2

Molecular Cloning

A LABORATORY MANUAL

SECOND EDITION

Sambrook • Fritsch • Maniatis

Molecular Cloning

Associate Author	**Nina Irwin**
Managing Editor	Nancy Ford
Editor	Chris Nolan
Associate Editor	Michele Ferguson
Illustrator	Michael Ockler

2

Molecular Cloning

A LABORATORY MANUAL

SECOND EDITION

J. Sambrook
UNIVERSITY OF TEXAS SOUTHWESTERN MEDICAL CENTER

E.F. Fritsch
GENETICS INSTITUTE

T. Maniatis
HARVARD UNIVERSITY

Cold Spring Harbor Laboratory Press
1989

Molecular Cloning
A LABORATORY MANUAL
SECOND EDITION

9 8 7 6 5

Book and cover design by Emily Harste

Cover: The electron micrograph of bacteriophage λ particles stained with uranyl acetate was digitized and assigned false color by computer. (Thomas R. Broker, Louise T. Chow, and James I. Garrels)

Cataloging in Publications data

Sambrook, Joseph
Molecular cloning : a laboratory manual / E.F. Fritsch, T. Maniatis—2nd ed.
p. cm.
Bibliography: p.
Includes index.
ISBN 0-87969-309-6
1. Molecular cloning—Laboratory manuals. 2. Eukaryotic cells--Laboratory manuals. I. Fritsch, Edward F. II. Maniatis, Thomas III. Title.
QH442.2.M26 1987
574.87'3224—dc19 87-35464

Researchers using the procedures of this manual do so at their own risk. Cold Spring Harbor Laboratory makes no representations or warranties with respect to the material set forth in this manual and has no liability in connection with the use of these materials.

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 10 Skyline Drive, Plainview, New York 11803. Phone: 1-800-843-4388. In New York (516) 367-8423. FAX: (516) 367-8432.

13

DNA Sequencing

Sequencing Techniques and Strategies 13.3

SANGER METHOD OF DIDEOXY-MEDIATED CHAIN TERMINATION 13.6
Reagents Used in the Sanger Method of DNA Sequencing 13.6
PRIMERS 13.6
TEMPLATES 13.7
DNA POLYMERASES 13.7
RADIOLABELED dNTPs 13.9
ANALOGS OF dNTPs 13.10

MAXAM-GILBERT CHEMICAL DEGRADATION OF DNA METHOD 13.11

SEQUENCING STRATEGIES 13.14
Confirmatory Sequencing 13.14
De Novo Sequencing 13.14
FACTORS AFFECTING THE CHOICE BETWEEN RANDOM AND DIRECTED STRATEGIES 13.18

Random Sequencing 13.21

GENERATION OF A LIBRARY OF RANDOMLY OVERLAPPING CLONES 13.24
Purification and Ligation of the Target DNA 13.24
Fragmentation of the Target DNA 13.26
SONICATION 13.26
DIGESTION WITH DNAase I IN THE PRESENCE OF MANGANESE IONS 13.28
Repair and Size Selection of DNA 13.30
Preparation of Vector DNA 13.31
Ligation to Vector DNA 13.33

Directed Sequencing 13.34

GENERATION OF NESTED SETS OF DELETION MUTANTS 13.34
Generation of Nested Sets of Deletions with Exonuclease III 13.39

Sequencing by the Sanger Dideoxy-mediated Chain-termination Method 13.42

SETTING UP DIDEOXY-MEDIATED SEQUENCING REACTIONS 13.42
Preparation of Single-stranded DNA 13.42
Preparation of Primers 13.42
Microtiter Plates 13.42
Chain-extension/Chain-termination Reaction Mixtures 13.43
STOCK SOLUTIONS OF dNTPs AND ddNTPs 13.44

DENATURING POLYACRYLAMIDE GELS 13.45
Preparation of Buffer-gradient Polyacrylamide Gels 13.47
Loading and Running Gradient Sequencing Gels 13.54
Autoradiography of Sequencing Gels 13.56
Reading the Sequence 13.58

DIDEOXY-MEDIATED SEQUENCING REACTIONS USING THE KLENOW FRAGMENT OF E. coli *DNA POLYMERASE I* *13.59*
Preparation *13.59*
PREPARATION OF WORKING SOLUTIONS OF dNTPs 13.60
PREPARATION OF WORKING SOLUTIONS OF ddNTPs 13.60
Sequencing Reactions *13.61*

DIDEOXY-MEDIATED SEQUENCING REACTIONS USING SEQUENASES *13.65*
Preparation *13.65*
Sequencing Reactions *13.67*

SEQUENCING DENATURED DOUBLE-STRANDED DNA TEMPLATES *13.70*
Sequencing of Plasmid DNAs Purified by Equilibrium Centrifugation in CsCl–Ethidium Bromide Gradients *13.71*
Removal of RNA from Minipreparations of Plasmid DNA by Precipitation with Lithium Chloride *13.72*

PROBLEMS THAT ARISE WITH DIDEOXY-MEDIATED SEQUENCING *13.73*
Template-specific Problems *13.73*
Systematic Problems *13.73*
Problems with Polyacrylamide Gels *13.74*

Sequencing by the Maxam-Gilbert Method ***13.78***

Asymmetric Labeling of Target DNA *13.78*
Preparation of Target DNA for Maxam-Gilbert Sequencing *13.83*
Reagents, Solutions, and Apparatuses *13.83*

THE TRADITIONAL METHOD OF MAXAM-GILBERT SEQUENCING *13.88*
Cleavage at G Residues *13.88*
Cleavage at Purine Residues (A + G) *13.90*
Cleavage at Pyrimidine Residues (C + T) *13.91*
Cleavage at C Residues *13.92*
Cleavage at A and C Residues (A > C) *13.93*
Treatment of Samples with Piperidine *13.94*

ALTERNATIVE METHODS OF MAXAM-GILBERT SEQUENCING *13.95*

TROUBLESHOOTING GUIDE FOR MAXAM-GILBERT SEQUENCING *13.95*
Reading Sequencing Gels *13.95*
Problems Commonly Encountered *13.95*

References ***13.102***

14

In Vitro Amplification of DNA by the Polymerase Chain Reaction

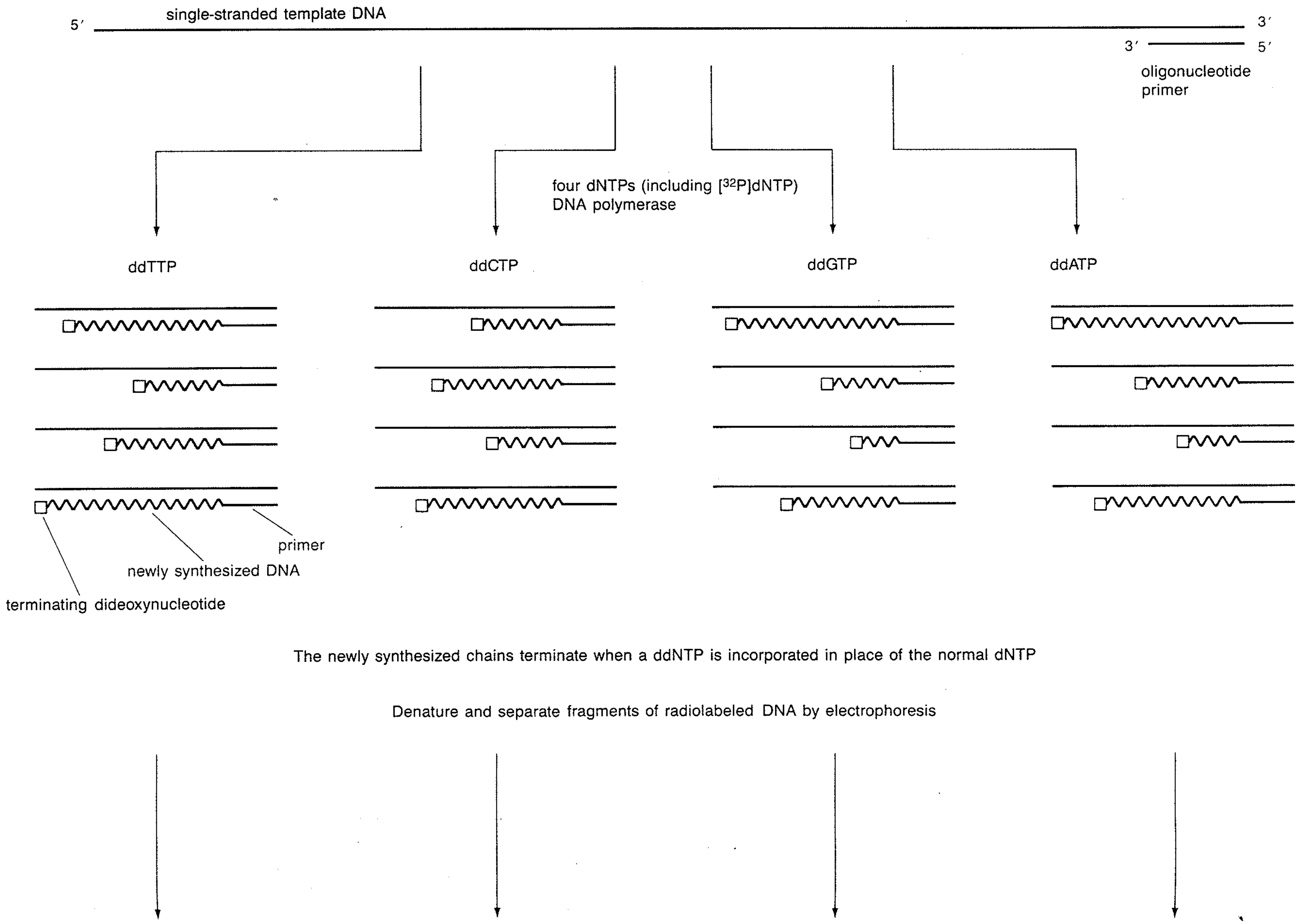
APPLICATIONS OF PCR AMPLIFICATION *14.5*
Generation of Specific Sequences of Cloned Double-stranded DNA for Use as Probes *14.6*

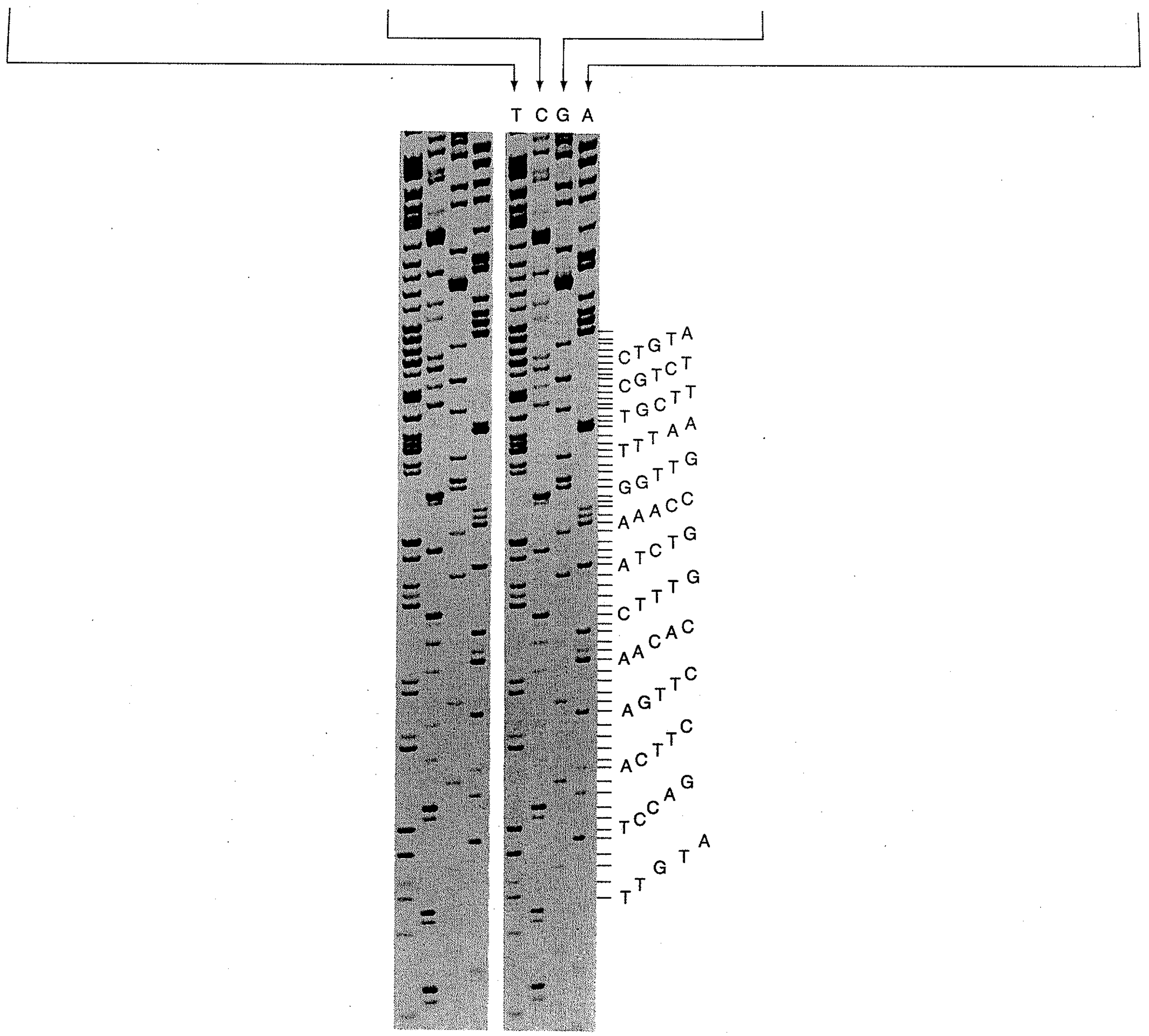
13
DNA Sequencing

In the mid-1970s, when molecular cloning techniques in general were rapidly improving, simple methods were also developed to determine the nucleotide sequence of DNA. These advances laid the foundation for the detailed analysis of the structure and function of large numbers of genes. The first attempts to sequence DNA mirrored techniques developed in the 1960s to sequence RNA (see Sanger et al. 1965; Brownlee et al. 1968; Brownlee 1972). These involved (1) specific cleavage of the DNA into smaller fragments by enzymatic digestion (endonuclease IV [Robertson et al. 1973; Ziff et al. 1973]) or chemical digestion (pyrimidine tract analysis [Robertson et al. 1973; Ziff et al. 1973]), (2) nearest neighbor analysis (Wu and Taylor 1971), and (3) the wandering spot method (Sanger et al. 1973; Tu and Wu 1980). Indeed, in some studies the DNA was transcribed into RNA with *Escherichia coli* RNA polymerase and then sequenced as RNA (Gilbert and Maxam 1973). It is a testimony to the success of DNA sequencing that today most protein sequences are deduced from the nucleotide sequences of genes or cDNAs.

Sequencing Techniques and Strategies

The two rapid sequencing techniques in current use are the enzymatic method of Sanger et al. (1977) and the chemical degradation method of Maxam and Gilbert (1977). Although very different in principle, these two methods both generate separate populations of radiolabeled oligonucleotides that begin from a fixed point and terminate randomly at a fixed residue or combination of residues. Because every base in the DNA has an equal chance of being the variable terminus, each population consists of a mixture of oligonucleotides whose lengths are determined by the location of a particular base along the length of the original DNA. These populations of oligonucleotides are then resolved by electrophoresis under conditions that can discriminate between individual DNAs that differ in length by as little as one nucleotide. When the populations are loaded into adjacent lanes of a sequencing gel, the order of nucleotides along the DNA can be read directly from an autoradiographic image of the gel (see, e.g., Figure 13.1).


5′
single-stranded template DNA
3′
3′
5′
oligonucleotide primer
four dNTPs (including [32P]dNTP)
DNA polymerase
ddTTP
ddCTP
ddGTP
ddATP
primer
newly synthesized DNA
terminating dideoxynucleotide
The newly synthesized chains terminate when a ddNTP is incorporated in place of the normal dNTP
Denature and separate fragments of radiolabeled DNA by electrophoresis



Sequencing by the Sanger dideoxy-mediated chain-termination method.

SANGER METHOD OF DIDEOXY-MEDIATED CHAIN TERMINATION

The current chain-termination method evolved from the +/− sequencing technique (Sanger and Coulson 1975), which first described (1) the use of a specific primer for extension by DNA polymerase, (2) base-specific chain termination, and (3) the use of polyacrylamide gels to discriminate between single-stranded DNA chains differing in length by a single nucleotide. Despite these advances, the +/− method was too inaccurate and clumsy to gain general acceptance, and it was not until the introduction of chain-terminating dideoxynucleoside triphosphates (ddNTPs) (Sanger et al. 1977) that enzymatic methods of DNA sequencing were used extensively.

2′,3′ ddNTPs differ from conventional dNTPs in that they lack a hydroxyl residue at the 3′ position of deoxyribose. They can be incorporated by DNA polymerases into a growing DNA chain through their 5′ triphosphate groups. However, the absence of a 3′-hydroxyl residue prevents formation of a phosphodiester bond with the succeeding dNTP. Further extension of the growing DNA chain is therefore impossible. Thus, when a small amount of one ddNTP is included with the four conventional dNTPs in a reaction mixture for DNA synthesis, there is competition between extension of the chain and infrequent, but specific, termination. The products of the reaction are a series of oligonucleotide chains whose lengths are determined by the distance between the terminus of the primer used to initiate DNA synthesis and the sites of premature termination. By using the four different ddNTPs in four separate enzymatic reactions, populations of oligonucleotides are generated that terminate at positions occupied by every A, C, G, or T in the template strand (see Figure 13.1, pages 13.4–13.5).

Reagents Used in the Sanger Method of DNA Sequencing

PRIMERS

In enzymatic sequencing reactions, priming of DNA synthesis is achieved by the use of a synthetic oligonucleotide complementary to a specific sequence on the template strand. In many cases, this template is obtained as a single-stranded DNA molecule by cloning the target DNA fragment into a bacteriophage M13 or phagemid vector. However, it is also possible to use the Sanger method to sequence denatured double-stranded DNA templates (e.g., denatured plasmid DNA) (see pages 13.70–13.72). In either case, the problem of obtaining primers that are complementary to an unknown sequence of DNA is then solved by using a "universal" primer that anneals to vector sequences that flank the target DNA. Universal primers used for the sequencing of bacteriophage M13 recombinant clones are typically 15–29 nucleotides in length and anneal to the sequences immediately adjacent to (1) the *Hin*dIII site in the polycloning region of bacteriophage M13mp18 and (2) the *Eco*RI site in the polycloning region of bacteriophage M13mp19. These primers, which can also be used for "double-stranded" sequencing of DNAs cloned into pUC plasmids, are available from a large number of commercial suppliers. In addition, several companies sell primers that have been designed to allow sequencing of target DNAs cloned into a variety of restriction sites in different plasmids.

TEMPLATES

As mentioned above, two types of DNA can be used as templates in the Sanger method of sequencing: pure single-stranded DNA and double-stranded DNA that has been denatured by heat or alkali. The best results are obtained from single-stranded DNA templates, which are usually isolated from recombinant bacteriophage M13 particles. When care is taken to optimize the ratio of single-stranded template to primer, it is possible to obtain several hundred nucleotides of sequence from each set of chain-termination reactions. Results of this quality are more difficult to obtain when denatured double-stranded DNA is used as a template. Despite the apparent simplicity and convenience of the method (Chen and Seeburg 1985), it has only recently been improved to the point where it can reliably yield unambiguous results from double-stranded DNA templates. Two factors are critical: the quality of the template DNA and the type of DNA polymerase that is used (see below). Minipreparations of plasmid DNA are always contaminated by small oligodeoxyribonucleotides and ribonucleotides, which serve as random primers, and by inhibitors of DNA polymerases. As a consequence, sequencing gels are frequently obscured by a variety of "ghost" bands, strong stops, and other artifacts. We therefore recommend that minipreparations of plasmid DNAs should not be used to determine the sequence of cloned segments of unknown DNA. However, such DNAs are often adequate templates for confirming a sequence that has already been determined by another method. Plasmid DNA that has been purified by equilibrium centrifugation in CsCl–ethidium bromide gradients yields far better results, although the labor and expense of preparing plasmid DNA in this way are considerable.

DNA POLYMERASES

Several different enzymes are commonly used for dideoxy-mediated sequencing, including the Klenow fragment of *E. coli* DNA polymerase I (Sanger et al. 1977), reverse transcriptase (see, e.g., Mierendorf and Pfeffer 1987), bacteriophage T7 DNA polymerases that have been modified to eliminate 3′→5′ exonuclease activity (Sequenase and Sequenase version 2.0) (Tabor and Richardson 1987), and the thermostable DNA polymerase isolated from *Thermus aquaticus* (*Taq* DNA polymerase). The properties of these DNA polymerases (see Table 13.1) differ greatly in ways that can considerably affect the quantity and quality of the DNA sequence obtained from chain-termination reactions.

Klenow fragment of E. coli *DNA polymerase I*

This enzyme was originally used to develop the Sanger method and is still used extensively for DNA sequencing. Two problems frequently arise:

- The low processivity of the enzyme causes the Klenow fragment to generate a high background of fragments that terminate, not by incorporation of a ddNTP, but by the random dissociation of the polymerase from the template. The inability of the enzyme to travel long distances along the template limits the length of seqquence that can be obtained from standard sequencing reactions using this enzyme. Typically, such reactions generate approximately 250 to 350 nucleotides of sequence. The amount of sequence can be doubled by carrying out the reaction in two steps—an initial labeling step containing low concentrations of dNTPs, followed by a chain-extension/chain-termination reaction containing ddNTPs and a high concentration of dNTPs (Johnston-Dow et al. 1987; Stambaugh and Blakesley 1988). However, even with these improvements, the Klenow enzyme does not routinely yield as much sequence as the more processive Sequenase enzymes (see below).

- The enzyme will not efficiently copy homopolymeric tracts or other regions of high secondary structure in the template. This problem can be alleviated, but not completely solved, by increasing the temperature of the polymerization reaction to 55°C (Gomer and Firtel 1985). dNTP analogs (e.g., dITP or 7-deaza-dGTP [see page 13.10]), which are sometimes used to obtain sequence information at regions of the template that form stable secondary structures, are less effective with the Klenow enzyme than with Sequenases, perhaps because they decrease still further the already low processivity of the enzyme.

In summary, the Klenow fragment of *E. coli* DNA polymerase I is the enzyme of choice for determining the sequence of tracts of DNA that lie within 250 bases of the 5′ terminus of the primer. It is not recommended for sequencing longer segments of DNA or DNAs with dyad symmetry and/or homopolymeric tracts.

Reverse transcriptase

Although not widely used for routine sequencing, this enzyme is occasionally employed to resolve problems caused by the presence of homopolymeric regions of A/T or G/C in the template DNA. Reverse transcriptases from both avian and murine sources appear to be slightly better in this respect than the Klenow enzyme (Karanthanasis 1982; Graham et al. 1986), although perhaps not as good as the Sequenases (Cameron-Mills 1988; Revak et al. 1988).

TABLE 13.1 Properties of DNA Polymerases Used in DNA Sequencing Reactions

Enzyme	Processivity[a]	Rate of polymerization[b]
Klenow fragment of *E. coli* DNA polymerase I	10–50	45
Reverse transcriptase (AMV)	n.d.	5
Sequenase and Sequenase version 2.0	~2000 ~3000	300
Taq DNA polymerase and AmpliTaq[c]	>7600	35–100

[a] Processivity is expressed as the average number of nucleotides synthesized before the enzyme dissociates from the template; n.d. indicates not determined.

[b] Rate of polymerization in nucleotides/second.

[c] *Taq* DNA polymerase, a highly processive DNA polymerase, is useful for determining the sequence of DNA templates that form stable secondary structures.

Sequenases

Sequenase^{TM} is a form of bacteriophage T7 DNA polymerase that has been chemically modified to eliminate much of the enzyme's powerful 3′→5′ exonuclease activity. Sequenase version 2.0 is a genetically engineered form of Sequenase that entirely lacks 3′→5′ exonuclease activity, is extremely stable, and has a threefold higher specific activity than the chemically modified enzyme. Sequenases are the enzymes of choice for determining the sequences of long tracts of DNA because of their very high processivity, their high rate of polymerization, and their wide tolerance for nucleotide analogs such as dITP and 7-deaza-dGTP that are used to resolve regions of compression in sequencing gels. Sequenases travel such long distances along the template that several hundred nucleotides of DNA sequence can often be determined from a single set of reactions. In fact, the amount of sequence is limited more by the resolving power of polyacrylamide gels than by the properties of the enzyme.

To take full advantage of the high processivity of Sequenases, a two-step sequencing reaction is set up. In the first stage, low concentrations of dNTPs and low temperature are used to limit the extent of synthesis and to ensure efficient incorporation of a radiolabeled dNTP. The products of this reaction are primers that have been extended by only 20–30 bases. The first reaction is then divided into the standard set of four reactions, each of which contains high concentrations of dNTPs and a single ddNTP. Polymerization then continues until a chain-terminating nucleotide is incorporated into the growing chain.

Taq *DNA polymerase*

Taq DNA polymerase is useful for determining the sequence of single-stranded DNA templates that form extensive stable secondary structures at 37°C. This is because *Taq* DNA polymerase works efficiently at 70–75°C, a temperature that precludes formation of secondary structure even in templates that are rich in G + C. When used as described by Innis et al. (1988), sequencing ladders produced by *Taq* DNA polymerase demonstrate a high uniformity of band intensity for several hundred nucleotides, suggesting that the enzyme has a high degree of processivity.

RADIOLABELED dNTPs

Until a few years ago, virtually all DNA sequencing was carried out with [α-^{32}P]dNTPs. However, the strong β particles emitted by ^{32}P created two problems. First, because of scattering, the bands on the autoradiograph were far larger and more diffuse than the bands of DNA in the gel. This affected the ability to read a sequence correctly (particularly from the upper part of the autoradiograph) and limited the number of nucleotides that could be read from a single gel. Second, decay of ^{32}P caused radiolysis of the DNA in the sample. Sequencing reactions radiolabeled with ^{32}P could therefore be stored for only 1 or 2 days before the DNA was so badly damaged that it generated indecipherable sequencing gels.

The introduction of [^{35}S]dATP (Biggin et al. 1983) greatly alleviated both of these problems. Because of the decreased scatter of the weaker β particles produced by decay of ^{35}S, there is little loss of resolution between the gel and the autoradiograph. This allows unambiguous determination of several hundred nucleotides of DNA sequence from a single reaction set. Furthermore, the lower energy of ^{35}S produces less radiolysis, allowing sequencing reactions to be stored for up to 1 week at −20°C without noticeable loss of resolution. Thus, if technical problems arise with a polyacrylamide gel, the sequencing reactions can simply be reanalyzed.

ANALOGS OF dNTPs

Regions of DNA with dyad symmetry (especially those with a high G + C content) can form intrastrand secondary structures that are not fully denatured during electrophoresis. This can cause an anomalous pattern of migration in which adjacent bands of DNA become compressed to the point where they are difficult to read. Compression is entirely dependent on the presence of secondary structures in DNA and cannot be alleviated by changing the type of DNA polymerase used in the sequencing reaction. However, compressed regions of gels can usually be resolved by using a nucleotide analog such as dITP (2′-deoxyinosine-5′-triphosphate) or 7-deaza-dGTP (7-deaza-2′-deoxyguanosine-5′-triphosphate). These analogs pair weakly with conventional bases and are good substrates for DNA polymerases such as the Sequenases and *Taq* DNA polymerase (Gough and Murray 1983; Mizusawa et al. 1986; Innis et al. 1988). Some compressions are not resolved by 7-deaza-dGTP; others (particularly those occurring in GC-rich regions) are not resolved by dITP. If it is necessary to use analogs, try dITP first (see pages 13.74–13.75). This analog, in contrast to 7-deaza-dGTP, does not affect the sharpness of the DNA bands in the sequencing gel. Any compression that is not resolved by either dITP or 7-deaza-dGTP can almost always be cleared up by determining the sequence of both strands of the DNA.

As discussed above, both forms of Sequenase and *Taq* DNA polymerase tolerate nucleotide analogs better than does the Klenow fragment of *E. coli* DNA polymerase I. In addition, the manufacturer claims that Sequenase version 2.0 is superior to the original enzyme when sequencing templates with strong secondary structure. Version 2.0 is more processive than Sequenase, having less tendency to pause, thereby eliminating “ghost” bands. Furthermore, version 2.0 appears to tolerate nucleotide analogs such as dITP better than does the original version.

MAXAM-GILBERT CHEMICAL DEGRADATION OF DNA METHOD

Unlike the chain-termination technique, which involves synthesis, the Maxam-Gilbert method involves chemical degradation of the original DNA. This method grew out of studies of the interaction between the *lac* repressor and the *lac* operator in vitro. To this day, the ability to probe DNA conformations and protein–DNA interactions remains a unique feature of the Maxam-Gilbert method.

In this procedure (Maxam and Gilbert 1980), a fragment of DNA radiolabeled at one end is partially cleaved in five separate chemical reactions, each of which is specific for a particular base or type of base. This generates five populations of radiolabeled molecules that extend from a common point (the radiolabeled terminus) to the site of chemical cleavage. Each population consists of a mixture of molecules whose lengths are determined by the locations of a particular base along the length of the original DNA. These populations are then resolved by electrophoresis through polyacrylamide gels, and the end-labeled molecules are detected by autoradiography (see Figure 13.2).

The Maxam-Gilbert method has remained relatively unchanged since its initial development. Although additional chemical cleavage reactions have been devised (for review, see Ambrose and Pless 1987), these are generally used to supplement the reactions originally described by Maxam and Gilbert (1977, 1980). The success of the method depends entirely on the specificity of these cleavage reactions, which are carried out in two stages. In the first stage, specific bases (or types of bases) undergo chemical modification; in the second stage, the modified base is removed from its sugar and the phosphodiester bonds 5′ and 3′ to the modified base are cleaved (see Table 13.2). In every case, these reactions are carried out under carefully controlled conditions to ensure that on average only one of the target bases in each DNA molecule is modified. Subsequent cleavage by piperidine at the 5′ and 3′ sides of the modified bases yields a set of end-labeled molecules whose lengths range from one to several hundred nucleotides. The DNA sequence can then be read from an autoradiograph of a sequencing gel by comparing the G, A + G, C + T, C, and A > C tracks. For a number of reasons (e.g., the use of ^{32}P as a radiolabel, the specific activity of end-labeled DNA, the statistical distribution of cleavage sites, and the limitations of gel technology), the range of the Maxam-Gilbert method is less than that of the Sanger method; the Maxam-Gilbert method works optimally for DNA sequences that lie less than 250 nucleotides from the radiolabeled end.

When the Maxam-Gilbert and Sanger methods were first developed in the 1970s, sequencing by chemical degradation was both more reproducible and more accessible to the average worker. The Sanger method required single-stranded templates, specific oligonucleotide primers, and access to high-quality preparations of the Klenow fragment of *E. coli* DNA polymerase I. The Maxam-Gilbert method used simple chemical reagents that were available to everyone. However, with the development of bacteriophage M13 and phagemid vectors, the ready availability of synthetic primers, and improvements to the sequencing reactions, the dideoxy-mediated chain-termination method is now used much more extensively than the Maxam-Gilbert method. Nevertheless, the chemical degradation approach has one clear advantage over the chain-termination method: Sequence is obtained from the original DNA molecule and not from an enzymatic copy. Therefore, with the Maxam-Gilbert method, one can sequence synthetic oligonucleotides, analyze DNA modifications such as methylation, and study both DNA secondary structure and the interaction of proteins with DNA by either chemical protection or modification interference experiments. However, because of its ease and rapidity, the Sanger technique is now the best choice for simple determination of DNA sequence. In fact, most of the current sequencing strategies have been designed for use with this method.

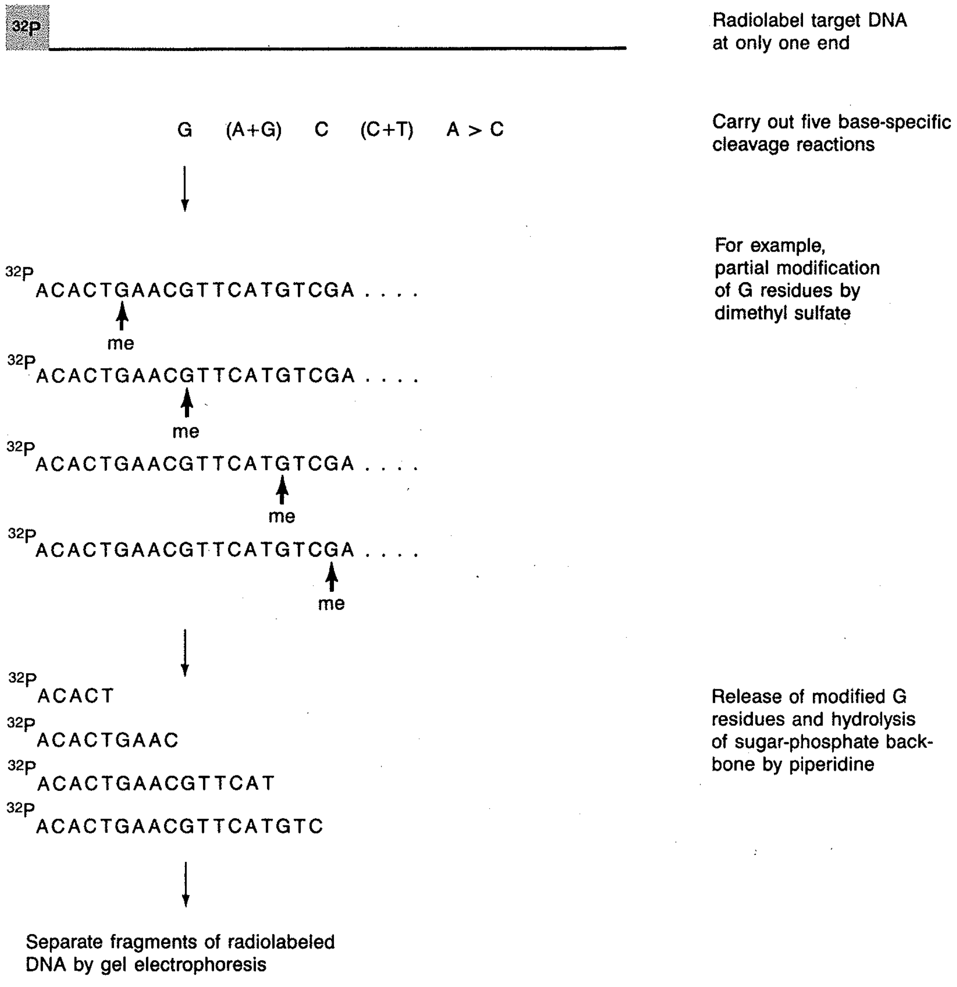

32P
Radiolabel target DNA at only one end
G (A+G) C (C+T) A > C
Carry out five base-specific cleavage reactions
32P ACACTGAACGTTCATGTCGA . . . .
me
32P ACACTGAACGTTCATGTCGA . . . .
me
32P ACACTGAACGTTCATGTCGA . . . .
me
32P ACACTGAACGTTCATGTCGA . . . .
me
For example, partial modification of G residues by dimethyl sulfate
32P ACACT
32P ACACTGAAC
32P ACACTGAACGTTCAT
32P ACACTGAACGTTCATGTC
Release of modified G residues and hydrolysis of sugar-phosphate backbone by piperidine
Separate fragments of radiolabeled DNA by gel electrophoresis

FIGURE 13.2
Sequencing by the Maxam-Gilbert chemical degradation of DNA method.

TABLE 13.2 Chemical Modifications Used in the Maxam-Gilbert Method

Base	Specific modification[a]
G	Methylation of N_7 with dimethyl sulfate at pH 8.0 makes the C_8—C_9 bond specifically susceptible to cleavage by base
A + G	Piperidine formate at pH 2.0 weakens the glycosidic bond of adenine and guanine by protonating nitrogen atoms in the purine rings resulting in depurination
C + T	Hydrazine opens pyrimidine rings, which recyclize in a five-membered form that is susceptible to removal
C	In the presence of 1.5 M NaCl, only cytosine reacts appreciably with hydrazine
A > C	1.2 N NaOH at 90°C results in strong cleavage at A and weaker cleavage at C

[a] Hot (90°C) piperidine (1 M in H_2O) is used to cleave the sugar-phosphate chain of DNA at the sites of chemical modifications.

SEQUENCING STRATEGIES

Before beginning to sequence, it is important to develop an overall strategy that takes into account the size of the region to be sequenced, the accuracy of the sequence required, and the facilities that are available. Only a minor proportion of projects involve the de novo accumulation of large tracts of virgin sequence. More often, sequencing is used to map and identify mutations (e.g., point mutations and deletions) and to verify the orientation and structure of recombinant DNA constructs. The approaches used for these two purposes are very different.

Confirmatory Sequencing

Confirmatory sequencing (e.g., sequencing of mutants generated by oligonucleotide-mediated mutagenesis) often requires no more than one set of reactions that generates the nucleotide sequence of a local region of one of the two strands of DNA. This can usually be achieved by sequencing an appropriate restriction fragment that has been subcloned into a bacteriophage M13 or phagemid vector. In many cases, the region of interest will lie within the sequencing range of a universal primer; if not, the best strategy is to synthesize a priming oligonucleotide 17–19 nucleotides long that is complementary to sequences located approximately 50–100 nucleotides from the region of interest. Whenever possible, the sequence of the homologous region of the wild-type gene should be determined at the same time as that of the mutant. Direct comparison of the sequences on the same autoradiograph greatly facilitates confirmation of the sequence of the altered region and clearly reveals any unexpected, additional differences between mutant and wild-type genes.

De Novo Sequencing

The aim of de novo sequencing is to provide the accurate nucleotide sequence of a virgin segment of DNA that may be many kilobases in length. This task requires careful planning because the maximum length of target DNA that can be sequenced accurately in a single set of sequencing reactions is approximately 400 bases. Target DNAs of this size can be sequenced by cloning in opposite orientations in each of two bacteriophage M13 vectors (e.g., M13mp18 and M13mp19). The entire sequence of each strand can then be determined in a single reaction set using a universal sequencing primer. To sequence longer target DNAs (e.g., several kilobases in length), one of two general strategies can be used:

- *A random approach (or shotgun sequencing),* in which sequence data are collected from subclones containing random segments of the target DNA. No attempt is made to determine where these subclones map in the target DNA or which strand of DNA is being sequenced. Instead, the accumulated data are stored and finally arranged in order by a computer (Staden 1986). This approach, which was pioneered by the M.R.C. Laboratory in Cambridge, has been used successfully to determine the sequences of human mitochondrial DNA (Anderson et al. 1981), human adenovirus DNA (Gingeras et al. 1982; Roberts et al. 1986), bacteriophage λ DNA (Sanger et al. 1982), and Epstein-Barr virus DNA (Baer et al. 1984).

- *Directed approaches,* in which sequences of the target DNA are obtained in a systematic fashion. For example, the entire sequence of the target DNA might be obtained by sequencing a nested set of deletion mutants that begin at a common point (usually at one end of the target DNA) and penetrate various distances into the target region. They therefore bring progressively more remote regions of the target DNA into range for sequencing by universal primer (see Figure 13.3A). Alternatively, virgin segments of target DNA can be sequenced in a step-wise fashion by using the nucleotide sequences obtained from one set of reactions to design a new oligonucleotide that is then used to prime the subsequent set of reactions. In this approach, therefore, DNA sequence is accumulated by moving the priming site along the DNA in a progressive fashion (see Figure 13.3B).

Although the choice between the random and directed strategies is usually dictated by the resources and expertise that are available in the laboratory, a number of additional factors that may also influence the final selection are discussed on pages 13.18–13.20.

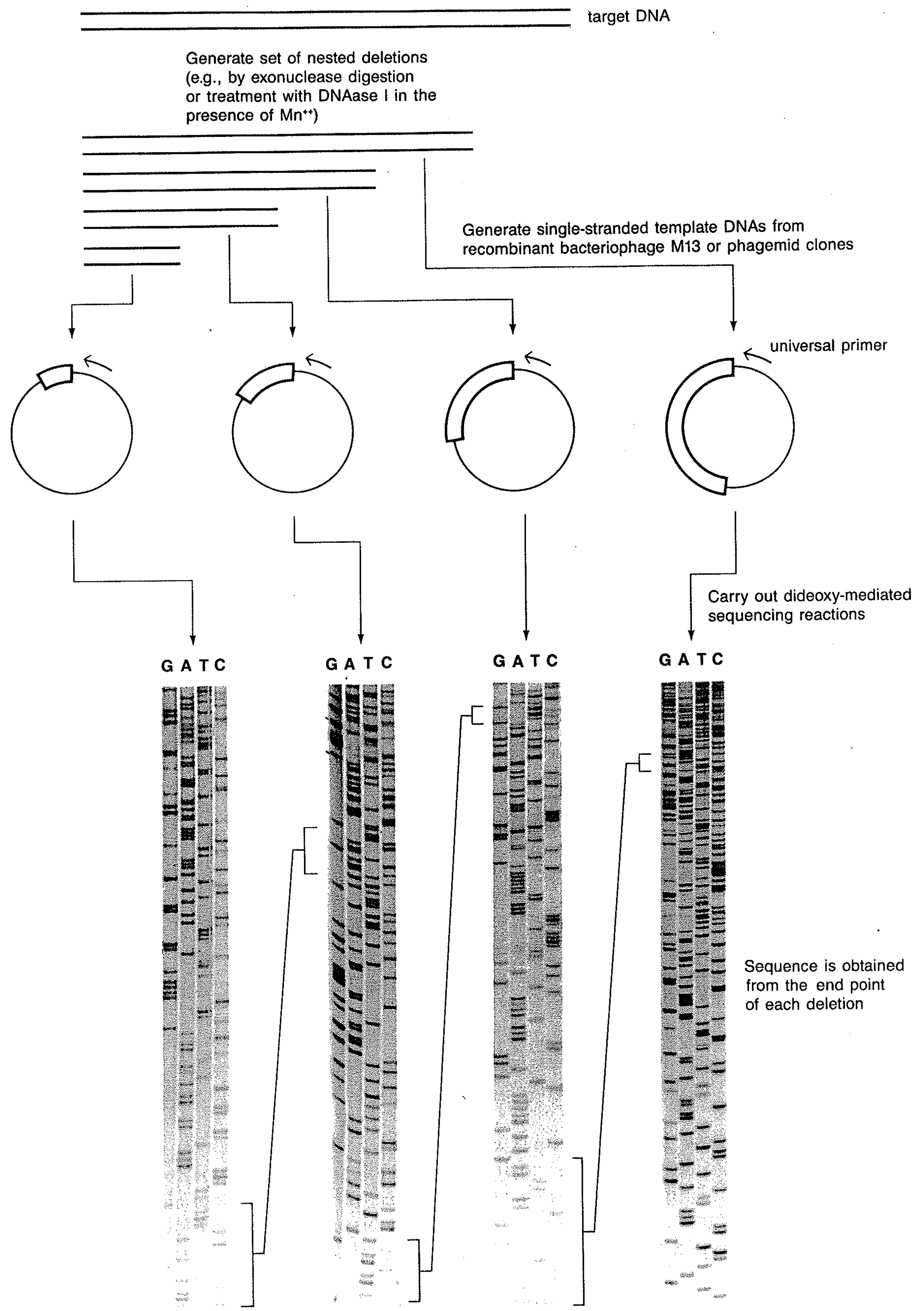
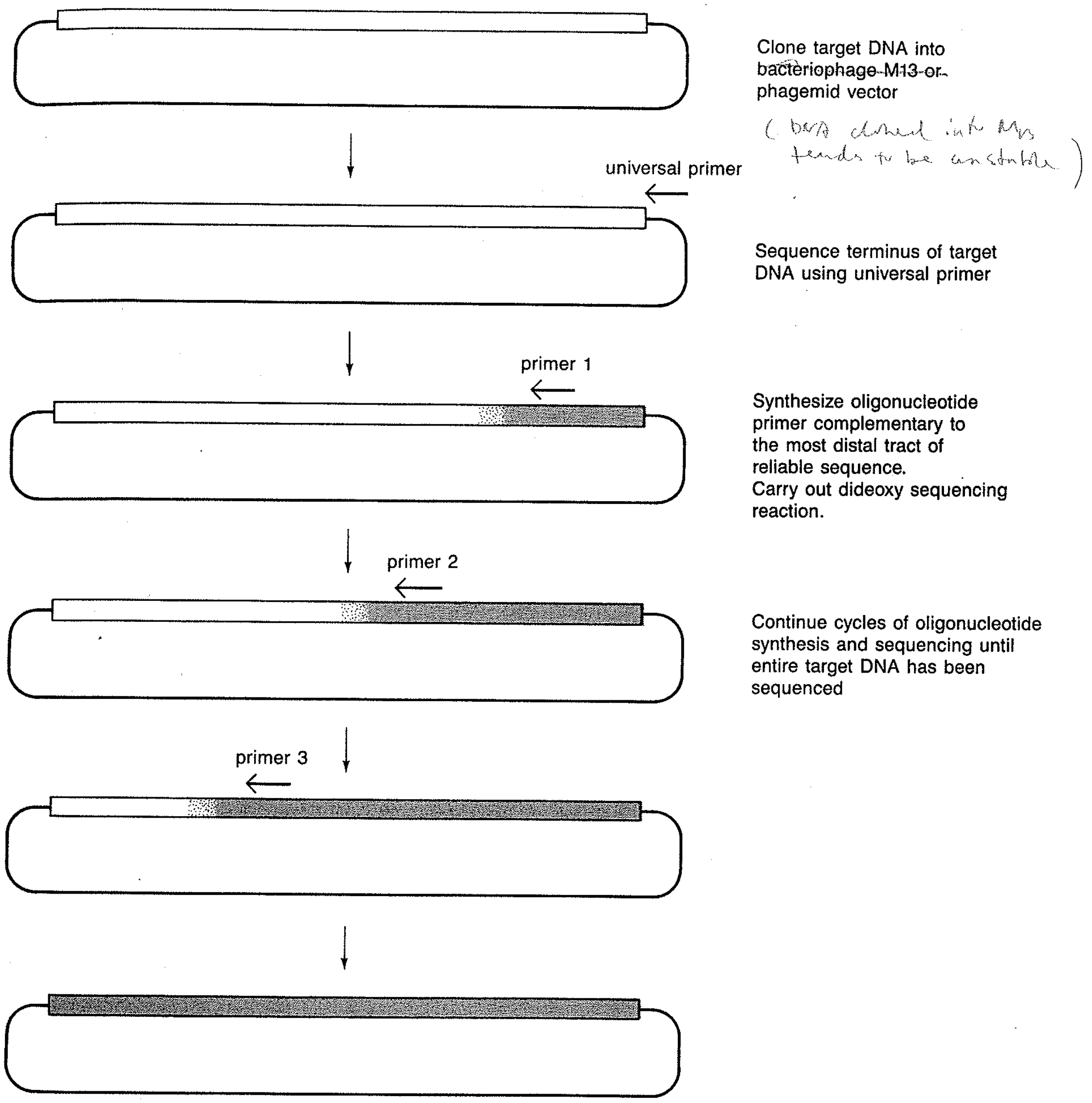

target DNA
Generate set of nested deletions (e.g., by exonuclease digestion or treatment with DNAase I in the presence of Mn++)
Generate single-stranded template DNAs from recombinant bacteriophage M13 or phagemid clones
universal primer
Carry out dideoxy-mediated sequencing reactions
G A T C
G A T C
G A T C
G A T C
Sequence is obtained from the end point of each deletion

FIGURE 13.3A
Directed sequencing with nested sets of deletion mutants.



Clone target DNA into ~~bacteriophage M13 or~~ phagemid vector

(DNA cloned into M13 tends to be unstable)

Sequence terminus of target DNA using universal primer

Synthesize oligonucleotide primer complementary to the most distal tract of reliable sequence.
Carry out dideoxy sequencing reaction.

Continue cycles of oligonucleotide synthesis and sequencing until entire target DNA has been sequenced

FIGURE 13.3B
Directed sequencing with progressive oligonucleotides.

FACTORS AFFECTING THE CHOICE BETWEEN RANDOM AND DIRECTED STRATEGIES

Computing facilities

Every large-scale sequencing project relies heavily on computer programs to sort and order the primary sequence data (Staden 1986). Access to adequate computing facilities becomes an overriding consideration when weighing the pros and cons of the random approach. If these facilities are not available, abandon any idea of using this strategy and turn instead to one of the two directed approaches discussed above.

Nature of the target DNA

If the target DNA is likely to contain dispersed repetitive sequences, then nested deletions should be constructed and used for sequencing. A computer may have difficulty sorting out the repetitive sequences, and oligonucleotide primers may anneal to multiple sites.

Time required to complete the project

The amount of work required to complete a sequencing project can be estimated from the following guidelines:

1. An average of 300–400 nucleotides of sequence can routinely be obtained from a single reaction set.
2. One person can comfortably handle 24–32 reaction sets in a day.
3. A typical week of sequencing, which might generate up to 15 kb of nucleotide sequence, therefore involves
    - 1 day to prepare single-stranded DNA templates
    - 1 day of DNA sequencing
    - 1 day of reading primary DNA sequences and aligning them
    - 2 days of repeating sequencing reactions and rerunning gels to resolve ambiguities and to obtain overlaps between clones

In the random approach, it is usually necessary to sequence about five to seven times more nucleotides than the actual length of target DNA contains. In most cases, a single contiguous sequence does not emerge until approximately 90% of the sequence of both strands has been determined. Because subclones for sequencing are selected at random, certain regions of the target DNA will be sequenced repeatedly before the entire region is covered, and there is no way to predict how long it will take to find and sequence the last subclones required to complete a sequence. Often, these subclones turn out to be underrepresented in the library, and it is then necessary to isolate them by screening with oligonucleotide probes corresponding to flanking sequences. These logistic problems can be alleviated by using restriction enzymes to subdivide large target DNAs into pieces of manageable size (4–5 kb). Each of these pieces is then sequenced independently by the random method.

The directed deletion approach sometimes requires a considerable investment of time to generate and characterize a complete set of nested deletions. However, once this step has been accomplished, informative DNA sequence is accumulated simultaneously from ordered regions of the target DNA. Typically, two sets of deletion mutants extending from opposite ends of the target fragment are required to determine the entire sequence of both strands of the DNA. Alternatively, a single set of deletion mutants can be used to obtain the sequence of a single strand of the target DNA. This information can then be used to synthesize a set of oligonucleotide primers that can be used to confirm the sequence of the opposite strand of DNA (see below).

Availability of an oligonucleotide synthesizer

Given unlimited access to an oligonucleotide synthesizer, custom-designed primers can be generated and synthesized rapidly and cheaply. Assuming that it takes 1–2 days to synthesize an oligonucleotide, virgin nucleotide sequence can be accumulated at a maximum rate of 600–800 nucleotides per week from a designated entry point to the target DNA. This rate of progress can be accelerated by the simultaneous use of several entry points, for example, by cloning the target fragment into bacteriophage M13mp18 and M13mp19 vectors and using a universal primer to begin sequencing from both termini simultaneously or by sequencing internal restriction fragments that have been subcloned into appropriate vectors.

When designing primers for DNA sequencing, observe the following rules:

1. Make sure that the oligonucleotide is complementary to the correct strand of the target DNA and to sequences that have been determined unambiguously to be present in the target DNA. This is particularly important when sequencing virgin DNA by the progressive oligonucleotide method (see Figure 13.3B).

   There is a natural tendency to design an oligonucleotide that is complementary to the furthest limits of known sequence. However, in most circumstances, this sequence has been obtained from closely spaced bands at the top of a sequencing gel, where reading errors frequently occur. It is therefore better to be conservative and to design the primer to be complementary to sequences that lie some distance behind the advancing front, within a region of the gel that can be read with confidence.

2. The primer should have a balanced base composition (40–55% G + C) and should be a least 18 nucleotides in length. If the %(G + C) lies outside these limits, design an oligonucleotide whose length is $18 + n/2$ nucleotides, where $n = 50 - \%(\mathrm{G} + \mathrm{C})$ for A/T-rich regions and $n = \%(\mathrm{G} + \mathrm{C}) - 50$ for G/C-rich regions.

3. Check the sequence of the proposed primer to ensure that

   a. It does not contain regions of dyad symmetry. Oligonucleotides that can self-hybridize to form hairpin or stem-loop structures are inefficient primers.

   b. It is not complementary either to the vector DNA or to regions of the target DNA that have already been sequenced. This greatly reduces the possibility that the oligonucleotide might prime DNA synthesis from more than one location on the template DNA. Most commercially available computer programs for the analysis of DNA have the ability to search sequences for regions that are complementary to synthetic oligonucleotides.

Accuracy of the sequence

When DNA sequencing is carried out carefully, the error rate is less than 0.1%. However, to achieve this high rate of accuracy, it is necessary to sequence both strands of the target DNA completely and to resolve all ambiguities and discrepancies. In this respect, random sequencing has an advantage, since the gradual accumulation of redundant primary sequences greatly improves the accuracy of the final assembled sequence. However, there may be regions of the target DNA that cannot be sequenced accurately by either the random method or directed methods. Resolving these difficult sequences often takes a surprisingly long time and sometimes requires the use of base analogs (to eliminate compressions) or Maxam-Gilbert sequencing.

Future direction of the project

Different sequencing strategies yield different types of material that can be used in later experiments. For example, nested sets of deletions generated for DNA sequencing can be used to study the domains within a promoter region or sets of oligonucleotides complementary to different regions of the target fragment can be used to sequence mutant forms of the target sequence. Random subclones created for shotgun sequencing provide a store of material that can subsequently be used for site-directed mutagenesis or for the generation of radiolabeled probes.