Chapter 10

Sequencing Genes and Genomes

Chapter contents

Part I of this book has shown how a skilfully performed cloning or PCR experiment can provide a pure sample of an individual gene, or any other DNA sequence, separated from all the other genes and DNA sequences in the cell. Now we can turn our attention to the ways in which cloning, PCR, and other DNA analysis techniques are used to study genes and genomes. We will consider three aspects of molecular biology research:

- The techniques used to obtain the nucleotide sequence of individual genes and entire genomes (this Chapter);
- The methods used to study the expression and function of individual genes (Chapter 11);
- The techniques that are used to study entire genomes (Chapter 12).

Probably the most important technique available to the molecular biologist is DNA sequencing, by which the precise order of nucleotides in a piece of DNA can be determined. DNA sequencing methods have been around for 40 years, and since the mid-1970s rapid and efficient sequencing has been possible. Initially these techniques were applied to individual genes, but since the early 1990s an increasing number of entire genome sequences have been obtained. In this chapter we will study the methodology used in DNA sequencing and then examine how these techniques are used in genome projects.

10.1 The methodology for DNA sequencing

There are several procedures for DNA sequencing, the most popular being the chain termination method first devised by Fred Sanger and colleagues in the mid-1970s. Chain

Gene Cloning and DNA Analysis: An Introduction. 6th edition. By T.A. Brown. Published 2010 by Blackwell Publishing.

termination sequencing has gained pre-eminence for several reasons, not least being the relative ease with which the technique can be automated. As we will see later in this chapter, in order to sequence an entire genome a huge number of individual sequencing experiments must be carried out, and it would take many years to perform all of these by hand. Automated sequencing techniques are therefore essential if a genome project is to be completed in a reasonable timespan.

Part of the automation strategy is to design systems that enable many individual sequencing experiments to be carried out at once. With the chain termination method, up to 96 sequences can be obtained simultaneously in a single run of a sequencing machine. This is still not enough to fully satisfy the demands of genome sequencing, and during the last few years an alternative method called pyrosequencing has become popular. Pyrosequencing, which was invented in 1998, forms the basis to a massively parallel strategy that enables hundreds of thousands of short sequences to be generated at the same time.

10.1.1 Chain termination DNA sequencing

Chain termination DNA sequencing is based on the principle that single-stranded DNA molecules that differ in length by just a single nucleotide can be separated from one another by polyacrylamide gel electrophoresis. This means that it is possible to resolve a family of molecules, representing all lengths from 10 to 1500 nucleotides, into a series of bands in a slab or capillary gel (Figure 10.1).

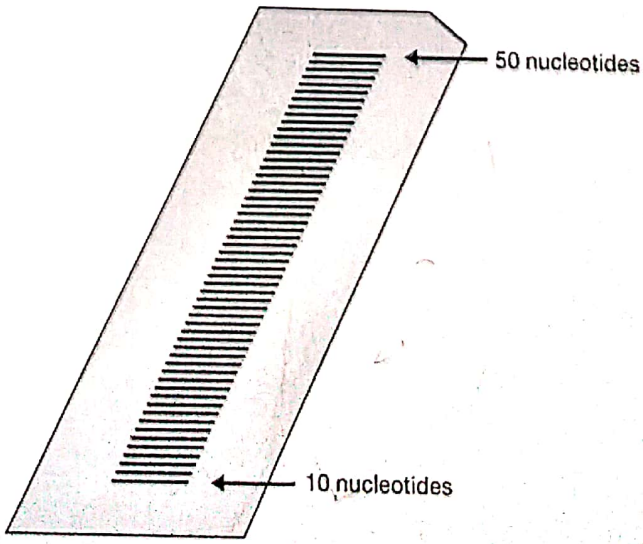
Chain termination sequencing in outline

The starting material for a chain termination sequencing experiment is a preparation of identical single-stranded DNA molecules. The first step is to anneal a short oligonucleotide to the same position on each molecule, this oligonucleotide subsequently acting as the primer for synthesis of a new DNA strand that is complementary to the template (Figure 10.2a).

The strand synthesis reaction, which is catalyzed by a DNA polymerase enzyme and requires the four deoxyribonucleotide triphosphates (dNTPs—dATP, dCTP, dGTP, and dTTP) as substrates, would normally continue until several thousand nucleotides had been polymerized. This does not occur in a chain termination sequencing experiment

Figure 10.1

Polyacrylamide gel electrophoresis can resolve single-stranded DNA molecules that differ in length by just one nucleotide. The banding pattern shown here is produced after separation of single-stranded DNA molecules by denaturing polyacrylamide gel electrophoresis. The molecules have been labeled with a radioactive marker and the bands visualized by autoradiography.



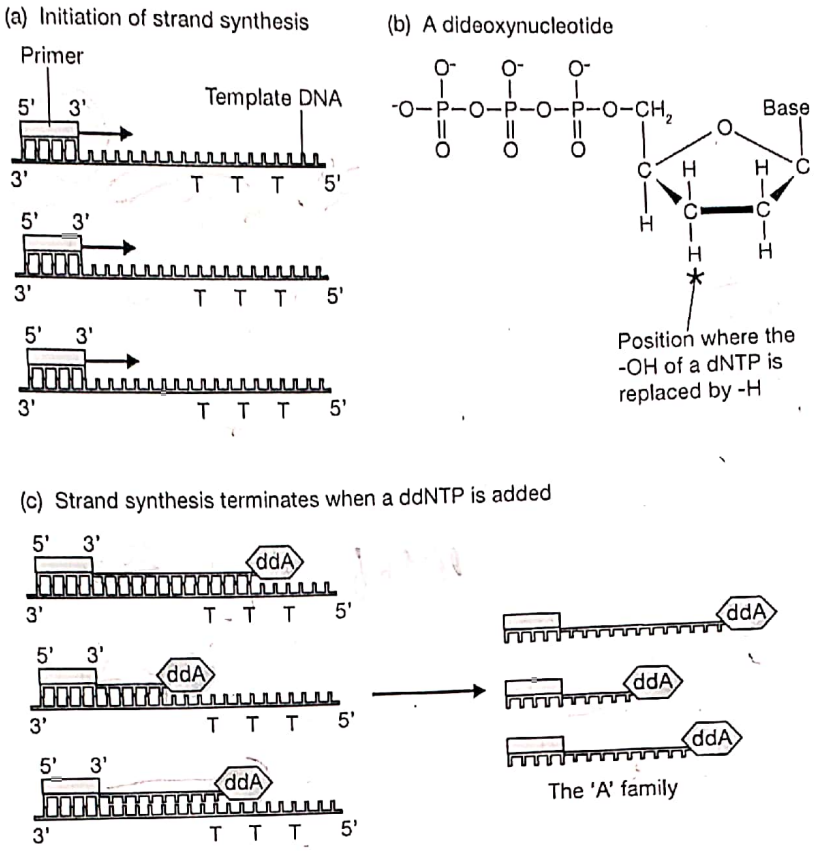


Figure 10.2
Chain termination DNA sequencing.

because, as well as the four deoxynucleotides, a small amount of each of four dideoxynucleotides (ddNTPs—ddATP, ddCTP, ddGTP, and ddTTP) is added to the reaction. Each of these dideoxynucleotides is labeled with a different fluorescent marker.

The polymerase enzyme does not discriminate between deoxy- and dideoxynucleotides, but once incorporated a dideoxynucleotide blocks further elongation because it lacks the 3′-hydroxyl group needed to form a connection with the next nucleotide (Figure 10.2b). Because the normal deoxynucleotides are also present, in larger amounts than the dideoxynucleotides, the strand synthesis does not always terminate close to the primer: in fact, several hundred nucleotides may be polymerized before a dideoxynucleotide is eventually incorporated. The result is a set of new molecules, all of different lengths, and each ending in a dideoxynucleotide whose identity indicates the nucleotide—A, C, G, or T—that is present at the equivalent position in the template DNA (Figure 10.2c).

To work out the DNA sequence, all that we have to do is identify the dideoxynucleotide at the end of each chain-terminated molecule. This is where the polyacrylamide gel comes into play. The mixture is loaded into a well of a polyacrylamide slab gel, or into a tube of a capillary gel system, and electrophoresis carried out to separate the molecules according to their lengths. After separation, the molecules are run past a fluorescent detector capable of discriminating the labels attached to the dideoxynucleotides (Figure 10.3a). The detector therefore determines if each molecule ends in an A, C, G, or T. The sequence can be printed out for examination by the operator (Figure 10.3b), or entered directly into a storage device for future analysis.

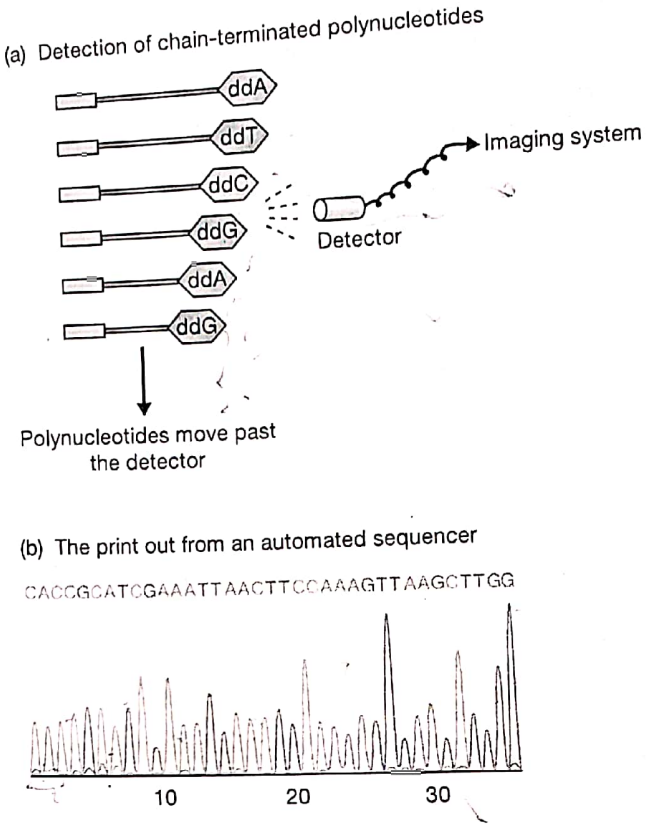


Figure 10.3

Reading the sequence generated by a chain termination experiment. (a) Each dideoxynucleotide is labeled with a different fluorochrome, so the chain-terminated polynucleotides are distinguished as they pass by the detector. (b) An example of a sequence print out.

Not all DNA polymerases can be used for sequencing

Any DNA polymerase is capable of extending a primer that has been annealed to a single-stranded DNA molecule, but not all polymerases can be used for DNA sequencing. This is because many DNA polymerases have a mixed enzymatic activity, being able to degrade as well as synthesize DNA (p. 48). Degradation can occur in either the 5′→3′ or 3′→5′ direction (Figure 10.4), and both activities are detrimental to accurate chain termination sequencing. The 5′→3′ exonuclease activity enables the polymerase to remove nucleotides from the 5′ ends of the newly-synthesized strands, changing the lengths of these strands so that they no longer run through the polyacrylamide gel in the appropriate order. The 3′→5′ activity could have the same effect, but more importantly

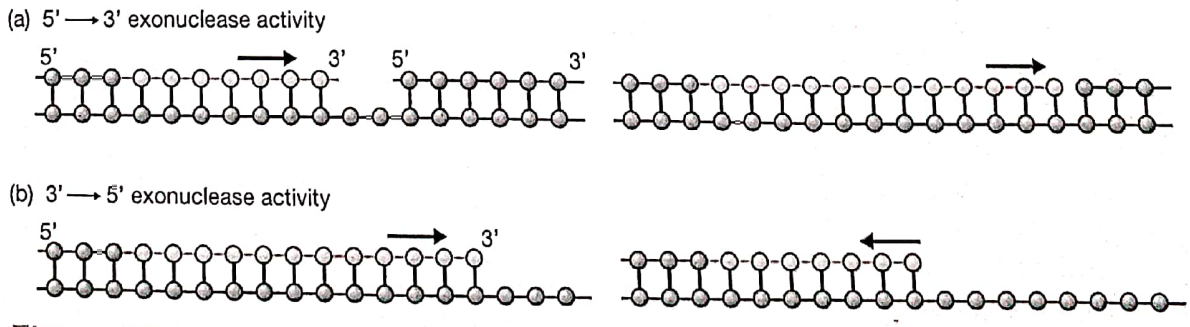


Figure 10.4

The exonuclease activities of DNA polymerases. (a) The 5′→3′ activity has an important role in DNA repair in the cell, as it enables the polymerase to replace a damaged DNA strand. In DNA sequencing this activity can result in the 5′ ends of newly-synthesized strands becoming shortened. (b) The 3′→5′ activity also has an important role in the cell, as it allows the polymerase to correct its own mistakes, by reversing and replacing a nucleotide that has been added in error (e.g., a T instead of a G). This is called proofreading. During DNA sequencing, this activity can result in removal of a dideoxynucleotide that has just been added to the newly-synthesized strand, so that chain termination does not occur.

will remove a dideoxynucleotide that has just been added at the 3′ end, preventing chain termination from occurring.

In the original method for chain termination sequencing, the Klenow polymerase was used as the sequencing enzyme. As described on p. 49, this is a modified version of the DNA polymerase I enzyme from *E. coli*, the modification removing the 5′→3′ exonuclease activity of the standard enzyme. However, the Klenow polymerase has low processivity, meaning that it can only synthesize a relatively short DNA strand before dissociating from the template due to natural causes. This limits the length of sequence that can be obtained from a single experiment to about 250 bp. To avoid this problem, most sequencing today makes use of a more specialized enzyme, such as Sequenase, a modified version of the DNA polymerase encoded by bacteriophage T7. Sequenase has high processivity and no exonuclease activity and so is ideal for chain termination sequencing, enabling sequences of up to 750 bp to be obtained in a single experiment.

Chain termination sequencing requires a single-stranded DNA template

The template for a chain termination experiment is a single-stranded version of the DNA molecule to be sequenced. One way of obtaining single-stranded DNA is to use an M13 vector, but the M13 system, although designed specifically to provide DNA for chain termination sequencing, is not ideal for this purpose. The problem is that cloned DNA fragments that are longer than about 3 kb are unstable in an M13 vector and can undergo deletions and rearrangements. This means that M13 cloning can only be used with short pieces of DNA.

Plasmid vectors, which do not suffer instability problems, are therefore more popular, but some means is needed of converting the double-stranded plasmid into a single-stranded form. There are two possibilities:

- Double-stranded plasmid DNA can be converted into single-stranded DNA by denaturation with alkali or by boiling. This is a common method for obtaining template DNA for DNA sequencing, but a shortcoming is that it can be difficult to prepare plasmid DNA that is not contaminated with small quantities of bacterial DNA and RNA, which can act as spurious templates or primers in the DNA sequencing experiment.
- The DNA can be cloned in a phagemid, a plasmid vector that contains an M13 origin of replication and which can therefore be obtained as both double- and single-stranded DNA versions (p. 96). Phagemids avoid the instabilities of M13 cloning and can be used with fragments up to 10 kb or more.

The need for single-stranded DNA can also be sidestepped by using a thermostable DNA polymerase as the sequencing enzyme. This method, called thermal cycle sequencing, is carried out in a similar way to PCR, but just one primer is used and the reaction mixture includes the four dideoxynucleotides (Figure 10.5). Because there is only one primer, only one of the strands of the starting molecule is copied, and the product accumulates in a linear fashion, not exponentially as is the case in a real PCR. The presence of the dideoxynucleotides in the reaction mixture causes chain termination, as in the standard methodology, and the family of resulting strands can be analyzed and the sequence read in the usual way. Thermal cycle sequencing can therefore be used with DNA cloned in any type of vector.

The primer determines the region of the template DNA that will be sequenced

In the first stage of a chain termination sequencing experiment, an oligonucleotide primer is annealed onto the template DNA (see Figure 10.2a). The main function of the

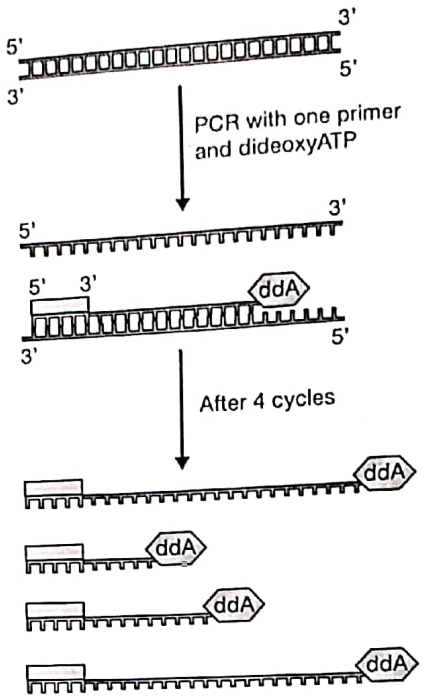


Figure 10.5
The basis to thermal cycle sequencing. A PCR is set up with just one primer and one of the dideoxynucleotides. One of the template strands is copied into a family of chain-terminated polynucleotides. ddA = dideoxyATP.

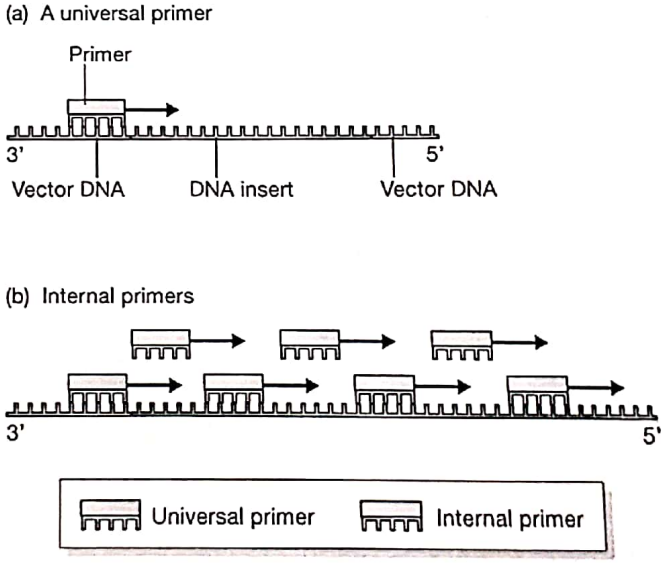


Figure 10.6
Different types of primer for chain termination sequencing.

primer is to provide the short double-stranded region that is needed in order for the DNA polymerase to initiate DNA synthesis. The primer also plays a second critical role in determining the region of the template molecule that will be sequenced.

For most sequencing experiments a universal primer is used, this being one that is complementary to the part of the vector DNA immediately adjacent to the point into which new DNA is ligated (Figure 10.6a). The 3′ end of the primer points toward the inserted DNA, so the sequence that is obtained starts with a short stretch of the vector and then progresses into the cloned DNA fragment. If the DNA is cloned in a plasmid vector, then both forward and reverse universal primers can be used, enabling sequences to be obtained from both ends of the insert. This is an advantage if the cloned DNA is more than 750 bp and hence too long to be sequenced completely in one experiment. Alternatively, it is possible to extend the sequence in one direction by synthesizing a non-universal primer, designed to anneal at a position within the insert DNA (Figure 10.6b).

An experiment with this primer will provide a second short sequence that overlaps the previous one.

10.1.2 Pyrosequencing

Pyrosequencing is the second important type of DNA sequencing methodology that is in use today. Pyrosequencing does not require electrophoresis or any other fragment separation procedure and so is more rapid than chain termination sequencing. It is only able to generate up to 150 bp in a single experiment, and at first glance might appear to be less useful than the chain termination method, especially if the objective is to sequence a genome. The advantage with pyrosequencing is that it can be automated in a massively parallel manner that enables hundreds of thousands of sequences to be obtained at once, perhaps as much as 1000 Mb in a single run. Sequence is therefore produced much more quickly than is possible by the chain termination method, which explains why pyrosequencing is gradually taking over as the method of choice for genome projects.

Pyrosequencing involves detection of pulses of chemiluminescence

Pyrosequencing, like the chain termination method, requires a preparation of identical single-stranded DNA molecules as the starting material. These are obtained by alkali denaturation of PCR products or, more rarely, recombinant plasmid molecules. After attachment of the primer, the template is copied by a DNA polymerase in a straightforward manner without added dideoxynucleotides. As the new strand is being made, the order in which the deoxynucleotides are incorporated is detected, so the sequence can be "read" as the reaction proceeds.

The addition of a deoxynucleotide to the end of the growing strand is detectable because it is accompanied by release of a molecule of pyrophosphate, which can be converted by the enzyme sulfurylase into a flash of chemiluminescence. Of course, if all four deoxynucleotides were added at once, then flashes of light would be seen all the time and no useful sequence information would be obtained. Each deoxynucleotide is therefore added separately, one after the other, with a nucleotidase enzyme also present in the reaction mixture so that if a deoxynucleotide is not incorporated into the polynucleotide then it is rapidly degraded before the next one is added (Figure 10.7). This procedure makes it possible to follow the order in which the deoxynucleotides are incorporated into the growing strand. The technique sounds complicated, but it simply requires that a repetitive series of additions be made to the reaction mixture, precisely the type of procedure that is easily automated.

Massively parallel pyrosequencing

The high throughput version of pyrosequencing usually begins with genomic DNA, rather than PCR products or clones. The DNA is broken into fragments between 300 and 500 bp in length (Figure 10.8a), and each fragment is ligated to a pair of adaptors (p. 65), one adaptor to either end (Figure 10.8b). These adaptors play two important roles. First, they enable the DNA fragments to be attached to small metallic beads. This is because one of the adaptors has a biotin label attached to its 5′ end, and the beads are coated with streptavidin, to which biotin binds with great affinity (p. 137). DNA fragments therefore become attached to the beads via biotin-streptavidin linkages (Figure 10.8c). The ratio of DNA fragments to beads is set so that, on average, just one fragment becomes attached to each bead.

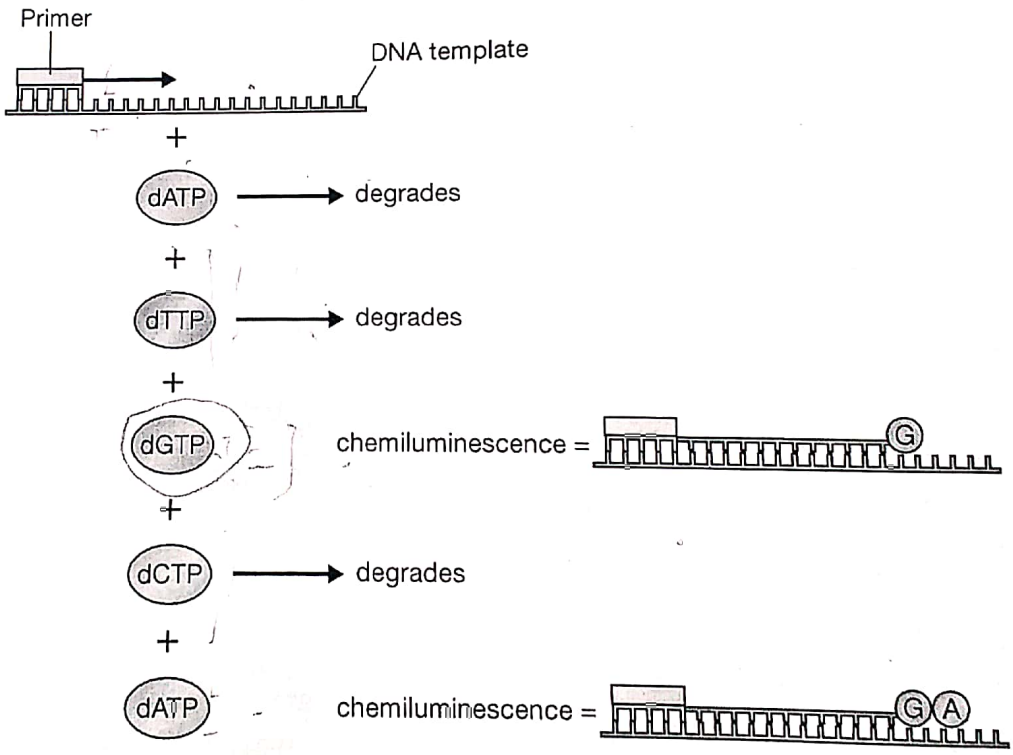


Figure 10.7
Pyrosequencing.