Molecular Biology in Epidemiology

Robert E. Hurst
and Jian Yu Rao

Introduction

The revolution in molecular biology began with perhaps one of the finest, most concise papers ever written: Crick and Watson’s seminal paper (1953) in which they demonstrated the double-stranded nature of DNA and how it was produced by base pairing. In a classical understatement, they stated, “It has not escaped our attention that the specific pairing of bases in DNA we have postulated immediately suggests a possible copying mechanism for the genetic material.” The 1960s saw the realization of results that flowed from this simple statement in the unraveling of the genetic code and the elucidation of the mechanisms of DNA transcription into messenger RNA (mRNA) and mRNA translation into proteins. In the 1970s, recombinant DNA technology and techniques to rapidly sequence nucleic acids emerged. Recombinant DNA technology allowed individual genes to be isolated and replicated; rapid sequencing techniques made it practical to sequence stretches of up to a few thousand base pairs with relatively little effort. The explosive radiation of this basic knowledge into almost every area of biology occurred in the 1980s, providing an understanding at the molecular level of the mechanisms and regulation of cell growth, the response to drugs and toxic chemicals, carcinogenesis, progression and metastasis, as well as many other biologic processes.

In this chapter, we provide an overview of the major concepts of molecular biology, presenting a brief outline of the major techniques and their strengths, weaknesses, and practical limitations. We present how these concepts can be used to map the origins of human disease at a molecular level and outline a scenario for integrating molecular biologic monitoring into
field epidemiologic studies. Of necessity, molecular biology is presented simply; our presentation assumes some knowledge on the part of the reader.

Basic Principles of Molecular Biology

The use of molecular biologic data in epidemiologic research requires a familiarity with how genes are organized and how they function. DNA constitutes the genetic material of the cell. Other than some rare methylated bases, DNA is composed of only four bases, the pyrimidines thymine (T) and cyto-

sine (C) and the purines adenine (A) and guanine (G). A segment of three base pairs is shown in Figure 2.1. The strands are arranged in the familiar double helix. The backbone consists of alternating sugars and phosphates; the bases are attached to the sugars at carbon 1. The numbering systems of sugars and bases are indicated. The strands have a polarity, indicated by the designations 5' and 3', as defined by the numbering system of the sugars. The strands are also complementary in structure because of the hydrogen bonding scheme in which A always bonds with T (or the closely related base ura- cil, U, found in RNA) and G always bonds with C. Thus, either strand can serve as the template for synthesis of the other. Note that A–T pairs have two hydrogen bonds whereas G–C pairs have three. The extra bond confers extra stability on G–C pairs. Information is stored in a hierarchy of structure. The lowest level is the individual nucleotide. A codon consists of 3 nucleotides and specifies a single amino acid. A gene is the functional unit that specifies a protein. A chromosome contains several thousand genes and is the smallest replicating unit. The genome is the complete set of genetic information that an organism contains.

The structure of a generic stretch of mammalian DNA is illustrated in Figure 2.2. This figure presents a solitary coding gene and a gene family separated by long stretches of noncoding DNA. In general, less than 30% of the total genome is ever transcribed into RNA, and only a fraction of that is ever translated into protein. At least 20 kilobase pairs (kb; 1000 base pairs) separate the solitary coding gene from other expressed elements. In general, many of these stretches consist of repetitive sequences that usually constitute over 70% of the total genome. The structure of the repetitive sequences can repeat on scales of every 5–10 base pairs (bp), every 100–300 bp, or even every 5000–6000 bp. Usually, the repetitive structures are not exact copies of each other; repeats of high to moderate sequence homology (similarity in sequence) are found. Of particular interest are Alu sequences, found in mammals, that repeat on a scale of 150–300 bp. These sequences have a species specificity and can be used to mark genes that are introduced into cells of

![FIGURE 2.1](image1.png) Structure of 3 complementary base pairs showing numbering systems for bases (unprimed) and deoxyribose (primed).

![FIGURE 2.2](image2.png) Structure of a large segment of mammalian DNA, including a gene family. Boxes indicate genes, which are sequences that are transcribed into RNA. Shown are a solitary coding gene and a gene family such as the globin family or keratin family. Untranscribed pseudogenes (θ) are also shown. A more detailed structure of a coding gene and RNA processing is shown in Fig. 2.3.
another species. Alu sequences are common in introns and may play a part in mRNA splicing. Also present in these noncoding regions are sequences that bind to various proteins and regulate the transcription of genes. These control regions can range upstream (in the 5' direction) up to several thousand bp from the start of the gene.

Downstream from the solitary coding gene in the figure is a gene family. Such gene families are common and code for proteins that are structurally related and whose expression is coordinated in some way. Rarely, however, are all the members expressed at the same time in the same cell. The globin genes, for example, are expressed at different times during embryogenesis. The keratins are expressed in different cell types; only a few of the over 20 genes are expressed in any one cell. Gene families also may contain pseudogenes, which are homologous in structure to the other members of the gene family but are not expressed. Gene duplication, with or without recombination of parts of other genes, is apparently the most common mechanism by which new proteins are produced. Unequal recombination, in which one strand receives both copies of a gene, is the mechanism for such duplication. How the expression of gene families is regulated and coordinated is an important unanswered question.

Figure 2.3 illustrates the structure of a gene on a finer scale and the transcription of its sequence into RNA together, in addition to the further processing that the RNA undergoes to produce mRNA used as the template for protein synthesis. The eukaryotic gene sequence almost invariably consists of sequences that are expressed (exons), that is, encode protein, and intervening sequences (introns) that do not encode protein. Regulatory sequences are found upstream of the initiation site and, less commonly, downstream. One regulatory sequence that is found very commonly at -20 to -30 bp is a "TATA box," so called because it contains the sequence TATAA. The TATA box orientates RNA polymerase II, the enzyme responsible for transcription of most protein genes; most eukaryotic genes seem to have such a sequence. In general, promoter elements are found in the regions between -500 and -100 and -70 and -30 bp, where 0 bp is the start of the transcribed region. These regulatory sequences bind regulatory proteins that enhance or inhibit the binding of RNA polymerase and constitute the mechanism by which the expression of a gene can be regulated or coordinated with the expression of other genes. Another sequence that commonly is found in these regions is CAAT, which whimsically has been named "CAAT box."

The first step in transcription is production of a "primary transcript" RNA by RNA polymerase II. This transcript is complementary to the template DNA sequence, ranges from the initiation sequence on the DNA to a termination sequence, and includes both exons and introns.

Transcription starts with the binding of RNA polymerase II. This binding generally requires a transcription factor. Transcription factors are proteins that recognize the region of DNA to be transcribed and enable polymerase binding. The RNA polymerase proceeds in a 3' → 5' direction along one strand of the DNA and synthesizes the complementary RNA sequence, including exons, introns, and untranslated sequences on both the 5' and 3' end. This molecule is referred to as a "primary transcript." A convention exists concerning the sense and complementarity of DNA and transcribed mRNA. The two strands of DNA are labeled "sense" and "antisense." Transcription is considered to proceed from the antisense strand in a 3' → 5' direction, which is 5' → 3' on the sense strand. Thus, the sense strand of the DNA and the RNA transcribed from the antisense strand have the same direction and sequence, except U in RNA substitutes for T in DNA.

The cellular nucleus contains a large variety of primary transcripts,
known collectively as “heterogeneous nuclear RNA” (hnRNA). The first step in RNA processing is to cap the 5’ end by adding a guanilate nucleotide to the 5’ end of the chain in a unique 5’ → 5’ arrangement. This G is then methylated. The RNA no longer has a free 5’-OH group. The next step in processing consists of removing and destroying sequences at the 3’ end that extend past the actual protein-coding sequence. A long repetitive sequence of adenylate nucleotides is then added to the 3’ end of the RNA. This “poly(A) tail” is approximately 250 nucleotides long in mammals. After this reaction, a complex series occurs during which the intron sequences are clipped out and the exon sequences are spliced together very precisely. The product of this reaction is the mature mRNA, which is exported to the cytoplasm where it is translated into protein by specialized organelles called ribosomes. Most proteins are further modified by one or more of several possible posttranslational modification steps, including proteolytic cleavage, addition of carbohydrate or lipid moieties, and modification of amino acids.

**TABLE 2.1 Terms Used in Molecular Biology**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>anneal</td>
<td>To incubate DNA at a temperature a few degrees below its melting point to allow complementary sequences to hybridize.</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA is derived from RNA by reverse transcriptase. If prepared from mRNA, the usual template for cDNA, it differs from genomic DNA in not containing introns.</td>
</tr>
<tr>
<td>clone</td>
<td>A genetic unit (cell, plasmid, vector) derived from a single progenitor.</td>
</tr>
<tr>
<td>codon</td>
<td>A three-base sequence that codes for a single amino acid.</td>
</tr>
<tr>
<td>denaturing gel</td>
<td>A gel for electrophoresis in which the molecules are not in their native conformation (single-stranded DNA, random coil proteins).</td>
</tr>
<tr>
<td>epitope</td>
<td>The specific portion of a protein recognized by an antibody.</td>
</tr>
<tr>
<td>exon</td>
<td>A section of DNA that is expressed in a protein sequence.</td>
</tr>
<tr>
<td>gene</td>
<td>The functional unit of genetic inheritance and a sequence of DNA that codes for a single polypeptide.</td>
</tr>
<tr>
<td>genome</td>
<td>The complete genetic information of an organization.</td>
</tr>
<tr>
<td>genomic DNA</td>
<td>DNA derived from the genome (vs cDNA).</td>
</tr>
<tr>
<td>hnRNA</td>
<td>Heterogeneous nuclear RNA. The primary transcript.</td>
</tr>
<tr>
<td>hybridoma</td>
<td>A hybrid cell derived from fusing an immortal cancer cell (myeloma cell) with an antibody-producing B-cell.</td>
</tr>
<tr>
<td>immunogen</td>
<td>A protein used to immunize an animal to produce antibodies.</td>
</tr>
<tr>
<td>intron</td>
<td>Intervening sequence between coding regions of DNA.</td>
</tr>
<tr>
<td>library</td>
<td>A collection of plasmids containing recombinant DNA fragments obtained from genomic or cDNA.</td>
</tr>
<tr>
<td>ladder gel</td>
<td>The gel resulting from DNA sequencing; so called because the result resembles a ladder.</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA. RNA that is used as a template to code for protein synthesis.</td>
</tr>
<tr>
<td>melt</td>
<td>To denature (dissociate into two single strands) double-stranded DNA by heating.</td>
</tr>
</tbody>
</table>

**TABLE 2.1 Terms Used in Molecular Biology (continued)**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>northern blot</td>
<td>A technique for RNA analysis in which RNAs are separated by electrophoresis, transferred (“blotted”) to a carrier, and visualized by hybridizing to radioactive probes.</td>
</tr>
<tr>
<td>oligonucleotide</td>
<td>A short (&lt;20) sequence of bases.</td>
</tr>
<tr>
<td>oncogene</td>
<td>A gene that codes for a protein involved in cell growth that unless properly regulated will result in inappropriate growth.</td>
</tr>
<tr>
<td>plasmid</td>
<td>A genetic element, usually circular, derived from bacteria and used to carry inserted genetic elements (see vector).</td>
</tr>
<tr>
<td>recombinant DNA</td>
<td>The techniques for recombining DNA from disparate sources into a single molecule.</td>
</tr>
<tr>
<td>restriction nuclease</td>
<td>An endonuclease that cleaves double-stranded DNA at precise sequences.</td>
</tr>
<tr>
<td>restriction fragment length polymorphism (RFLP)</td>
<td>A difference in the sequences of DNA from the two chromosomal copies such that cleavage by restriction nuclease produces different sized fragments from each copy.</td>
</tr>
<tr>
<td>restriction site</td>
<td>A DNA sequence recognized by a restriction nuclease.</td>
</tr>
<tr>
<td>sequence homology</td>
<td>The similarity in sequence between different nucleic acids. Homology is usually expressed as a percentage.</td>
</tr>
<tr>
<td>stringency</td>
<td>Solution conditions and temperature for hybridization that regulate the degree of homology required for hybridization.</td>
</tr>
<tr>
<td>Southern blot</td>
<td>A technique of DNA analysis in which DNA fragments are separated by electrophoresis, transferred (“blotted”) to a carrier, and visualized by hybridizing to radioactive probes.</td>
</tr>
<tr>
<td>tumor suppressor gene</td>
<td>A regulatory gene that prevents inappropriate cell growth.</td>
</tr>
<tr>
<td>vector</td>
<td>A genetic element, usually a plasmid or virus, used to carry exogenous DNA inserts for replication and manipulation.</td>
</tr>
<tr>
<td>western blot</td>
<td>A technique of protein analysis in which proteins are separated by electrophoresis, transferred (“blotted”) to a carrier, and visualized using an antibody against the protein and immunochemical visualization.</td>
</tr>
</tbody>
</table>
Molecular biology has developed truly amazing techniques to probe these systems for genes of interest from a large collection of possible genes. The mammalian genome has been estimated to consist of 100,000 genes; some 5,000–10,000 normally are expressed in any given cell type. Much of molecular biology is finding a “needle in the haystack,” where the “needle” is a gene sequence or protein of interest in a “haystack” of other sequences or proteins. Molecular epidemiology seeks to apply these techniques to understanding the causes of disease and to developing markers that can chart the progress of disease. These techniques can advance infectious disease epidemiology by defining and detecting infectious organisms with accuracy and sensitivity not possible with conventional serologic techniques. In addition, molecular techniques used to examine the mRNAs in a given cell or tissue directly identify the genes that are expressed and how that expression may change with exposure to infectious or toxic agents. Also these procedures elucidate how other agents may modify these responses. Finally, using monoclonal antibodies and other protein chemistry techniques, the levels of specific protein markers that ultimately govern the behavior of cells and define their phenotypes can be measured with high precision.

Core Techniques of Molecular Biology

This section reviews the general principles and some applications to epidemiology of the major methods of molecular biology. Specific experimental details can be found in several excellent laboratory texts cited at the end of this chapter.

Restriction Endonucleases

The DNA of even a small chromosome is enormous, consisting of millions of base pairs. Such a large molecule cannot be manipulated conveniently. One of the first breakthroughs of molecular biology was the identification of restriction endonucleases, which are enzymes found in bacteria that cleave double-stranded DNA at precise sequences that are often palindromic. A palindrome reads the same backward and forward. The classic word palindrome is “Madam, I’m Adam.” A palindromic DNA sequence is CGATCG because its complementary sequence GCTAGC reads the same from right to left (5’ to 3’). These DNA restriction enzymes serve as bacterial defenses against foreign viral DNA, since their own DNA is protected by methylation of one or more bases in the restriction site, or sequence of bases defining the specificity of the restriction enzyme. Restriction endonucleases are named for the organism of origin; a roman numeral indicates the order of discovery. Thus, EcoRI is the first enzyme isolated from the R strain of Escherichia coli; HpaII is the second enzyme to be isolated from Haemophilus aphrophilus. Restriction endonucleases allow precise cleavage of DNA into fragments of reproducible size that can be separated by electrophoresis or other techniques. The size of the fragments produced generally is a function of the number of bases in the restriction site; the larger the site, the less frequent it is, and the fewer there are in a genome. “Four cutters,” which detect tetrameric sequences, usually cleave genomic DNA into small fragments, whereas “eight cutters” usually produce gene-sized fragments because of their lower frequency. Interestingly, a site recognized by the eight cutter NotI frequently is found at or near the beginning of many mammalian genes and rarely anywhere else. Restriction endonucleases can leave “blunt” ends, as illustrated for BalI in Figure 2.4, or “sticky” or “overhanging” ends, as illustrated for EcoRI. As discussed later, these properties are useful in recombinant DNA technology for ligating DNA fragments into vectors that can be amplified or multiplied.

![Restriction Endonucleases](image)

**FIGURE 2.4** Actions of restriction nucleases. BalI yields blunt ends, EcoRI yields “sticky” ends.
Hybridization

A fundamental property of nucleic acid is base complementarity or base pairing, the formation of hydrogen bonds that allows one strand of a nucleic acid to specify a complementary strand. Adenine (A) always pairs with thymine (T) or uracil (U) and guanine (G) always pairs with cytosine (C). This formation of complementary sequences stabilized by hydrogen bonds is the major principle by which molecular biology finds the “needle in the haystack.” Hybridization is one of the cornerstones of molecular biology. A probe consisting of a segment of known sequence is isolated or synthesized and is used to “fish out” the complementary sequence from a large sample of heterogeneous material.

Although the principle of hybridization is simply the recognition of complementary sequences, the technology depends on an understanding of the conditions that stabilize the DNA double helix and how these conditions are manipulated to disrupt, or melt, the DNA into two single strands or to allow the double-stranded helix to form again, or anneal. The first step in any hybridization experiment is to convert double-stranded DNA into single-stranded DNA to allow the probe to bind to complementary sequences. Failure to properly meet the conditions required for this process will lead to incorrect results.

The melting of a double-stranded helix is a sudden event that occurs over a narrow temperature range, as illustrated in Figure 2.5. As the temperature is raised, a point is reached at which the thermal energy begins to overcome the bond energy of the hydrogen bonds that stabilize the helix, so it begins to come apart. Because the stabilization of the helix is cooperative, that is, the formation of one base pair increases the probability that a neighboring base pair will form, the transition to single strands occurs suddenly. The melting temperature, or Tm, represents the temperature at which half the bonds are broken. DNAs have different Tm's as determined by the percentage of G−C pairs, among other factors. Because G−C pairs form three hydrogen bonds and A−T pairs only two, the former are more stable; therefore, Tm increases with GC content.

After dissociation of the DNA into its constituent single strands, the probe is added and double strands are allowed to reform. This process is called annealing and depends on several factors, including the concentration of DNA, the time allowed for annealing, the temperature, and the concentration of salts in the solution. The latter two variables are referred to collectively as stringency. The process of annealing is statistical and depends on successive formation and breaking of hydrogen bonds, always leading to successive minimization of potential energy, which corresponds to formation of the maximum number of hydrogen bonds. If the annealing temperature is too low, insufficient thermal energy will be available to support this process. The binding that occurs tends to be nonspecific. Rapid cooling on ice tends to suppress hydrogen bond formation and preserves single strands. Annealing at too high a temperature will melt the strands continuously.

Manipulating stringency allows the experimenter to control the degree of allowable mismatch between the probe and its target sequence. Stringent conditions are those of relatively high temperature and low salt concentrations which favor helix formation only between exactly complementary sequences. By carefully maintaining the temperature just under Tm, it is sometimes possible to distinguish single base pair mutations in small DNA fragments. Under lower stringency, some mismatches are allowed. Less stringent conditions can be used to probe for homologous genes or to use probes from another species in conserved genes. Thus, the entire family of hemoglobin or ras genes, for example, can be detected using a probe under lower stringency, that is, exactly complementary to one member of the family. Likewise, a readily available mouse probe sometimes may be substituted for an unavailable human probe. The concentrations of reactants and the time allowed for reaction are also important in hybridization; longer times lead to more complete hybridization. Mathematical models, which are beyond the scope of this chapter, have been derived to describe hybridization. These factors are discussed at much greater depth in laboratory manuals.

The simplest use of hybridization is the dot or slot blot. In this procedure, a drop of single-stranded DNA-containing solution is spotted onto a nitrocellulose or nylon membrane and immobilized by one of several techniques. The membrane is then “prehybridized” with sheared, heterologous DNA (e.g. a salmon sperm DNA) or RNA (e.g., tRNA) to reduce nonspecific binding of radioactive probe. The radioactive probe is then added and allowed to hybridize, usually overnight. The membrane is washed under conditions of appropriate stringency until no further radioactivity is detectable.
in the wash solution. The radioactivity remaining on the membrane is bound specifically. The blot is developed by laying a piece of X-ray film over the membrane and placing the “sandwich” in the freezer. Disintegrations of the radioactive tracer will expose the film. Freezer temperatures slow the reactions in the film emulsion, increasing resolution. Thus, the presence of a particular sequence and an estimate of the amount can be obtained from the density of the spot.

**Electrophoresis**

Another important technique of molecular biology is electrophoresis, in which proteins or nucleic acids are separated by size and charge. In practice, all electrophoretic techniques are carried out using a supporting gel of controlled pore size. This gel minimizes diffusion, which tends to degrade resolution. Usually gels are prepared from polyacrylamide or agarose. Polyacrylamide is used to separate smaller molecules, whereas agarose is used for larger ones. Although the theory of electrophoresis shows that the charge actually governs migration of molecules in an electric field, most separations are by size and governed by size. This sieving occurs because the holes in the gel are approximately the size of the molecules being separated; larger molecules are held up and migrate more slowly than smaller ones. Moreover, with DNA, each nucleotide has one charge, so all DNA molecules have very nearly the same charge density, or ratio of charge to molecular weight. For proteins this is not true, since the charge of proteins is highly dependent on its amino acid composition. However, by adding the charged detergent sodium dodecyl sulfate (SDS), which binds strongly and in large numbers to proteins, the native charge of the molecules is overwhelmed by the large numbers of SDS molecules, leading to a separation of proteins mainly by size as well. The size of the pores in the gel is regulated by the composition of the gel. Polyacrylamide gels are formed by adding a chemical cross-linking agent to acrylamide; the higher the concentration of cross-linker, the tighter the gel. Porosity of agarose gels is controlled by the amount of agarose added to the gel. The gel can be denaturing or non-denaturing, depending on whether agents such as urea or SDS are added. High concentrations of urea disrupt hydrogen bonds, thereby dissociating double-stranded DNA into single strands. For proteins, a sulfhydryl reagent such as mercaptoethanol or diethiothreitol is added to reduce secondary disulfide bonds that help maintain the native conformations of many proteins, allowing the SDS to denature the protein completely.

Polyacrylamide gels are prepared by pouring the liquid gel between two cleaned glass plates separated by a spacer. “Combs,” devices used to form wells into which to inject sample at the top of the gel, are placed in the gel and the gel is allowed to polymerize. Gels can be operated in the vertical position, which is usual for DNA sequencing gels or protein gels, or in the horizontal position, which is usual for DNA separations. The sample usually is dissolved in a loading buffer that contains a “tracking dye,” which migrates more rapidly than the sample and is used to track the progress of the separation. The loading buffer solution also is more dense than the running buffer surrounding the gel, preventing the sample from floating away when it is applied with a small pipet directly into the well. The separation is started by applying a voltage ranging from several tens of volts for a small gel to 1000 or more volts for a large gel. Nucleic acids and proteins in SDS gels, both being negatively charged, migrate toward the positive electrode (anode). After an appropriate time, the electrophoresis is stopped and the DNA or protein is visualized by one of several techniques.

The simplest technique for visualizing DNA is to stain the gel with ethidium bromide, a dye that binds tightly to double-stranded DNA and fluoresces brightly under ultraviolet light. The ethidium bromide can be added prior to electrophoresis or the entire gel can be bathed in a solution containing the dye; the unbound dye is removed by washing. Blotting techniques can be used also.

The Southern blot, named for Edward Southern who developed the technique in 1975, is used to visualize DNA sequences; an analogous technique, the Northern blot, is used to visualize RNA sequences. In the typical Southern blot, the gel is laid atop a piece of nylon or nitrocellulose membrane. Several techniques are available for eluting the DNA from the gel, including electroelution, which uses an electric field, or simple capillary action generated by pieces of filter paper overlaid by paper towels, causing a flow of buffer through the gel onto the membrane. Transfer of fragments larger than 5 kb becomes progressively poorer with increasing size. Southern blotting is used to identify the number and sizes of DNA fragments that contain a particular sequence. The technique is used, for example, to identify fragments containing a particular sequence in restriction fragment length polymorphism (RFLP) analysis (see subsequent text) and to analyze complementary DNA (cDNA) (see subsequent text).

The Northern blot, so named because it detects the nucleic acid “opposite” that detected by the Southern blot, is an important technique for studying gene expression. The total RNA fraction is isolated from a cell sample or tissue, as described subsequently, and 10–20 μg are loaded into a lane on a 1% agarose gel containing 2.2 M formaldehyde to destroy secondary structure in the RNA. After electrophoresis, the RNA is transferred to a membrane and probed. The bands are used to identify the size of the primary transcript as well as alternative splicings of the primary transcript into different messages. Mammalian messenger RNA is frequently assembled by differential splicing, leading to different sized messages. The Northern blot also is used to determine whether a particular tissue or cell expresses a certain gene and to estimate the level of expression.

An analogous technique exists for proteins. Proteins are separated by
RNA has been isolated. Usually the mRNA fraction rather than total RNA is desired. mRNA is isolated by chromatography on oligo(dT)-cellulose or poly(U)-Sepharose, taking advantage of base pairing between the poly(A) tail on the mRNA and the oligo(dT) or poly(U). A single oligo(dT) step usually will remove most tRNA and reduce rRNA by about 50%; a second step generally reduces the level of rRNA to about 10% of the original. RNA isolation techniques are continuously being improved. Commercially available kits include a “one-step” isolation of mRNA involving lysis in guanidine isothiocyanate and oligo(dT) chromatography without the intermediate step of purification of total RNA, on the use of oligo(dT)-coated magnetic beads for “30 minute” isolation of highly purified mRNA.

Synthesis of cDNA

The conversion of unstable mRNA into stable DNA through the use of reverse transcriptase originally isolated from retroviruses is the first step in many procedures. In essence, an oligonucleotide primer (a short sequence of bases, generally less than 20) complementary to a short sequence on the mRNA is used to initiate the reaction. Reverse transcriptase then synthesizes a complementary copy in DNA from deoxyribonucleoside triphosphates in the reaction mixture. Because of its complementary nature, the product is referred to as cDNA. Although the process is conceptually simple, several choices must be made and a few shortcomings must be understood. Reactions must be carried out carefully to maximize the synthesis of full-length cDNA and minimize the synthesis of small fragments and a second strand. Second strand synthesis can occur when the cDNA forms a hairpin loop by doubling back on itself. Two reverse transcriptases are available, one from Moloney murine leukemia virus (MMLV) and the other from avian leukemia virus (ALV). The two enzymes are not identical, responding differently to agents such as pyrophosphate, spermidine, and actinomycin D that are added to increase full-length cDNA synthesis and inhibit hairpin loop formation. The choice of primer is also important. When oligo(dT) is used to prime the reverse transcriptase, it is possible to obtain full-length cDNAs. In general, yields of full-length cDNA are less than 50% relative to the mRNA template, introducing a distinct bias toward the 3’ end of the mRNA. A cDNA more representative of the entire mRNA can be synthesized at the expense of obtaining full-length cDNAs by priming with a random mixture of hexamers. A third alternative is to prime with a specific primer that contains a restriction site, an approach that is useful in cloning.

Polymerase Chain Reaction

Polymerase chain reaction (PCR) techniques have made certain experiments possible on a time scale of hours that otherwise would have taken weeks or
months. With PCR, cDNAs in the 300–1000 bp size range can be duplicated explosively to microgram quantities from femtogram quantities and, in theory at least, from a single DNA or RNA molecule. The principle of PCR is illustrated in Figure 2.6. The original template can be single or double stranded. The key to PCR is the selective use of specific oligonucleotide primers to prime synthesis only of sequences complementary to the primers. Two primers are needed, one for each strand of the double helix. The primers

![Diagram of PCR cycle](image)

1. Primer binding
2. Elongation
3. Denaturation, second primer bind
4. Second elongation

FIGURE 2.6 Principles of the polymerase chain reaction. Chains are melted, cooled, and primers added. The antisense primer (ASP) binds to the sense strand, while the sense primer (SP) binds to the antisense strand. After elongation, 2 complementary chains are produced. The A1 chain is antisense but delimited on the 3' end by ASP, while the converse is true for the S1 chain. In the second round, the S and A chains again serve as templates for A1 and S1 chains, but the A1 and S1 chains now serve as S2 and A2 chains delimited on both ends. S2 and A2 chains then serve as templates for other A2 and S2 chains, which then multiply exponentially.

usually are custom synthesized on automated synthesizers at a cost of a few hundred dollars. The enzyme used in PCR is one of several DNA polymerases isolated from thermophilic bacteria native to hot springs or oceanic vents. Nucleoside triphosphates and magnesium ion also are added to the reaction mixture. If the template is single stranded, the first step consists of synthesis of a complementary second strand by the polymerase. If the template is a double-stranded DNA, this step is skipped. The temperature is increased to above 95°C for 1–2 min to melt the two strands. The reaction mixture is cooled rapidly to near 50°C for 1–2 min to allow the primers to anneal, then raised to 70°C for 30 sec to 10 min to allow the DNA polymerase to synthesize a second complementary strand. The two primers delimit the size of the sequence, as illustrated in Figure 2.6. This program is repeated 20 to 40 times, producing an enormous amplification of the original template. If doubling were perfect each time, this process would produce $1.1 \times 10^{12}$ copies from each template. In fact, the efficiency is usually considerably less; practical applications usually yield $10^7$–$10^8$ copies.

The primers need not match at the 5' end completely. The ability to introduce sequences at the 5' end of the molecule is one of the strengths of PCR. The experimenter can take advantage of this to, for example, introduce restriction sites than can be used to clone the products. As an example, a primer synthesized to be complementary to 20 bases is synthesized with a CTAG sequence at the 5' end. This sequence is not complementary to the target, but is the sequence recognized by EcoRI. Because each sequence of the amplified DNA contains one primer sequence, each of the amplified product sequences will contain an EcoRI restriction site. Another use is to introduce specific mutations into the product at a specific site. Thus, to introduce an A → G mutation at nucleotide 77 of a particular gene, the primer would incorporate nucleotides 60–80, but G would be substituted for A at position 77. Each of the amplified sequences will now contain a G at position 77 rather than the wild-type A. This approach of site-directed mutagenesis can be used to investigate the structure and function of proteins by introducing specific mutations at specific sites.

The particular strength of PCR is that the amplification can be highly selective. For example, the ras gene can be amplified selectively in the presence of the remainder of the genome or total RNA isolated from a cell. The amplified ras genes can be probed for mutants as a way of determining the ratio of mutant to wildtype genes. Even more exciting, as described in the applications, mutant sequences can be amplified selectively in the presence of normal sequences. PCR amplification also tends to favor relatively short amplified sequences, usually less than 1000 bp. In producing cDNA from mRNA, there usually is a distribution of lengths of cDNAs; the full-length product constitutes 50% or less of the total. There seems to be a limit to the amount of total amplified DNA that can be produced, particularly after 20 cycles or so; shorter sequences seem to have a competitive advantage over
longer sequences. Thus, libraries produced from PCR-amplified cDNA derived from mRNA will contain few full-length sequences. However, such libraries are nonetheless useful by producing cDNA probes that can be used later to extract full-length cDNA or to isolate the genomic sequence.

In practice, PCR, like all techniques, has considerable “art” associated with it, as well as several limitations. The selectivity lies almost entirely in the choice of primers. Since many genes share regions of homology, selectivity is rarely as perfect as the experimenter desires. Primers also should be selected carefully to avoid internal complementarity and palindromes, particularly at the 3' end. Primer complementarity produces “primer dimers,” often as the main reaction product. Also, contamination from exogenous DNA in the laboratory can be a serious problem. Often, PCRs take place even when no DNA was added to the reaction tube. This potential for contamination requires careful attention to technique. The PCR laboratory should be kept as separate as possible from amplified DNA or DNA plasmid solutions. Separate pipets should be maintained, as should careful attention to control experiments. Non-specific priming can be a problem also; nucleic acid preparations should be purified carefully by precipitation to remove small fragments that could prime non-specifically. The melting and annealing temperatures are also important considerations. Sequences high in GC content may not melt completely and sequences high in AT may not anneal completely, or may anneal at a temperature so low it supports non-specific priming. Unlike many claims, PCR is an exacting technique that must be optimized carefully and carried out with elaborate positive and negative controls to be successful.

**Ligase chain reaction (LCR),** a modification of the PCR approach, is much better suited to detecting single point mutations. As in PCR, repetitive cycles of denaturation, annealing, and hybridization are used. In contrast to PCR, in which the two primers delimit the fragment to be amplified, LCR uses two primers that completely span the region of interest; if they match the sequence of the template DNA exactly, the two primers will lie flat on the template, leaving only an incomplete sugar-phosphate bond gap between them. A thermostable DNA ligase now joins the exactly matching primers. On the other hand, if they do not match exactly due to, for example, a point mutation or deletion, then DNA ligase will not join the two ends. Repetitive cycling will lead to accumulation of joined oligonucleotide primers if there is an exact match. In the absence of an exact match, no amplification will be seen.

Various techniques for detection have been developed. One of the simplest is biotinylation of one of the primers and addition of a fluorescent or radioactive label to the other. Passing the reaction mixture through oligo(dT) an avidin-coated polymer column will remove the biotinylated primers. If they have been joined to the labeled primers, label will be retained on the column; otherwise it will pass through the column into the effluent. This technique is a very simple screen for point mutations and deletions.

**Cloning**

The objective in cloning is to isolate and expand a single cDNA obtained from a single mRNA or fragment of genomic DNA and permanently maintain that cDNA in a form that can be duplicated. In general, this problem is approached by making a library of cDNAs or genomic fragments, inserting all the cDNAs or genomic fragments in the mixture into vectors. The process is illustrated in Figure 2.7. A vector is a DNA molecule, usually circular, that can incorporate the target DNA and be propagated stably in a suitable host cell. Many plasmid DNAs are circular because they have complementary ends that hybridize with each other. Vectors usually are prepared from plasmids or viruses. It is relatively simple to insert foreign sequences, such as promoters or a variety of restriction sites, into plasmids, but viruses generally exhibit a limit to the amount of additional genetic material that can be inserted. The cDNA produced by reverse transcription must be treated further to produce a double-stranded DNA. The RNA that served as the template is degraded partially, leaving the single-stranded cDNA intact. Fragments of

---

**FIGURE 2.7** Cloning. A cDNA is made from mRNA. RNase H cleaves the mRNA, leaving a few fragments to act as primers for DNA polymerase. DNA ligase ligates the junctions, and the Klenow fragment of DNA polymerase and T4-polymerase fill in any unmatched sequences at the ends. Adapters containing restriction sites are added to the blunt ends, making the “insert,” which is now ligated into the vector. When these are inserted into bacteria, and a single colony containing a single insert is selected and expanded, the original cDNA is said to be cloned.
The original RNA usually serve as primers for second strand synthesis. The two steps usually are carried out in a single reaction by mixing RNase H to fragment the original RNA, DNA polymerase to synthesize the second strand (using the RNA fragments as primers and the first strand as the template), and DNA ligase to fill in the breaks. The double-stranded cDNA must now have linkers or adapters attached to the ends to produce ends that are complementary to a restriction site, unless the cDNA was synthesized by PCR using primers that contained such sites. Inserts produced by PCR are treated with the appropriate exonuclease to cleave any artificial adenine groups that are sometimes added upstream of the primers during amplification.

The currently favored route for cloning an insert that does not contain restriction sites is ligation (attachment) of the blunt-ended double-stranded cDNA to appropriate oligonucleotide adaptors using the enzyme T4 ligase. First, the cDNA is blunt-ended, either using the Klenow fragment of DNA polymerase, which fills in any overhangs on the 3' end, or using T4 polymerase, which fills in 5' overhanging termini. The vector contains one or more restriction sites that allow it to be cleaved and that will be complementary to a similar restriction site in the insert DNA. After the treatment with the appropriate restriction enzyme, the insert and opened vector both contain complementary single-stranded ends, so they hybridize. The enzyme DNA ligase closes the breaks, creating a stable circular DNA that contains the insert. The vector is introduced into bacteria, a process called transfection. The bacteria grow and duplicate the vector. Unfortunately, the library thus formed is a random collection of fragments with no classification system. It must be screened to identify the clones of interest.

Cloning can be directional or nondirectional, depending on whether the same or different restriction sites are on the two ends of the insert. If the same restriction site is on both ends, then the orientation of the insert will be random; half the inserts will match (+) and (-) strands of the vector with the insert and half will be in the opposite orientation. Orientation is most significant when expression of the insert is desired.

The screening strategies for the two types of vectors, plasmids and viruses are different. To screen a plasmid, bacteria must be selected that have received the plasmid. This selection is achieved by using plasmids that encode antibiotic resistance; plating the bacteria on a medium that contains the antibiotic selects only those bacteria that received a plasmid. These cells will survive and grow as a clonal colony. Each colony will contain a single cDNA insert.

If the vector contains an initiation site and a promoter, some of the inserts will be in the correct reading frame to produce protein, which can be detected with antibodies. Alternatively, the colonies can be screened by hybridization to a radioactive nucleic acid probe. In either case, the plate containing the colonies is touched to a hybridization filter and the bacteria are lysed. Screening with antibodies is performed similarly to a Western blot, whereas screening with nucleic acids is performed similarly to a Southern blot. When phage vectors are used, they are mixed with a large excess of bacteria. The entire mixture is then plated out. Each bacterium that contains a phage will serve as a focus for further growth and lysis, resulting in clear plaques on an otherwise dense lawn of bacteria. The plate is probed as described using blotting techniques.

When particular clones containing cDNAs of interest are found, they can be grown in almost limitless quantities. The inserts can be manipulated further by, for example, removing the insert with the appropriate restriction nuclease and inserting it into another vector, perhaps one that supports protein synthesis. This manipulation allows production of so-called "recombinant" protein. If placed in a vector active in mammalian cells, the inserted gene can be expressed and translated, a technique that allows gene function to be studied under controlled conditions. Placing an inducible promoter, for example, the androgen- or estrogen-sensitive promoters, in front of the recombinant gene allows it to be turned on and off at will by adding or removing the inducer. Additionally, a strand antisense to the one that is transcribed will bind with the natural sense RNA, thereby selectively stopping the expression of a selected gene.

**Sequencing DNA**

Sequence analysis using deoxy chain termination is, by far, the most popular current method because of its relative simplicity. This sequencing technique is dependent on having appropriate polymerase promoter sites built into the cloning vector at or near the insertion site. Usually two sites are built into such vectors, one on each strand. Two popular promoter sites are T7 and Sp6. An appropriate primer complementary to either T7 or Sp6 promoter site is used to initiate synthesis from one strand; another primer is used to initiate synthesis on the other strand, as illustrated in Figure 2.8. Four reactions are set up for each priming reaction; each contains polymerase and all four nucleotides needed for cDNA synthesis. However, each of the reactions also contains a small amount of one radio-labeled deoxy base analog. A deoxy nucleotide is capable of being added to the 5' end of a growing chain but, because the sugar does not contain any free OH groups, cannot serve as a substrate for further extension. Thus, the chain is terminated. Because the deoxy nucleotide represents a small fraction of the total nucleotide added, the statistical probability that any single elongation step will be terminated is small. Therefore, there will be a wide distribution of cDNAs of different lengths, ranging from termination at the first step to, in theory, a full-length cDNA. Next, all four reaction mixtures are subjected to electrophoresis in adjacent lanes in a long sequencing gel. The electrophoretic step separates the DNAs by size; autoradiography is used to develop the gel. Because the deoxy sugars are radiolabeled, a dark band will be noted where
a dideoxy sugar was incorporated. The result is a ladder gel, so called because of its resemblance to a ladder. A straight edge is laid across all four lanes. Because each reaction contained only a single dideoxy nucleotide, a dark line appears in only one of the four lanes and identifies the base present at that sequence. The ruler is moved up one rung at a time. The sequence is simply read from bottom to top in this manner. The technique is simple, requiring only pipetting of four reaction mixtures, incubation for 1 hr, electrophoresis, and autoradiography. A sequencing reaction can be set up on one day and the sequence read the next. The practical limit of this technology is about 200 bases. However, because primers are used to initiate cDNA synthesis from both ends, a total of about 400 bases can be read. Obtaining complete sequences for longer fragments requires sequencing “nested” or overlapping fragments.

**Monoclonal Antibodies**

The molecular revolution has affected protein chemistry as well as nucleic acid chemistry. In the usual immune response, multiple B-cell lineages will respond to a single antigen. As the antigenic protein circulates, it encounters B cells that present their antibody on the surface. Some small fraction of the B cells will have some affinity for the protein or **immunogen**. Each such B cell may recognize different portions, or **epitopes**, of the molecule. An epitope is a single antigenic determinant or site that selects and induces a specific antibody. Even among those B cells that recognize the same epitope, each one is likely to have a different affinity for, or strength of interaction with, the epitope. The resulting stimulated B cells now begin to reproduce rapidly in the spleen or bone marrow and produce antibodies. These polyclonal antibodies represent a multiplicity of antibodies directed at more than one epitope and with a range of affinities for the antigen.

In 1975, Köhler and Milstein discovered how to obtain almost limitless quantities of a single antibody. A mouse is immunized with the target protein, preferably in purified form. After the immune response of the mouse has reached its peak, spleen cells are harvested and, in the presence of polyethylene glycol to weaken cell membranes, fused with immortal *myeloma* cells. (Myeloma cells are cancerous lymphocytes capable of continuous immortal growth.) The fused cells are grown in a medium that selects for hybrids so neither the unfused myeloma line nor the spleen cells are able to grow, but the hybrids are able to grow. Of the hybrids, or **hybridomas**, a small fraction will have incorporated the immunoglobulin genes of the spleen cell into the immortal and rapidly replicating myeloma cell. Again we encounter the familiar “needle in a haystack” problem of molecular biology.

After fusion, the hybridomas are distributed in low density into 96-well plates. The supernatants, which contain the antibodies, are screened against the original antigen using an enzyme-linked immunosorbent assay (ELISA), which will detect the desired antibodies. Cells from wells that contain antibody that reacts with the original antigen are cloned. The contents of the well are diluted into another set of 96-well plates so each well contains no more than a single cell. Wells containing cells are tested to insure that the antibodies produced react with the target protein. Each such cell produces a monoclonal antibody against a single epitope of the protein. The hybridomas can
be grown in almost limitless quantities, producing large amounts of the monoclonal antibody. A simpler technique has been developed for producing monoclonal antibodies, in which the DNA of spleen cells is isolated and transfected directly into myeloma cells.

Monoclonal antibodies are screened further to identify the epitope against which the antibody reacts and to determine the affinity. In general, 2–10 different clones producing antibody against the antigen of interest will result from one fusion reaction. Some of these antibodies will be against the same epitope, but are likely to differ in affinity. On the other hand, a set of monoclonals against different epitopes can be extraordinarily useful. For example, glycoproteins usually will yield antibodies against both the protein and the carbohydrate portions of the molecule. Also, antibodies against different protein epitopes can be useful in mapping protein structure.

Monoclonal antibodies have revolutionized protein chemistry because of their specificity and their ready availability in reproducible form. It is now sometimes possible to isolate monoclonal antibodies against proteins that have never been isolated and are known only from a gene sequence by synthesizing a peptide from the coding sequence and using the peptide as the immunogen to produce a monoclonal antibody specific to that peptide.

Applications

Restriction Fragment Length Polymorphism Analysis and the Genetics of Kidney Cancer

RFLP analysis is a technique that is being used widely in biology to study gene distribution in populations and to identify certain mutations. Sometimes a mutation occurs that will either create or destroy a restriction site, or a significant deletion is generated within a fragment excised by a restriction nuclease that will alter the size of the fragment. If these changes exist in different forms throughout a population for a single region of DNA, some individuals are heterozygous for the two forms, the gene is said to display a RFLP. The analysis for RFLPs is illustrated in Figure 2.9. Genomic DNA is digested with restriction enzymes and separated by electrophoresis. The genomic DNA can be derived from tumors, normal tissue, or leukocytes from peripheral blood. The electrophoretic gel is Southern blotted with a radioactive probe to the target DNA region. When a RFLP has been identified, the probe will bind to two sizes of DNA fragments, a “long” form and a “short” form. A large number of RFLPs have been cataloged and probes are available for a small fee from the American Type Culture Collection. The technique is widely used in genetic studies of populations.

Linehan and his group (Anglard, et. al., 1991) have used this approach to investigate allelic loss in renal cell carcinoma. Ten different RFLPs mapped to chromosome 3 were determined in tumor tissue and normal tissue from 60 patients with renal cell carcinoma. Normal DNA from 58 of the 60 patients was heterozygous at one or more locations. Heterozygosity was identified by the presence of both allelic forms in the same individual. One possible confounding variable in studies of this kind is that tumors frequently contain infiltrating lymphocytes and connective tissue elements that are not altered genetically, either of which can overwhelm the observation of a deletion of one allele in the tumor cells. To account for this possibility the investigators isolated tumor cells from some negative samples; the purified tumor cells showed loss of heterozygosity. If one allele is lost due to complete deletion of the RFLP from one of the two chromosome pairs, the analysis will show the loss of either the long or the short band from the electrophoreogram, as shown in Figure 2.9. Linehan and his collaborators showed loss of heterozygosity at one or more of the 10 RFLP loci tested in 51 patients; loss of heterozygosity in all the loci was seen in 39 patients. The other 12 patients...
retained the loci at one end or the other of the large stretch of DNA defined
by the short arm of chromosome 3.

The significance of these findings is that they strongly support the exis-
tence of a tumor suppressor gene located in the region of the deletions. Loss
of this suppressor gene seems to be a primary event in the development of
renal cell carcinoma. Interestingly, the investigators also found a constitu-
tional deletion in the same region in von Hippel–Lindau syndrome, a het-
irable disorder characterized by renal cell carcinoma.

**Human Papilloma Virus and Cervical Cancer**

Papilloma viruses are small viruses with a DNA genome some 7.9 kb in
length that infect keratinocytes. DNA sequence homology represents the
main, if not the exclusive tool for classification of these viruses. A papilloma
virus is considered to be a new type if it shares less than 50% homology with
each of the other known types, as established by hybridization under stringent
conditions. On this basis, at least 60 different types have been identified
to date. The papilloma viruses are extremely selective about which keratin-
ocyes they infect; HPV-1 and HPV-2 form warts on the plantar and palmar
epithelia, respectively, and nowhere else, whereas HPV-16 and HPV-18 in-
fect the genital epithelium, causing genital warts (condyloma), but no other
epithelial tissue.

The increase in prevalence of sexually transmitted diseases that has oc-
curred in recent years has provided information that suggests that genital
warts may not be entirely benign and that the causative papilloma viruses
may play a causal role in human cervical cancer (reviewed by Schlegel, 1990).
More than 90% of human cervical carcinomas contain and express HPV;
70% are either HPV-16 or HPV-18. Moreover, the epidemiology of human
cervical cancer is similar to that of other sexually transmitted diseases, that
is, it is extremely rare in women who are members of celibate religious orders
and its prevalence is correlated with the number of sexual partners of both
the woman and her primary partner.

These discoveries suggest a mechanism for cervical carcinogenesis by
certain types of human papilloma but not by others. The genome of papil-
oma viruses consist of 9 or 10 genes. The genes are subdivided into 7–8
“early” genes and 2 large “late” genes. The late genes are capsid proteins that
encapsulate viral DNA in productive lesions such as warts or condyloma.
The early genes are expressed in cervical carcinomas and derived cell lines
and represent functions of cell transformation, control of RNA transcription,
and viral DNA replication. Of particular interest are the E6 and E7 genes.
Detailed comparison of the structure of the E6 and E7 proteins from non-
ocogenic papilloma viruses such as HPV-6 and HPV-11 and oncogenic
forms such as HPV-16 and HPV-18 shows that, although there are large
areas of homology, important differences exist as well. The question is
whether any of these differences in structure could be associated with differ-
ences in function that might explain the oncogenic properties of the latter
two types.

Evidence points to the importance of key regulatory genes known ge-
nerically as tumor suppressor genes in maintaining a normal cell phenotype.
The Rb (retinoblastoma) and p53 genes have been identified as important
members of this family. The observation that the E6 and E7 proteins of oncog-
enic papilloma viruses bind to the p53 and Rb proteins, respectively,
whereas the corresponding proteins from nononcogenic HPV-6 or HPV-11
do not suggest that carcinogenesis may proceed by inactivation of p53 or Rb
by binding to the E6 and E7 proteins. Further evidence for this model was
obtained in studies in which inducible plasmid vectors that express antisen-
se E6/E7 mRNA were transfected into cervical carcinoma cells. Antisense RNA
is complementary in sequence and binds to normal mRNA. The vectors use
a promoter that is active in the presence of dexamethasone and inactive in its
absence. This feature allows the expression of the insert, in this case antisense
E6 or E7, to be controlled reversibly. When the inserted antisense RNA was
expressed actively, the cervical carcinoma cells expressed a normal pheno-
type in culture, but when the antisense RNA was not expressed they reverted
to their usual transformed phenotype.

Diagnosis of HPV infection and its typing is becoming an increasingly
important issue in studies of cervical cancer. Until recently, the standard
for such typing was the Southern blot. However, the Southern approach
suffers from a crucial shortcoming: The tissue must be lysed to extract the
DNA, which destroys anatomic structure and mixes normal and malignant
tissue together. Of increasing use are one or more techniques of in situ
hybridization.

The principle of in situ hybridization is simple. A tissue section that has
been treated to be permeable is treated with a labeled probe that is comple-
mentary to the sequence being sought. Most often the probe is labeled radio-
actively, but new technologies using fluorescent or luminescent probes are
being developed. When a radioactive probe is used, the tissue is overlaid with
a silver-containing gel. Radioactive disintegrations produce grains of silver
metal that can be seen easily with the light microscope at high power. The
particular power of in situ techniques is that they allow the investigator to
view the signal and the morphology of the object that yielded the hybridiza-
tion signal. Stoler, et. al (1992) used in situ hybridization to investigate the
distribution of HPV-16 genes through a range of pathologic lesions ranging
from dysplasias to squamous carcinoma and adenocarcinoma lesions.

**Selective Amplification Techniques and ras Genes**

The ras oncogene was one of the first to be discovered. The discovery that
this oncogene was found in viral tumors and naturally occurring tumors,
as shown by transfection studies, firmly established the oncogene theory as a general theory of carcinogenesis (reviewed by Bos, 1989). The ras gene family includes three well-characterized genes that have a high degree of homology and encode proteins of 21,000 molecular weight, referred to generically as p21 proteins. Although the exact role of the p21 proteins is not clear, their function seems to be to modulate the activity of transmembrane receptor–tyrosine kinase complexes. In general, cellular growth is controlled exogenously through the binding of growth factors or growth peptides to external receptors. (The difference between a growth factor and a growth peptide is that a factor has not been identified with a particular chemical structure.) The receptors span the cell membrane; when they are occupied by the peptide, the associated tyrosine kinase on the inner membrane becomes active, initiating a program of intracellular biochemical changes. Mutations in the p21 proteins are theorized to lead to permanent activation of the tyrosine kinase in the absence of growth peptide. The most frequent sites for mutations are codons 12, 13, and 61.

Originally the ras genes were identified by transfection into 3T3 mouse fibroblasts. This technique is relatively insensitive. Genomic DNA from naturally occurring tumors was digested with a restriction endonuclease to produce conveniently sized fragments. These fragments were introduced into cultured 3T3 cells. Transfection can be achieved simply by coprecipitating DNA with calcium phosphate. The small crystals are taken up by cells and, in a small percentage, the DNA is incorporated into the cel genome. Foci of transformed cells indicate the presence of a transforming gene, which in this case was identified eventually as the ras gene.

The sensitivity of detecting ras mutations has been increased markedly by using specific amplification techniques, first with a thermostable DNA polymerase, later with the more stable polymerases now in use. A nonspecific PCR technique that amplified mutant and wild type sequences equally was used to amplify the region between codons 1–61 of the ras gene roughly 10³ fold in DNA prepared from tumors obtained at thoracotomy (Rodenhuis, et al., 1988). The amplified DNA was probed with oligonucleotides that were complementary to each of the commonly identified mutations under conditions of high stringency. Under such conditions, a single base mismatch would inhibit hybridization. Of interest was the finding that tumors from nonsmokers or those who had quit smoking at least 5 years earlier tended not to contain K-ras mutations (9/10 had normal K-ras) whereas tumors in smokers tended to show mutated K-ras genes (13/32 had mutant K-ras).

One of the weaknesses of such an approach is that tumors contain normal tissue and, moreover, represent many steps of selection. Thus, there is no guarantee that finding mutations in one or more genes means that these mutations are causative events. They may represent the effects of cumulative mutation and selection after emergence of the primary tumor. Bert Vogelstein and several colleagues, (1988), investigated the origins of human colon can-
cer to determine whether ras mutations were causative. Colon adenocarcinoma seems to develop in three stages. First, small tubular adenomas form that sometimes develop in the colon and do not progress, in most cases. Occasionally, these simple adenomas progress to a larger, more villous type that frequently contains patches of developing adenocarcinoma. Finally, an invasive adenocarcinoma develops. These investigators concentrated on adenomas, which are early noncancerous premalignant lesions, and early adenocarcinomas that were not likely to have undergone extensive additional genetic rearrangement. Tissue sections were microdissected carefully to isolate only cancer cells. The investigators were interested particularly in separating adenocarcinomatous areas that had developed within adenomas. PCR was used to amplify the ras genes from the malignant and premalignant areas selectively. In the smaller tubular adenomas, the incidence of mutated ras genes was less than 20% but, in the villous adenomas and the adenocarcinomas, the incidence had risen to about 50%; a large fraction of those lesions contained both adenoma and adenocarcinoma and the mutation was found in both. These results suggested that a ras mutation was a very early event associated with the initial progression from tubular to villous adenoma.

A number of techniques have been developed to enhance the sensitivity of the detection of mutated genes. One technique is use of a primer that has a mismatch with the wild-type sequence at the 3’ end (illustrated in Figure 2.10A for selective amplification of ras). In the wild-type, the mismatch at the 3’ end effectively inhibits PCR amplification of the normal sequence, but permits amplification of a specific mutant sequence. A primer selective for each mutant sequence must be included to detect the specific mutation. Another approach involves introducing a restriction endonuclease-sensitive site at the point of the mutation, as illustrated in Figure 2.10B. Sometimes this will occur spontaneously, but more often a slightly mismatched primer must be used. A one- or two-base mismatch will not preclude amplification, as long as it does not occur at the 3’ end of the primer. In the illustration, the 6-base sequence encompassing the potential mutation site and the previous codon almost match the sequence for a BstI site in the wild-type. Only the 3 bases on the 5’ end of the primer are shown. The mismatch in the second base does not preclude amplification, but all the amplified DNA will contain a TGG sequence rather than the wild-type TCG. The amplification products of the wild-type DNA are sensitive to BstI, but those derived from the mutant DNA are not, since the fourth base from the 5’ end is T instead of C. Digestion of the amplified product with BstI will yield RFLPs because mutant sequences will be longer than wild-type sequences. This approach does not require a radioactive probe because staining the gel with ethidium bromide will reveal the different sized fragments.

PCR also can be used to detect gene amplification. Genes in cancers are amplified frequently. The N-myc gene frequently is amplified in neuroblastomas, epidermal growth factor receptor is amplified in gliomas, and neu is
gans possibly represents immunologically related forms with common or related epitopes.

Northern blotting was used to determine whether prostatic acid phosphatase was synthesized by tissues or cancers other than prostate (Solin et al., 1990). The mRNA of prostatic acid phosphatase had been shown earlier to be 3.3 kb in size whereas lysosomal acid phosphatase mRNA was 2.4 kb in size. Using probes specific to the two mRNAs, differentiated prostate cancers were shown to synthesize both prostatic and lysosomal acid phosphatases, as evidenced by bands at 3.3 and 2.4 kb, respectively. However, undifferentiated prostate cancer cells yielded only the 2.4-kb band, confirming the loss of the prostatic acid phosphatase. Wide variations in the amount of prostatic acid phosphatase were seen in cancers and benign prostatic hyperplasia. In addition, when other cancers and other tissues were probed, none were found to produce a 3.3-kb mRNA. Although prostatic acid phosphatase appears to be a gene product specific to differentiated secretory prostatic cells, the amount can vary over wide ranges, and antibodies or enzymatic techniques may not be entirely reliable in measuring the true level of expression.

**Mutations in the p53 Suppressor Gene**

The p53 gene is, so far, the most commonly found deletion of a suppressor gene associated with human cancers. The p53 gene encompasses 16–20 kb of DNA on the short arm of human chromosome 17. The gene contains 11 exons. Mutations of p53, including point mutations, allelic loss, rearrangements, or deletions, have been found in most human tumors examined so far, including colon, lung, liver, breast, esophageal, bladder, and stomach cancers, as well as lymphomas and leukemias. At least 300 different mutations have been identified, most in exons 5–8. Cancers from different geographical areas have been suggested to display different mutational spectra, an observation that, if confirmed, is of importance to molecular epidemiology. In common with other suppressor genes, germ line p53 mutations are associated with development of cancers relatively early in life, whereas somatic deletions are associated with “sporadic” incidence due to environmental exposures.

The variety of mutations has limited hybridization methods. Two approaches are being used; single-strand conformational polymorphism (SSCP) and direct sequencing of amplified segments of the exons most frequently involved in mutations. The SSCP technique is relatively simple. Either genomic DNA or mRNA is used as a template for PCR. The product is then melted, quickly cooled to prevent renaturation, and separated by electrophoresis as single-stranded DNA. The single-stranded DNA usually forms extensive secondary structure, but the presence of a single point mutation in a relatively short fragment (up to 100 bases) usually will alter this secondary structure and, hence, the electrophoretic mobility. Thus, the bands corre-

Amplifier does not have 3' mismatch

**A**

<table>
<thead>
<tr>
<th>5'</th>
<th>CTA</th>
<th>GTG</th>
<th>GAG</th>
<th>GTT</th>
<th>3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>GAT</td>
<td>CAA</td>
<td>CTC</td>
<td>CA</td>
<td>5'</td>
</tr>
</tbody>
</table>

---

**B**

**WILD TYPE SEQUENCE**

almost matches *Bgl* site

5' — TGGCCA — 3' — TGGCCA —

**Mutations in the 3'-codon (CCA) are being sought**

<table>
<thead>
<tr>
<th>5'</th>
<th>TGGCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>ACC</td>
</tr>
</tbody>
</table>

**Wild type**

<table>
<thead>
<tr>
<th>5'</th>
<th>TGGCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>ACC</td>
</tr>
</tbody>
</table>

**Mutant**

<table>
<thead>
<tr>
<th>5'</th>
<th>TGGCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>ACC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5'</th>
<th>TGGTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>ACC</td>
</tr>
</tbody>
</table>

**Not recognized by Bgl**
exposure. Traditional epidemiology has been concerned with groups of individuals with given group characteristics, and risk has been defined in terms of those group characteristics. Molecular epidemiology offers the possibility of producing much more specific estimates of risk, based on knowledge of events at the gene level. With this knowledge will come much more specific methods for monitoring the risk of individuals and of groups. Theoretically it may be possible eventually to predict which 5% of smokers will develop lung cancer or to detect very early the signs that, in a given individual, the process of carcinogenesis is proceeding toward an irreversible outcome. However, with this increased ability to predict individual risk come other risks resulting from societal decisions. The essence of insurance is group risk, not individual risk; society soon will have to deal with such difficult issues.

Suggested Further Reading

Textbooks


Laboratory Manuals


References


3

Validation

Paul A. Schulte
and Frederica P. Perera

The contribution of molecular epidemiology to etiologic research, risk assessment, or disease prevention and control depends on the use of valid biologic markers. The use of invalid markers can lead to wrong conclusions and costly programs. Validity is the best approximation of the truth or falsehood of a marker (Cook and Campbell, 1979). Validity is a sense of degree rather than an all-or-none state. To validate the use of biologic measurement as a marker, it is necessary to understand the relationship between the marker and the event or condition of interest. Biologic markers can be validated against exposure, disease, or susceptibility events. To date, most molecular epidemiologic research has involved validating biologic markers (as opposed to using them for etiologic and intervention research or risk assessment). When the validity, reliability, and practicality of a marker have not been demonstrated, pilot studies are useful. Perera (1987), among others, has demonstrated a strategy and approaches for these pilot studies: characterize the marker in known high-dose groups such as chemotherapy patients, proceed to highly exposed groups such as occupational groups, then use less exposed occupational and environmental groups. The goal of these studies is to determine characteristics of markers that must be known prior to their use in large population studies. These characteristics include a dose–response relationship, marker persistence, inter- and intraperson variation, correlation between markers, and correlation with clinical response. For example, Perera et al. (1992) studied cancer patients treated with cis-platinum (cis-DDP)-based chemotherapy and found posttreatment differences in a battery of biologic markers, including increased levels of platinum–protein and platinum–DNA adducts and increased incidence of sister chromatid exchanges, micronuclei, and gene mutation at the glycoporphin A locus.

In many, but not all, of the studies (conducted in the 1980s) using biomarkers, particularly genetic and molecular markers, the validity of findings

Note Added in Proof

A number of different methods have been used to quantify the interaction of xenobiotics and DNA, each having particular advantages and disadvantages. Detection of carcinogen-DNA adducts by 32P-postlabelling has been used in many human studies. The method involves hydrolysis of DNA to 3'-phosphonucleosides and labelling with carrier-free 32P at the 5' position. Bulky adducts will have different chromatographic mobility than the unadducted nucleotides and appear as a distinct peak. This procedure has proven useful because of high sensitivity, relatively small DNA requirements, and because it is sensitive to a variety of chemical and naturally occurring carcinogens, several of which can be distinguished by their chromatographic behavior.

Fluorescence methods, which are concentration dependent, have also been used in several cases, but have been limited by relatively great tissue requirements. However, high sensitivity is possible if the molecules are confined to a very small volume, such as the nucleus of a cell. Coupled with specific fluorescence-labelled immunological probes, microscopic photometry might detect as few as 300 adduct molecules per cell.