4

Fluorescence-based DNA Sequencing

RICHARD K. WILSON AND ELAINE R. MARDIS

DNA sequencing is a fundamental technique in genome analysis. With the progression of the Human Genome Project and the large-scale identification of new genes, its importance is growing at an ever-increasing pace. Besides its use for complete characterization of large genomic regions—and eventually the entire genomes of many organisms—DNA sequencing is used to characterize alleles associated with disease, to diagnose hereditary and infectious disorders, to aid physical and genetic mapping studies, to explore the expression and regulation of genes, and to study the mechanisms of chromosomal rearrangement, translocation, and mutation. Although basic DNA sequencing techniques have been described in several molecular biology laboratory manuals, methods and technologies that make large-scale genomic sequencing more feasible have been developed very recently. This chapter focuses on the basic approaches used for sequencing genomic segments cloned into plasmid, bacteriophage λ, cosmid, fosmid (a low-copy-number cosmid-based vector that contains an F factor origin of replication), PAC, bacteriophage P1, BAC, and YAC vectors and for sequencing cDNA clones and PCR products. Emphasis is placed on fluorescence-based procedures and instruments that can be used for automated collection of DNA sequencing data, since they have proved to be the most effective methods and technologies for generating large amounts of genomic or cDNA sequencing data.

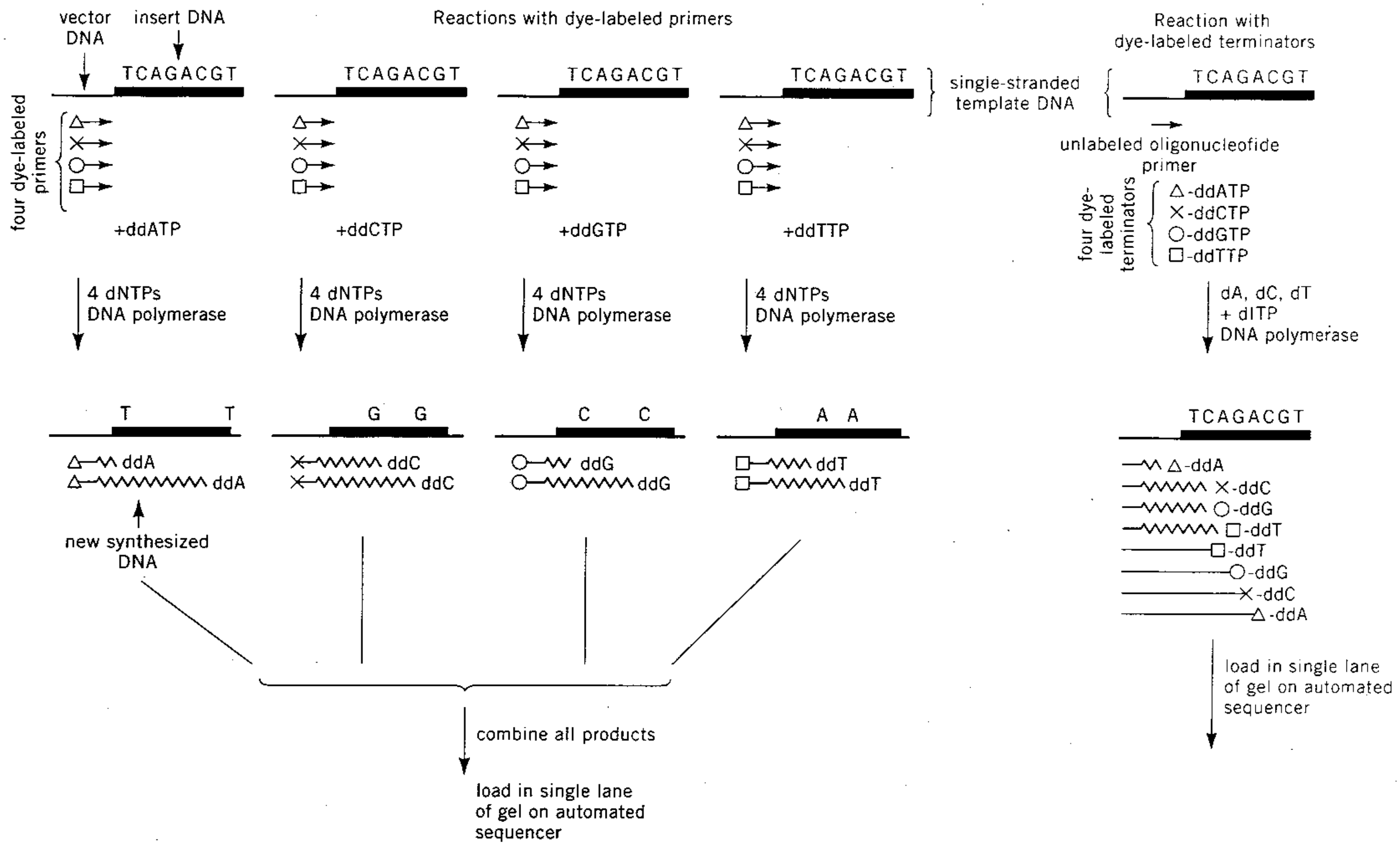


Figure 1 Fluorescence-based DNA sequencing. Fluorescence-based DNA sequencing reactions use a set of four spectrally resolvable fluorescent dyes instead of the radioisotopic labels used in the Sanger method. The two approaches for this type of DNA sequencing are based on whether the dyes are attached to the sequencing primers (5′-dye-labeled oligonucleotide primers) or to the terminators (dye-labeled 2′,3′-ddNTPs). In either approach, the products of the four base-specific sequencing reactions are analyzed in a single lane of a sequencing gel (instead of in four adjacent lanes as in the Sanger method). The four dyes are then discriminated by a laser-based optical system.

Overview of DNA Sequencing Methods

Two basic methods are used for DNA sequencing—the ddNTP-mediated chain-termination method of Sanger et al. (1977) and the chemical degradation method of Maxam and Gilbert (1977). Traditionally, DNA sequence analysis has been performed manually by using the standard Sanger method with radioisotopic labels (e.g., ^{32}P or ^{35}S) attached either to the incorporated dNTPs or to the starting oligonucleotide primer. These basic sequencing techniques have been described comprehesively by Sambrook et al. (1989) and Ausubel et al. (1987). As the scale of DNA sequencing increases (i.e., as the size of the genomic region and/or the number of samples to be analyzed increases), more efficient sequencing methods are required. By using high-throughput robotic or manual protocols for template DNA preparation and processing of sequencing reactions, large numbers of samples can be analyzed daily (Zimmerman et al. 1988; Mardis and Roe 1989; Koop et al. 1990; Wilson et al. 1990b; Wilson and Hood 1991). However, the subsequent tasks of data entry, management, and analysis quickly become overwhelming. For example, data entry is unequivocally the most time-consuming task associated with high-throughput radioisotopic sequencing. Since effective automated systems for reading autoradiographs have not been developed, this tedious step must still be performed manually.

Over the past few years, novel DNA sequencing methods and new types of automated detection technologies have been devised. These greatly reduce the effort required for DNA sequencing and data entry. The most notable advances involve the replacement of radioisotopic labels with fluorescent dyes in the Sanger method (Figure 1) and the development of instruments that perform in situ imaging of fluorescent dye-labeled DNA fragments during electrophoresis (Smith et al. 1986; Ansorge et al. 1987; Prober et al. 1987; Brumbaugh et al. 1988). One such instrument, the 373A DNA Sequencer (Perkin-Elmer [Applied Biosystems Division]), is diagramed in Figure 2. This instrument uses continuous real-time laser-based detection of fluorescent dye-labeled reaction products on the gel and enters sequencing data directly into a computer, thus eliminating the tedious task of reading autoradiographs. A second generation instrument from Perkin-Elmer, the 377 DNA Sequencer, uses slightly different technology that allows more rapid electrophoresis and data collection than the 373A DNA Sequencer does.

Although other instruments that allow continuous real-time detection have been developed (e.g., the Pharmacia ALF, the LI-COR 4000, and the Millipore BaseStation), the Perkin-Elmer instruments are the only devices with multispectral detection systems that can discriminate four fluorescent dyes, each of which corresponds to one of the four purine/pyrimidine bases. This type of detection system allows the four sequencing reactions to be analyzed in a single lane of a gel. Instead of an autoradiograph, the automated DNA sequencer produces a digital two-dimensional color trace (a chromatogram) that can be stored electronically. Since these traces can be viewed and analyzed within sequence assembly programs, the editing tasks necessary to complete large DNA sequencing projects are greatly facilitated (Staden 1980, 1982; Bonfield et al. 1995). Instruments with multispectral detection also provide a fourfold increase in sample capacity over radioisotopic methods or instruments with unispectral detection, which require four lanes per sample (i.e., one lane per base-specific sequencing reaction) instead of one.

In general, the fluorescent dyes currently in use for instruments with multispectral detection are selected because they share a common excitation wavelength and exhibit minimally overlapping emission wavelengths. Rhodamine- or fluorescein-based dyes are typically used for multispectral detection. These dyes are excited by a 488/512-nm argon ion laser. Instruments with unispectral detection use either fluorescein-based dyes or, in the case of the LI-COR instrument, infrared-emitting dyes that are excited by a diode laser.

A key advance that made fluorescence-based DNA sequencing truly practical for widespread use was the development of a sequencing method that used dye-labeled primers and linear amplification of template DNA in the presence of ddNTPs (i.e., cycle sequencing; Figure 3) (Craxton 1991). This method eliminated a rate-limiting step in DNA sequencing by coupling a simple reaction setup (one addition of base-specific reaction mixture and one addition of template DNA) to automatic thermal cycling of the reaction mixture. Previous sequencing methods required multiple steps for addition of

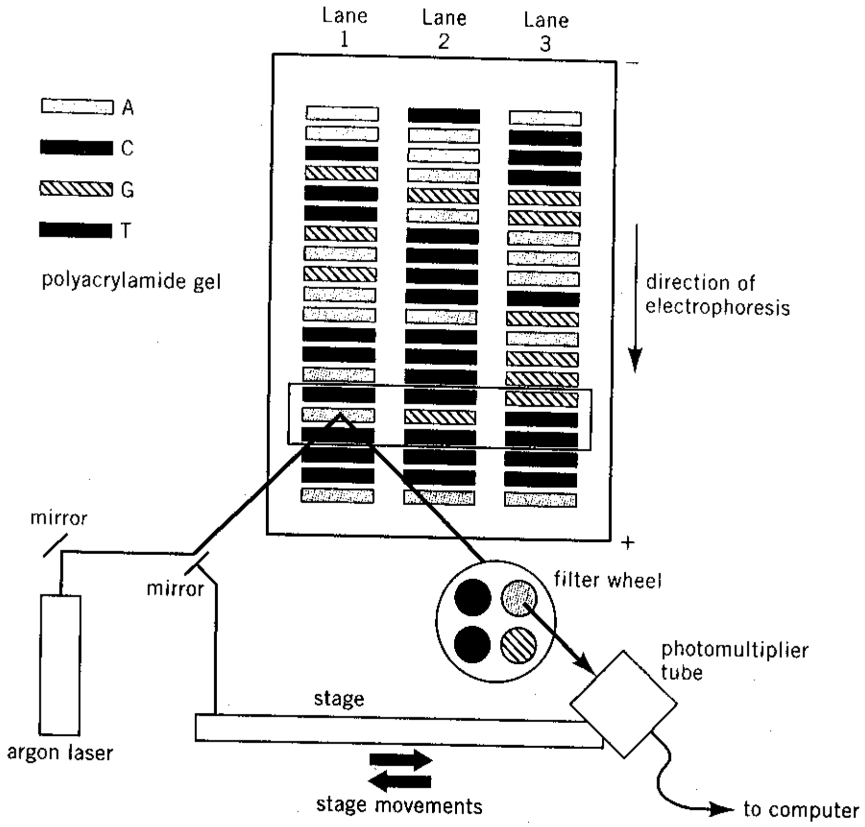


Figure 2 Instrument for multispectral fluorescence-based DNA sequencing. The 373A (originally the 370A) DNA Sequencer from Perkin-Elmer (Applied Biosystems Division) shown here was the first instrument of its kind to become commercially available. This instrument and the next generation model, the 377 DNA Sequencer, are the only instruments available for multispectral fluorescence-based DNA sequencing. Each gel lane corresponds to a single template and contains the products of the four base-specific sequencing reactions. These reaction products migrate through the gel until they reach a fixed point (the rectangle shows the point of detection) where the dye molecules they carry are excited by a scanning laser. The emitted light is directed through a filter wheel to a photomultiplier tube. The stage moves to allow the laser to scan the width of the gel. As the laser makes each pass across the gel, the filter wheel advances by one color. This allows discrimination of the four dyes used for sequencing and thus allows each base in the sample to be identified.

reactants as well as user-mediated incubations. Cycle sequencing is essentially a single-primer linear amplification by PCR (instead of an exponential amplification) in which four reactions—one for A, C, G, and T—are performed for each template. This method can be used for sequencing many types of templates, including single-stranded bacteriophage M13 DNA, double-stranded plasmid DNA, or PCR products. Furthermore, significantly less template DNA is needed to obtain good results with this method than with fluorescence-based sequencing methods that use the Klenow fragment of *E. coli* DNA polymerase I or bacteriophage T7 DNA polymerase (for methods, see Johnston-Dow et al. 1987; Koop et al. 1990).

Fluorescence-based sequencing methods are constantly changing, especially with the recent introduction of new DNA polymerases and labeling approaches (Ju et al. 1995; Tabor and Richardson 1995). In general, high-throughput methods for DNA sequencing are aimed at producing the best possible data in the most rapid and streamlined

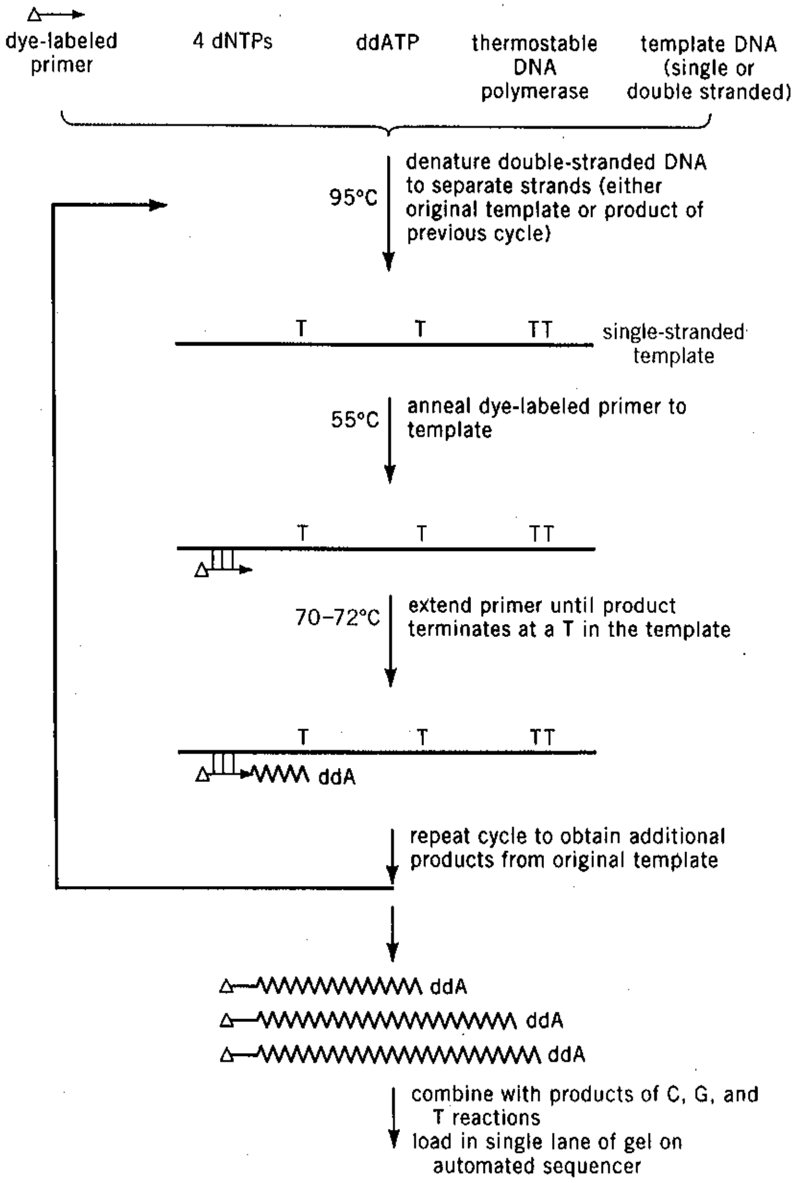

dye-labeled primer
4 dNTPs
ddATP
thermostable DNA polymerase
template DNA (single or double stranded)
95°C
denature double-stranded DNA to separate strands (either original template or product of previous cycle)
T
T
TT
single-stranded template
55°C
anneal dye-labeled primer to template
70–72°C
extend primer until product terminates at a T in the template
ddA
repeat cycle to obtain additional products from original template
ddA
ddA
ddA
combine with products of C, G, and T reactions
load in single lane of gel on automated sequencer

Figure 3 Linear amplification (cycle sequencing). A cycle sequencing reaction with a dye-labeled primer is shown for the A reaction. See text for details.

manner during the production steps of a large-scale sequencing project. This largely translates into batch processing reactions and templates in 96- or 384-well arrays from the time the subclones are grown to the time the samples are loaded on the automated sequencer. This type of approach is used in several of the protocols in this chapter. This chapter also includes more specialized methods that are aimed at solving specific problems (e.g., resolving compressions by sequencing using dye-labeled terminators or extending the sequence by a primer-directed approach) and discussions of various practical considerations for sequencing (e.g., primer choice and purification methods, preparation and casting of sequencing gels, and miscellaneous methods related to sequencing).

Sequencing Primers

The primers used in fluorescence-based sequencing are of two basic types: those that are labeled with a fluorescent moiety and those that are not. A labeled primer typically has a fluorescent or infrared dye attached at its 5′ end via a succinimide ester bond. The dyes' emission wavelengths are appropriately selected for the detection system of the sequencing instrument.

In a newer class of primers known as energy transfer fluorescent primers, each primer contains two fluors—a common donor fluor (typically a 5′ fluorescein) that efficiently accepts the excitation wavelength and one of four acceptor fluors that emit the wavelength (Ju et al. 1995). The acceptor fluors are typically internal, approximately 8–10 bases from the donor fluor, with each of the four dyes corresponding to a different base—JOE or R6G for A, FAM for C, TAMRA for G, and ROX for T (available from Molecular Probes, except JOE). When excited by the laser, the donor fluor passes most of its fluorescent excitation energy to the acceptor fluor by radiationless energy transfer, thereby boosting the emitted fluorescence of the acceptor and increasing its emission wavelength. This provides several benefits over the use of single-fluor primers (as well as single-fluor terminators). The enhanced detection of fragments labeled with energy transfer primers offers improvements in the length of sequence that can be read and in peak resolution (Ju et al. 1995).

There is no need to double components in the G and T reaction mixtures to compensate for the lower efficiencies of the single TAMRA and ROX fluors typically used to label these base-specific reaction products. This fact and the ability to reduce the absolute amount of primer per reaction to 0.1 pmole for A and C and 0.2 pmole for G and T contribute to overall lower costs for fluorescence-based sequencing with energy transfer primers than with single-fluor primers. (For the sequences and T_ms of some of the major fluorescent sequencing primers that are available commercially, see Chapter 5 and Figure 2.)

An unlabeled primer typically contains a custom sequence that is usually synthesized for a specific experiment or project. Several publicly available programs (e.g., the Oligo Selection Program [Hillier and Green 1991]) can be used to design sequencing primers. The parameters involved in designing sequencing primers for synthesis include the following:

- Choose a primer sequence that is longer than 15 nucleotides.
- Choose a primer sequence with a roughly equivalent number of A, C, G, and T bases and a G + C content of approximately 40% to 50%.
- Avoid runs containing more than three of the same nucleotide.
- Avoid regions with a potential for secondary structure formation or self-complementarity.
- Choose a primer sequence with a C or G at the 3′ end. This provides a higher degree of hydrogen bonding and thus ensures good annealing of the 3′ end.
- Choose a primer sequence from a DNA segment for which high-quality sequencing data are available.
- Choose a primer with a T_m of 50–55°C, especially if it will be used for cycle sequencing. If this is not possible because of the criteria above or if a longer primer with a higher T_m is to be used, the annealing temperature for thermal cycling should be adjusted to the primer T_m. The T_m of the primer can be estimated by using the following equation:

  $$T_m = 69.3 + (0.4)(\%\ G{+}C) - (650/L)$$

  where L is the length of the primer in nucleotides.

Purifying primers for DNA sequencing reactions is much more straightforward than in the past, primarily because of the high efficiency of nucleotide incorporation in automated synthesizers. This high efficiency contributes to fewer failures in synthesizing sequences and hence less worry about the presence of products that are shorter than those desired. In general, a primer that has been deprotected by treatment with ammonium hydroxide can be prepared sufficiently for inclusion in a sequencing reaction by placing a 10-µl aliquot in a 1.5-ml microcentrifuge tube, drying the aliquot completely in a SpeedVac Concentrator, and then dissolving the dried primer in 5 µl of H_2O. For each sequencing reaction, use 1 µl of the primer. The volumes used in this preparation can be adjusted according to the required number of reactions.

FLUORESCENCE-BASED DNA SEQUENCING REACTIONS

Five procedures for performing fluorescence-based DNA sequencing reactions are presented here. These procedures use either 5′-end-labeled primers or ddNTPs labeled with fluors and encompass most of the reaction types commonly used for performing large-scale fluorescence-based DNA sequencing. All of these procedures are designed for use with Perkin-Elmer (Applied Biosystems Division) DNA sequencers. The non-cycle sequencing procedure and the cycle sequencing protocols using dye-labeled primers can be modified for use with LI-COR 4000 or Pharmacia ALF instruments as specified on pp. 345, 349, and 354. Dye-labeled terminators are not available for the LI-COR and Pharmacia instruments.

Non-cycle Sequencing Reactions using Dye-labeled Primers and a Thermolabile DNA Polymerase

The procedure on pp. 342–345 uses Sequenase DNA polymerase and 5′-end-labeled primers. Unlike the other procedures here, it does not involve thermal cycling. Although this procedure is not as well suited as cycle sequencing for processing large numbers of samples in a high-throughput fashion, it has proved to be effective for sequencing difficult regions with repeats that cause premature "stops" in cycle sequencing reactions (for further discussion of resolving this problem, see pp. 376–377).

Cycle Sequencing Reactions using Dye-labeled Primers and a Thermostable DNA Polymerase

The cycle sequencing procedures on pp. 346–354 use dye-labeled primers and a thermostable DNA polymerase. These procedures are useful for either random sequencing or deletional sequencing in which large numbers of templates are being analyzed.

The first of these procedures uses ThermoSequenase DNA polymerase and either energy transfer dye-labeled primers or single-fluor dye-labeled primers (for a further discussion of these types of primers, see p. 339). This enzyme is a mutated thermostable DNA polymerase in which a single-base mutation (a phenylalanine-to-tyrosine amino acid substitution) abolishes the enzyme's ability to discriminate between incorporating dNTPs and ddNTPs (Tabor and Richardson 1995). Since ddNTPs and dNTPs are incorporated with the same reaction kinetics, the dNTP:ddNTP ratio determines the fragment lengths produced. Use of ThermoSequenase results in a fluorescent sequencing trace with very even peak heights, which are especially noticeable in runs of the same nucleotide. This result, in turn, enhances the ability of the base-calling software to call a base correctly from the automated sequencer's data, which can lead to less time required for sequence editing and more accurate assembly of the sequence contigs.

The second procedure is similar to the one on pp. 346–349 except that a nonmutated thermostable DNA polymerase such as SequiTherm is used (Wilson and Fulton 1994). This procedure uses nucleotide mixtures in which the ratios of dNTPs and ddNTPs are adjusted to yield longer DNA fragments.

Cycle Sequencing Reactions using Dye-labeled Terminators and a Thermostable DNA Polymerase

In the cycle sequencing procedures on pp. 355–358, a dye-labeled ddNTP is incorporated into the fragment that is being produced by the polymerase. Both procedures use the type of mutated thermostable DNA polymerase mentioned above. The first protocol uses individual dye-labeled terminator stock solutions that are commercially available. The second uses a commercially available premixture containing dye-labeled terminators, dNTPs, and DNA polymerase that requires only the addition of template, primer, and H_2O before thermal cycling can begin.

Use of these newer enzymes (e.g., ThermoSequenase DNA polymerase or AmpliTaq DNA polymerase FS) in sequencing reactions with dye-labeled terminators has several advantages over use of the first generation of thermostable DNA polymerases (e.g., SequiTherm DNA polymerase or AmpliTaq DNA polymerase CS). Since ddNTPs are more readily incorporated when the mutated enzymes are used, large amounts of dye-labeled terminators are no longer needed to drive the reaction to a high level of completion. Any remaining unincorporated dye-labeled terminators must still be removed, since they can lead to high background and "blobs" on the gel that may obscure the underlying data, but decreasing the amount of terminators allows removal by precipitation with ethanol instead of by the more time-consuming procedures provided on pp. 359–362. Purification on gel-filtration media as described on pp. 359–362 can be reserved for cases where the minor amount of dye-labeled terminator that remains after precipitation is undesirable (e.g., when it is critical to read a sequence that is very near the priming site). In addition, these mutated polymerases require fewer cycles and shorter primer-extension periods, thereby reducing the time required for thermal cycling by approximately half.

Besides the reduced requirement for terminators, the faster removal of unincorporated terminators, and the reduced cycling time associated with the newer DNA polymerases, there are advantages to using dye-labeled terminators instead of dye-labeled primers in a sequencing project. Since each of the four types of dye-labeled terminators has a specific fluorescent label, only one sequencing reaction must be set up for each template—not four. Primers with custom sequences can be used. Use of dye-labeled terminators provides the ability to read through most compressions (probably because the fluor on the 3′ end modifies or eliminates in-gel secondary structure that causes compressions). These final two advantages also mean that procedures involving dye-labeled terminators are useful for primer-directed sequencing and for closing gaps or resolving ambiguities during the final phase of a shotgun sequencing project or a deletional sequencing project.

Troubleshooting for Fluorescence-based DNA Sequencing

As in radioisotopic sequencing, several common problems in fluorescence-based DNA sequencing can result in ambiguous or poor-quality data. Examples of some of these problems, along with suggested solutions, are presented here. For additional discussion of resolving repeats and sequence ambiguities caused by DNA structure, see Chapter 5.

PROBLEMS CAUSED BY SOFTWARE

Incorrect Location of the Primer Peak

Occasionally, the Perkin-Elmer (Applied Biosystems Division) analysis software incorrectly determines the location of the dye-labeled primer peak that precedes the sequencing data. The manifestation of this error is shown in Figure 5, where a long stretch with no signal is followed by the primer peak and then the sequencing data. This error can also occur with dye-labeled terminator reactions and looks similar to the example in Figure 5. In addition, a missed primer peak typically shows that analysis started at an *X* coordinate of 0 instead of one further into data collection.

In most instances, this problem can be corrected by determining the position of the *X* coordinate just beyond the true dye-labeled primer peak and then reanalyzing the data. The proper position can be found by examining the raw data for a large peak and then clicking the computer mouse to hold the cursor to the right of this peak. The *X* coordinate shown on the screen is the primer peak position (scan number 560 in this trace) that

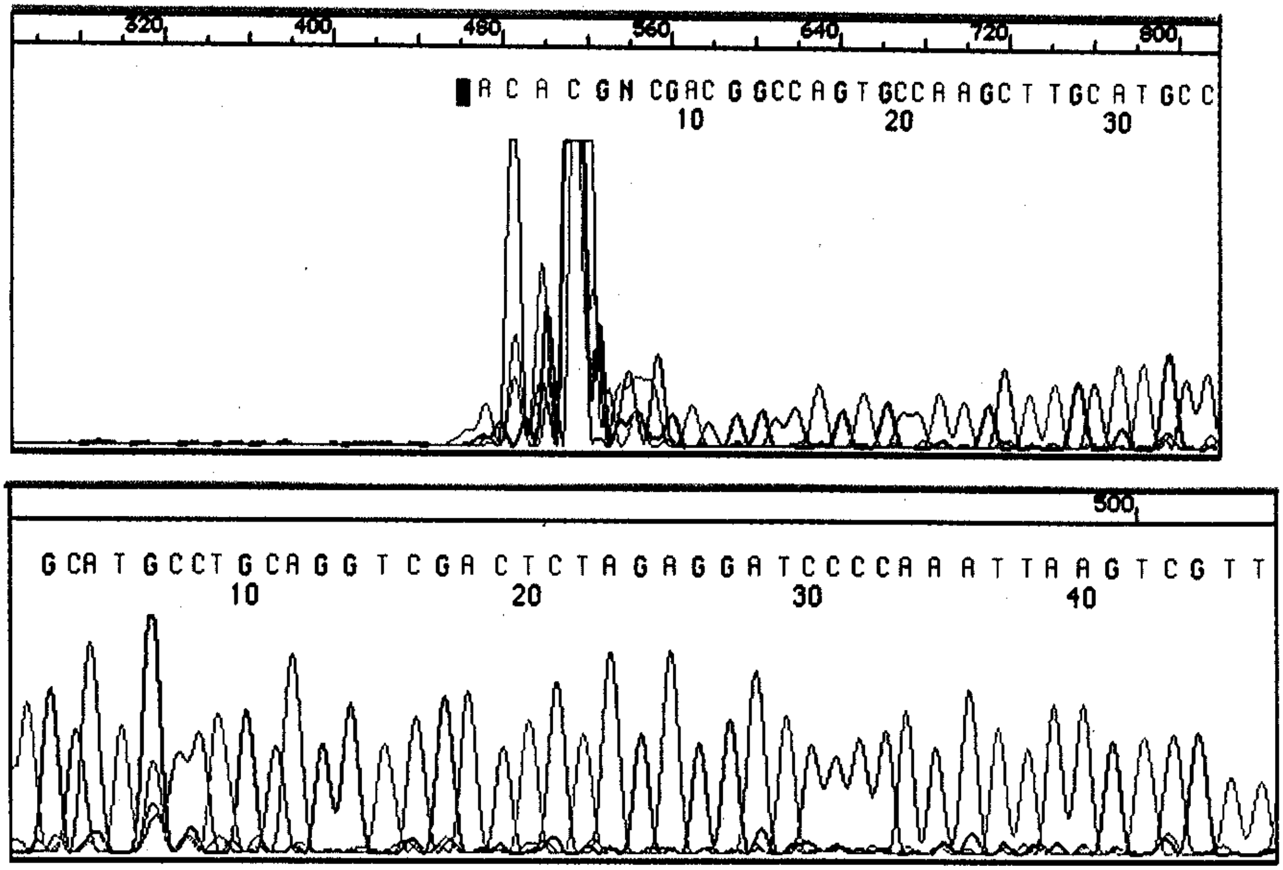


Figure 5 Incorrect location of the primer peak. (*Top*) The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. See text for details. (*Bottom*) A normal trace is shown for comparison.

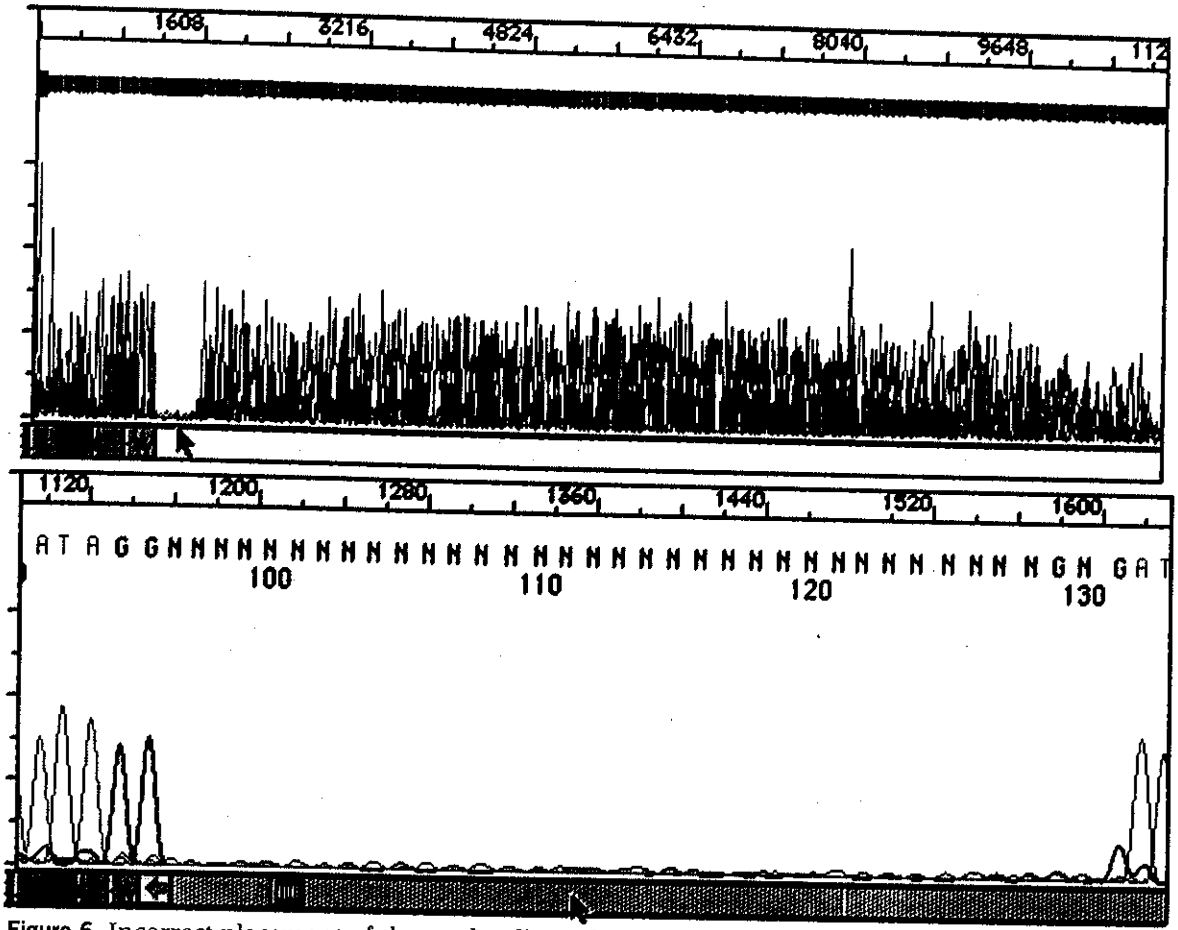

1608
3216
4824
6432
8040
9648
112
1120
1200
1280
1360
1440
1520
1600
A T A G G N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N G N G A T
100
110
120
130

Figure 6 Incorrect placement of the tracker line. The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. The bottom panel provides an expanded scale for a portion of the data shown in the top panel. The arrow in each points to the area of improper placement of the tracker line. See text for details.

should be used as the initial *X* value during reanalysis of the data.

Incorrect Placement of the Tracker Line

Incorrect placement of a data tracker line can occur during the initial (automatic) or secondary (manual) analysis of a gel image. Optimal tracking involves either automatically or manually placing tracker lines directly on the areas with the highest signal in a given sample lane. With either placement method, segments of the tracker line may inadvertently be placed adjacent to the sample lane but not close enough to the signal. An example of the result of this occurrence is shown in Figure 6. The problem is easily overcome by determining the sample lane number with the error and adjusting the corresponding tracker line's placement in the problem area. Reanalysis of the sample should result in the inclusion of the previously omitted data (i.e., the data in the region where the tracker line was improperly placed).

Incorrect Matrix Subtraction File

Overlap of the emissions from the four fluorescent dyes used on the Perkin-Elmer (Applied Biosystems Division) DNA sequencers necessitates the application of subtraction matrices that are specific for the instrument and the type of label (i.e., single-fluor primer or energy transfer fluorescent primer or dye-labeled terminator). To apply subtraction files to samples during initial data analysis, the user defines settings for the instrument and the type of label in the data collection program. The matrix subtraction files remove the overlaps in the emission spectra, thus allowing the true peaks to be identified by the base-calling software. An example of specifying the wrong matrix subtraction file for a sample is shown in Figure 7. This error results in high nonspecific background throughout the trace, which can interfere with base-calling. Simply reanalyzing the data with the appropriate matrix file corrects the problem. For reanalysis with a different matrix file, consult the manufacturer's instructions.

Incorrect Mobility Correction File

Each of the four fluorescent dyes alters the electrophoretic mobility of a dye-labeled DNA fragment in a slightly different fashion. Mobility also varies throughout the course of electrophoresis and according to the gel matrix used for separation. A mobility correction file should be applied to each sample during analysis by specifying the gel matrix percentage and the type of label (i.e., single-fluor primer or energy transfer fluorescent primer or dye-labeled terminator). An example of applying the wrong mobility correction file to a sample is shown in Figure 8. The peaks in this trace are not evenly spaced, especially at the beginning and end of the trace, and base-calling is adversely affected. Reanalyzing the data with the appropriate mobility correction file solves the problem. For further information, consult manufacturer's instructions.

PROBLEMS CAUSED BY TEMPLATES OR BY PIPETTING ERRORS

Low Signal Strength

One of the reasons that fluorescence-based DNA sequencing occasionally yields data of poor quality is low signal strength. An example of this is shown in Figure 9. Possible causes of this weak signal and remedies for each include the following:

- *Presence of an inadequate amount of DNA in the reaction.* First, analyze the sample by agarose gel electrophoresis or by measuring the OD_{260} (see Appendix) to make sure DNA was isolated and to determine the DNA concentration. Next, run the reaction again using more template DNA.
- *Failure of the primer to anneal.* First, make sure the correct primer was used for the sequencing vector of the subclone. Next, check the concentration of the primer. Finally, resequence with and without a change in annealing conditions (typically, a lower temperature or a longer incubation at the original temperature).
- *Use of improper thermal cycling conditions.* Typically, conditions for primer annealing were suboptimal, so poor binding of the primer caused few synthesized strands to be polymerized. To solve this, change the length and/or temperature of the cycles.
- *Use of a poor quality DNA template.* Sequence a control template of known purity in parallel with the sample.
- *Failure to add all of the required reaction components.* Run the reaction again, being careful to pipette all components accurately.

In general, the type of label no longer affects the signal intensity. Use of a mutated DNA polymerase makes the signal from a dye-labeled terminator approximately equal to that of a single-fluor dye-labeled primer. Energy transfer dye-labeled primers have the highest signals but may not be appropriate for all situations as discussed earlier.

Contaminated Template

The process of harvesting bacteriophage plaques or bacterial colonies can contribute to low-quality sequencing data if the template is prepared from a culture containing more than a single subclone. An example of the data produced in such a case is shown in Figure 10. Here, the sequence quality is good through the multiple cloning site that is just downstream from the primer, but it deteriorates rapidly beyond the cloning site (*Sma*I in this example). In such cases, the trace is simply not used. Other examples of contamination that can interfere with data analysis are RNA, protein, or host genomic DNA.

Errors in Pooling Reaction Products

Since Perkin-Elmer (Applied Biosystems Division) DNA sequencers are capable of analyzing all four base-specific reactions in one gel lane, the four

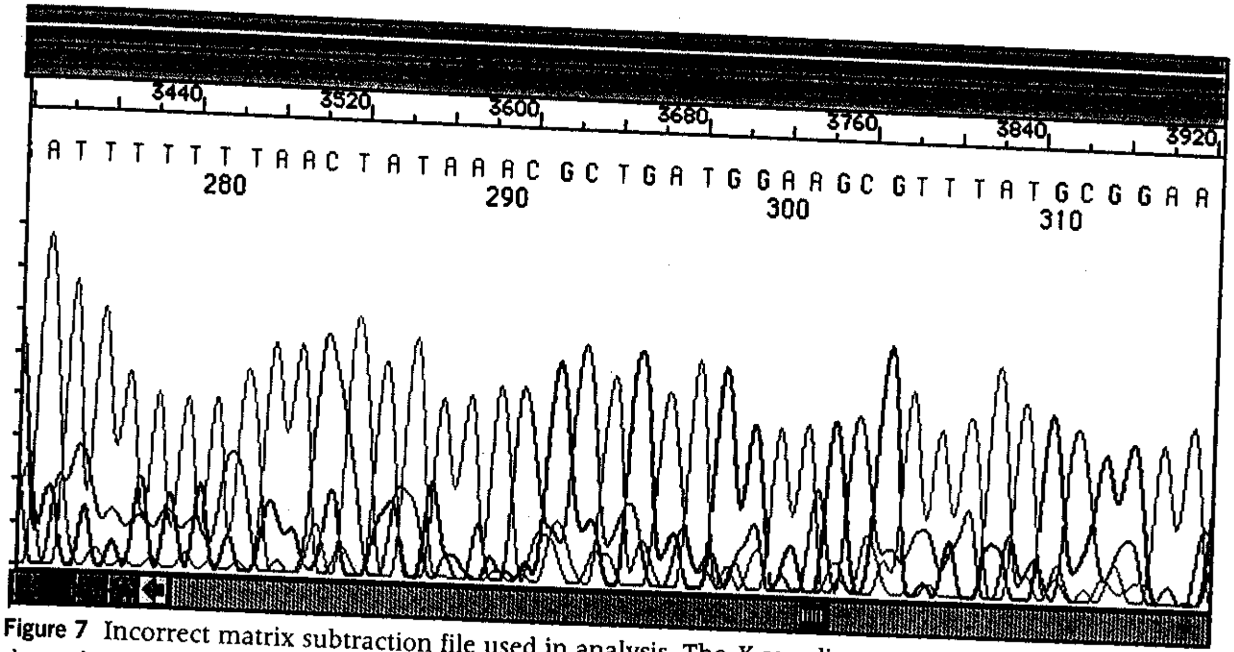


Figure 7 Incorrect matrix subtraction file used in analysis. The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. See text for details.

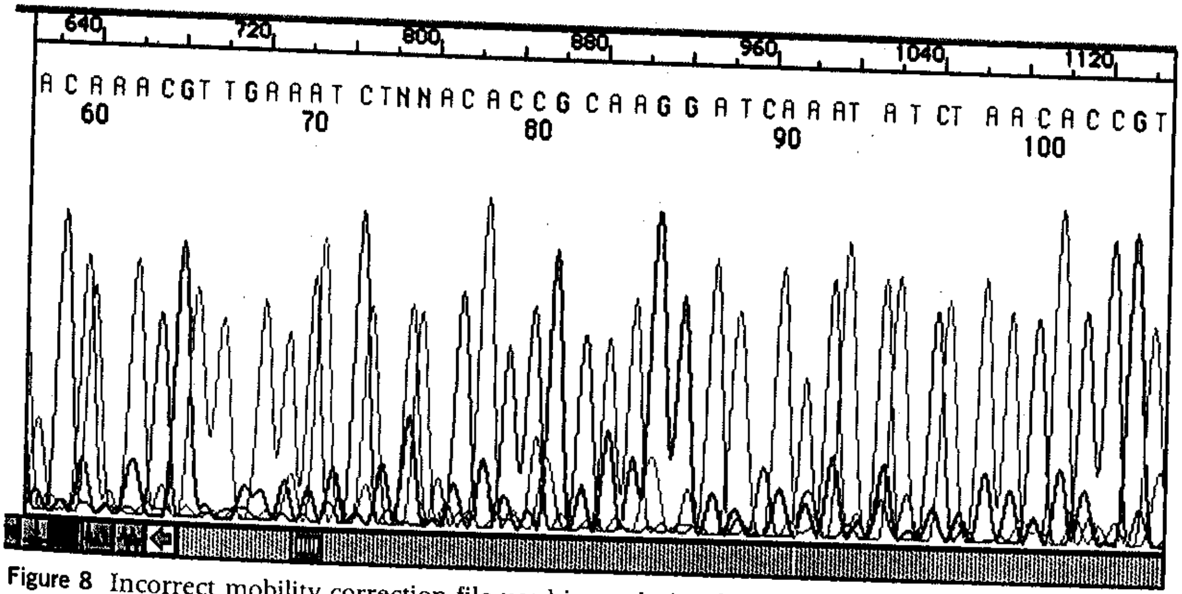
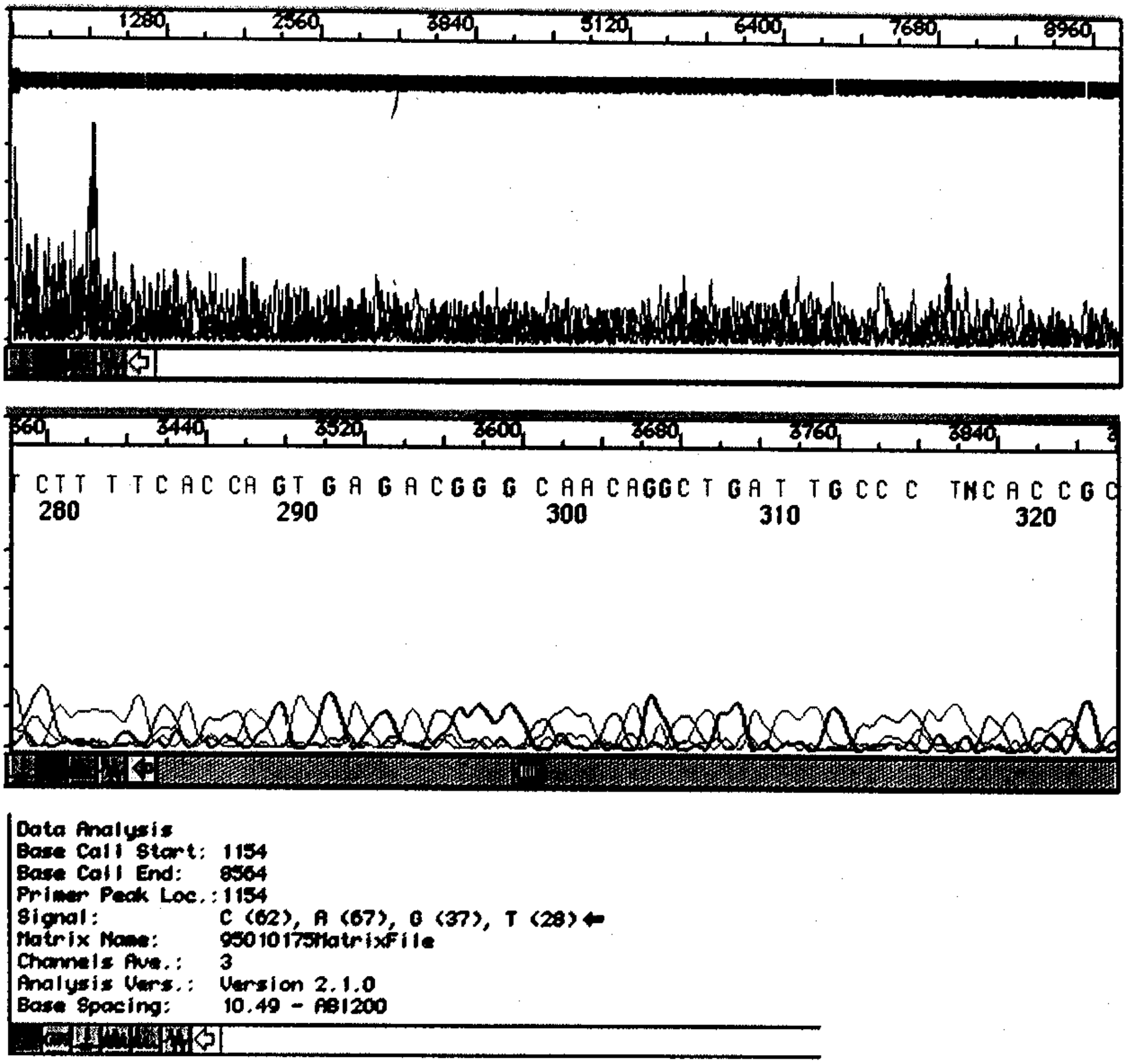


Figure 8 Incorrect mobility correction file used in analysis. The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. See text for details.

reaction products are typically pooled and precipitated with ethanol before they are loaded on the gel. Pooling represents a potential source of error, especially when large numbers of reactions are being processed. One example of incorrect pooling is shown in Figure 11, where the A, C, and G reac-


1280
2560
3840
5120
6400
7680
8960
3360
3440
3520
3600
3680
3760
3840
T CTT T T C A C CA GT G A G A C G G G C A A C A G G C T G A T T G C C C T N C A C C G C
280
290
300
310
320

```
Data Analysis
Base Call Start: 1154
Base Call End:   8564
Primer Peak Loc.:1154
Signal:          C (62), A (67), G (37), T (28)←
Matrix Name:     95010175MatrixFile
Channels Ave.:   3
Analysis Vers.:  Version 2.1.0
Base Spacing:     10.49 - AB1200
```

Figure 9 Low signal strength. The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. The bottom panel provides an expanded scale for a portion of the data shown in the top panel. See text for details.

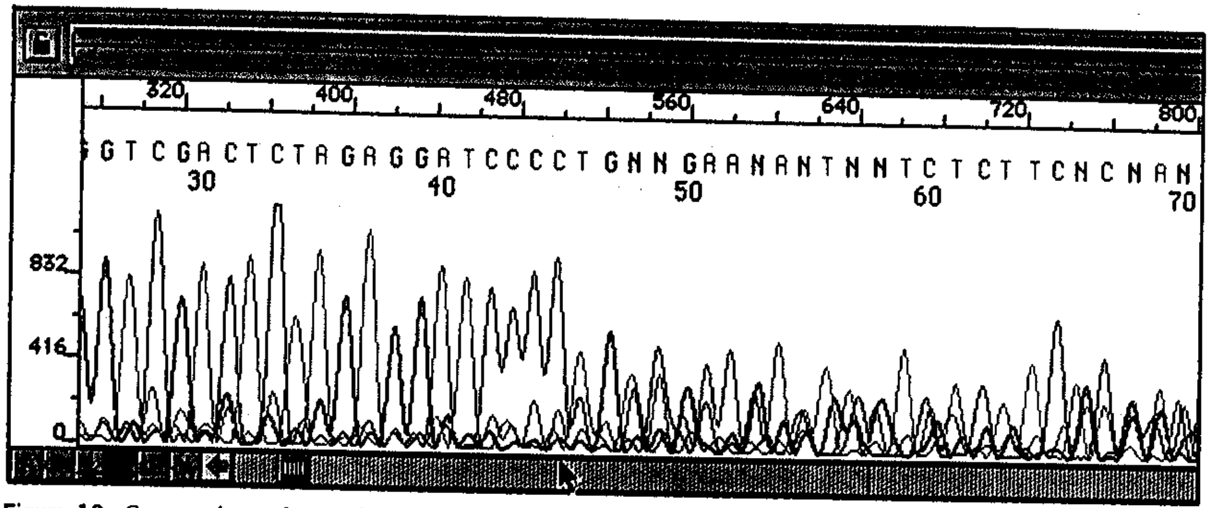


Figure 10 Contaminated template. The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. The arrow points to the end of the cloning site. See text for details.

tions for one template have been pooled and mistakenly mixed with the T reaction for another template. The resulting trace shows T peaks in anomalous locations throughout the trace as well as gaps in the sequence where T peaks would have occurred for the correct template. The only solution for this problem is resequencing the template.

Omission of a Base-specific Reaction

Figure 12 shows an example of a four-color fluorescent trace in which the products of a single base-specific reaction have been omitted. A reaction can be omitted for a variety of reasons, including errors in pooling, errors in pipetting reaction

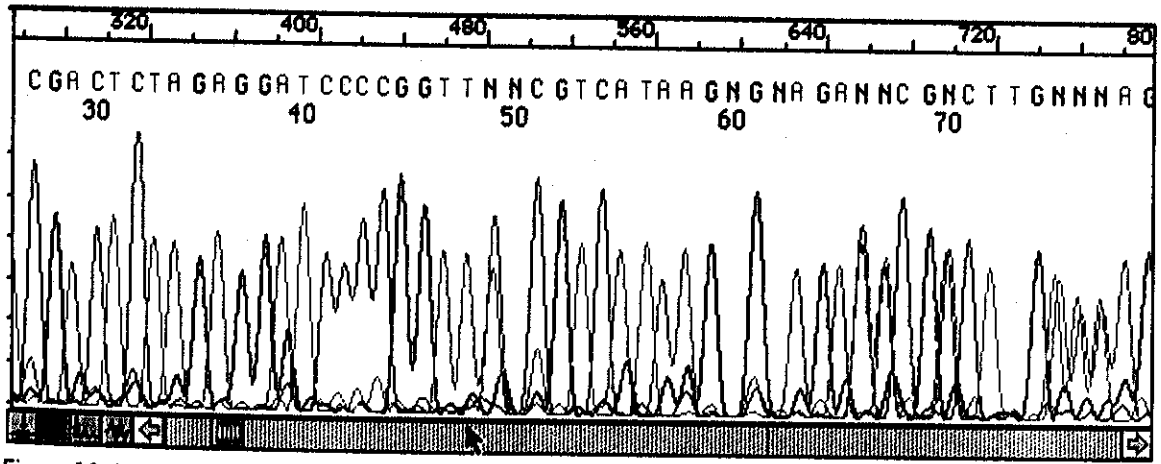
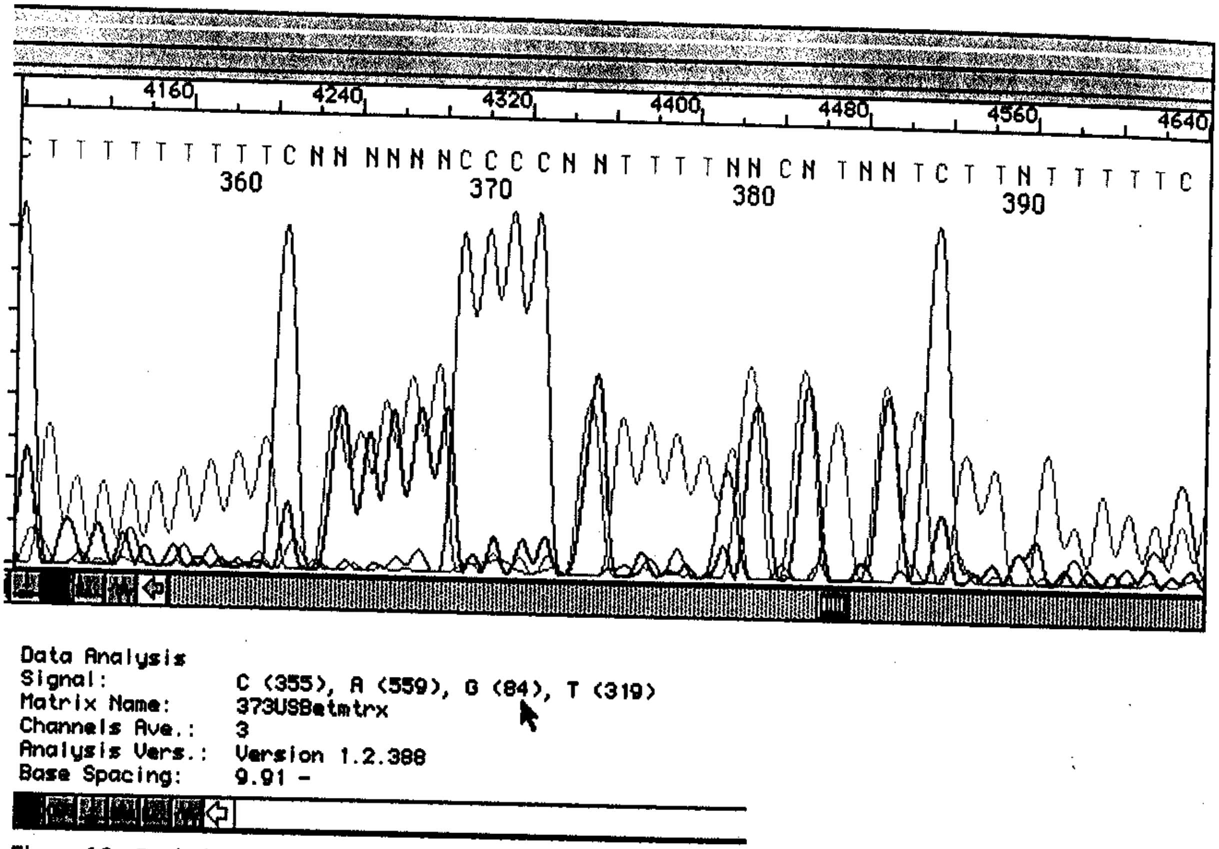


Figure 11 Errors in pooling reaction products. The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. The arrow points to the cloning site and the first anomalous T peak. See text for details.


4160
4240
4320
4400
4480
4560
4640
C T T T T T T T T T T C N N N N N N C C C C N N T T T T N N C N T N N T C T T N T T T T T C
360
370
380
390

```
Data Analysis
Signal:           C (355), A (559), G (84), T (319)
Matrix Name:      373USBetmtrx
Channels Ave.:    3
Analysis Vers.:   Version 1.2.388
Base Spacing:     9.91 -
```

Figure 12 Omission of a base-specific reaction. The *X* coordinate corresponds to the scan number throughout the time of data collection. The *Y* axis is a relative height scale for peak intensity. Traces for A residues are shown in green, C in blue, G in black, and T in red. The arrow indicates the low signal strength for G. See text for details.

components, or evaporation of the sample during thermal cycling. The missing reaction is indicated by an unusually low signal strength relative to those from the other bases. In the trace itself, the missing reaction is exhibited both as a high background signal under the peaks of other bases and as a missing peak. This problem must be solved by repeating the experiment.

PROBLEMS CAUSED BY DNA STRUCTURE

Compressions Caused by Anomalous Mobility

Traces can reveal areas of simple ("foldback") compression, where intrastrand base pairing (usually involving one or more GC dinucleotides) causes the anomalous mobility of one or more fragments. This anomalous mobility, in turn, results in errors in the automatic base-calling. Typically, sequencing the region on the opposite strand allows the compression to be resolved. However, this is not always the case. An example of a simple compression in which sequencing data generated from opposite strands with a dye-labeled primer failed to provide well-separated, unambiguous peaks is shown in the upper two panels of Figure 13. In cases such as these, using a dye-labeled terminator procedure to sequence one of the subclones for the region often resolves the compression. This is demonstrated by the trace shown in the bottom panel.

Homopolymeric Sequences and Stops

Homopolymeric sequences present problems for fluorescence-based DNA seqиencing. Runs of C or

G residues often produce artifacts such as the one shown in the top panel of Figure 14. These artifacts are referred to as "stops" since primer extension by the DNA polymerase is apparently arrested at this point by strong base pairing between two strands of a double-stranded template or by a secondary structure within the template strand alone that cannot be displaced by the polymerase. Interestingly, these artifacts are also observed with single-stranded templates for which cycle sequencing has been used (as shown in the three panels in Figure 14). In the example shown in Figure 14, sequencing the opposite strand (center panel) helped define the C run, but the data distal to the run (the left side of the figure) is of poor quality. This is probably due to offset annealing of the template and product strands during the cycle sequencing reaction. (Offset annealing occurs during cycling when a nonterminated sequencing fragment reanneals within a repeat, but not at the correct position within the repeat. Extension of these strands then causes background noise in the data in the region of the repeat since the fidelity of primer binding along the strand has effectively been lost.) A dye-labeled terminator reaction (bottom panel) failed to resolve the problem. For short (10–20 bp) homopolymeric sequences composed of C or G residues, the addition of DMSO at a final concentration of 5–10% often helps resolve these regions (Burgett and Rosteck 1994). Higher denaturation, annealing, and extension temperatures can also help. Longer C or G runs (e.g., the one shown in Figure 14) are more difficult (or impossible) to resolve. The non-cycle sequencing procedure on pp. 342–345 has also proved to be effective for sequencing difficult regions with repeats that cause premature stops in cycle sequencing reactions.

Runs of A or T residues (or dinucleotide repeats) are not generally problematic unless they are quite long (e.g., >30 bp). Traces for such regions are typically characterized by a "stuttering" pattern after the run (Figure 15). This is probably due to offset annealing of template and product strands during the cycle sequencing reaction. The best way to resolve a long A or T homopolymeric run is to identify and sequence subclones that can be read into the region from both ends.

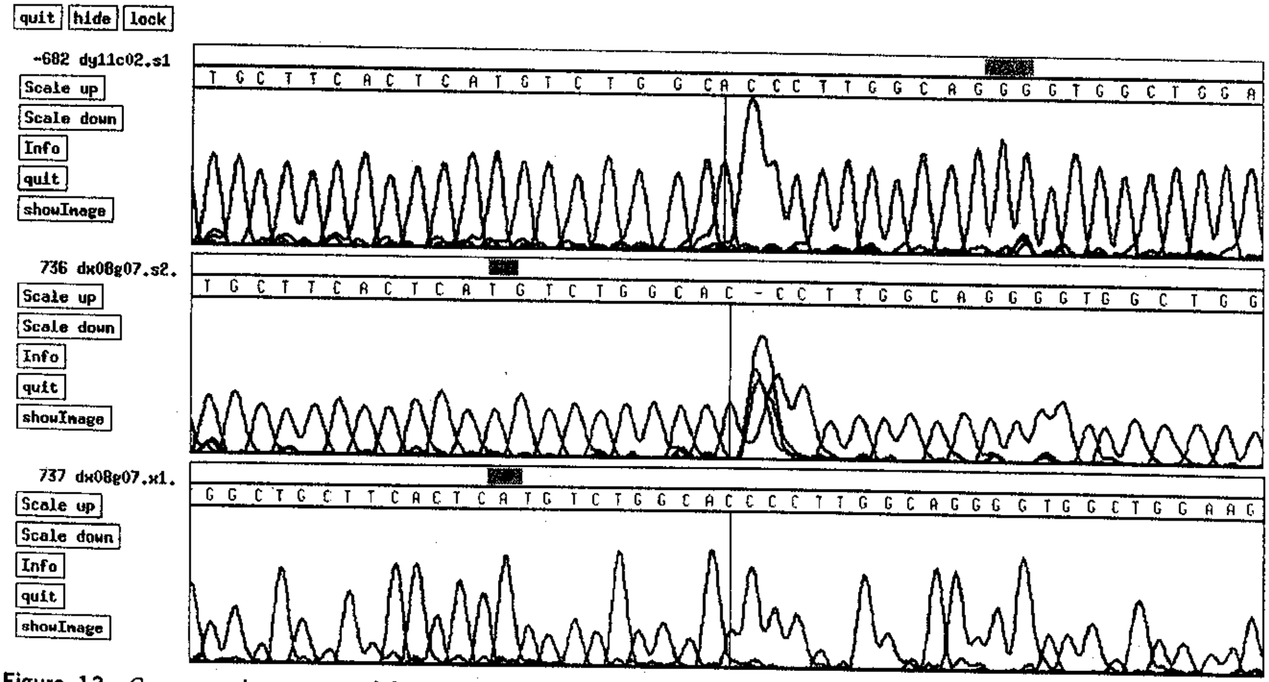

quit
hide
lock
-682 dy11c02.s1
Scale up
Scale down
Info
quit
showImage
T G C T T C A C T C A T G T C T G G C A C C C T T G G C A G G G G T G G C T G G A
736 dx08g07.s2.
Scale up
Scale down
Info
quit
showImage
T G C T T C A C T C A T G T C T G G C A C - C C T T G G C A G G G G T G G C T G G
737 dx08g07.x1.
Scale up
Scale down
Info
quit
showImage
G G C T G C T T C A C T C A T G T C T G G C A C C C C T T G G C A G G G G T G G C T G G A A G

Figure 13 Compressions caused by anomalous mobility. The upper two panels show failure to resolve a compression by sequencing each of the strands with a dye-labeled primer. The bottom panel shows resolution of the compression in sequence generated with a dye-labeled terminator. Traces for A residues are shown in green, C in blue, G in black, and T in red. (The shaded boxes above the sequence in each panel are scroll bars from the XBAP program.) See text for details.

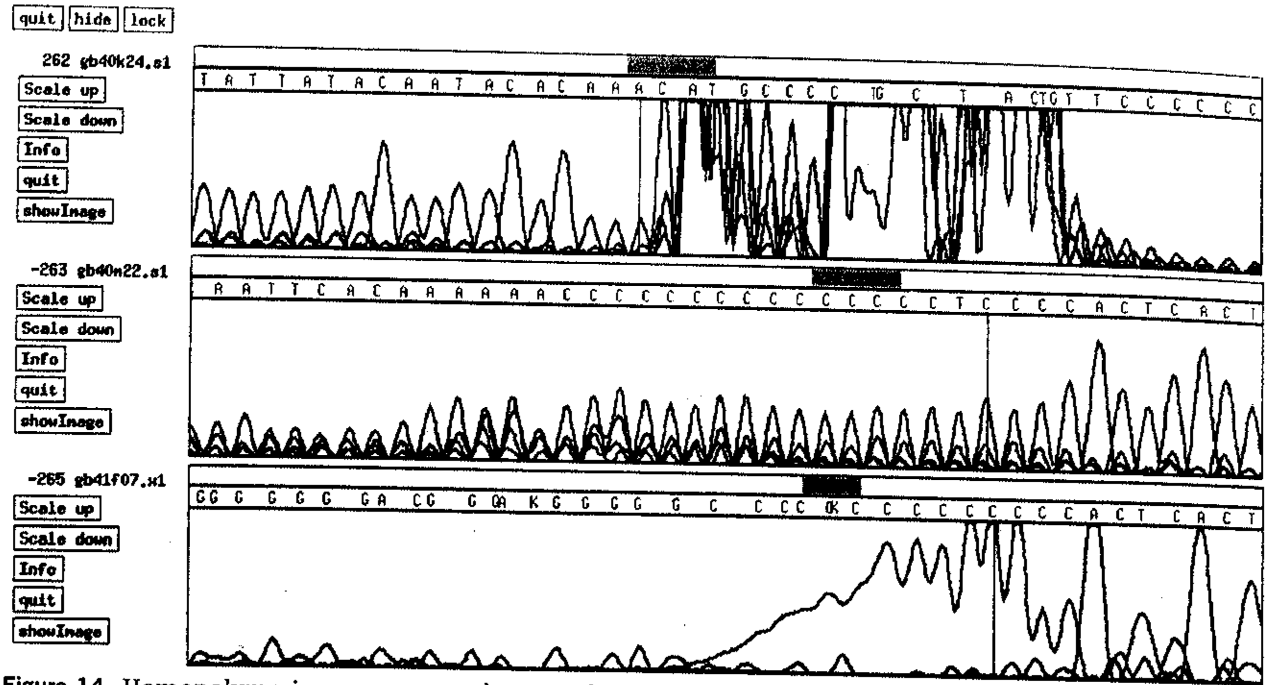

quit
hide
lock
262 gb40k24.s1
Scale up
Scale down
Info
quit
showImage
T A T T A T A C A A T A C A C A A A C A T G C C C C TG C T A CTGT T C C C C C C
-263 gb40n22.s1
Scale up
Scale down
Info
quit
showImage
A A T T C A C A A A A A A C C C C C C C C C C C C C C T C C C C A C T C A C T
-265 gb41f07.s1
Scale up
Scale down
Info
quit
showImage
GG G G G G GA CG G GA K G G G G G C C C C CK C C C C C C C C C C A C T C A C T

Figure 14 Homopolymeric sequences and stops. The upper two panels show the results of sequencing each of the strands with a dye-labeled primer in a region with a run of C residues. The bottom panel shows sequence generated with a dye-labeled terminator. Traces for A residues are shown in green, C in blue, G in black, and T in red. (The shaded boxes above the sequence in each panel are scroll bars from the XBAP program.) See text for details.

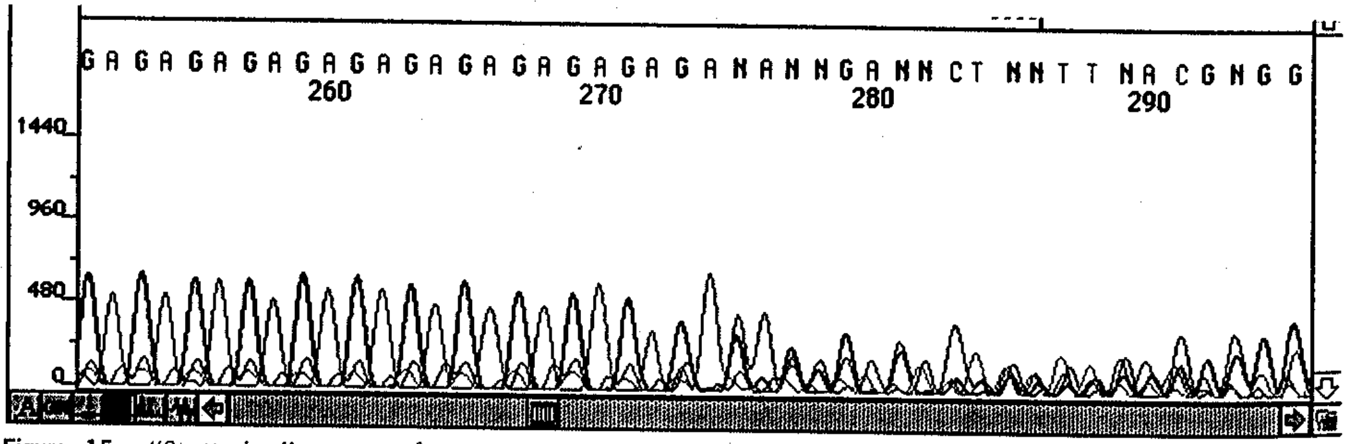


Figure 15 "Stuttering" pattern for GA repeats. This trace shows the stuttering pattern caused by dinucleotide repeats. A similar effect occurs after long homopolymeric runs. Traces for A residues are shown in green, C in blue, G in black, and T in red. See text for details.

Strategies for Determining Contiguous Segments of DNA Sequence

TYPES OF STRATEGIES

The two major DNA sequencing strategies are (1) a random strategy, in which the starting points for sequencing are random, and (2) a directed strategy, in which specific starting points are used for sequencing. Some of the approaches used are shown in Figure 16.

The approaches for generating random starting points for sequencing include the following:

- breaking the target DNA into fragments (using either physical shearing or enzymatic digestion) and subcloning the fragments into a bacteriophage M13 or plasmid vector (for shotgun sequencing, see Figure 16a and Chapter 5)
- using a transposon to randomly insert a univer sal priming site into the target DNA (for transposon-mediated sequencing without prior mapping of insertions, see Figure 16b and Chapter 6)

After the random starting points are generated, a universal primer can be used to produce the primary data for computer assembly of the sequence of the target DNA. Once the desired coverage has been achieved, gap closure and further "finishing" steps can be done with one of the directed approaches below.

The directed approaches for sequencing from defined starting points include the following:

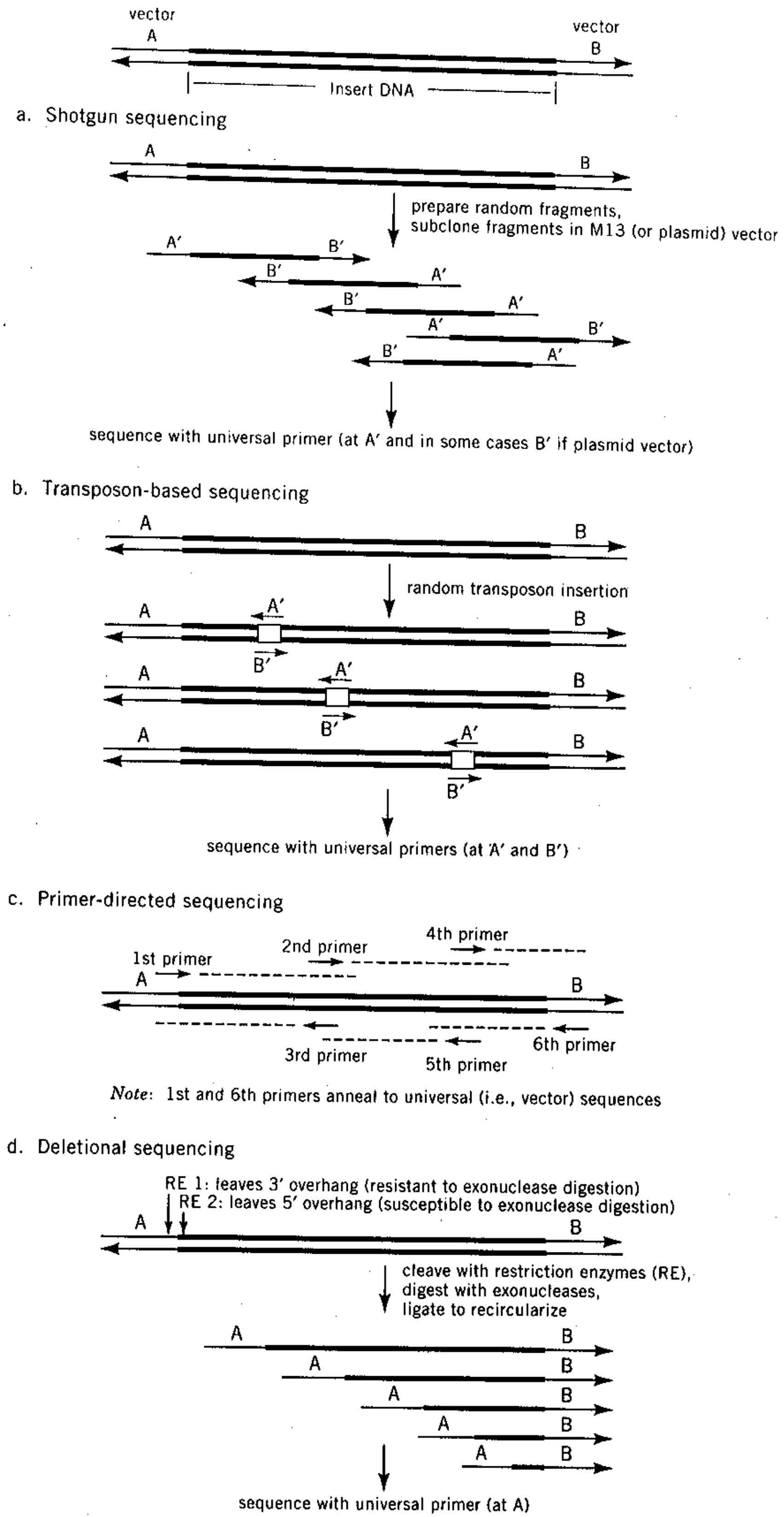
- choosing specific partially characterized subclones or transposon-insertion sites (e.g., those mapped by initial shotgun sequencing) and then resequencing with a reverse primer, a different type of sequencing procedure (i.e., one using dye-labeled terminators versus primers), or gel conditions that permit longer sequences to be read (for subclone-directed sequencing, see Chapter 5, p. 448; for transposon-mediated sequencing, see Chapter 6)
- subcloning specific restriction fragments (e.g., those mapped by initial shotgun sequencing) and then sequencing the fragment(s) by a random or directed method
- using specific synthetic oligonucleotide primers (e.g., synthesize a primer and walk down a stretch of DNA; repeat if necessary) (for primer-directed sequencing, see Figure 16c and p. 384)
- creating and mapping deletions in the target DNA by enzymatically digesting the cloned target DNA with exonuclease III to delete portions from either or both ends (for deletional sequencing, see Figure 16d and pp. 385–391)
- using a transposon to create deletions at either or both ends (for transposon-mediated deletion, see Wang et al. 1993)

CHOOSING A STRATEGY

Before a sequencing strategy can be chosen, it is important to consider the source of the DNA to be sequenced and the goals of the sequencing project (i.e., the desired level of accuracy and contiguity for the final sequence). Successful sequencing requires implementation of the most appropriate approach for a particular region of DNA. In many cases, the primary strategy will be sufficient to achieve the project's goal. However, the most effective approach for complete and efficient sequencing of large regions of genomic DNA is to use a random strategy and directed strategy in concert. Combining these strategies can make the most effective use of resources and also minimize the time and cost of completing a sequence.

The largest DNA sequencing projects completed to date have relied on an initial phase in which a random strategy is used. In a random sequencing strategy, random starting points are generated either by fragmenting the target DNA and subcloning the fragments into a common vector or by inserting a small transposable element into the target DNA. The number of random starting points (i.e., subclones or insertion sites) needed for obtaining most of the target DNA's sequence is directly proportional to the size of the target DNA (see Chapter 5 and Figure 1). However, biological and technical factors often complicate matters. For example, resolving repeats (e.g., two adjacent *Alu* elements with the opposite orientation) often requires a directed strategy as described in Chapter 5.

Methods for classic shotgun sequencing and transposon-mediated sequencing are described in Chapters 5 and 6, respectively. In this chapter, methods are provided for directed sequencing. These directed methods can be used to achieve closure (i.e., completion of a contiguous sequence for a segment of DNA) after use of an initial random sequencing strategy or to sequence small clones and PCR products.



a. Shotgun sequencing

b. Transposon-based sequencing

c. Primer-directed sequencing

Note: 1st and 6th primers anneal to universal (i.e., vector) sequences

d. Deletional sequencing

vector
A
vector
B
Insert DNA
A
B
prepare random fragments,
subclone fragments in M13 (or plasmid) vector
A′
B′
B′
A′
B′
A′
A′
B′
B′
A′
sequence with universal primer (at A′ and in some cases B′ if plasmid vector)
A
B
random transposon insertion
A
A′
B
B′
A
A′
B
B′
A
A′
B
B′
sequence with universal primers (at A′ and B′)
4th primer
2nd primer
1st primer
A
B
3rd primer
5th primer
6th primer
RE 1: leaves 3′ overhang (resistant to exonuclease digestion)
RE 2: leaves 5′ overhang (susceptible to exonuclease digestion)
A
B
cleave with restriction enzymes (RE),
digest with exonucleases,
ligate to recircularize
A
B
A
B
A
B
A
B
A
B
sequence with universal primer (at A)

Figure 16 (*See facing page for legend.*)

Source of the Target DNA

The appropriate sequencing strategy to select depends on the cloning vector and the size and complexity of the target DNA. In this manual, the following sources of target DNA are considered: circular bacterial clones (plasmid, cosmid, fosmid, BAC, PAC, and bacteriophage P1 clones), linear bacteriophage λ clones, linear YAC clones, and linear DNA fragments (restriction fragments and PCR products).

Circular Bacterial Clones

To date, the most common source of DNA for large-scale genomic sequencing is cosmid clones, which typically accept 35–45-kb inserts. One potential disadvantage of sequencing genomic DNA in cosmids (as opposed to using low-copy-number clones such as plasmid, fosmid, BAC, PAC, or bacteriophage P1 clones) is the problem of deletion or rearrangement during growth. It is therefore imperative that cosmid clones be carefully sized by gel electrophoresis before they are used for constructing subclone libraries and that a means be available for confirming the final cosmid sequence with the corresponding genome sequence. Sequences can be confirmed by obtaining sequencing data from overlapping cosmids in the region of interest or by designing and testing a series of PCR assays across the insert sequence to prove colinearity.

Most investigators prefer a shotgun approach for sequencing cosmids, at least for the initial sequencing phase. Since the cosmid vector represents only about 10% of the clone, it is not necessary to purify the insert DNA before fragmentation and subcloning; the entire cosmid clone can simply be fragmented and sequenced en masse. The vector sequence can then be identified by computer analysis and eliminated before the insert sequence is assembled. The vector sequence can also be assembled separately and used as a measure of sequencing accuracy. This initial shotgun phase provides most of the final sequencing data required for the project. The initial effort to assemble the sequence from the primary data obtained during this phase also provides a high-resolution "map" of the subclones. Completion of the sequence (i.e., achieving closure) requires directed sequencing of subclones strategically selected from this map. Other large bacterial clones (e.g., bacteriophage P1, PAC, and BAC clones) can be handled in the same manner as cosmid clones, although a larger number of subclones must be sequenced to provide sufficient coverage of larger inserts (80–200 kb).

For plasmid clones, insert size is a particularly important consideration in choosing a sequencing strategy. Large plasmid inserts (5–10 kb) can be sequenced by using a shotgun approach similar to that used for cosmids. One of the deletional sequencing approaches can be used for plasmid inserts smaller than 10 kb (with an optimal range of 1–5 kb). If appropriate unique restriction sites are present, these inserts can be sequenced by using a deletional approach in which a series of subclones with nested deletions are produced from the target DNA by digestion with exonuclease III (see pp. 385–391). Since the restriction sites that would ensure unidirectional digestion and subsequent religation are not always present, this is not a generally applicable approach. An alternative directed sequencing approach for these inserts uses a transposon-mediated system to produce a series of subclones with deletions (Wang et al. 1993). A disadvantage of this approach is that the target DNA must first be recloned into a specialized vector.

Figure 16 General approaches for sequencing cloned DNA. Four different approaches are shown for sequencing a cloned DNA fragment with flanking vector segments A and B. (*a*) In a shotgun approach, the target DNA is fragmented, and the fragments are subcloned to provide templates for sequencing reactions. The new flanking vector segments A′ and B′ in the subclones provide known primer sequences. (*b*) In a transposon-mediated approach, randomly inserted transposons provide universal, bidirectional primer sequences. (*c*) In a primer-directed approach, the initial sequence is read from a vector priming site. New oligonucleotide primers are then synthesized and used to extend the sequence. Primers for each successive round of sequencing are designed to anneal near the end of the sequence read previously. (*d*) In a deletional sequencing approach (see Figure 17 for details), digestion with exonuclease III is used to create a series of subclones with nested deletions. Since digestion is unidirectional when appropriate restriction sites are present, the adjacent universal priming site (A) is preserved.

If a random approach is preferred, small plasmid inserts (1–5 kb) should be excised, purified, and concatenated by ligation to avoid overrepresentation of insert ends (see Chapter 5, p. 418, step 1) and then fragmented for shotgun sequencing (Bankier and Barrell 1983). An alternative random sequencing approach for a 1–5-kb plasmid insert uses transposon insertion to provide random starting points (see Chapter 6). To minimize the sequencing effort, the transposon-insertion sites in multiple clones can be mapped by PCR and an appropriate set of clones with minimal overlap can be chosen for sequencing.

Some investigators advocate the use a primer-directed approach for walking across the plasmid insert in both directions. Although this approach is problematic for cosmids and other large genomic clones (>5 kb) because of the presence of repeated sequences, it is practical for smaller plasmid clones (1–3 kb) when oligonucleotides can be synthesized conveniently and at a reasonable cost. A library of hexamers may also provide a lower-cost alternative to unique 18–20-mers for this approach (Lodhi and McCombie 1996).

Linear Bacteriophage λ Clones

Bacteriophage λ clones are a common source of starting material for both genomic and cDNA sequencing. A major drawback to the use of bacteriophage λ clones is that the vector can account for more than 50% of the DNA. Thus, steps should be taken to avoid sequencing the vector (or at least to minimize the fraction of vector repeatedly sequenced). This can be accomplished by one of several approaches: (1) If unique flanking restriction sites are present, the insert can be excised from the vector and purified by gel electrophoresis. The purified insert DNA can then be randomly fragmented and subcloned into a plasmid or bacteriophage M13 vector to produce templates for shotgun sequencing. Alternatively, the purified bacteriophage λ insert can be subcloned directly into a plasmid vector for either random or directed sequencing. Insert DNA cloned into bacteriophage λ vectors with an autoexcision mechanism for the insert (e.g., Lambda ZAP) can be transferred into a plasmid vector simply by plating the bacteriophage λ clone on a different *E. coli* host strain (Short et al. 1988). At this point, the 10–20-kb plasmid can be sequenced as described for cosmids. (2) A less desirable approach is to fragment the entire bacteriophage λ clone, subclone the fragments into plasmid or bacteriophage M13 vectors, and then screen out vector-containing subclones by hybridization to a radiolabeled probe corresponding to the bacteriophage λ vector. This can be done either by probing plaques/colonies of the subclone library that have been transferred onto filters or by probing dot blots made with samples of purified subclone DNA (Sambrook et al. 1989). Nonhybridizing subclones can then be sequenced by using a shotgun approach. (3) Bacteriophage λ DNA can be used as the template for primer-directed sequencing, although the presence of repeated sequences in these large inserts often causes problems.

Bacteriophage λ clones containing smaller inserts (1–7 kb; e.g., cDNAs) can be sequenced by using one of a variety of methods: (1) If the insert DNA can be excised and purified to remove the vector DNA, it can be randomly fragmented and sequenced by using a shotgun approach. Andersson et al. (1996) described an effective method for sequencing multiple full-length cDNAs by first concatenating several such insert DNAs. Once concatenated by DNA ligase, the DNA is randomly fragmented, subcloned into bacteriophage M13, and sequenced. After shotgun sequencing and computer assembly, each cDNA will be present as a separate sequence contig. (2) If appropriate restriction sites are present, inserts in bacteriophage λ clones can be sequenced by using the deletional approach (see pp. 385–391). (3) A primer-directed approach can be used to walk across the insert (with an optimal range of 1–3 kb for inserts), although repeated sequences may interfere with genomic DNA sequencing (p. 384). (4) Inserts that are 5 kb in size or smaller can often be amplified by PCR and then purified for either subcloning followed by sequencing or sequencing in an uncloned form using any strategy.

Linear YAC Clones

YAC clones can be sequenced by using a random strategy (see Chapter 5) although there are difficulties to overcome. Isolating a sufficient amount of purified YAC DNA is a limiting step. This step can be accomplished by preparative PFGE (see Chapter 5, pp. 410–413) (Vaudin et al. 1995), but yeast DNA is always present in the isolated YAC DNA. The amount of yeast DNA can often be reduced by performing a second round of preparative PFGE. Sequence assembly and editing tasks are also considerable because of the size of the inserts. An alternative approach that has been used successfully for BACs and small (100–250 kb) YACs combines ordered, two-stage shotgun sequencing and some mapping elements (Chen et al. 1993).

Linear DNA Fragments

Linear DNA such as restriction fragments and PCR products can be sequenced completely by using a number of strategies. The choice of strategy mainly depends on the size of the fragment and the sequencing goal. For example, if the fragment is short (≤1 kb) and sequencing primers are available for both ends (e.g., amplification primers), a walking approach will efficiently provide the entire sequence of the fragment. For longer fragments (>1 kb) where an accurate final sequence is the goal, a shotgun approach (see Chapter 5) or transposon-mediated approach (see Chapter 6) can be used. If appropriate ends are present on the fragment, the nested deletion approach (see pp. 385–391) is also an alternative.

Restriction fragments are essentially handled as described for inserts from circular bacterial clones (see pp. 381–382). If a shotgun approach is chosen, restriction fragments can be concatenated by ligation before they are sheared (see Chapter 5, p. 418, step 1). This helps avoid overrepresentation of the original fragment ends in the resulting subclones (Bankier and Barrell 1983).

PCR products that are longer than 1 kb can be directly fragmented and sequenced by using a shotgun approach. Concatenation is not necessary as long as the PCR product is not 5′ phosphorylated (i.e., because of the lack of 5′ phosphate groups on the amplification primers). Alternatively, PCR products can be cloned by using one of several methods (see Chapter 3) and then sequenced as described above for plasmids. The advantage of cloning the PCR product is that sequence data of higher quality can typically be produced. Disadvantages include the extra effort required for the cloning step and the potential for misincorporation errors and end modification caused by *Taq* DNA polymerase. It should be noted that since misincorporation errors occur randomly during PCR, this problem can be overcome simply by sequencing at least three independent clones.

Many procedures are available for direct sequencing of uncloned PCR products. To obtain good results, a homogeneous PCR product that is free of unincorporated primers and dNTPs must be used. Procedures for purifying PCR products (as well as methods for preparing single-stranded template from the PCR product and thus improving the quality of the sequence data that can be obtained) are provided on pp. 321–338.

Sequencing Goals

Once the source of the target DNA has been selected, the overall goal for sequencing should be considered (i.e., the desired level of accuracy and contiguity for the final sequence).

De Novo DNA Sequencing

For de novo DNA sequencing (i.e., sequencing a previously uncharacterized region of DNA), the goal is typically to produce a contiguous sequence with a very high degree of accuracy. The best way to achieve this goal depends on the size of the region to be sequenced and the type of clones available. The use of a combination of strategies may even be necessary. For complete sequencing of large genomic clones (>40 kb; e.g., cosmid or BAC clones), an initial shotgun phase to rapidly produce most of the sequence plus a secondary directed phase to achieve closure and resolve local ambiguities has proved to be highly effective. Since the presence of repeated sequences often hampers primer-directed approaches (see p. 384) in larger, more complex clones (e.g., >5 kb), an initial random strategy is preferred for most de novo sequencing.

Because of its simplicity, a random strategy is also typically preferred for de novo sequencing of smaller clones (e.g., cDNAs or genomic segments of <10 kb), but a directed strategy (e.g., the enzymatic deletional sequencing approach, transposon insertion, or transposon-mediated deletion) can also be used. For de novo sequencing of small segments of large genomic clones (<1 kb; e.g., the ends of large-insert clones), a primer-directed approach is often possible.

Although a random strategy is preferred for de novo sequencing of cosmid clones, a deletional approach can often be used for smaller regions of cosmids that are difficult to sequence (e.g., a region of an insert that contains a series of highly conserved tandem repeats [Chen et al. 1993]). The most common approach for deletional sequencing uses digestion with exonuclease III to produce subclones with nested deletions (see pp. 385–391), which can be sized by gel electrophoresis. Deletional sequencing has the advantages that it always proceeds from a known location (i.e., the universal primer) and it avoids the high degree of "oversequencing" (redundancy) characteristic of shotgun sequencing. The deletional sequencing approach also has several drawbacks. The techniques used to produce subclones with nested deletions are tedious. The results vary greatly with different enzyme lots and DNA preparations. Furthermore, when digestion with exonuclease III is used, the DNA fragment must contain an appropriate restriction site for unidirectional digestion and subsequent religation. This is a major disadvantage when genomic clones are being sequenced unless the vector has been specifically engineered to contain the appropriate rarely occurring restriction site(s).

Confirmatory DNA Sequencing

Confirmatory sequencing (i.e., resequencing) of a small region of interest (<1 kb) can often be performed by using cloned DNA or a PCR product as the template for a primer-directed approach. Although primer-directed sequencing of uncloned PCR products does not always produce data of high quality, it is useful for applications such as generating markers for physical mapping, confirming plasmid constructs, screening the products of site-directed mutagenesis experiments, or examining sequences associated with genetic disease. Since the target region has often been characterized, the experiment can be designed by synthesizing an appropriate primer so that the sequence of interest is within 50 to 100 bases of the sequencing primer and thus within the region where the highest sequencing accuracy can be obtained.

Primer-directed Sequencing

In a primer-directed approach (primer "walking"), custom oligonucleotide primers are used to read sequence directly from the ends or other small regions (1–3 kb) of genomic or cDNA clones. First, approximately 500 bp are read from a universal priming site. Next, a new primer is synthesized. This primer typically should anneal 50–100 bp from the 3′ end of the sequence that was read initially, so that there is some overlap with the original sequence. Finally, the sequence read from the original template DNA is extended by using the same template and the new primer. These steps are reiterated several times to walk through the sequence.

Primer-directed sequencing is very useful as the primary approach for sequencing small regions of DNA or as a secondary approach for achieving closure after an initial random sequencing strategy (see Chapter 5, pp. 449–450). In principle, a primer-directed approach (Ansorge et al. 1987) can also be used to walk down the entire length of both strands of a large genomic clone (e.g., a cosmid or BAC clone). However, the success of this approach is limited by the amount of template DNA required, the stability of the template, the presence of repeated sequences, and the effort and organization required to efficiently perform the iterative process for such a project.

Procedures for using custom oligonucleotide primers with dye-labeled terminators in fluorescence-based DNA sequencing are described on pp. 355–362. To ensure the highest success rate with primer-directed sequencing, follow the guidelines on p. 339 for designing the custom primers. When the primers are designed carefully, 80–90% of primer-directed sequencing reactions are successful. Factors that can produce failures include (1) a repeated sequence in the template that results in the presence of multiple sites for primer annealing and (2) the secondary structure (e.g., hairpin loops) of either the primer or the template. Although the reason for failure is difficult to determine in many cases, it may be possible to overcome. For example, when primer-directed sequencing is used as a closure method in a random sequencing project, closure is often achieved by simply using the primer that originally failed with a second appropriate subclone to provide the necessary data. In such a case, the first subclone may have contained a small repeat that caused the initial failure, but the second subclone did not. If other suitable subclones are not available or do not solve the problem, a different primer can be designed and synthesized. Before the new primer is synthesized, the candidate primer sequence should be carefully scanned to avoid small repeats. In addition, the annealing site for the new primer should be different from that of the first primer in case the problem was caused by local secondary structure in the template DNA.