Chapter 10

Sequencing Genes and Genomes

Chapter contents

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).

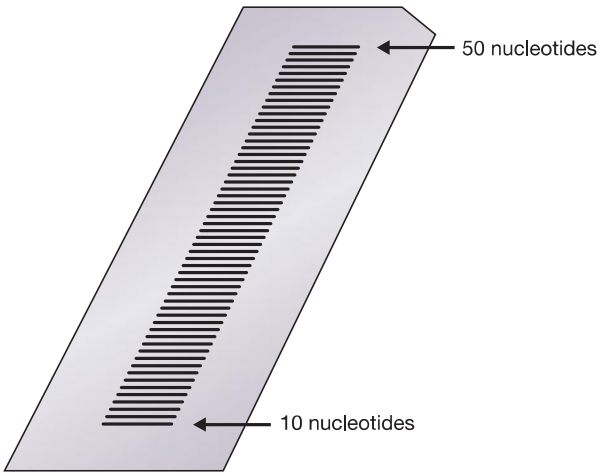
Chain termination sequencing in outline

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.



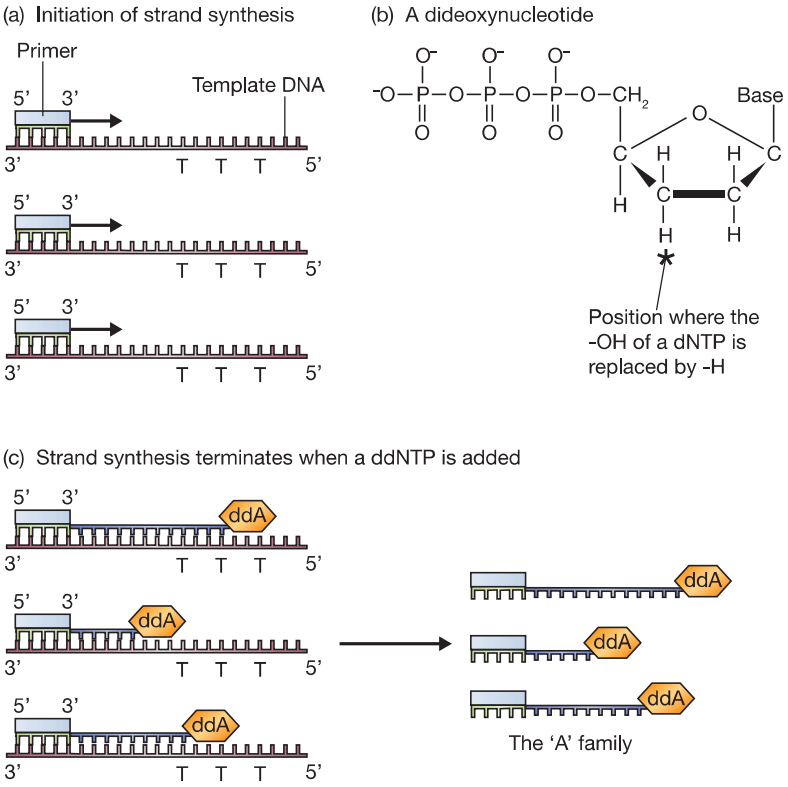


Figure 10.2
Chain termination DNA sequencing.

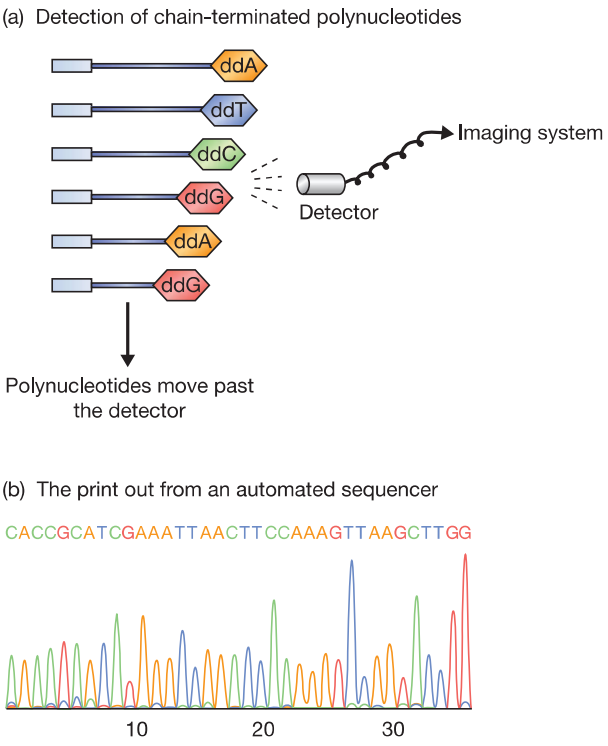
because, as well as the four deoxynucleotides, a small amount of each of four **dideoxynucleotides** (ddNTPs—ddATP, ddCTP, ddGTP, and ddTTP) is added to the reaction. Each of these dideoxynucleotides is labeled with a different fluorescent marker.

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

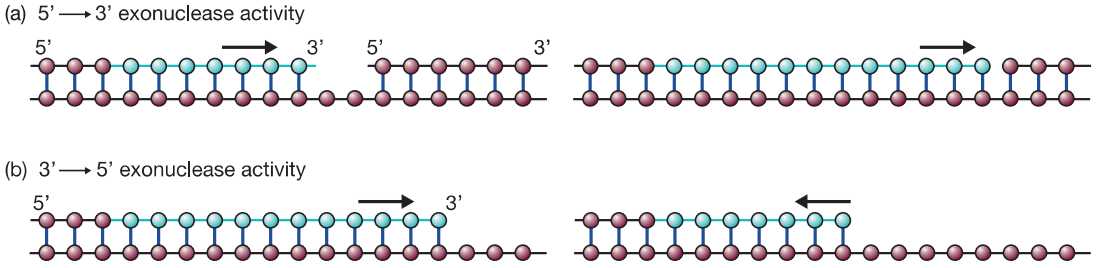


Figure 10.4

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.

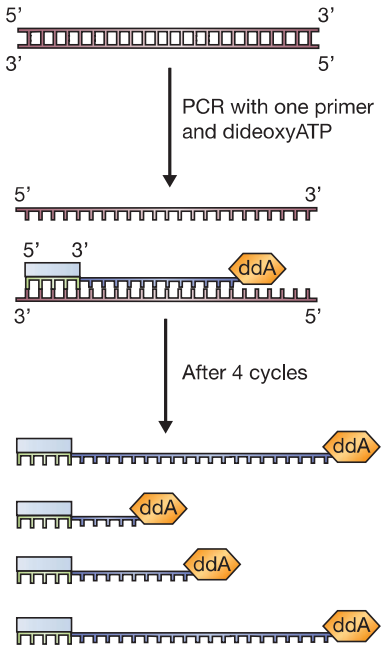
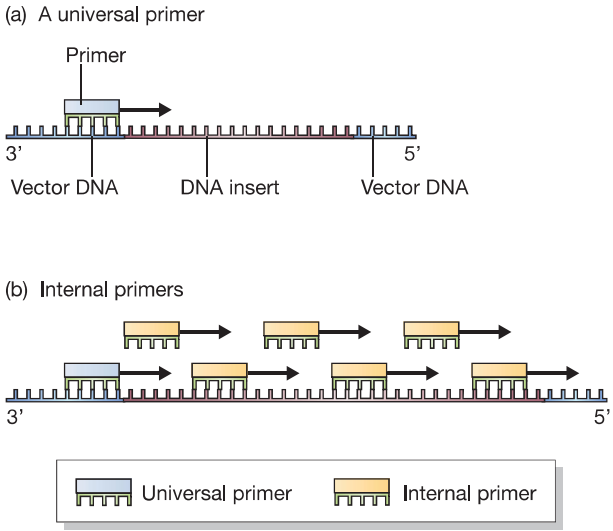


Figure 10.6

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.

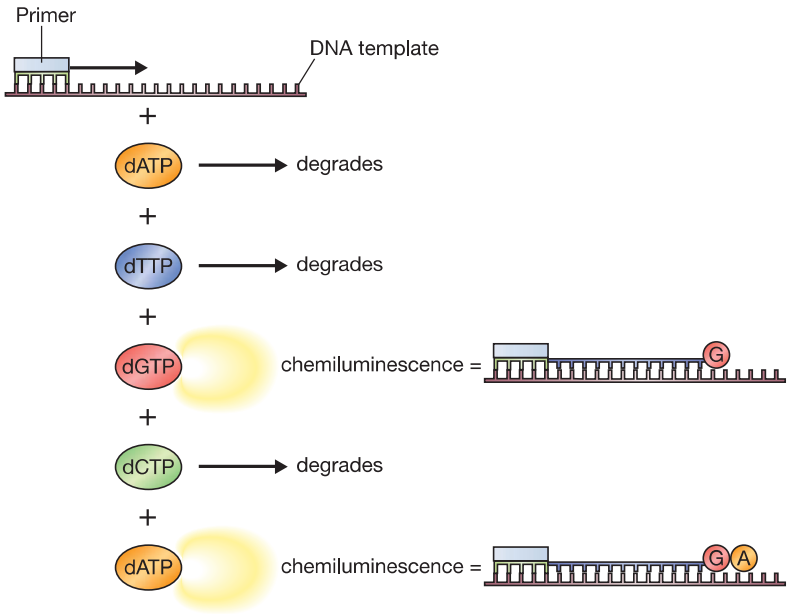
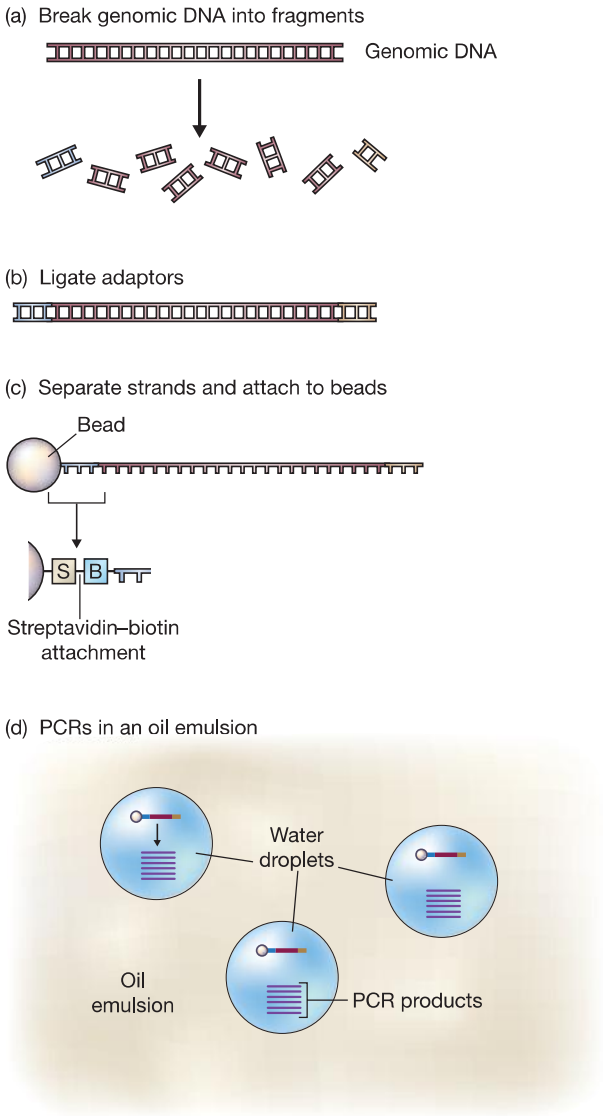


Figure 10.8
One method for massively parallel DNA sequencing. B = biotin, S = streptavidin.



Each DNA fragment will now be amplified by PCR so that enough copies are made for sequencing. The adaptors now play their second role as they provide the annealing sites for the primers for this PCR. The same pair of primers can therefore be used for all the fragments, even though the fragments themselves have many different sequences. If the PCR is carried out immediately then all we will obtain is a mixture of all the products, which will not enable us to obtain the individual sequences of each one. To solve this problem, PCR is carried out in an oil emulsion, each bead residing in its own aqueous droplet within the emulsion (Figure 10.8d). Each droplet contains all the reagents needed for PCR, and is physically separated from all the other droplets by the barrier provided by the oil component of emulsion. After PCR, the aqueous droplets are transferred into wells on a plastic strip so there is one droplet and hence once PCR product per well, and the pyrosequencing reactions are carried out in each well.

10.2 How to sequence a genome

The first DNA molecule to be completely sequenced was the 5386 nucleotide genome of bacteriophage ϕX174, which was completed in 1975. This was quickly followed by sequences for SV40 virus (5243 bp) in 1977 and pBR322 (4363 bp) in 1978. Gradually sequencing was applied to larger molecules. Professor Sanger's group published the sequence of the human mitochondrial genome (16.6 kb) in 1981 and of bacteriophage λ (49 kb) in 1982. Nowadays sequences of 100–200 kb are routine and most research laboratories have the necessary expertise to generate this amount of information.

The pioneering projects today are the massive genome initiatives, each aimed at obtaining the nucleotide sequence of the entire genome of a particular organism. The first chromosome sequence, for chromosome III of the yeast *Saccharomyces cerevisiae*, was published in 1992, and the entire yeast genome was completed in 1996. There are now complete genome sequences for the worm *Caenorhabditis elegans*, the fly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, the human *Homo sapiens*, and over 1000 other species. It is even possible to obtain genome sequences for extinct species such as mammoths and Neanderthals.

A single chain termination sequencing experiment produces about 750 bp of sequence, and a single pyrosequence yields up to 150 bp. But the total size of a fairly typical bacterial genome is 4,000,000 bp and the human genome is 3,200,000,000 bp (Table 10.1). Clearly a large number of sequencing experiments must be carried out in

Table 10.1

Sizes of representative genomes.

SPECIES	TYPE OF ORGANISM	GENOME SIZE (Mb)
Mycoplasma genitalium	Bacterium	0.58
Haemophilus influenzae	Bacterium	1.83
Escherichia coli	Bacterium	4.64
Saccharomyces cerevisiae	Yeast	12.10
Caenorhabditis elegans	Nematode worm	97.00
Drosophila melanogaster	Insect	180.00
Arabidopsis thaliana	Plant	125.00
Homo sapiens	Mammal	3200.00
Triticum aestivum	Plant (wheat)	16,000.00

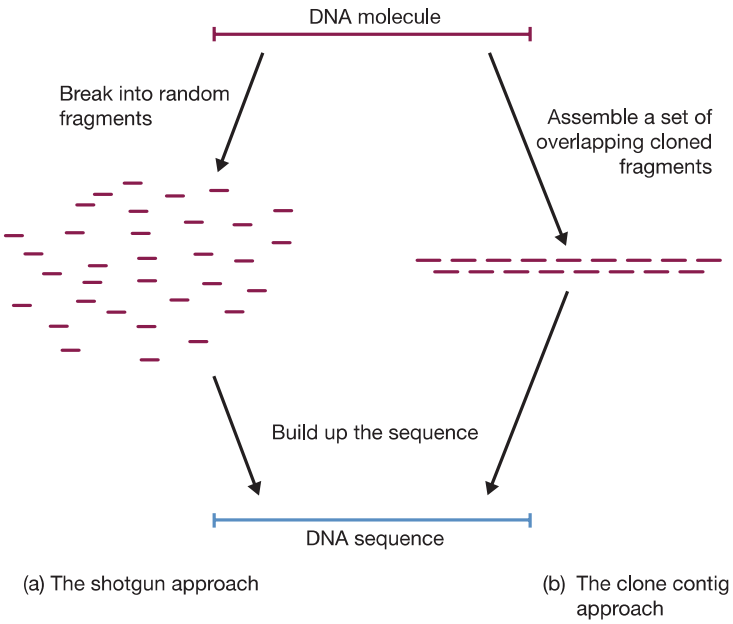


Figure 10.9

Strategies for assembly of a contiguous genome sequence: (a) the shotgun approach; (b) the clone contig approach.

order to determine the sequence of an entire genome. In practice, thanks to automated systems, the generation of sufficient sequence data is one of the more routine aspects of a genome project. The first real problem that arises is the need to assemble the thousands or perhaps millions of individual sequences into a contiguous genome sequence. Two different strategies have been developed for sequence assembly (Figure 10.9):

- The **shotgun approach**, in which the genome is randomly broken into short fragments. The resulting sequences are examined for overlaps and these are used to build up the contiguous genome sequence.
- The **clone contig approach**, which involves a pre-sequencing phase during which a series of overlapping clones is identified. This contiguous series is called a **contig**. Each piece of cloned DNA is then sequenced, and this sequence placed at its appropriate position on the contig map in order to gradually build up the overlapping genome sequence.

10.2.1 The shotgun approach to genome sequencing

The key requirement of the shotgun approach is that it must be possible to identify overlaps between all the individual sequences that are generated, and this identification process must be accurate and unambiguous so that the correct genome sequence is obtained. An error in identifying a pair of overlapping sequences could lead to the genome sequence becoming scrambled, or parts being missed out entirely. The probability of making mistakes increases with larger genome sizes, so the shotgun approach has been used mainly with the smaller bacterial genomes.

The Haemophilus influenzae *genome sequencing project*

The shotgun approach was first used successfully with the bacterium *Haemophilus. influenzae*, which was the first free-living organism whose genome was entirely

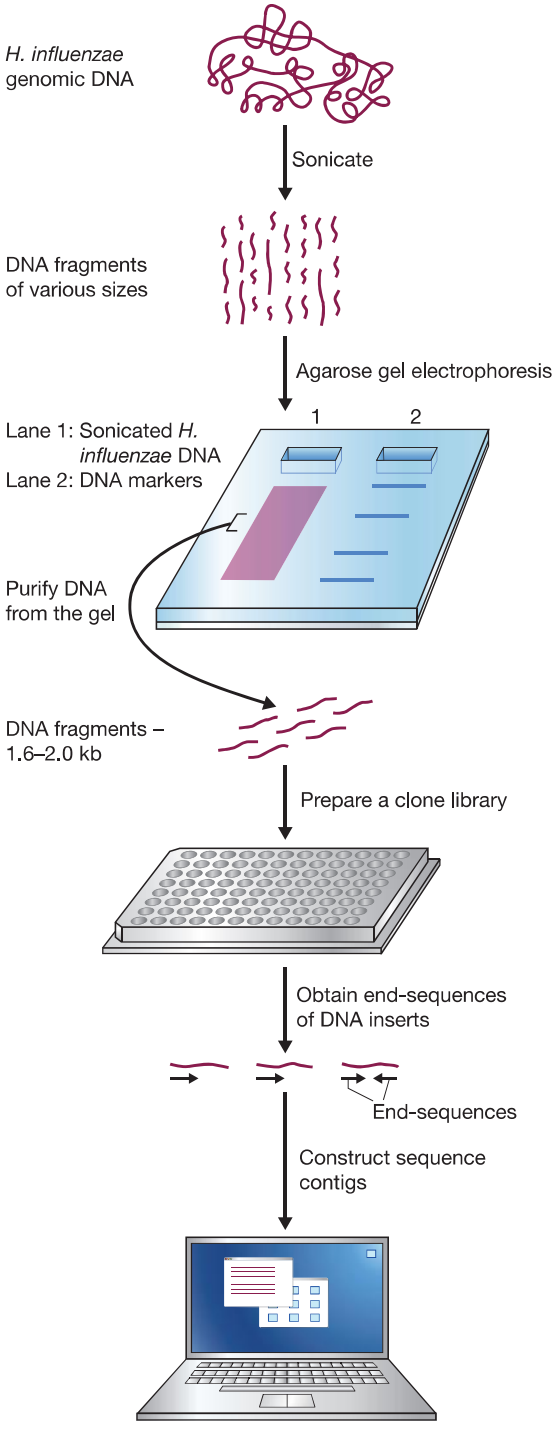


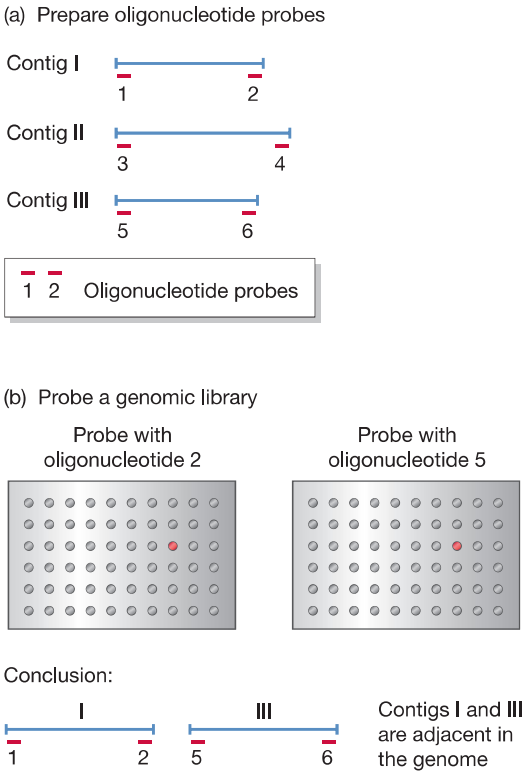
Figure 10.10
A schematic of the key steps in the *H. influenzae* genome sequencing project.

sequenced, the results being published in 1995. The first step was to break the 1830 kb genome of the bacterium into short fragments, which would provide the templates for the sequencing experiments (Figure 10.10). A restriction endonuclease could have been used but **sonication** was chosen because this technique cleaves DNA in a more random fashion and hence reduces the possibility of gaps appearing in the genome sequence.

It was decided to concentrate on fragments of 1.6–2.0 kb because these could yield two DNA sequences, one from each end, reducing the amount of cloning and DNA

Figure 10.11

Using oligonucleotide hybridization to close gaps in the *H. influenzae* genome sequence. Oligonucleotides 2 and 5 both hybridize to the same λ clone, indicating that contigs I and III are adjacent. The gap between them can be closed by sequencing the appropriate part of the λ clone.



preparation that was required. The sonicated DNA was therefore fractionated by agarose gel electrophoresis and fragments of the desired size purified from the gel. After cloning, 28,643 chain termination sequencing experiments were carried out with 19,687 of the clones. A few of these sequences—4339 in all—were rejected because they were less than 400 bp in length. The remaining 24,304 sequences were entered into a computer, which spent 30 hours analyzing the data. The result was 140 contiguous sequences, each a different segment of the *H. influenzae* genome.

It might have been possible to continue sequencing more of the sonicated fragments in order eventually to close the gaps between the individual segments. However, 11,631,485 bp of sequence had already been generated—six times the length of the genome—suggesting that a large amount of additional work would be needed before the correct fragments were, by chance, sequenced. At this stage of the project the most time-effective approach was to use a more directed strategy in order to close each of the gaps individually. Several approaches were used for gap closure, the most successful of these involving hybridization analysis of a clone library prepared in a λ vector (Figure 10.11). The library was probed in turn with a series of oligonucleotides whose sequences corresponded with the ends of each of the 140 segments. In some cases, two oligonucleotides hybridized to the same λ clone, indicating that the two segment ends represented by those oligonucleotides lay adjacent to one another in the genome. The gap between these two ends could then be closed by sequencing the appropriate part of the λ clone.

Problems with shotgun sequencing

Shotgun sequencing has been successful with many bacterial genomes. Not only are these genomes small, so the computational requirements for finding sequence overlaps are not too great, but they contain little or no repetitive DNA sequences. These are

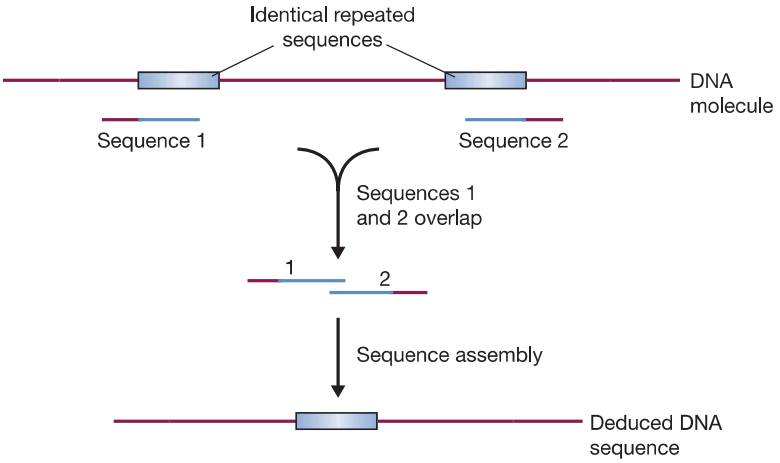


Figure 10.12

One problem with the shotgun approach. An incorrect overlap is made between two sequences that both terminate within a repeated element. The result is that a segment of the DNA molecule is absent from the DNA sequence.

sequences, from a few base pairs to several kilobases, which are repeated at two or more places in a genome. They cause problems for the shotgun approach because when sequences are assembled those that lie partly or wholly within one repeat element might accidentally be assigned an overlap with the identical sequence present in a different repeat element (Figure 10.12). This could lead to a part of the genome sequence being placed at the incorrect position or left out entirely. For this reason, it has generally been thought that shotgun sequencing is inappropriate for eukaryotic genomes, as these have many repeat elements. Later in this chapter (p. 183) we will see how this limitation can be circumvented by using a genome map to direct assembly of sequences obtained by the shotgun approach.

10.2.2 The clone contig approach

The clone contig approach does not suffer from the limitations of shotgun sequencing and so can provide an accurate sequence of a large genome that contains repetitive DNA. Its drawback is that it involves much more work and so takes longer and costs more money. The additional time and effort is needed to construct the overlapping series of cloned DNA fragments. Once this has been done, each cloned fragment is sequenced by the shotgun method and the genome sequence built up step by step (see Figure 10.9).

The cloned fragments should be as long as possible in order to minimize the total number needed to cover the entire genome. A high capacity vector is therefore used. The first eukaryotic chromosome to be sequenced—chromosome III of *Saccharomyces cerevisiae* – was initially cloned in a cosmid vector (p. 101) with the resulting contig comprising 29 cloned fragments. Chromosome III is relatively short, however, and the average size of the cloned fragments was just 10.8 kb. Sequencing of the much longer human genome required 300,000 bacterial artificial chromosome (BAC) clones (p. 103). Assembling all of these into chromosome-specific contigs was a massive task.

Clone contig assembly by chromosome walking

One technique that can be used to assemble a clone contig is **chromosome walking**. To begin a chromosome walk a clone is selected at random from the library, labeled, and

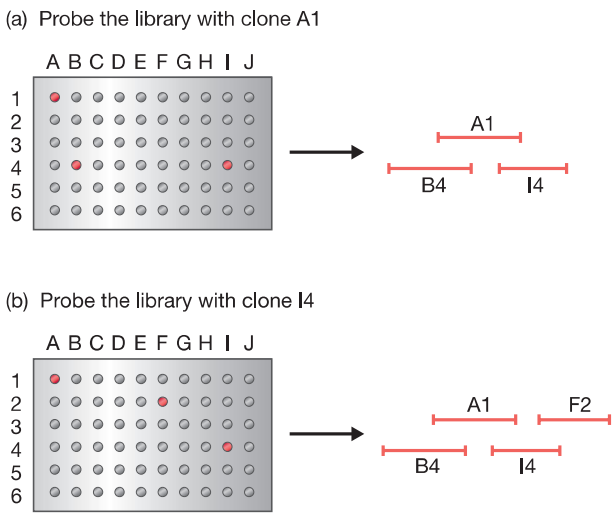


Figure 10.13
Chromosome walking.

used as a hybridization probe against all the other clones in the library (Figure 10.13a). Those clones that give hybridization signals are ones that overlap with the probe. One of these overlapping clones is now labeled and a second round of probing carried out. More hybridization signals are seen, some of these indicating additional overlaps (Figure 10.13b). Gradually the clone contig is built up in a step-by-step fashion. But this is a laborious process and is only attempted when the contig is for a short chromosome and so involves relatively few clones, or when the aim is to close one or more small gaps between contigs that have been built up by more rapid methods.

Rapid methods for clone contig assembly

The weakness of chromosome walking is that it begins at a fixed starting point and builds up the clone contig step by step, and hence slowly, from that fixed point. The more rapid techniques for clone contig assembly do not use a fixed starting point and instead aim to identify pairs of overlapping clones: when enough overlapping pairs have been identified the contig is revealed (Figure 10.14). The various techniques that can be used to identify overlaps are collectively known as **clone fingerprinting**.

Clone fingerprinting is based on the identification of sequence features that are shared by a pair of clones. The simplest approach is to digest each clone with one or more restriction endonucleases and to look for pairs of clones that share restriction fragments of the same size, excluding those fragments that derive from the vector rather than

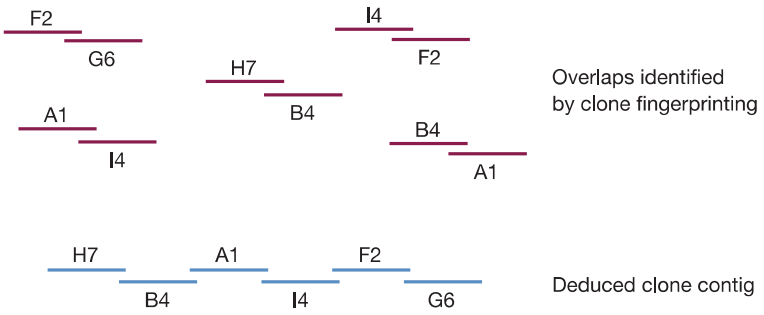


Figure 10.14
Building up a clone contig by a clone fingerprinting technique.

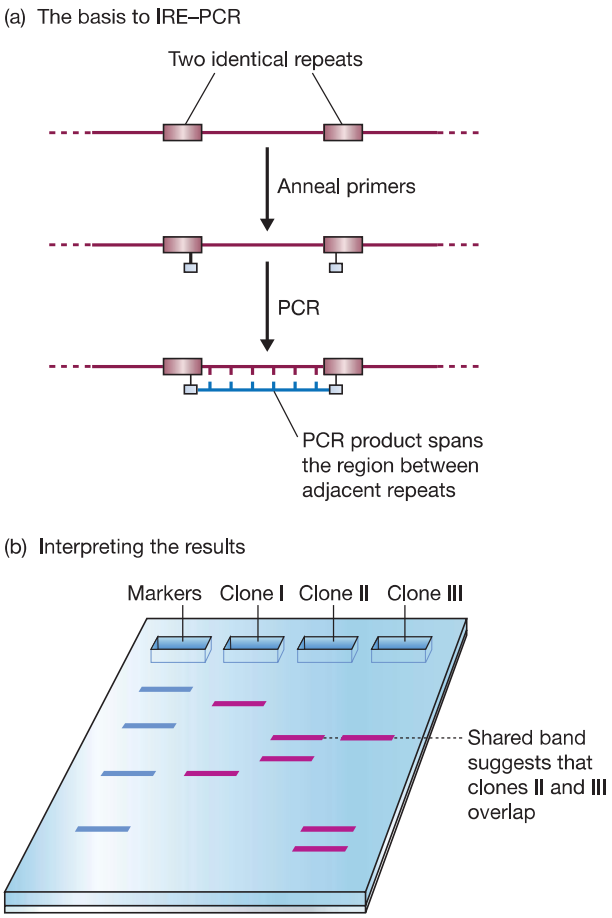


Figure 10.15
Interspersed repeat element PCR (IRE–PCR).

the inserted DNA. This technique might appear to be easy to carry out, but in practice it takes a great deal of time to scan the resulting agarose gels for shared fragments. There is also a relatively high possibility that two clones that do not overlap will, by chance, share restriction fragments whose sizes are indistinguishable by agarose gel electrophoresis.

More accurate results can be obtained by repetitive DNA PCR, also known as interspersed repeat element PCR (IRE–PCR). This type of PCR uses primers that are designed to anneal within repetitive DNA sequences and direct amplification of the DNA between adjacent repeats (Figure 10.15). Repeats of a particular type are distributed fairly randomly in a eukaryotic genome, with varying distances between them, so a variety of product sizes are obtained when these primers are used with clones of eukaryotic DNA. If a pair of clones gives PCR products of the same size, they must contain repeats that are identically spaced, possibly because the cloned DNA fragments overlap.

Clone contig assembly by sequence tagged site content analysis

A third way to assemble a clone contig is to search for pairs of clones that contain a specific DNA sequence that occurs at just one position in the genome under study. If two clones contain this feature, then clearly they must overlap (Figure 10.16). A sequence of this type is called a sequence tagged site (STS). Often an STS is a gene that has been sequenced in an earlier project. As the sequence is known, a pair of PCR primers can be designed that are specific for that gene and then used to identify which members of a clone library contain the gene. The STS does not have to be a gene and can be any

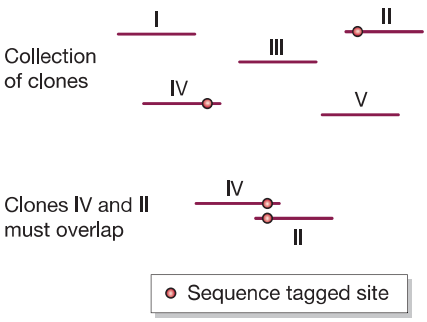


Figure 10.16
The basis to STS content mapping.

short piece of DNA sequence, the only requirement being that it occurs just once in the genome.

10.2.3 Using a map to aid sequence assembly

Sequence tagged site content mapping is a particularly important method for clone contig assembly, because often the positions of STSs within the genome will have been determined by genetic mapping or physical mapping. This means that the STS positions can be used to anchor the clone contig onto a genome map, enabling the position of the contig within a chromosome to be determined. We will now look at how these maps are obtained.

Genetic maps

A genetic map is one that is obtained by genetic studies using Mendelian principles and involving directed breeding programmes for experimental organisms or pedigree analysis for humans. In many cases the loci that are studied are genes, whose inheritance patterns are followed by monitoring the phenotypes of the offspring produced after a cross between parents with contrasting characteristics (e.g., tall and short for the pea plants studied by Mendel). The inheritance patterns reveal the extent of genetic linkage between genes present on the same chromosome, enabling the relative positions of those genes to be deduced and a genetic map to be built up.

More recently, techniques have been devised for genetic mapping of DNA sequences that are not genes but which still display variability in the human population. The most important of these DNA markers are:

- Single nucleotide polymorphisms (SNPs), which are positions in a genome where either of two different nucleotides can occur (Figure 10.17). Some members of the species have one version of the SNP and some have the other version. SNPs are usually typed with short oligonucleotide probes that hybridize to the alternative forms and hence distinguish which is present.
- Restriction fragment length polymorphisms (RFLPs) are special types of SNPs, ones which result in a restriction site being changed. When digested with a restriction endonuclease the loss of the site is revealed because two fragments remain joined together. Originally, RFLPs were typed by Southern hybridization

Figure 10.17
Two versions of an SNP.

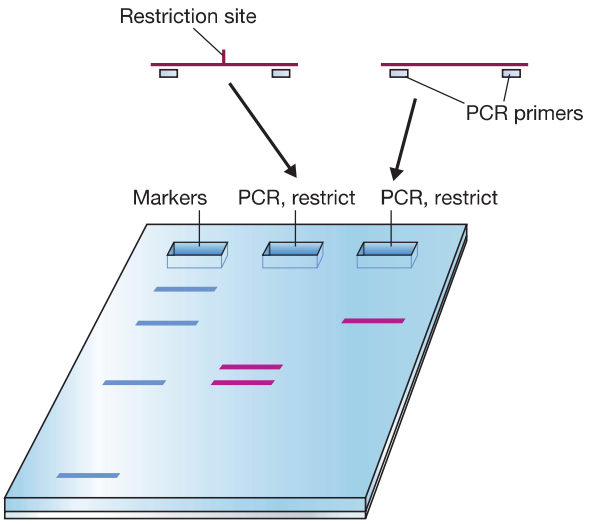


Figure 10.18
Typing a restriction site polymorphism by PCR. In the middle lane the PCR product gives two bands because it is cut by treatment with the restriction enzyme. In the right-hand lane there is just one band because the template DNA lacks the restriction site.

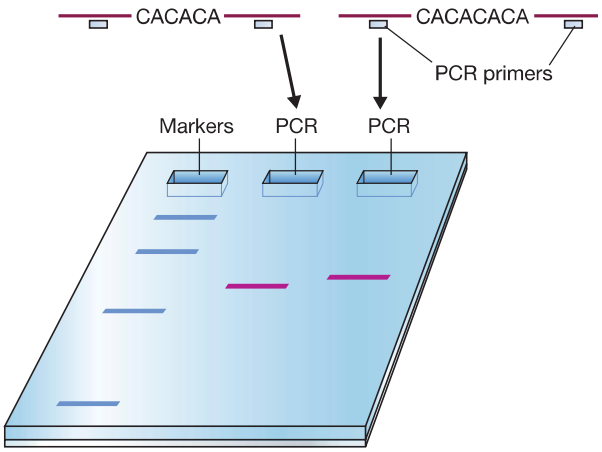


Figure 10.19
Typing an STR by PCR. The PCR product in the right-hand lane is slightly longer than that in the middle lane, because the template DNA from which it is generated contains an additional CA unit.

of restricted genomic DNA, but this is a time-consuming process, so nowadays the presence or absence of the restriction site is usually determined by PCR (Figure 10.18).

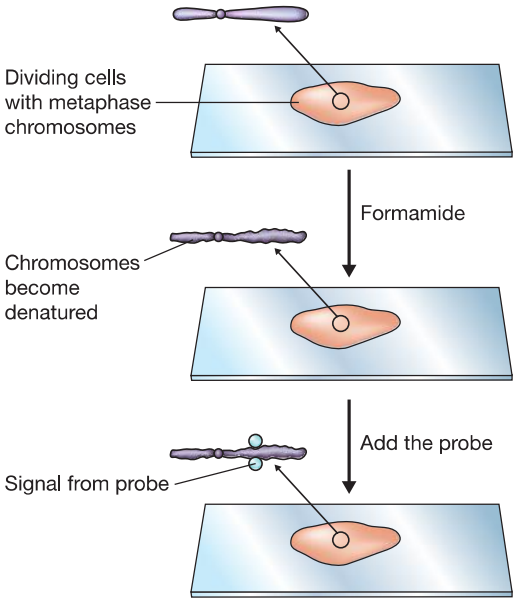
- **Short tandem repeats (STRs)**, also called **microsatellites**, are made up of short repetitive sequences of 1–13 nucleotides in length, linked head to tail. The number of repeats present in a particular STR varies, usually between 5 and 20. The number can be determined by carrying out a PCR using primers that anneal either side of the STR, and then examining the size of the resulting product by agarose or polyacrylamide gel electrophoresis (Figure 10.19).

All of these DNA markers are variable and so exist in two or more allelic forms. Their inheritance patterns can be determined by analysis of DNA prepared from the parents and offspring from a genetic cross, and the data used to place the DNA markers on a genetic map, in exactly the same way as genes are mapped.

Physical maps

A physical map is generated by methods that directly locate the positions of specific sequences on a chromosomal DNA molecule. As in genetic mapping, the loci that are studied can be genes or DNA markers. The latter might include **expressed sequence tags**

Figure 10.20
Fluorescent *in situ* hybridization.



(ESTs), which are short sequences obtained from the ends of complementary DNAs (cDNAs) (p. 133). Expressed sequence tags are therefore partial gene sequences, and when used in map construction they provide a quick way of locating the positions of genes, even though the identity of the gene might not be apparent from the EST sequence.

Two types of technique are used in physical mapping:

- Direct examination of chromosomal DNA molecules, for example by fluorescence *in situ* hybridization (FISH). In this technique, a cloned DNA fragment is labeled with a fluorescent marker and then hybridized to a preparation of chromosomes immobilized on a glass slide. The physical position of the DNA fragment within the chromosome is then revealed simply by examining the preparation with a microscope (Figure 10.20). If FISH is carried out simultaneously with two DNA probes, each labeled with a different fluorochrome, the relevant positions on the chromosome of the two markers represented by the probes can be visualized. Special techniques for working with extended chromosomes, whose DNA molecules are stretched out rather than tightly coiled as in normal chromosomes, enable markers to be positioned with a high degree of accuracy.
- Physical mapping with a mapping reagent, which is a collection of overlapping DNA fragments spanning the chromosome or genome that is being studied. Pairs of markers that lie within a single fragment must be located close to each other on the chromosome: how close can be determined by measuring the frequency with which the pair occurs together in different fragments in the mapping reagent (Figure 10.21). The mapping reagent could be a clone library, possibly one that is also being assembled into a contig prior to DNA sequencing. Radiation hybrids are a second type of mapping reagent and were particularly important in the Human Genome Project. These are hamster cell lines that contain fragments of human chromosomes, prepared by a treatment involving irradiation (hence their name). Mapping is carried out by hybridization of marker probes to a panel of cell lines, each one containing a different part of the human genome.

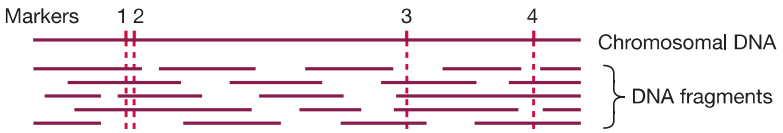


Figure 10.21

The principle behind the use of a mapping reagent. It can be deduced that markers 1 and 2 are relatively close because they are present together on four DNA fragments. In contrast, markers 3 and 4 must be relatively far apart because they occur together on just one fragment.

The importance of a map in sequence assembly

It is possible to obtain a genome sequence without the use of a genetic or physical map. This is illustrated by the *H. influenzae* project that we followed on p. 174, and many other bacterial genomes have been sequenced without the aid of a map. But a map is very important when a larger genome is being sequenced because it provides a guide that can be used to check that the genome sequence is being assembled correctly from the many short sequences that emerge from the automated sequencer. If a marker that has been mapped by genetic and/or physical means appears in the genome sequence at an unexpected position, then an error in sequence assembly is suspected.

Detailed genetic and/or physical maps have been important in the Human Genome Project, as well as those for yeast, fruit fly, *C. elegans*, and *A. thaliana*, all of which were based on the clone contig approach. Maps are also being used to direct sequence assembly in projects that use the shotgun approach. As described on p. 176, the major problem when applying shotgun sequencing to a large genome is the presence of repeated sequences and the possibility that the assembled sequence "jumps" between two repeats, so part of the genome is misplaced or left out (see Figure 10.12). These errors can be avoided if sequence assembly makes constant reference to a genome map. Because it avoids the laborious construction of clone contigs, this directed shotgun approach is becoming the method of choice for sequencing large genomes.

Further reading

FURTHER READING

Adams, M.D., Celnicker, S.E., Holt, R.A. et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, 287, 2185–2195. [A clear description of this genome project.]

Brown, T.A. (2007) *Genomes*, 3rd edn. Garland Science, Abingdon. [Gives details of techniques for studying genomes, including genetic and physical mapping.]

Fleischmann, R.D., Adams, M.D., White, O. et al. (1995) Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269, 496–512. [The first complete bacterial genome sequence to be published.]

Heiskanen, M., Peltonen, L. & Palotie, A. (1996) Visual mapping by high resolution FISH. *Trends in Genetics*, 12, 379–382.

Margulies, M., Egholm, M., Altman, W.E. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376–380. [Massively parallel pyrosequencing.]