DNA. Precautions against contamination need to be rigorous, especially when PCR is used for diagnosis of disease, eg, in testing for the presence of viral nucleic acid in blood samples.

6.2. Applications of PCR. The ability of PCR to specifically amplify DNA sequences leads to a wide variety of applications. It is possible to identify microorganisms in pathological samples without the need for culturing them [eg, in testing for the human immunodeficiency virus (HIV) that causes AIDS], identify mutant genes in analyzing genetic disease, construct mutant genes for structure–function studies, find specific clones without the need for screening (sib selection), "genetically fingerprint" DNA from forensic samples, synthesize proteins by coupled transcription and translation *in vitro*, and clone rare messages from mammalian cells by amplification of specific sequences. In analysis of complex genomes, specific PCR primers form Sequence Tagged Sites for physical mapping. Protocols for many of these uses of PCR have been collected (5,6,20) and new ones appear in the literature regularly.

7. Array Methodologies

The availability of complete genomic sequences for a number of organisms have opened up new possibilities for the analysis of all the genes in an organism. For

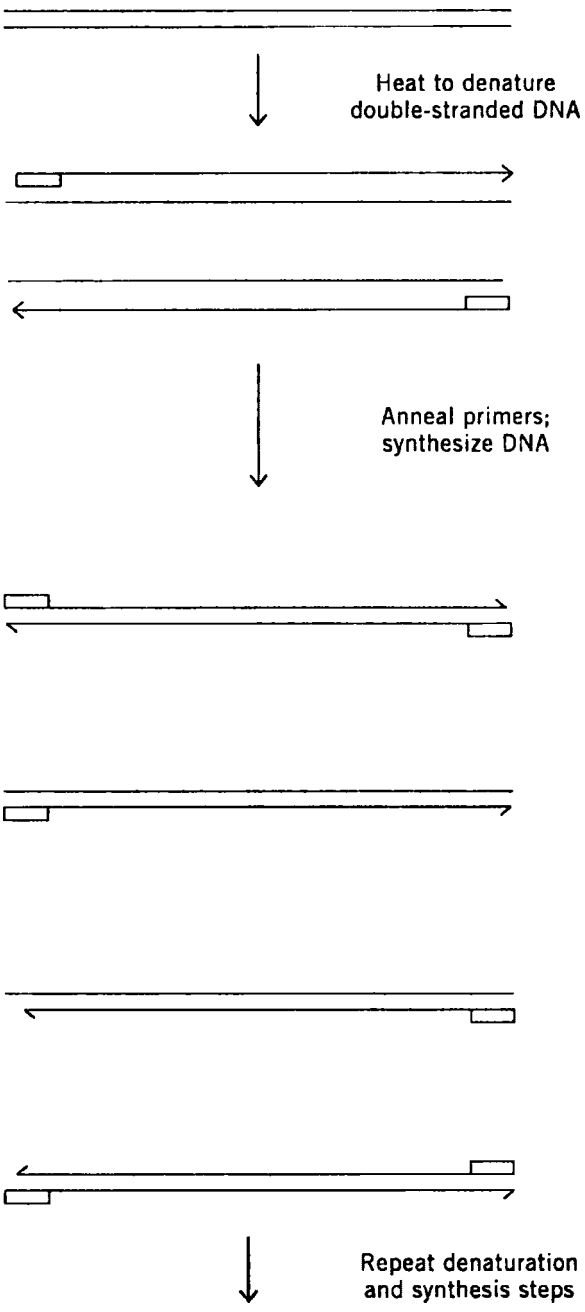

Fig. 7. The PCR reaction, showing amplification of a sequence located between two primers where (□) represents a primer, (—) a single strand of DNA, and the arrow the 3′-end. A double-stranded DNA is replicated fourfold.

example, the yeast genome contains ~12 million bp and ~6100 genes. Each of these genes contains unique sequences that can be used as PCR primers or hybridization probes so it is possible to make a probe that is diagnostic for each individual gene. The probes are spotted by a printer head onto a glass slide in a dense array (each spot is 50–200 μm in diameter). The array of all these printed arrays occupies a small area in to (<4 cm^2).

The array represents a global-scale Southern blot. In a prototype experiment. mRNA is isolated from cancer cells grown under different conditions and reverse transcribed. The deoxynucleoside triphosphates in each reverse-transcription reaction contain a different fluorescent dye, so that the cDNAs can be distinguished in a single hybridization reaction. The cDNAs are hybridized and the different dyes are identified by fluorescent microscopy. The relative intensities of the two dyes, hence, the amount of the mRNA, are determined and analyzed by computer. The software ranks the spots to identify those that show the greatest difference in expression level (21).

The simultaneous determination of so many expression levels has led to recognition of previously unsuspected relationships among genes. Similar logic can be used to identify genotypes and determine allelic differences among individuals. In this case, the hybridization probe set is the DNA complement of an individual. The array is a set of Sequence-Tagged Sites (STS) that serve as physical chromosomal markers similar to the use of RFLPs. Each STS is a PCR product amplified from chromosomal DNA. Thus, eg, it is possible to determine the gene expression profiles of tumor subtypes and correlate these with the clinical progression of the disease (21).

7.1. Combinatorial or Affinity-Selected Libraries. A recent application of PCR is the construction of new types of libraries consisting of randomized, synthetic sequences that are then selected for some biochemical property. The selected sequences are amplified by PCR and rescreened. The end result is a set of one or more sequences whose members share the biochemical property. After cloning, the sequences of individual members of the set are determined. An example is shown in Figure 8. A hypothetical protein binds to a recognition sequence six nucleotides long. In an effort to determine the sequences required for this binding, a library is made consisting of synthetic DNA oligonucleotides randomized at six positions and having a common sequence in all other positions. The complexity of the initial library is therefore $4^6 = 4096$. The oligonucleotide library is then bound to purified protein *in vitro*. Bound oligonucleotides are separated from the unbound sequences by physical means, such as electrophoresis or entrapment on a nitrocellulose filter. Because DNA-binding proteins will normally bind to a number of sequences with different affinities the initial population of selected DNAs is somewhat heterogeneous, consisting of 20 members in this example. These oligonucleotides are then amplified by the PCR reaction to yield a population of DNAs representative of the initially selected sequence.

DNA sequencing of the selected population at this stage would show bands in more than one lane of the sequencing gel indicating that several positions in the sequence population consisted of more than one residue. Reiteration of the selection and amplification process would yield the sequences in the initially selected population that had highest affinity for the binding protein. In the example shown, this would consist of two members. If the finally selected sequence

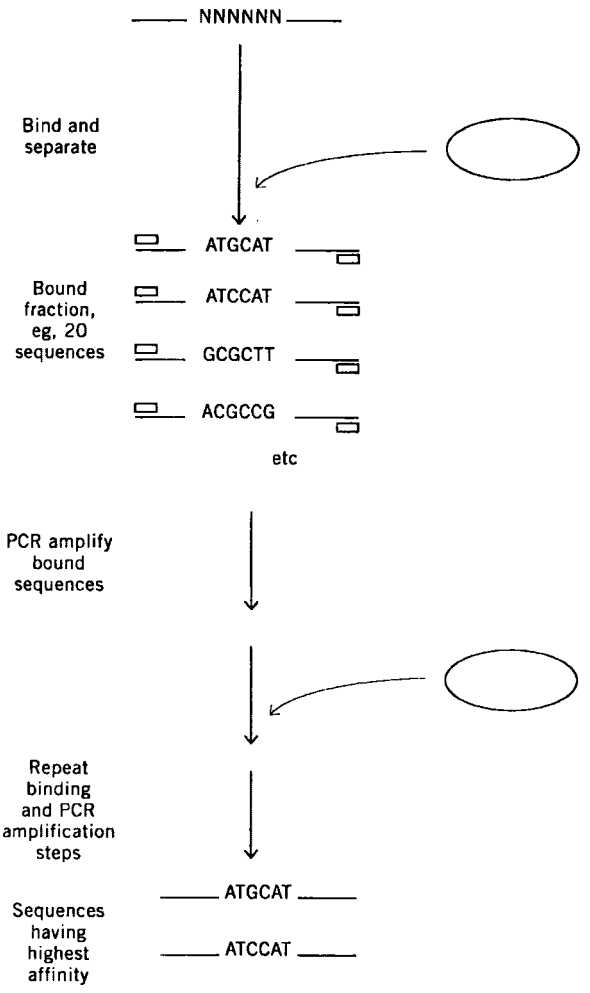

Fig. 8. A simplified combinatorial approach to identify the binding sequence for a (○) protein of interest, within 4096 sequences where (□) represents the PCR primers, N a nucleotide, ie, A, G, C, or T. An essential aspect of this experiment is the ability to separate protein-bound from free DNA prior to amplification. See text.

population were cloned and the DNA sequences of individual clones determined, two sequences were found to be equally represented and to have identical affinities. Because selection is stochastic, the representation of different sequences in the final, mixed population would reflect the affinity of the sequences for the protein with sequences having lowest K_d values being represented most often (22–24).

By similar logic, protein affinity libraries have been constructed to identify protein–protein combining sites, as in antibody–antigen interaction (24) and recombinant libraries have been made that produce a repertoire of antibodies in *E. coli* (25). In another case, a potential DNA-based therapeutic strategy has been studied (26). DNAs from a partially randomized library were selected

to bind thrombin *in vitro*. Oligonucleotides (termed "aptamers") that bound thrombin shared a conserved sequence 14–17 nucleotides long.

Combinatorial libraries are limited by the number of sequences that can be synthesized. For example, a library consisting of one molecule each of a 60-nucleotide sequence randomized at each position, would have a mass of $>10^{14}$ g, well beyond the capacity for synthesis and manipulation. Thus, even if nucleotide addition is random at all the steps during synthesis of the oligonucleotide only a minority of the sequences will be present in the output from a laboratory-scale chemical DNA synthesis reaction. In analyzing these random but incomplete libraries, the protocol is efficient enough to allow selection of aptamers with lowest dissociation constants from the mixture after a small number of repetitive selection and amplification cycles. Once a smaller population of oligonucleotides is amplified, the aptamer sequences can be used as the basis for constructing a less complex library for further selection.

8. Expression of Genes in a Heterologous Host

In many cases, it is possible to synthesize the product of a gene in a different organism, eg, bacteria, yeast, or higher eukaryote. The details of these methods are covered in other articles in this section. Recombinant DNAs directing the synthesis of the gene product must contain information specifying a number of biochemical processes:

8.1. Replication of the Recombinant DNA. In bacteria, this is provided by origin sequences, derived usually from plasmids indigenous to the host. Often, an *ori* sequence from one bacterium will not function in another; therefore the vector can contain two origins of replication, each functioning in a different host. These vectors are termed shuttle vectors. In some cases, the origin is provided by integrating the foreign DNA into the chromosome of the host.

8.2. Selection of Recombinants. This is provided either by a gene specifying antibiotic resistance or the ability to allow growth of recombinants in the absence of a particular nutrient.

8.3. Transcription of the Foreign Gene. Promoters are sequences preceding the start of transcription that direct RNA polymerase action. In general, they are specific to an organism and must be supplied, eg, when expressing a mammalian DNA sequence in bacteria. Often, expression of a heterologous gene is deleterious to the growth of the host; in these cases, strategies are available for the conditional transcription of the gene. For example, transcription of the gene might occur only at 42°.

8.4. Translation of the Foreign Gene. The translation of a mRNA into a protein is governed by the presence of appropriate initiation sequences that specify binding of the mRNA to the ribosome. In addition, not all the codons of the genetic code are used equally frequently by all organisms. Efficient translation depends on matching the preferred pattern of host codon usage in the heterologous gene.

8.5. Stability and Purification of the Recombinant Protein. There are no hard and fast rules specifying, eg, whether a recombinant protein is available in a soluble state in the cell. In some cases, the expression system must be engineered by *in vitro* mutagenesis to optimize overall yield of the protein.

9. Mutagenesis of Cloned DNA

Genetics begins with mutants; indeed, the primary definition of a gene is a unit of mutation. Mutational analysis of a cloned gene is often essential for identifying structure–function relationships in its expression or in the protein encoded by the cloned gene. Alternatively, expression of a recombinant protein is often dependent on the codon usage optimal for the host. A number of techniques are available for mutagenesis. Randomized treatment of DNA with a chemical mutagen continues to be useful. In addition, a short synthetic DNA can be made with specific or random mutations introduced during synthesis. This altered information can then be incorporated into the cloned gene. A few of the more general approaches are described here.

9.1. Mutagenesis by Synthetic DNA. This technique is the basis for many others and was historically the first one realized (Fig. 9, Ref. 27). A synthetic oligonucleotide with one specific alteration is hybridized to a single-stranded DNA (usually from a subcloning the gene to be mutagenized into a specialized bacteriophage). The oligonucleotide is then used to prime synthesis of a complementary strand. The double-stranded recombinant is then transformed or electroporated into a recipient cell. If the phage DNA is prepared from a mutant *E. coli* host that inserts deoxyuracil rather than thymidine into its DNA during growth, then subsequent growth in a wild-type host will discriminate against the template and mutant clones will be obtained with high efficiency. An alternative procedure uses two complementary mutagenic oligonucleotides and amplifies the rest of the sequence by the PCR (28).

It is possible to make a number of mutations in a small sequence by synthesizing the mutagenic oligonucleotide with a small amount of incorrect nucleotide at each position, such that each oligonucleotide contains an average of one or two incorrect bases (so-called "dirty bottle" synthesis). The oligonucleotide population is then hybridized to single-stranded DNA and primed synthesis is carried out as described above.

9.2. Linker-Scanning Mutagenesis. In this technology (29) small sequences of DNA are removed and replaced with a synthetic restriction fragment (linker). This technique is commonly used in analysis of promoters and other control sequences in DNA, while preserving the spatial relationship between the sequences.

9.3. Mutagenic PCR. Recently, methods have been developed to use the PCR reaction to randomly mutagenize a defined sequence (30). The *Taq* polymerase used in PCR will misincorporate nucleotides in a random fashion if $MnCl_2$ is included in the reaction buffer during PCR. The library of mutagenized PCR products can be screened for the desired phenotype.

10. Protein Pharmaceutical Products

Development of recombinant proteins for pharmaceutical use has grown exponentially since 1982, when recombinant human insulin received approval from the United States Food and Drug Administration. Paralleling the development of small-molecule pharmaceuticals, previously approved proteins are being

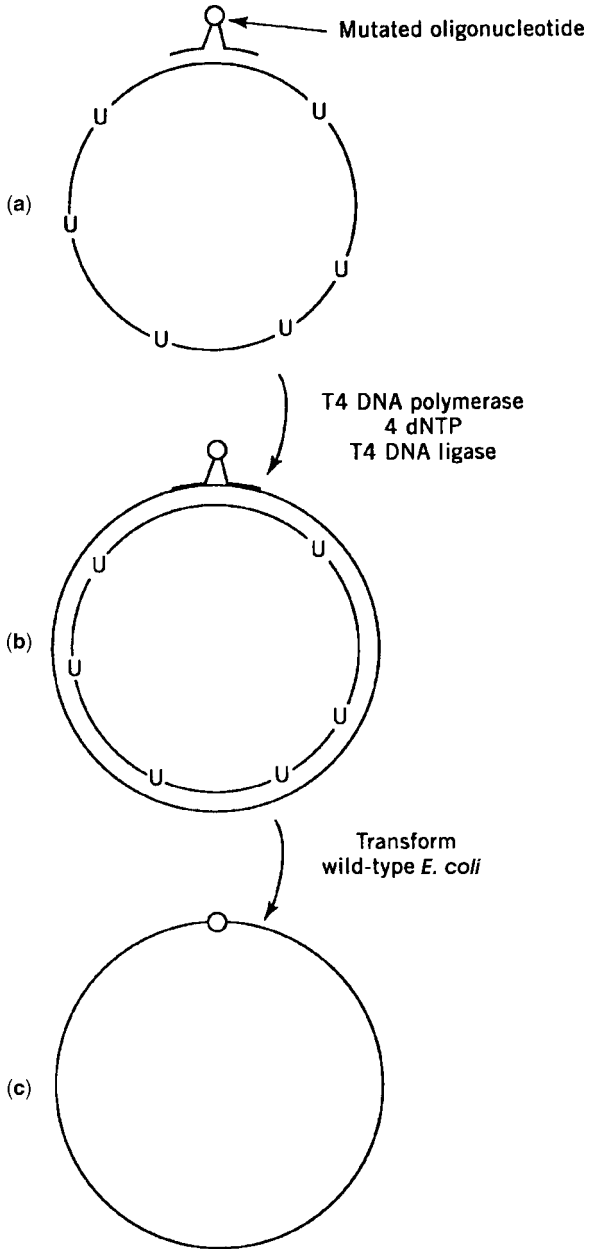

Fig. 9. Mutagenesis by a synthetic oligonucleotide of a cloned sequence available in single-stranded form: (**a**) single-stranded M13 template containing uracil (U) residues; (**b**) double-stranded product, uracil residues are not mutagenic; (**c**) strong selection for M13 phages containing mutation of interest (23).

tested for new applications. As of 2004, >200 recombinant proteins have been approved by the FDA for clinical use and an approximately equal number of small-molecule therapeutics have been developed by the biotechnology industry with the use of recombinant proteins as tool compounds.

11. Regulatory and Safety Issues

In the early days of recombinant DNA research, there were serious concerns about crossing assumed "species barriers" that prevented the exchange of DNA between different bacteria. Since that time, however, the exchange of DNA by conjugation has been demonstrated between numerous bacteria, including between Gram-negative and Gram-positive genera. Similar conjugal transfer has been shown to occur between bacteria and yeast. Thus, the DNA information in the biosphere can best be thought of as a continuum. In this background, safety regulations have been modified to recognize that no new hazards will be created by recombinant DNA research, eg, DNA introduction will not make a pathogen out of a nonpathogen (31).

12. Experimental Protocols

The experimental manipulations involved in construction and analysis of recombinant DNAs are well within the facility of a trained chemist, but they do require attention to details. For example, reactions may be exquisitely sensitive to temperature or to the presence of compounds used as enzyme stabilizers. In many cases, optimized reaction buffers, nucleotides, and enzymes are packaged in kits along with detailed protocols by the manufacturers. Beyond the information provided by reagent manufacturers, laboratories need a collection of experimental protocols (5,6). One such collection is distributed on a subscription basis and updated quarterly (5).

BIBLIOGRAPHY

"Genetic Engineering" in *ECT* 3rd ed., Vol. 11, pp. 730–745, by A. M. Chakrabarty, University of Illinois at the Medical Center, Chicago; in Suppl. Vol., pp. 495–513, by E. Jaworski and D. Tiemeier, Monsanto Co.; "Procedures", under "Genetic Engineering" in *ECT* 4th ed. Vol. 12, pp. 440–464, by Francis J. Schmidt, University of Missouri, Columbia; "Genetic Engineering, Procedures" in *ECT* (online), posting date: December 4, 2000, by Francis J. Schmidt, University of Missouri, Columbia, Mo.

1. R. J. Roberts and D. Macelis, *Nucleic Acids Res.* **20**, 2167 (1992).
2. G. Chu, D. Vollrath, and R. W. Davis, *Science* **234**, 1582 (1986).
3. J. E. Richards, T. C. Gilliam, J. L. Cole, M. L. Drumm, J. J. Wasmuth, J. F. Gusella, and F. S. Collins, *Proc. Natl. Acad. Sci.* **85**, 6437 (1988).
4. A. J. Jeffreys, A. MacLeod, K. Tamaki, D. L. Neil, and D. G. Monckton, *Nature (London)* **354**, 204 (1991).

5. F. M. Ausubel and co-authors eds., *Current Protocols in Molecular Biology*, Wiley-Interscience, New York.
6. J. Sambrook, E. F. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor N.Y., 1989.
7. J. Norrander, X. Kempe, and J. Messing, *Gene* **26**, 101 (1983).
8. L. Clarke and J. Carbon, *Cell* **9**, 91 (1976).
9. C. A. Adams, J. A. Fornwald, F. J. Schmidt, M. Rosenberg, and M. E. Brawner, *J. Bacteriol.* **170**, 203 (1988).
10. J. Sulston and co-workers, *Nature (London)* **356**, 37 (1992).
11. D. T. Burke and M. V. Olson, *Methods Enzymol.* **194**, 251 (1991).
12. M. C. Ianuzzi, M. Dean, M. L. Drumm, N. Hidaka, J. L. Cole, A. Perry, C. Stewart, B. Gerrard, and F. S. Collins, *Am. J. Hum. Genet.* **44**, 695 (1989).
13. F. Sanger, S. Nicklen, and A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
14. A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1980).
15. T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood, *Science* **254**, 59 (1991).
16. E. S. Lander and co-workers, *Nature (London)* **409**, 860 (2001).
17. J. C. Venter and co-workers, *Science* **291**, 1304 (2001).
18. W. R. Pearson, *Methods Enzymol.* **183**, 63 (1990).
19. R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin, *Nucleic Acids Res.* **29**, 22 (2001).
20. M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J. White, *PCR Protocols: A Guide to Methods and Applications*, Academic Press, New York, 1990.
21. J. Lapointe, C. Li, J. P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A. M. DeMarzo, R. Tibshirani, D. Botstein, P. O. Brown, J. D. Brooks, and J. R. Pollack, *Acad. Sci. U.S.A.* **101**, 811 (2004).
22. A. D. Ellington and J. W. Szostak, *Nature (London)* **346**, 818 (1990).
23. C. Tuerk and L. Gold, *Science* **249**, 505 (1990).
24. J. K. Scott and G. P. Smith, *Science* **249**, 386 (1990).
25. A. Plueckthun, *Immunol. Rev.* **130**, 151 (1992).
26. L. C. Bock, L. C. Griffin, J. A. Latham, E. H. Vermass, and E. H. Toole, *Nature (London)* **355**, 564 (1992).
27. T. A. Kunkel, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 488 (1985).
28. D. H. Jones and B. H. Howard, *Biotechniques* **10**, 62 (1991).
29. X. McKnight and R. Kingsbury, *Science* **217**, 316 (1982).
30. R. C. Cadwell and G. F. Joyce, *PCR Methods Appl.* **2**, 28 (1992).
31. B. D. Davis, ed., *The Genetic Revolution*, Johns Hopkins University Press, Baltimore, Md., 1991.

FRANCIS J. SCHMIDT
University of Missouri-Columbia