6 Recombinant DNA and genetic analysis

R. RAPLEY

6.1 INTRODUCTION

The considerable advances made in microarray, sequencing technologies and bioinformatics analysis are now beginning to provide true insights into the development and maintenance of cells and tissues. Indeed areas of analysis such as metabolomics, transcriptomics and systems biology are now well established and allow analysis of vast numbers of samples simultaneously. This type of large-scale parallel analysis is now the main driving force of biological discovery and analysis. However, the techniques of molecular biology and genetic analysis have their foundations in methods developed a number of decades ago. One of the main cornerstones on which molecular biology analysis was developed was the discovery of restriction endonucleases in the early 1970s which not only led to the possibility of analysing DNA more effectively but also provided the ability to cut different DNA molecules so that they could later be joined together to create new recombinant DNA fragments. The newly created DNA molecules heralded a new era in the manipulation, analysis and exploitation of biological molecules. This process, termed **gene cloning**, has enabled numerous discoveries and insights into gene structure, function and regulation. Since their

initial use the methods for the production of gene libraries have been steadily refined and developed. Although microarray analysis and the polymerase chain reaction (PCR) have provided short cuts to gene analysis there are still many cases where gene cloning methods are not only useful but are an absolute requirement. The following provides an account of the process of gene cloning and other methods based on recombinant DNA technology.

6.2 CONSTRUCTING GENE LIBRARIES

6.2.1 Digesting genomic DNA molecules

Following the isolation and purification of genomic DNA it is possible to specifically fragment it with enzymes termed **restriction endonucleases**. These enzymes are the key to molecular cloning because of the specificity they have for particular DNA sequences. It is important to note that every copy of a given DNA molecule from a specific organism will give the same set of fragments when digested with a particular enzyme. DNA from different organisms will, in general, give different sets of fragments when treated with the same enzyme. By digesting complex genomic DNA from an organism it is possible to reproducibly divide its genome into a large number of small fragments, each approximately the size of a single gene. Some enzymes cut straight across the DNA to give flush or blunt ends. Other restriction enzymes make staggered single-strand cuts, producing short single-stranded projections at each end of the digested DNA. These ends are not only identical, but complementary, and will base-pair with each other; they are therefore known as cohesive or sticky ends. In addition the 5′ end projection of the DNA always retains the phosphate groups.

Over 600 enzymes, recognising more than 200 different restriction sites, have been characterised. The choice of which enzyme to use depends on a number of factors. For example, the recognition sequence of 6 bp will occur, on average, every 4096 (46) bases assuming a random sequence of each of the four bases. This means that digesting genomic DNA with *Eco*R1, which recognises the sequence 5′-GAATTC-3′, will produce fragments each of which is on average just over 4 kb. Enzymes with 8 bp recognition sequences produce much longer fragments. Therefore very large genomes, such as human DNA, are usually digested with enzymes that produce long DNA fragments. This makes subsequent steps more manageable, since a smaller number of those fragments need to be cloned and subsequently analysed (Table 6.1).

6.2.2 Ligating DNA molecules

The DNA products resulting from restriction digestion to form sticky ends may be joined to any other DNA fragments treated with the same restriction enzyme. Thus, when the two sets of fragments are mixed, base-pairing between sticky ends will result in the annealing together of fragments that were derived from different starting DNA. There will, of course, also be pairing of fragments derived from the same starting DNA

Table 6.1 **Numbers of clones required for representation of DNA in a genome library**

Species	Genome size (kb)	No. of clones required	
		17 kb fragments	35 kb fragments
Bacteria (*E. coli*)	4 000	700	340
Yeast	20 000	3 500	1 700
Fruit fly	165 000	29 000	14 500
Man	3 000 000	535 000	258 250
Maize	15 000 000	2 700 000	1 350 000

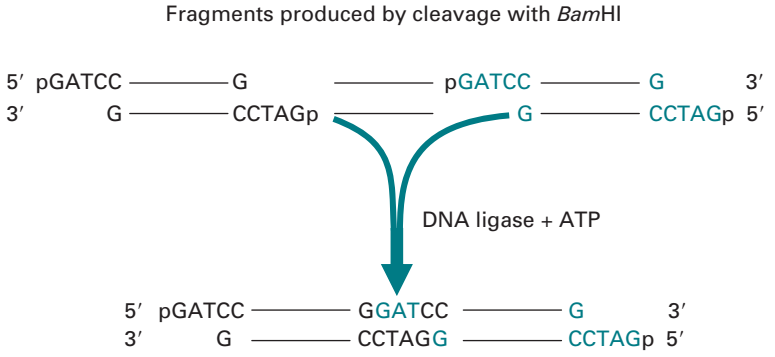

Fig. 6.1 Ligation molecules with cohesive ends. Complementary cohesive ends base-pair, forming a temporary link between two DNA fragments. This association of fragments is stabilised by the formation of 3′ to 5′ phosphodiester linkages between cohesive ends, a reaction catalysed by DNA ligase.

molecules, termed **reannealing**. All these pairings are transient, owing to the weakness of hydrogen bonding between the few bases in the sticky ends, but they can be stabilised by use of an enzyme termed **DNA ligase** in a process termed **ligation**. This enzyme, usually isolated from bacteriophage T4 and termed **T4 DNA ligase**, forms a covalent bond between the 5′ phosphate at the end of one strand and the 3′ hydroxyl of the adjacent strand (Fig. 6.1). The reaction, which is ATP dependent, is often carried out at 10 °C to lower the kinetic energy of molecules, and so reduce the chances of base-paired sticky ends parting before they have been stabilised by ligation. However, long reaction times are needed to compensate for the low activity of DNA ligase in the cold. It is also possible to join blunt ends of DNA molecules, although the efficiency of this reaction is much lower than sticky-ended ligations.

Since ligation reconstructs the site of cleavage, recombinant molecules produced by ligation of sticky ends can be cleaved again at the 'joins', using the same restriction enzyme that was used to generate the fragments initially. In order to propagate digested DNA from an organism it is necessary to join or ligate that DNA with

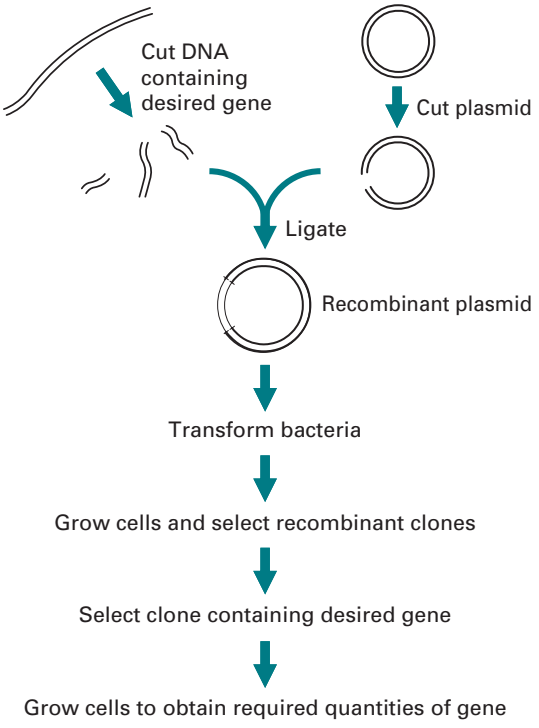

Fig. 6.2 Outline of gene cloning.

a specialised DNA carrier molecule termed a **vector** (Section 6.3). Thus each DNA fragment is inserted by ligation into the vector DNA molecule, which allows the whole recombined DNA to then be replicated indefinitely within microbial cells (Fig. 6.2). In this way a DNA fragment can be cloned to provide sufficient material for further detailed analysis, or for further manipulation. Thus, all of the DNA extracted from an organism and digested with a restriction enzyme will result in a collection of clones. This collection of clones is known as a gene library.

6.2.3 Aspects of gene libraries

There are two general types of gene library: a **genomic library** which consists of the total chromosomal DNA of an organism and a **cDNA library** which represents only the mRNA from a particular cell or tissue at a specific point in time (Fig. 6.3). The choice of the particular type of gene library depends on a number of factors, the most important being the final application of any DNA fragment derived from the library. If the ultimate aim is understanding the control of protein production for a particular gene or the analysis of its architecture, then genomic libraries must be used. However, if the goal is the production of new or modified proteins, or the determination of the tissue-specific expression and timing patterns, cDNA libraries are more appropriate. The main consideration in the construction of genomic or cDNA libraries is therefore

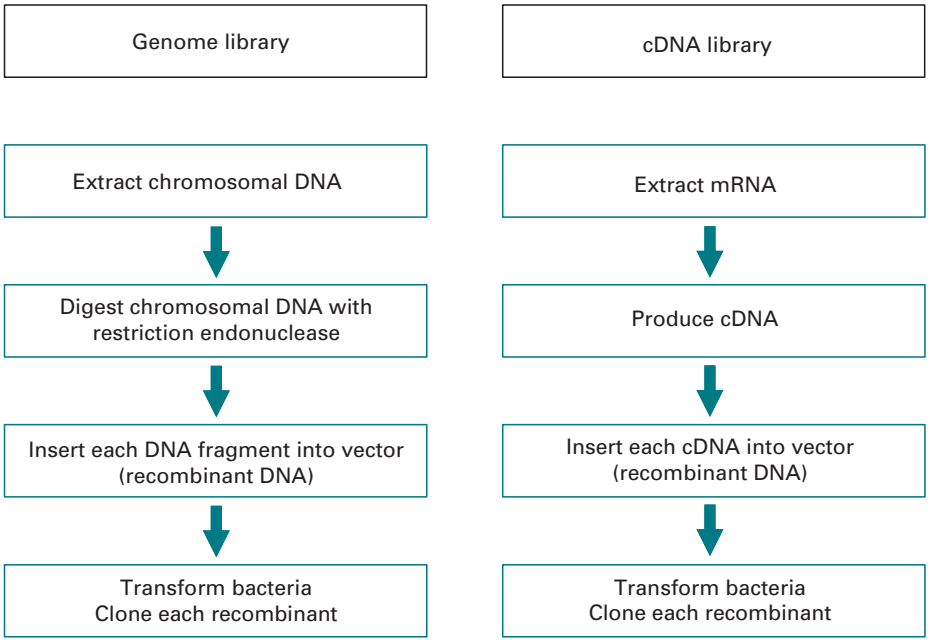

Fig. 6.3 Comparison of the general steps involved in the construction of genomic and complementary DNA (cDNA) libraries.

the nucleic acid starting material. Since the genome of an organism is fixed, chromosomal DNA may be isolated from almost any cell type in order to prepare genomic libraries. In contrast, however, cDNA libraries only represent the mRNA being produced from a specific cell type at a particular time. Thus, it is important to consider carefully the cell or tissue type from which the mRNA is to be derived in the construction of cDNA libraries.

There are a variety of cloning vectors available, many based on naturally occurring molecules such as bacterial plasmids or bacteria-infecting viruses. The choice of vector depends on whether a genomic library or cDNA library is constructed. The various types of vectors are explained in more detail in Section 6.3.

6.2.4 Genomic DNA libraries

Genomic libraries are constructed by isolating the complete chromosomal DNA from a cell, then digesting it into fragments of the desired average length with restriction endonucleases. This can be achieved by partial restriction digestion using an enzyme that recognises tetranucleotide sequences. Complete digestion with such an enzyme would produce a large number of very short fragments, but if the enzyme is allowed to cleave only a few of its potential restriction sites before the reaction is stopped, each DNA molecule will be cut into relatively large fragments. Average fragment size will depend on the relative concentrations of DNA and restriction enzyme, and in particular, on the conditions and duration of incubation (Fig. 6.4). It is also possible to produce fragments of DNA by physical shearing although the ends of the fragments

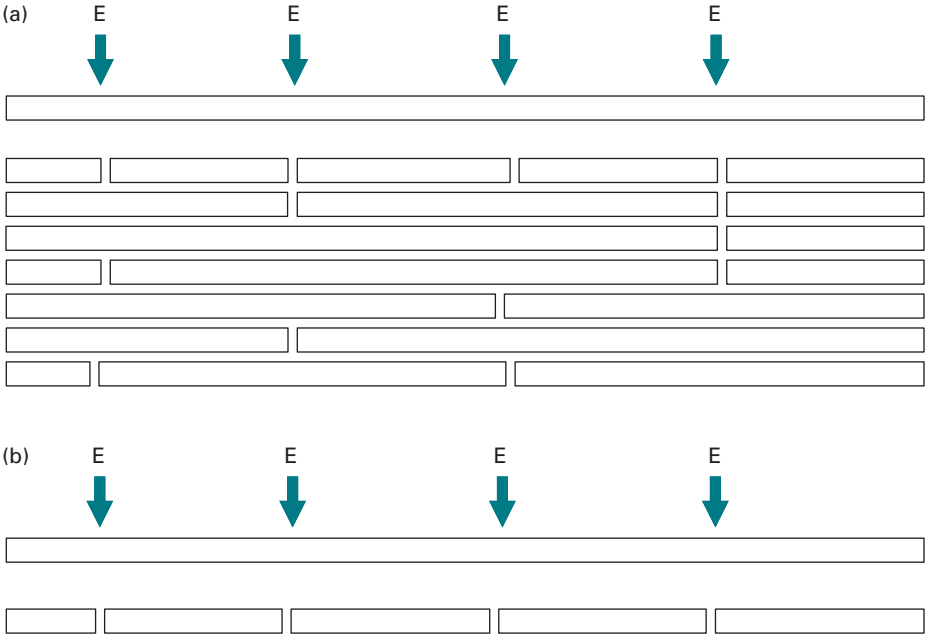

Fig. 6.4 Comparison of (a) partial and (b) complete digestion of DNA molecules at restriction enzyme sites (E).

may need to be repaired to make them flush-ended. This can be achieved by using a modified DNA polymerase termed Klenow polymerase. This is prepared by cleavage of DNA polymerase with subtilisin, giving a large enzyme fragment which has no 5′ to 3′ exonuclease activity, but which still acts as a 5′ to 3′ polymerase. This will fill in any recessed 3′ ends on the sheared DNA using the appropriate dNTPs.

The mixture of DNA fragments is then ligated with a vector, and subsequently cloned. If enough clones are produced there will be a very high chance that any particular DNA fragment such as a gene will be present in at least one of the clones. To keep the number of clones to a manageable size, fragments about 10 kb in length are needed for prokaryotic libraries, but the length must be increased to about 40 kb for mammalian libraries. It is possible to calculate the number of clones that must be present in a gene library to give a probability of obtaining a particular DNA sequence. This formula is:

$$N = \frac{\ln(1-P)}{\ln(1-f)}$$

where N is the number of recombinants, P is the probability and f is the fraction of the genome in one insert. Thus for the *E. coli* DNA chromosome of 5×10^6 bp and with an insert size of 20 kb the number of clones needed (N) would be 1×10^3 with a probability of 0.99.

6.2.5 cDNA libraries

There may be several thousand different proteins being produced in a cell at any one time, all of which have associated mRNA molecules. To identify any one of those

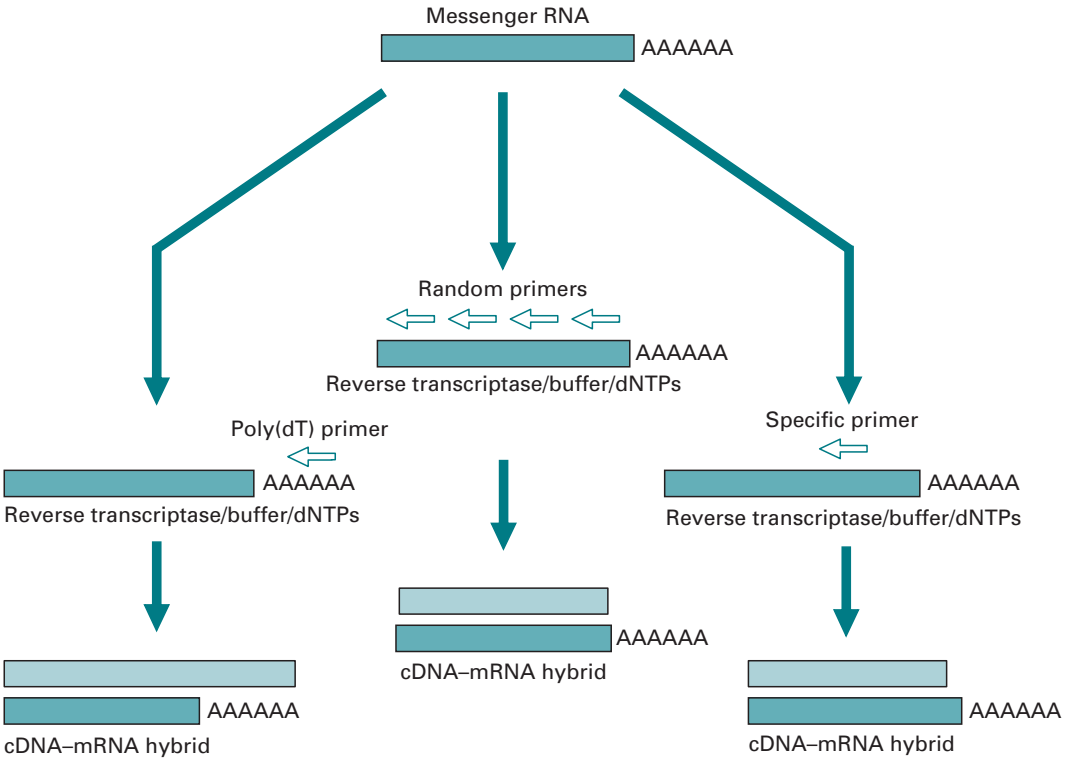

Fig. 6.5 Strategies for producing first-strand cDNA from mRNA.

mRNA molecules the clones of each individual mRNA have to be synthesised. Libraries that represent the mRNA in a particular cell or tissue are termed cDNA libraries. mRNA cannot be used directly in cloning since it is too unstable. However it is possible to synthesise complementary DNA molecules (cDNAs) to all the mRNAs from the selected tissue. The cDNA may be inserted into vectors and then cloned. The production of cDNA (complementary DNA) is carried out using an enzyme termed **reverse transcriptase** which is isolated from RNA-containing retroviruses.

Reverse transcriptase is an RNA-dependent DNA polymerase, and will synthesise a first-strand DNA complementary to an mRNA template, using a mixture of the four dNTPs. There is also a requirement (as with all polymerase enzymes) for a short oligonucleotide primer to be present (Fig. 6.5). With eukaryotic mRNA bearing a poly(A) tail, a complementary oligo(dT) primer may be used. Alternatively random hexamers may be used which randomly anneal to the mRNAs in the complex. Such primers provide a free 3′ hydroxyl group which is used as the starting point for the reverse transcriptase. Regardless of the method used to prepare the first-strand cDNA one absolute requirement is high-quality undegraded mRNA (Section 5.7.2). It is usual to check the integrity of the RNA by gel electrophoresis (Section 5.7.4). Alternatively a fraction of the extract may be used in a cell-free translation system, which, if intact mRNA is present, will direct the synthesis of proteins represented by the mRNA molecules in the sample (Section 6.7).

Following the synthesis of the first DNA strand, a poly(dC) tail is added to its 3′ end, using terminal transferase and dCTP. This will also, incidentally, put a poly(dC) tail on

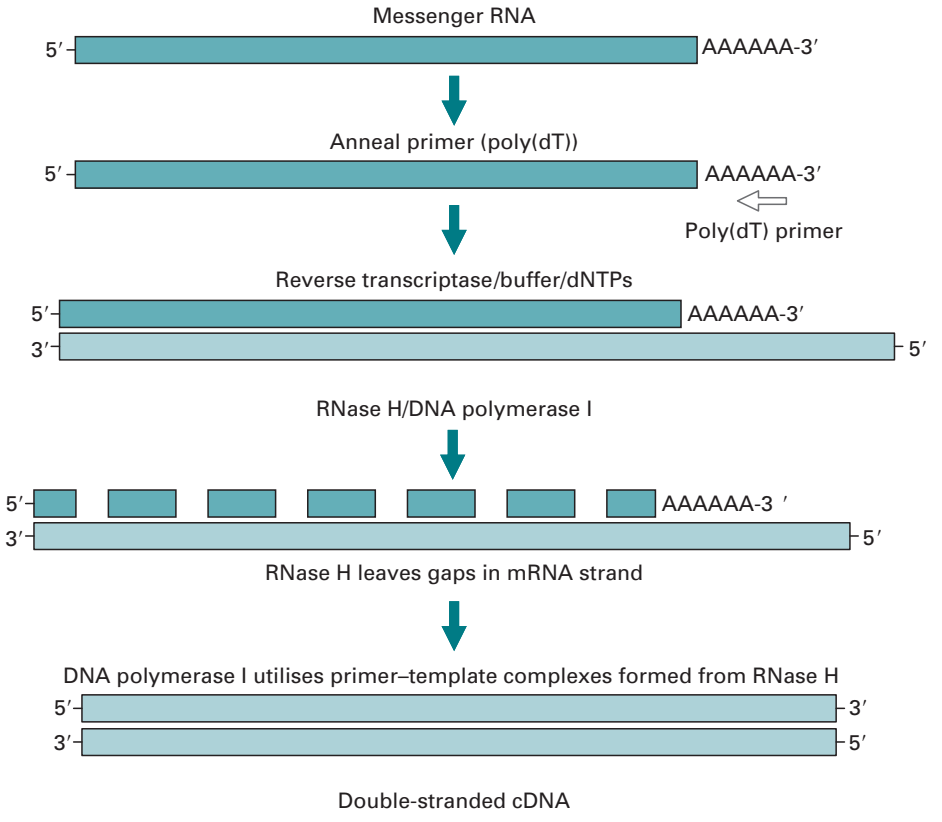

Fig. 6.6 Second-strand cDNA synthesis using the RNase H method.

the poly(A) of mRNA. Alkaline hydrolysis is then used to remove the RNA strand, leaving single-stranded DNA which can be used, like the mRNA, to direct the synthesis of a complementary DNA strand. The second-strand synthesis requires an oligo(dG) primer, base-paired with the poly(dC) tail, which is catalysed by the Klenow fragment of DNA polymerase I. The final product is double-stranded DNA, one of the strands being complementary to the mRNA. One further method of cDNA synthesis involves the use of RNase H. Here the first-strand cDNA is carried out as above with reverse transcriptase but the resulting mRNA–cDNA hybrid is retained. RNase H is then used at low concentrations to nick the RNA strand. The resulting nicks expose 3′ hydroxyl groups which are used by DNA polymerase as a primer to replace the RNA with a second strand of cDNA (Fig. 6.6).

6.2.6 Treatment of blunt cDNA ends

Ligation of blunt-ended DNA fragments is not as efficient as ligation of sticky ends, therefore with cDNA molecules additional procedures are undertaken before ligation with cloning vectors. One approach is to add small double-stranded molecules with one internal site for a restriction endonuclease, termed **nucleic acid linkers**, to the cDNA. Numerous linkers are commercially available with internal restriction sites for

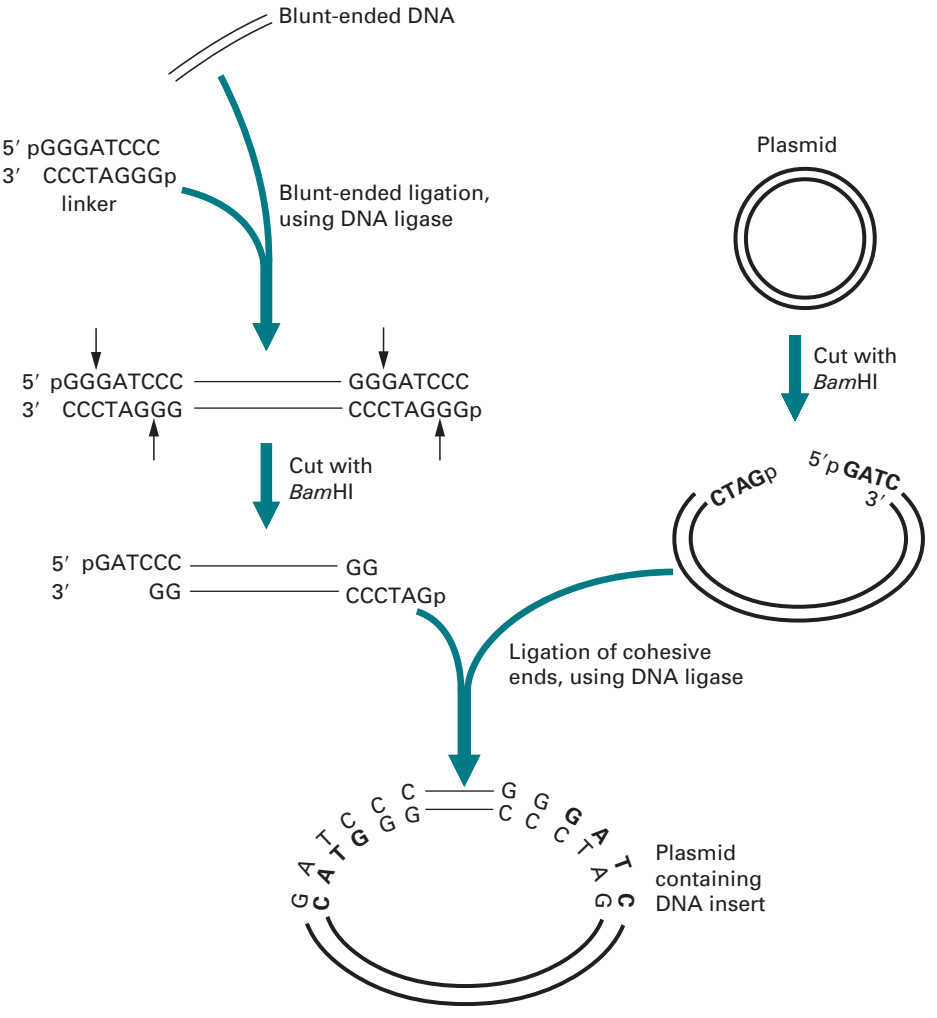

Fig. 6.7 Use of linkers. In this example, blunt-ended DNA is inserted into a specific restriction site on a plasmid, after ligation to a linker containing the same restriction site.

many of the most commonly used restriction enzymes. Linkers are blunt-end ligated to the cDNA but since they are added much in excess of the cDNA the ligation process is reasonably successful. Subsequently the linkers are digested with the appropriate restriction enzyme which provides the sticky ends for efficient ligation to a vector digested with the same enzyme. This process may be made easier by the addition of adaptors rather than linkers which are identical except that the sticky ends are preformed and so there is no need for restriction digestion following ligation (Fig. 6.7).

6.2.7 Enrichment methods for RNA

Frequently an attempt is made to isolate the mRNA transcribed from a desired gene within a particular cell or tissue that produces the protein in high amounts. Thus if the

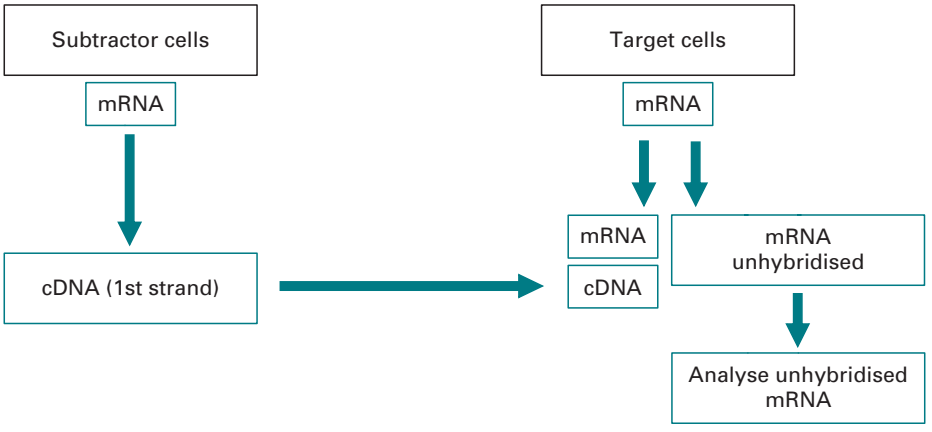

Fig. 6.8 Scheme of analysing specific mRNA molecules by subtractive hybridisation.

cell or tissue produces a major protein of the cell a large fraction of the total mRNA will code for the protein. An example of this are the B cells of the pancreas, which contain high levels of pro-insulin mRNA. In such cases it is possible to precipitate polysomes which are actively translating the mRNA, by using antibodies to the ribosomal proteins; mRNA can then be dissociated from the precipitated ribosomes. More usually the mRNA required is only a minor component of the total cellular mRNA. In such cases total mRNA may be fractionated by size using sucrose density gradient centrifugation. Then each fraction is used to direct the synthesis of proteins using an *in vitro* translation system (Section 6.7).

6.2.8 Subtractive hybridisation

It is often the case that genes are transcribed in a specific cell type or differentially activated during a particular stage of cellular growth, often at very low levels. It is possible to isolate those mRNA transcripts by **subtractive hybridisation**. Usually the the mRNA species common to the different cell types are removed, leaving the cell type or tissue-specific mRNAs for analysis (Fig. 6.8). This may be undertaken by isolating the mRNA from the so-called **subtractor cells** and producing a first-strand cDNA (Section 6.2.5). The original mRNA from the subtractor cells is then degraded and the mRNA from the target cells isolated and mixed with the cDNA. All the complementary mRNA–cDNA molecules common to both cell types will hybridise leaving the unbound mRNA which may be isolated and further analysed. A more rapid approach of analysing the differential expression of genes has been developed using the PCR. This technique, termed **differential display**, is explained in greater detail in Section 6.8.1.

6.2.9 Cloning PCR products

While PCR has to some extent replaced cloning as a method for the generation of large quantities of a desired DNA fragment there is, in certain circumstances, still

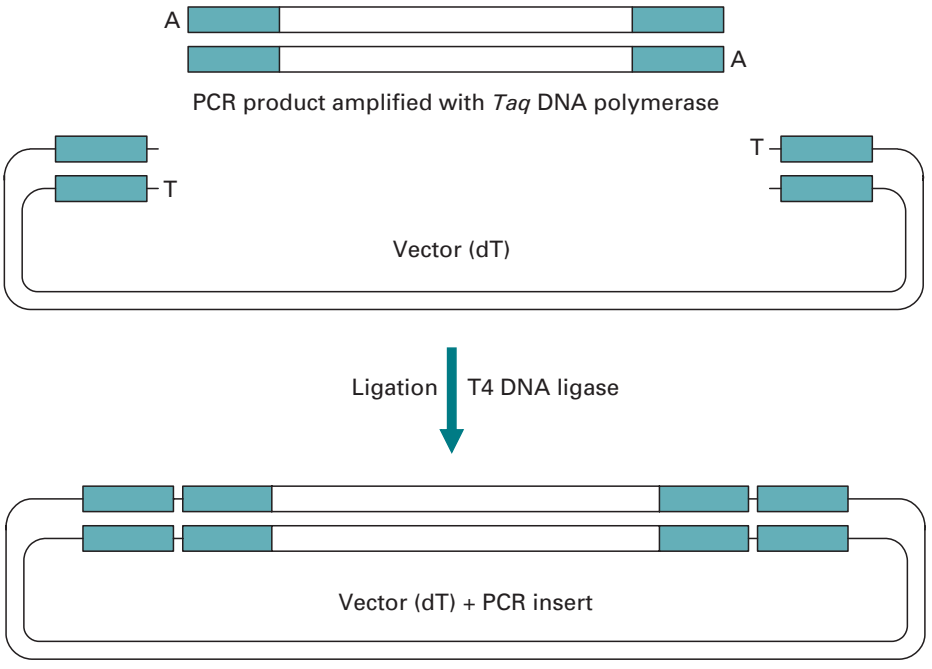

Fig. 6.9 Cloning of PCR products using dA : dT cloning.

a requirement for the cloning of PCR-amplified DNA. For example certain techniques such as *in vitro* protein synthesis are best achieved with the DNA fragment inserted into an appropriate plasmid or phage cloning vector (Section 6.7.1). Cloning methods for PCR follow closely the cloning of DNA fragments derived from the conventional manipulation of DNA. The techniques with which this may be achieved are through one of two ways, blunt-ended or cohesive-ended cloning. Certain thermostable DNA polymerases such as *Taq* DNA polymerase and *Tth* DNA polymerase give rise to PCR products having a 3′ overhanging A residue. It is possible to clone the PCR product into dT vectors termed dA : dT cloning. This makes use of the fact that the terminal additions of A residues may be successfully ligated to vectors prepared with T residue overhangs to allow efficient ligation of the PCR product (Fig. 6.9). The reaction is catalysed by DNA ligase as in conventional ligation reactions (Section 6.2.2).

It is also possible to carry out cohesive ended cloning with PCR products. In this case **oligonucleotide primers** are designed with a restriction endonuclease site incorporated into them. Since the complementarity of the primers needs to be absolute at the 3′ end the 5′ end of the primer is usually the region for the location of the restriction site. This needs to be designed with care since the efficiency of digestion with certain restriction endonuclease decreases if extra nucleotides, not involved in recognition, are absent at the 5′ end. In this case the digestion and ligation reactions are the same as those undertaken for conventional reactions (Section 6.2.1).

6.3 CLONING VECTORS

For the cloning of any molecule of DNA it is necessary for that DNA to be incorporated into a **cloning vector**. These are DNA elements that may be stably maintained and propagated in a host organism for which the vector has replication functions. A typical host organism is a bacterium such as *E. coli* which grows and divides rapidly. Thus any vector with a replication origin in *E. coli* will replicate (together with any incorporated DNA) efficiently. Thus, any DNA cloned into a vector will enable the amplification of the inserted foreign DNA fragment and also allow any subsequent analysis to be undertaken. In this way the cloning process resembles the PCR although there are some major differences between the two techniques. By cloning, it is possible to not only store a copy of any particular fragment of DNA, but also produce unlimited amounts of it (Fig. 6.10).

The vectors used for cloning vary in their complexity, their ease of manipulation, their selection and the amount of DNA sequence they can accommodate (the insert

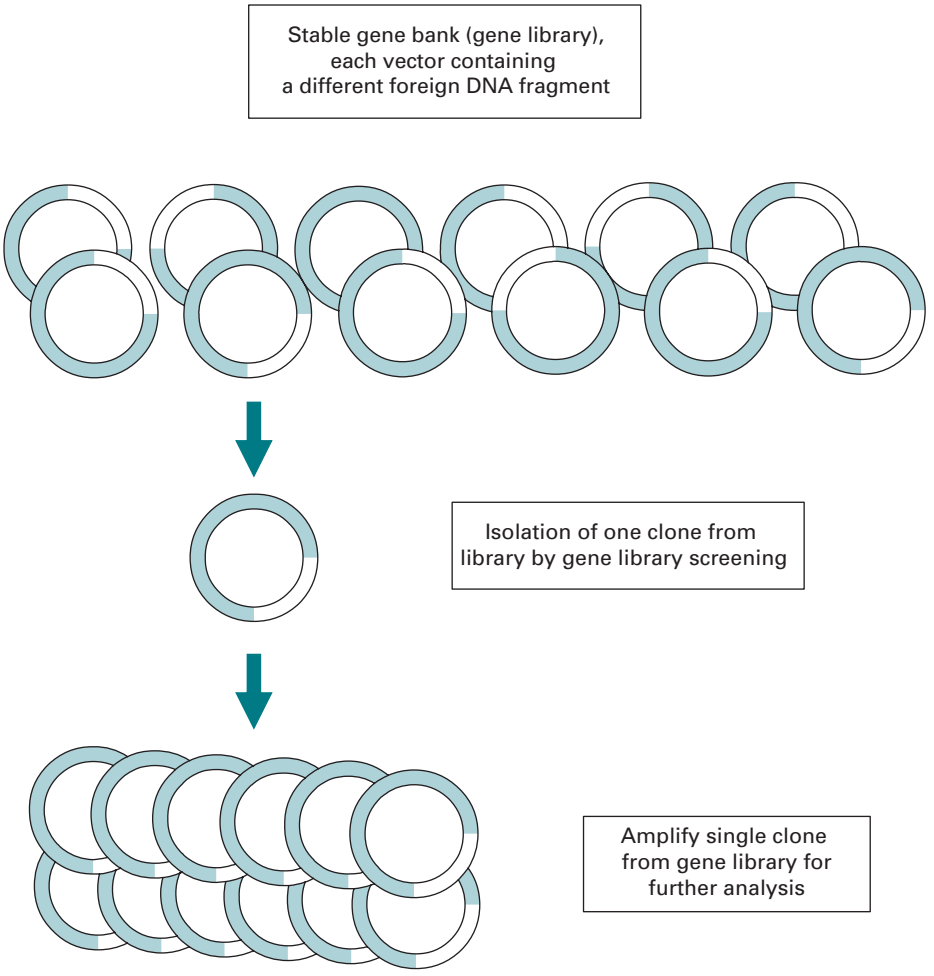

Fig. 6.10 Production of multiple copies of a single clone from a stable gene bank or library.

Table 6.2 **Comparison of vectors generally available for cloning DNA fragments**

Vector	Host cell	Vector structure	Insert range (kb)
M13	*E. coli*	Circular virus	1–4
Plasmid	*E. coli*	Circular plasmid	1–5
Phage λ	*E. coli*	Linear virus	2–25
Cosmids	*E. coli*	Circular plasmid	35–45
BACs	*E. coli*	Circular plasmid	50–300
YACs	*S. cerevisiae*	Linear chromosome	100–2000

Notes: BAC, bacterial artificial chromosome; YAC, yeast artificial chromosome.

capacity). Vectors have in general been developed from naturally occurring molecules such as bacterial plasmids, bacteriophages or combinations of the elements that make them up, such as **cosmids** (Section 6.3.4). For gene library constructions there is a choice and trade-off between various vector types, usually related to ease of the manipulations needed to construct the library and the maximum size of foreign DNA insert of the vector (Table 6.2). Thus, vectors with the advantage of large insert capacities are usually more difficult to manipulate, although there are many more factors to be considered, which are indicated in the following treatment of vector systems.

6.3.1 Plasmids

Many bacteria contain an extrachromosomal element of DNA, termed a **plasmid**, which is a relatively small, covalently closed circular molecule, carrying genes for antibiotic resistance, conjugation or the metabolism of 'unusual' substrates. Some plasmids are replicated at a high rate by bacteria such as *E. coli* and so are excellent potential vectors. In the early 1970s a number of natural plasmids were artificially modified and constructed as cloning vectors, by a complex series of digestion and ligation reactions. One of the most notable plasmids, termed pBR322 after its developers Bolivar and Rodriguez (pBR), was widely adopted and illustrates the desirable features of a cloning vector as indicated below (Fig. 6.11).

- The plasmid is much smaller than a natural plasmid, which makes it more resistant to damage by shearing, and increases the efficiency of uptake by bacteria, a process termed **transformation**.
- A bacterial origin of DNA replication ensures that the plasmid will be replicated by the host cell. Some replication origins display stringent regulation of replication, in which rounds of replication are initiated at the same frequency as cell division. Most plasmids, including pBR322, have a relaxed origin of replication, whose activity is not tightly linked to cell division, and so plasmid replication will be

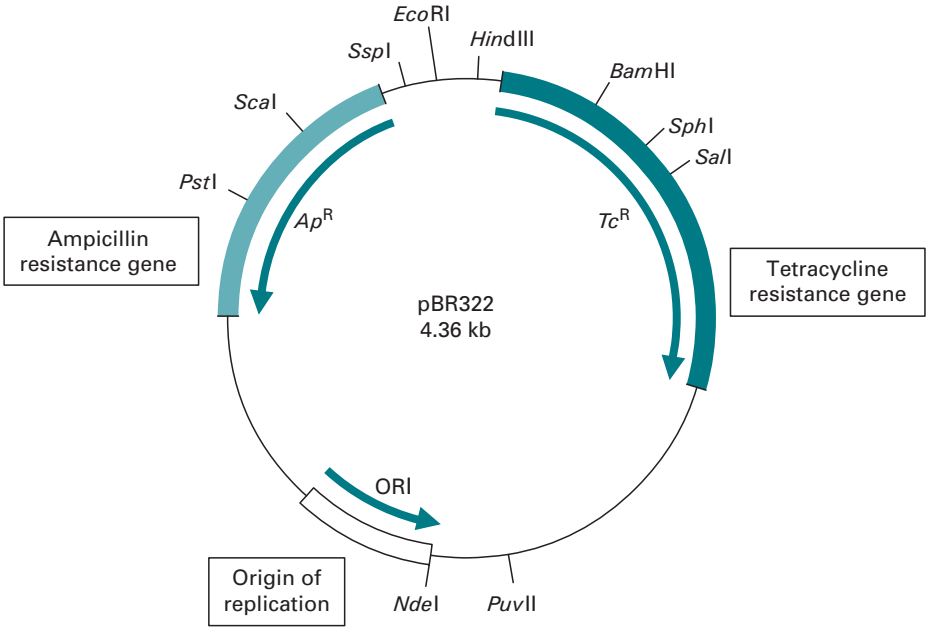

Fig. 6.11 Map and important features of pBR322.

initiated far more frequently than chromosomal replication. Hence a large number of plasmid molecules will be produced per cell.

- Two genes coding for resistance to antibiotics have been introduced. One of these allows the selection of cells which contain plasmid: if cells are plated on medium containing an appropriate antibiotic, only those that contain plasmid will grow to form colonies. The other resistance gene can be used, as described below, for detection of those plasmids that contain inserted DNA.
- There are single recognition sites for a number of restriction enzymes at various points around the plasmid, which can be used to open or linearise the circular plasmid. Linearising a plasmid allows a fragment of DNA to be inserted and the circle closed. The variety of sites not only makes it easier to find a restriction enzyme that is suitable for both the vector and the foreign DNA to be inserted, but, since some of the sites are placed within an antibiotic resistance gene, the presence of an insert can be detected by loss of resistance to that antibiotic. This is termed **insertional inactivation**.

Insertional inactivation is a useful selection method for identifying recombinant vectors with inserts. For example, a fragment of chromosomal DNA digested with *Bam*H1 would be isolated and purified. The plasmid pBR322 would also be digested at a single site, using *Bam*H1, and both samples would then be deproteinised to inactivate the restriction enzyme. *Bam*H1 cleaves to give sticky ends, and so it is possible to obtain ligation between the plasmid and digested DNA fragments in the presence of T4 DNA ligase. The products of this ligation will include plasmid containing a single fragment of the DNA as an insert, but there will also be unwanted products, such as

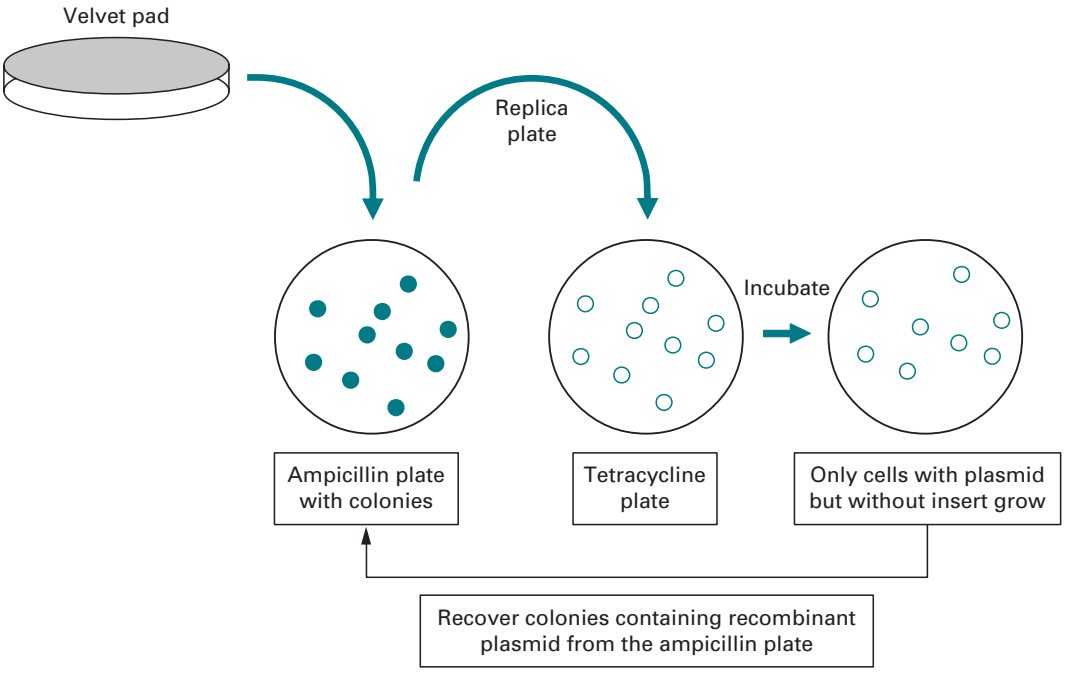

Fig. 6.12 Replica plating to detect recombinant plasmids. A sterile velvet pad is pressed onto the surface of an agar plate, picking up some cells from each colony growing on that plate. The pad is then pressed on to a fresh agar plate, thus inoculating it with cells in a pattern identical with that of the original colonies. Clones of cells that fail to grow on the second plate (e.g. owing to the loss of antibiotic resistance) can be recovered from their corresponding colonies on the first plate.

plasmid that has recircularised without an insert, dimers of plasmid, fragments joined to each other, and plasmid with an insert composed of more than one fragment. Most of these unwanted molecules can be eliminated during subsequent steps. The products of such reactions are usually identified by agarose gel electrophoresis (Section 5.7.4).

The ligated DNA must now be used to transform *E. coli*. Bacteria do not normally take up DNA from their surroundings, but can be induced to do so by prior treatment with Ca^{2+} at 4 °C; they are then termed **competent**, since DNA added to the suspension of competent cells will be taken up during a brief increase in temperature termed **heat shock**. Small, circular molecules are taken up most efficiently, whereas long, linear molecules will not enter the bacteria.

After a brief incubation to allow expression of the antibiotic resistance genes the cells are plated onto medium containing the antibiotic, e.g. ampicillin. Colonies that grow on these plates must be derived from cells that contain plasmid, since this carries the gene for resistance to ampicillin. It is not, at this stage, possible to distinguish between those colonies containing plasmids with inserts and those that simply contain recircularised plasmids. To do this, the colonies are replica plated, using a sterile velvet pad, onto plates containing tetracycline in their medium. Since the *Bam*HI site lies within the tetracycline resistance gene, this gene will be inactivated by the presence of insert, but will be intact in those plasmids that have merely recircularised (Fig. 6.12). Thus colonies that grow on ampicillin but not on tetracycline must contain

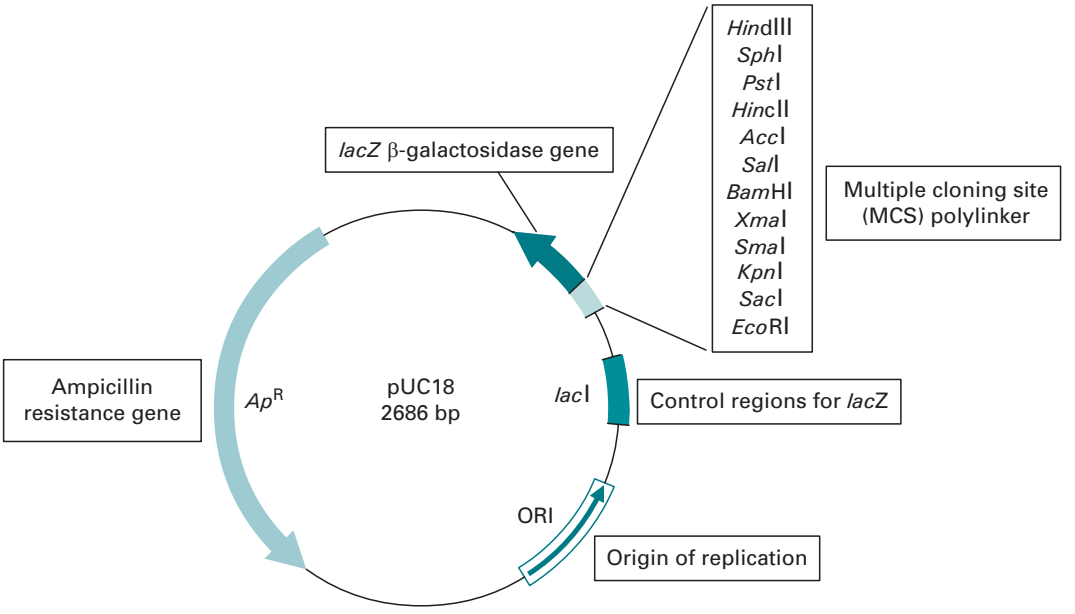

Fig. 6.13 Map and important features of pUC18.

plasmids with inserts. Since replica plating gives an identical pattern of colonies on both sets of plates, it is straightforward to recognise the colonies with inserts, and to recover them from the ampicillin plate for further growth. This illustrates the importance of a second gene for antibiotic resistance in a vector.

Although recircularised plasmid can be selected against, its presence decreases the yield of recombinant plasmid containing inserts. If the digested plasmid is treated with the enzyme alkaline phosphatase prior to ligation, recircularisation will be prevented, since this enzyme removes the 5′ phosphate groups that are essential for ligation. Links can still be made between the 5′ phosphate of insert and the 3′ hydroxyl of plasmid, so only recombinant plasmids and chains of linked DNA fragments will be formed. It does not matter that only one strand of the recombinant DNA is ligated, since the nick will be repaired by bacteria transformed with these molecules.

The valuable features of pBR322 have been enhanced by the construction of a series of plasmids termed pUC (produced at the University of California) (Fig. 6.13). There is an antibiotic resistance gene for tetracycline and origin of replication for *E. coli*. In addition the most popular restriction sites are concentrated into a region termed the **multiple cloning site** (MCS). In addition the MCS is part of a gene in its own right and codes for a portion of a polypeptide called β-galactosidase. When the pUC plasmid has been used to transform the host cell *E. coli* the gene may be switched on by adding the inducer IPTG (isopropyl-β-D-thiogalactopyranoside). Its presence causes the enzyme β-galactosidase to be produced (Section 5.5.5). The functional enzyme is able to hydrolyse a colourless substance called X-gal (5-bromo-4-chloro-3-indolyl-β-galactopyranoside) into a blue insoluble material (5,5′-dibromo-4,4′–dichloro indigo) (Fig. 6.14). However if the gene is disrupted by the insertion of a foreign

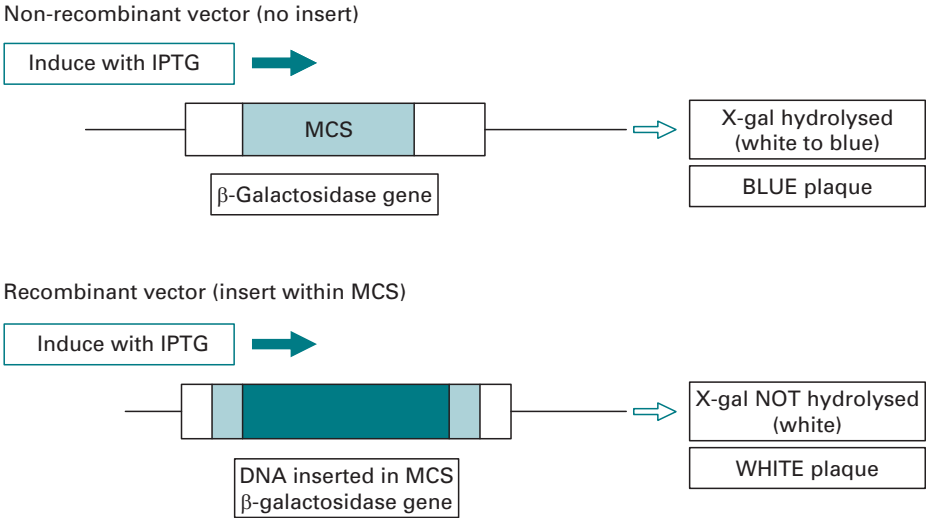

Fig. 6.14 Principle of blue/white selection for the detection of recombinant vectors.

fragment of DNA, a non-functional enzyme results which is unable to carry out hydrolysis of X-gal. Thus, a recombinant pUC plasmid may be easily detected since it is white or colourless in the presence of X-gal, whereas an intact non-recombinant pUC plasmid will be blue since its gene is fully functional and not disrupted. This elegant system, termed **blue/white selection**, allows the initial identification of recombinants to be undertaken very quickly and has been included in a number of subsequent vector systems. This selection method and insertional inactivation of antibiotic resistance genes do not, however, provide any information on the character of the DNA insert, just the status of the vector. To screen gene libraries for a desired insert hybridisation to gene probes is required and this is explained in Section 6.5.

6.3.2 Virus-based vectors

A useful feature of any cloning vector is the amount of DNA it may accept or have inserted before it becomes unviable. Inserts greater than 5 kb increase plasmid size to the point at which efficient transformation of bacterial cells decreases markedly, and so bacteriophages (bacterial viruses) have been adapted as vectors in order to propagate larger fragments of DNA in bacterial cells. Cloning vectors derived from **λ bacteriophage** are commonly used since they offer an approximately 16-fold advantage in cloning efficiency in comparison with the most efficient plasmid cloning vectors.

Phage λ is a linear double-stranded phage approximately 49 kb in length (Fig. 6.15). It infects *E. coli* with great efficiency by injecting its DNA through the cell membrane. In the wild-type phage λ the DNA follows one of two possible modes of replication. Firstly the DNA may either become stably integrated into the *E. coli* chromosome where it lies dormant until a signal triggers its excision. This is termed the **lysogenic life cycle**. Alternatively, it may follow a **lytic life cycle** where the DNA is replicated

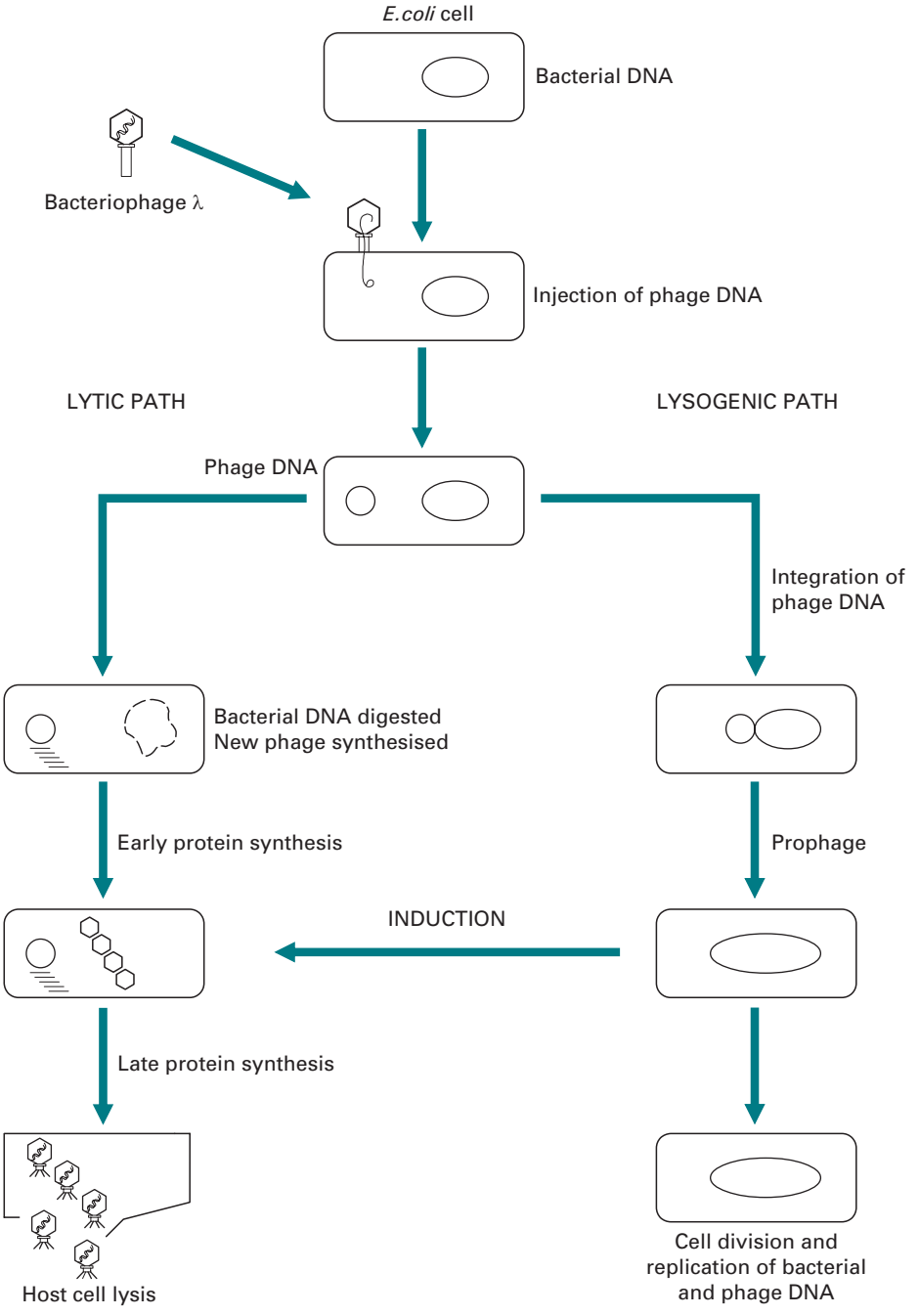

Fig. 6.15 The lysogenic and lytic cycles of bacteriophage λ.

upon entry to the cell, phage head and tail proteins synthesised rapidly and new functional phage assembled. The phage are subsequently released from the cell by lysing the cell membrane to infect further *E. coli* cells nearby. At the extreme ends of the phage λ are 12 bp sequences termed **cos (cohesive) sites**. Although they are

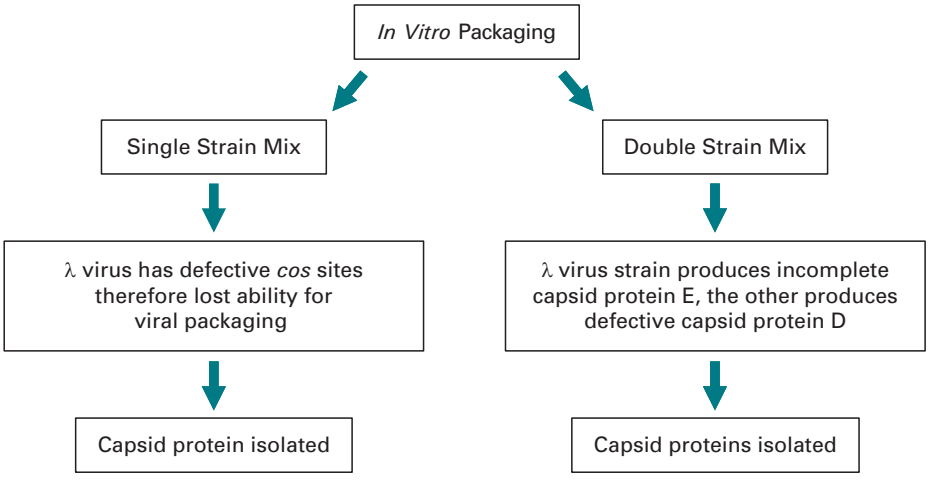

Fig. 6.16 Two strategies for producing *in vitro* packaging extracts for bacteriophage λ.

asymmetric they are similar to restriction sites and allow the phage DNA to be circularised. Phage may be replicated very efficiently in this way, the result of which are concatemers of many phage genomes which are cleaved at the cos sites and inserted into newly formed phage protein heads.

Much use of phage λ has been made in the production of gene libraries mainly because of its efficient entry into the *E. coli* cell and the fact that larger fragments of DNA may be stably integrated. For the cloning of long DNA fragments, up to approximately 25 kb, much of the non-essential λ DNA that codes for the lysogenic life cycle is removed and replaced by the foreign DNA insert. The recombinant phage is then assembled into pre-formed viral protein particles, a process termed *in vitro* packaging. These newly formed phage are used to infect bacterial cells that have been plated out on agar (Fig. 6.16).

Once inside the host cells, the recombinant viral DNA is replicated. All the genes needed for normal lytic growth are still present in the phage DNA, and so multiplication of the virus takes place by cycles of cell lysis and infection of surrounding cells, giving rise to plaques of lysed cells on a background, or **lawn**, of bacterial cells. The viral DNA including the cloned foreign DNA can be recovered from the viruses from these plaques and analysed further by restriction mapping (Section 5.9.1) and agarose gel electrophoresis (Section 5.7.4).

In general two types of λ phage vectors have been developed, **λ insertion vectors** and **λ replacement vectors** (Fig. 6.17). The λ insertion vectors accept less DNA than the replacement type since the foreign DNA is merely inserted into a region of the phage genome with appropriate restriction sites; common examples are λgt10 and λcharon16A. With a replacement vector a central region of DNA not essential for lytic growth is removed (a stuffer fragment) by a double digestion with for example *Eco*RI and *Bam*HI. This leaves two DNA fragments termed right and left arms. The central stuffer fragment is replaced by inserting foreign DNA between the arms to form a functional recombinant λ phage. The most notable examples of λ replacement vectors are λEMBL and λZap.

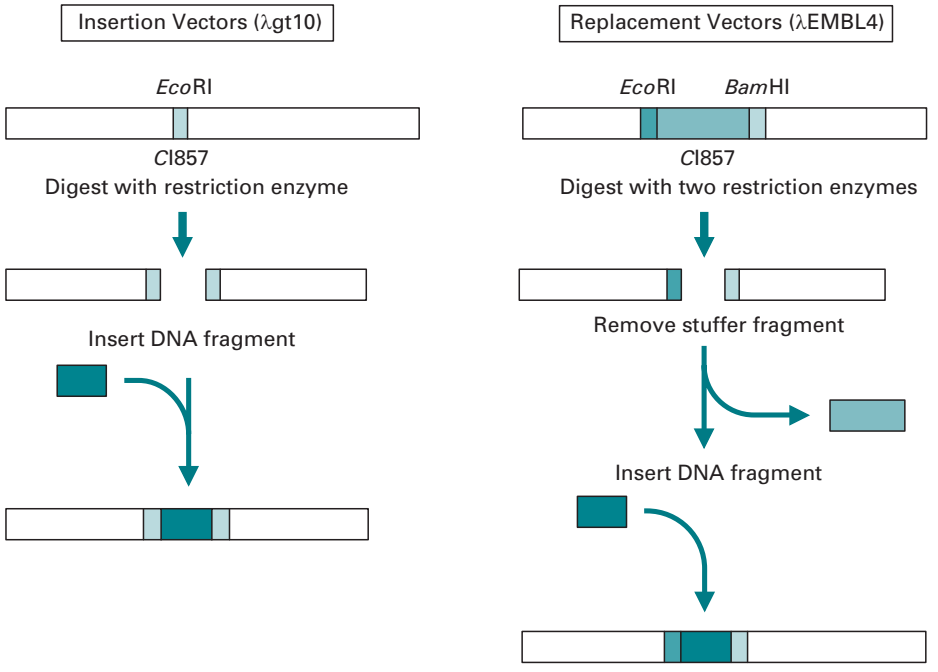

Fig. 6.17 General schemes used for cloning in λ insertion and λ replacement vectors. *C*I857 is a temperature-sensitive mutation that promotes lysis at 42 °C after incubation at 37 °C.

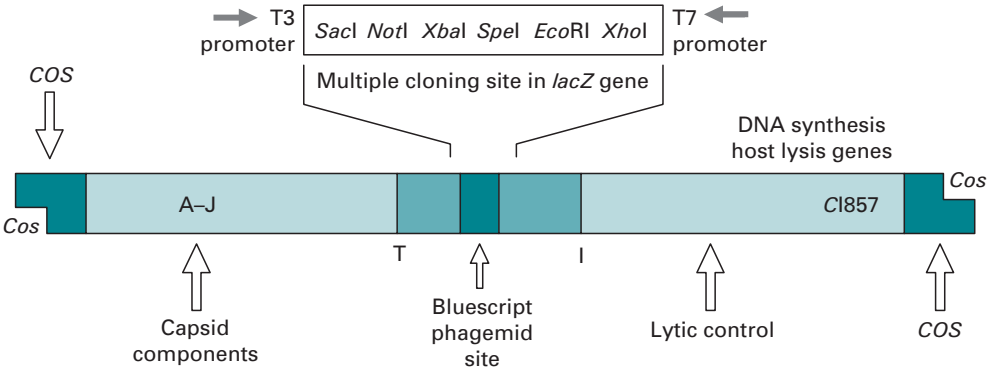

Fig. 6.18 General map of λZap cloning vector, indicating important areas of the vector. The multiple cloning site is based on the *lacZ* gene, providing blue/white selection based on the β-galactosidase gene. In between the initiator (I) site and terminator (T) site lie sequences encoding the phagemid Bluescript.

λZap is a commercially produced cloning vector that includes unique cloning sites clustered into a multiple cloning site (MCS) (Fig. 6.18). Furthermore the MCS is located within a *lacZ* region providing a blue/white screening system based on insertional inactivation. It is also possible to express foreign cloned DNA from this vector. This is a very useful feature of some λ vectors since it is then possible to screen for protein

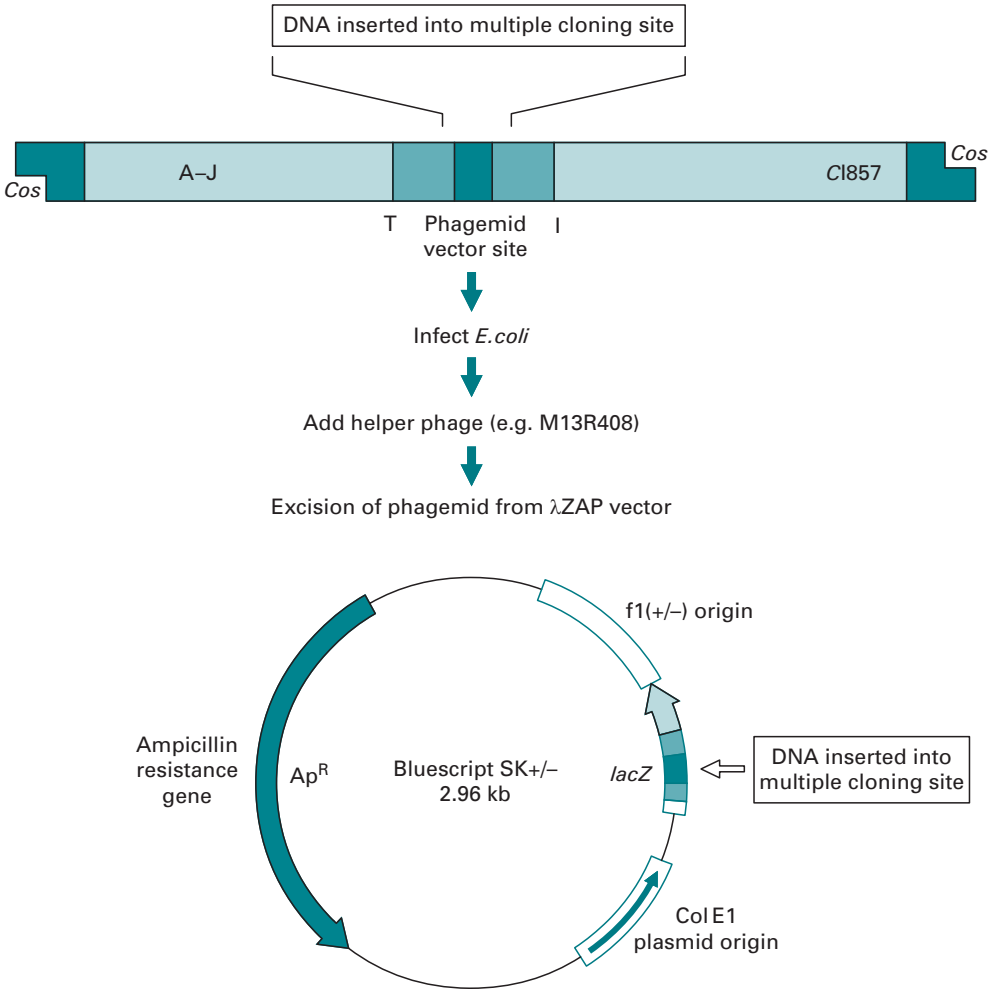

Fig. 6.19 Single-stranded DNA rescue of phagemid from λZap. The single-stranded phagemid pBluescript SK may be excised from λZap by addition of helper phage. This provides the necessary proteins and factors for transcription between the I and T sites in the parent phage to produce the phagemid with the DNA cloned into the parent vector.

product rather than the DNA inserted into the vector. This screening is therefore undertaken with antibody probes directed against the protein of interest (Section 6.5.4). Other features that make this a useful cloning vector are the ability to produce RNA transcripts termed cRNA or **riboprobes**. This is possible because two promoters for RNA polymerase enzymes exist in the vector, a T7 and a T3 promoter which flank the MCS (Section 6.4.2).

One of the most useful features of λZap is that it has been designed to allow automatic excision *in vivo* of a small 2.9 kb colony-producing vector termed a phagemid, pBluescript SK (Section 6.3.3). This technique is sometimes termed **single-stranded DNA rescue** and occurs as the result of a process termed **superinfection** where helper phage are added to the cells which are grown for an additional period of approximately 4 h (Fig. 6.19).

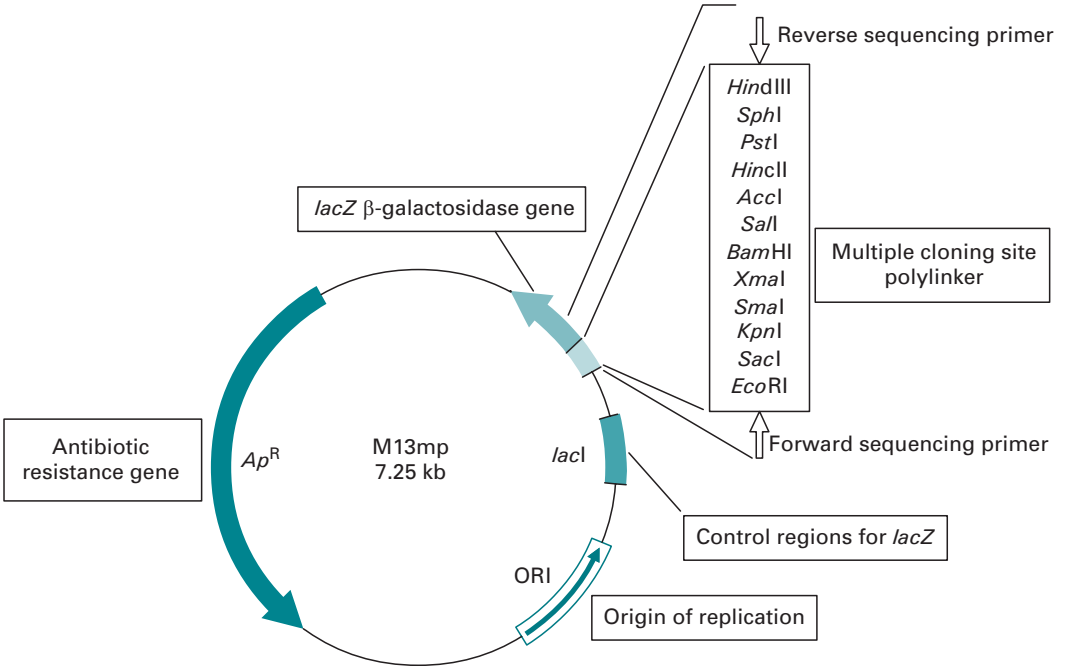

Fig. 6.20 Genetic map and important features of bacteriophage vector M13.

The helper phage displaces a strand within the λZap which contains the foreign DNA insert. This is circularised and packaged as a filamentous phage similar to M13 (Section 6.3.3). The packaged phagemid is secreted from the *E. coli* cell and may be recovered from the supernatant. Thus the λZap vector allows a number of diverse manipulations to be undertaken without the necessity of recloning or subcloning foreign DNA fragments. The process of subcloning is sometimes necessary when the manipulation of a gene fragment cloned in a general purpose vector needs to be inserted into a more specialised vector for the application of techniques such as *in vitro* mutagenesis or protein production (Section 6.6).

6.3.3 M13 and phagemid-based vectors

Much use has been made of single-stranded bacteriophage vectors such as M13 and vectors which have the combined properties of phage and plasmids, termed phagemids. M13 is a filamentous coliphage with a single-stranded circular DNA genome (Fig. 6.20). Upon infection of *E. coli*, the DNA replicates initially as a double-stranded molecule but subsequently produces single-stranded virions for infection of further bacterial cells (lytic growth). The nature of these vectors makes them ideal for techniques such as chain termination sequencing (Section 6.6.1) and *in vitro* mutagenesis (Section 6.6.3) since both require single-stranded DNA.

M13 or phagemids such as pBluescript SK infect *E. coli* harbouring a male-specific structure termed the F-pilus (Fig. 6.21). They enter the cell by adsorption to this structure and once inside the phage DNA is converted to a double-stranded replicative form or

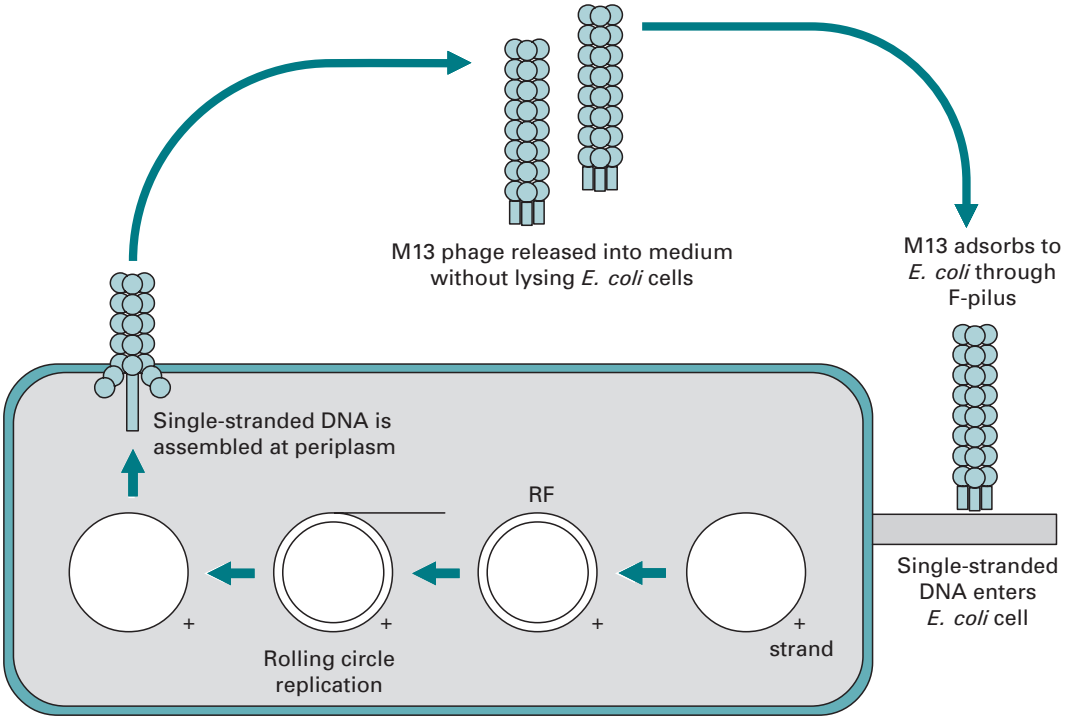

Fig. 6.21 Life cycle of bacteriophage M13. The bacteriophage virus enters the *E. coli* cell through the F-pilus. It then enters a stage where the circular single strands are converted to double strands. Rolling-circle replication then produces single strands, which are packaged and extruded through the *E. coli* cell membrane.

RF DNA. Replication then proceeds rapidly until some 100 RF molecules are produced within the *E. coli* cell. DNA synthesis then switches to the production of single strands and the DNA is assembled and packaged into the capsid at the bacterial periplasm. The bacteriophage DNA is then encapsulated by the major coat protein, gene VIII protein, of which there are approximately 2800 copies with three to six copies of the gene III protein at one end of the particle. The extrusion of the bacteriophage through the bacterial periplasm results in a decreased growth rate of the *E. coli* cell rather than host cell lysis and is visible on a bacterial lawn as an area of clearing. Approximately 1000 packaged phage particles may be released into the medium in one cell division.

In addition to producing single-stranded DNA the coliphage vectors have a number of other features that make them attractive as cloning vectors. Since the bacteriophage DNA is replicated as a double-stranded RF DNA intermediate a number of regular DNA manipulations may be performed such as restriction digestion, mapping and DNA ligation. RF DNA is prepared by lysing infected *E. coli* cells and purifying the super-coiled circular phage DNA with the same methods used for plasmid isolation. Intact single-stranded DNA packaged in the phage protein coat located in the supernatant may be precipitated with reagents such as polyethylene glycol, and the DNA purified with phenol/chloroform (Section 5.7.1). Thus the bacteriophage may act as a plasmid under certain circumstances and at other times produce DNA in the fashion of a virus. A family of vectors derived from M13 are currently widely used termed M13mp8/9,

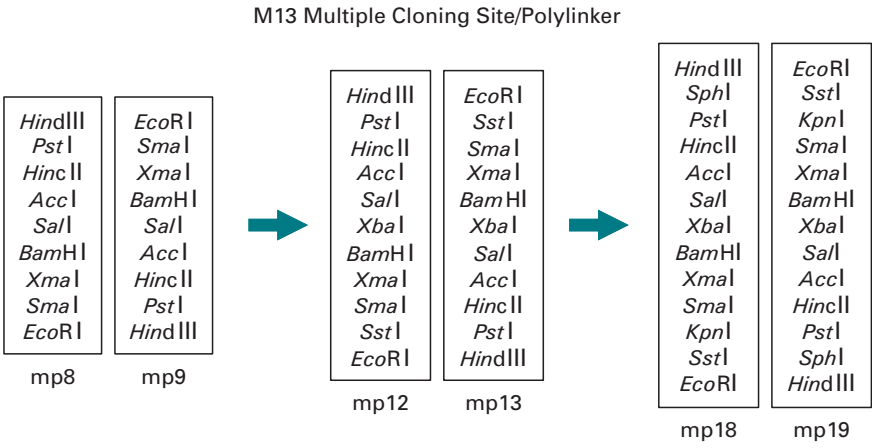

Fig. 6.22 Design and orientation of polylinkers in M13 series. Only the main restriction enzymes are indicated.

mp18/19, etc., all of which have a number of highly useful features. All contain a synthetic MCS, which is located in the *lacZ* gene without disruption of the reading frame of the gene. This allows efficient selection to be undertaken based on the technique of blue/white screening (Section 6.3.1). As the series of vectors were developed the number of restriction sites was increased in an asymmetric fashion. Thus M13mp8, mp12, mp18 and sister vectors which have the same MCS but in reverse orientation, M13mp9, mp13 and mp19 respectively have more restriction sites in the MCS making the vector more useful since greater choice of restriction enzymes is available (Fig. 6.22). However, one problem frequently encountered with M13 is the instability and spontaneous loss of inserts that are greater than 6 kb.

Phagemids are very similar to M13 and replicate in a similar fashion. One of the first phagemid vectors, pEMBL, was constructed by inserting a fragment of another phage termed f1 containing a phage origin of replication and elements for its morphogenesis into a pUC8 plasmid. Following superinfection with helper phage the f1 origin is activated allowing single-stranded DNA to be produced. The phage is assembled into a phage coat extruded through the periplasm and secreted into the culture medium in a similar way to M13. Without superinfection the phagemid replicates as a pUC type plasmid and in the replicative form (RF) the DNA isolated is double-stranded. This allows further manipulations such as restriction digestion, ligation and mapping analysis to be performed. The pBluescript SK vector is also a phagemid and can be used in its own right as a cloning vector and manipulated as if it were a plasmid. It may, like M13, be used in nucleotide sequencing and site-directed mutagenesis, and it is also possible to produce RNA transcripts that may be used in the production of labelled cRNA probes or riboprobes (Section 6.4.2).

6.3.4 Cosmid based vectors

The way in which the phage λ DNA is replicated is of particular interest in the development of larger insert cloning vectors termed cosmids (Fig. 6.23). These are

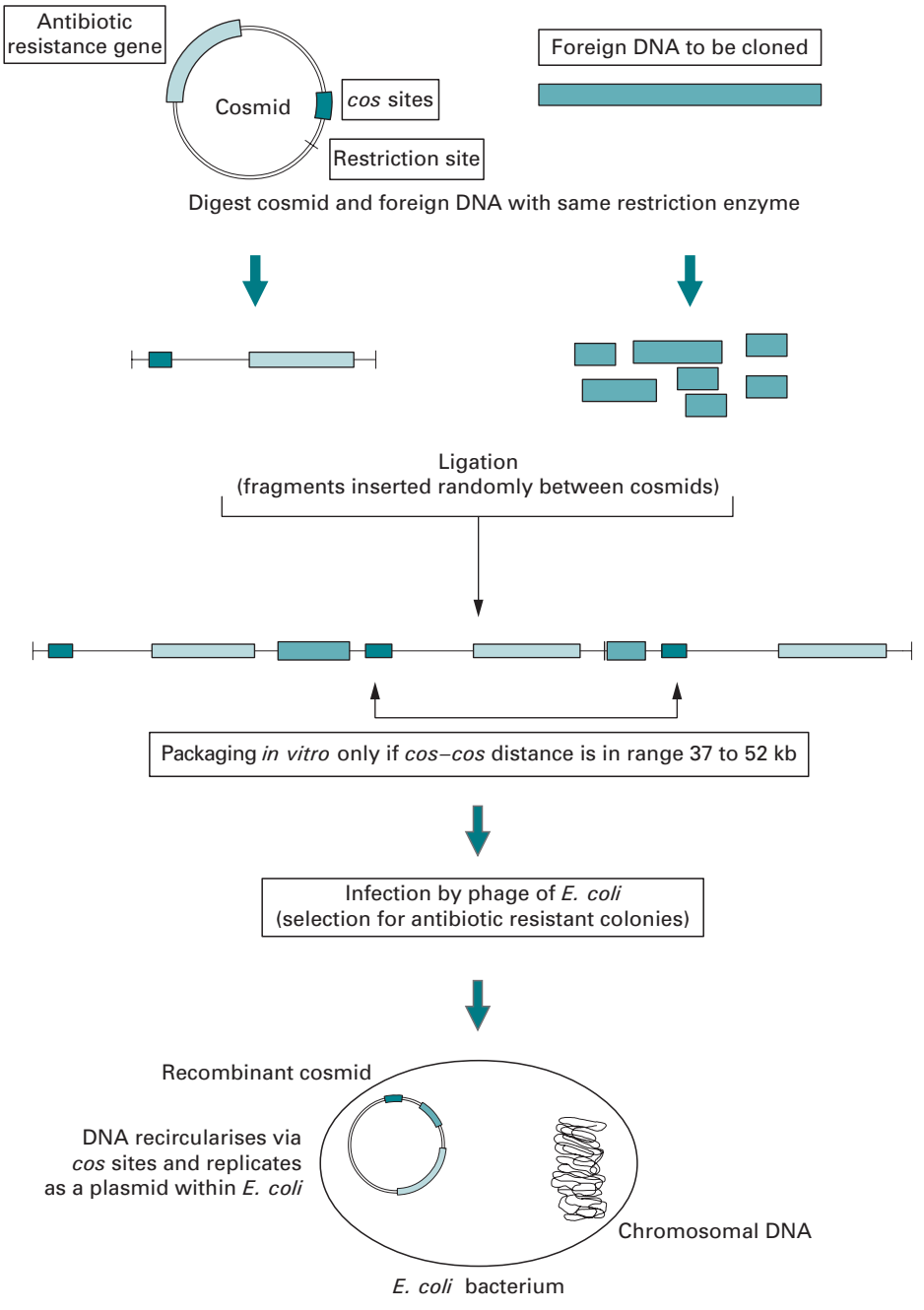

Fig. 6.23 Scheme for cloning foreign DNA fragments in cosmid vectors.

especially useful for the analysis of highly complex genomes and are an important part of various genome mapping projects (Section 6.9).

The upper limit of the insert capacity of phage λ is approximately 21 kb. This is because of the requirement for essential genes and the fact that the maximum length

between the *cos* sites is 52 kb. Consequently cosmid vectors have been constructed that incorporate the *cos* sites from phage λ and also the essential features of a plasmid, such as the plasmid origin of replication, a gene for drug resistance, and several unique restriction sites for insertion of the DNA to be cloned. When a cosmid preparation is linearised by restriction digestion, and ligated to DNA for cloning, the products will include concatamers of alternating cosmid vector and insert. Thus the only requirement for a length of DNA to be packaged into viral heads is that it should contain *cos* sites spaced the correct distance apart; in practice this spacing can range between 37 and 52 kb. Such DNA can be packaged *in vitro* if phage head precursors, tails and packaging proteins are provided. Since the cosmid is very small, inserts of about 40 kb in length will be most readily packaged. Once inside the cell, the DNA recircularises through its *cos* sites, and from then onwards behaves exactly like a plasmid.

6.3.5 Large insert capacity vectors

The advantage of vectors that accept larger fragments of DNA than phage λ or cosmids is that fewer clones need to be screened when searching for the foreign DNA of interest. They have also had an enormous impact in the mapping of the genomes of organisms such as the mouse and are used extensively in the human genome mapping project (Section 6.9.3). Recent developments have allowed the production of large insert capacity vectors based on human artificial chromosomes, bacterial artificial chromosomes (BACs), mammalian artificial chromosomes (MACs) and on the virus P1 (PACs), P1 artificial chromosomes. However, perhaps the most significant development are vectors based on yeast artificial chromosomes (YACs).

6.3.6 Yeast artificial chromosome (YAC) vectors

Yeast artificial chromosomes (YACs) are linear molecules composed of a centromere, telomere and a replication origin termed an ARS element (**autonomous replicating sequence**). The YAC is digested with restriction enzymes at the SUP4 site (a suppressor tRNA gene marker) and *Bam*HI sites separating the telomere sequences (Fig. 6.24). This produces two arms and the foreign genomic DNA is ligated to produce a functional YAC construct. YACs are replicated in yeast cells; however, the external cell wall of the yeast needs to be removed to leave a spheroplast. These are osmotically unstable and also need to be embedded in a solid matrix such as agar. Once the yeast cells are transformed only correctly constructed YACs with associated selectable markers are replicated in the yeast strains. DNA fragments with repeat sequences are sometimes difficult to clone in bacterial-based vectors but may be successfully cloned in YAC systems. The main advantage of YAC-based vectors, however, is the ability to clone very large fragments of DNA. Thus the stable maintenance and replication of foreign DNA fragments of up to 2000 kb have been carried out in YAC vectors and they are the main vector of choice in the various genome mapping and sequencing projects (Section 6.9).

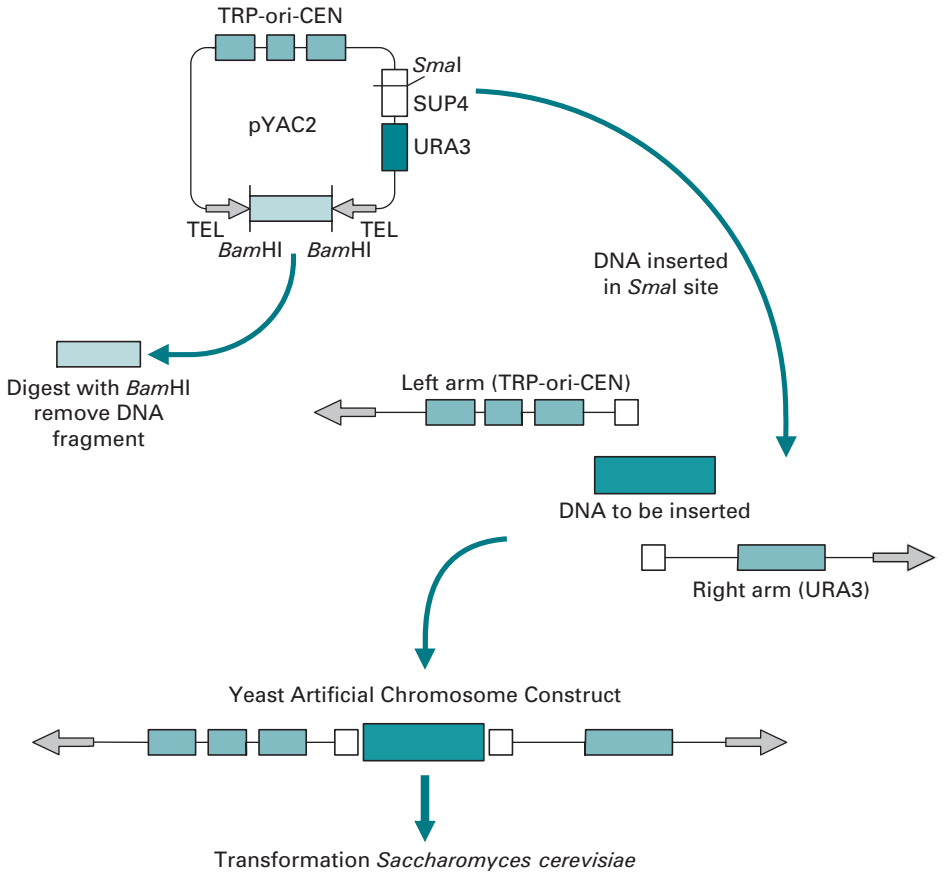

Fig. 6.24 Scheme for cloning large fragments of DNA into YAC vectors.

6.3.7 Vectors used in eukaryotes

The use of *E. coli* for general cloning and manipulation of DNA is well established; however, numerous developments have been made for cloning in eukaryotic cells. Plasmids used for cloning DNA in eukaryotic cells require a eukaryotic origin of replication and marker genes that will be expressed by eukaryotic cells. At present the two most important applications of plasmids to eukaryotic cells are for cloning in yeast and in plants.

Although yeast has a natural plasmid, called the 2μ circle, this is too large for use in cloning. Plasmids such as the yeast episomal plasmid (YEp) have been created by genetic manipulation using replication origins from the 2μ circle, and by incorporating a gene which will complement a defective gene in the host yeast cell. If, for example, a strain of yeast is used which has a defective gene for the biosynthesis of an amino acid, an active copy of that gene on a yeast plasmid can be used as a selectable marker for the presence of that plasmid. Yeast, like bacteria, can be grown rapidly, and it is therefore well suited for use in cloning. Of particular use has been the creation of shuttle vectors which have origins of replication for yeast and bacteria such as *E. coli*.

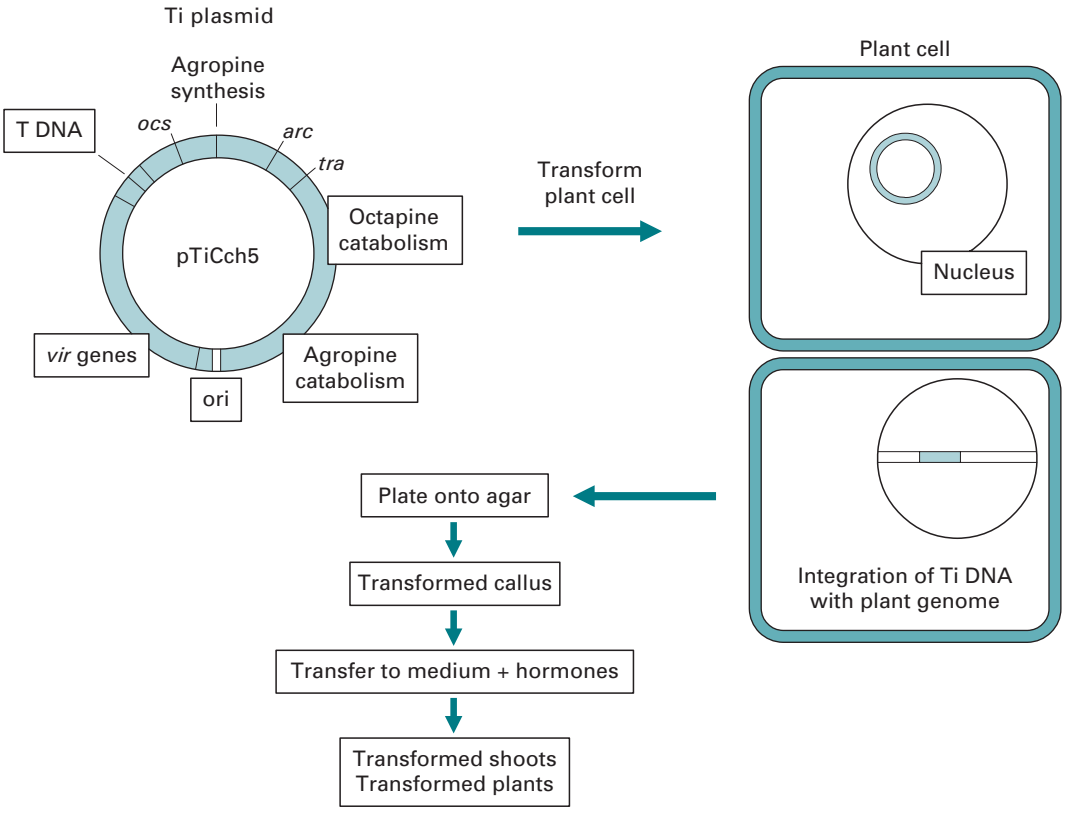

Fig. 6.25 Scheme for cloning in plant cells using the Ti plasmid.

This means that constructs may be prepared rapidly in the bacteria and delivered into yeast for expression studies.

The bacterium *Agrobacterium tumefaciens* infects plants that have been damaged near soil level, and this infection is often followed by the formation of plant tumours in the vicinity of the infected region. It is now known that *A. tumefaciens* contains a plasmid called the Ti plasmid, part of which is transferred into the nuclei of plant cells which are infected by the bacterium. Once in the nucleus, this DNA is maintained by integrating with the chromosomal DNA. The integrated DNA carries genes for the synthesis of opines (which are metabolised by the bacteria but not by the plants) and for tumour induction (hence 'Ti'). DNA inserted into the correct region of the Ti plasmid will be transferred to infected plant cells, and in this way it has been possible to clone and express foreign genes in plants (Fig. 6.25). This is an essential prerequisite for the genetic engineering of crops.

6.3.8 Delivery of vectors into eukaryotes

Following the production of a recombinant molecule, the so-called **constructs** are subsequently introduced into cells to enable it to be replicated a large number of times as the cells replicate. Initial recombinant DNA experiments were performed in bacterial

cells, because of their ease of growth and short doubling time. Gram-negative bacteria such as *E. coli* can be made **competent** for the introduction of extraneous plasmid DNA into cells (Section 6.3.1). The natural ability of bacteriophage to introduce DNA into *E. coli* has also been well exploited and results in 10–100-fold higher efficiency for the introduction of recombinant DNA compared to transformation of competent bacteria with plasmids. These well-established and traditional approaches are the reason why so many cloning vectors have been developed for *E. coli*. The delivery of cloning vectors into eukaryotic cells is, however, not as straightforward as that for the bacterium *E. coli*.

It is possible to deliver recombinant molecules into animal cells by **transfection**. The efficiency of this process can be increased by first precipitating the DNA with Ca^{2+} or making the membrane permeable with divalent cations. High-molecular-weight polymers such as DEAE-dextran or polyethylene glycol (PEG) may also be used to maximise the uptake of DNA. The technique is rather inefficient although a selectable marker that provides resistance to a toxic compound such as neomycin can be used to monitor the success. Alternatively, DNA can be introduced into animal cells by **electroporation**. In this process the cells are subjected to pulses of a high-voltage gradient, causing many of them to take up DNA from the surrounding solution. This technique has proved to be useful with cells from a range of animal, plant and microbial sources. More recently the technique of **lipofection** has been used as the delivery method. The recombinant DNA is encapsulated by a core of lipid-coated particles which fuse with the lipid membrane of cells and thus release the DNA into the cell. Microinjection of DNA into cell nuclei of eggs or embryos has also been performed successfully in many mammalian cells.

The ability to deliver recombinant molecules into plant cells is not without its problems. Generally the outer cell wall of the plant must be stripped, usually by enzymatic digestion, to leave a protoplast. The cells are then able to take up recombinants from the supernatant. The cell wall can be regenerated by providing appropriate media. In cases where protoplasts have been generated transformation may also be achieved by electroporation. An even more dramatic transformation procedure involves propelling microscopically small titanium or gold pellet microprojectiles coated with the recombinant DNA molecule, into plant cells in intact tissues. This **biolistic** technique involves the detonation of an explosive charge which is used to propel the microprojectiles into the cells at a high velocity. The cells then appear to reseal themselves after the delivery of the recombinant molecule. This is a particularly promising technique for use with plants whose protoplasts will not regenerate whole plants.

6.4 HYBRIDISATION AND GENE PROBES

6.4.1 Cloned DNA probes

The increasing accumulation of DNA sequences in nucleic acid databases coupled with the availability of custom synthesis of oligonucleotides has provided a relatively straightforward means to design and produce gene probes and primers for PCR. Such

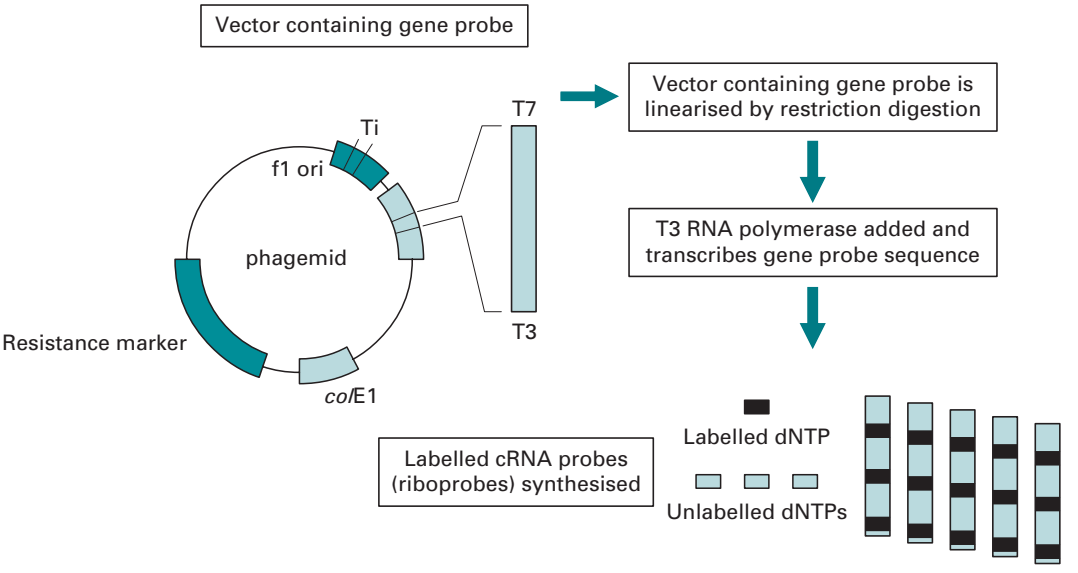

Fig. 6.26 Production of cRNA (riboprobes) using T3 RNA polymerase and phagemid vectors.

probes and primers are usually designed with bioinformatics software using sequence information from nucleic acid databases. Alternatively, gene family related sequences as indicated in Section 5.11 may also be successfully employed. However, there are many gene probes that have traditionally been derived from cDNA or from genomic sequences and which have been cloned into plasmid and phage vectors. These require manipulation before they may be labelled and used in hybridisation experiments. Gene probes may vary in length from 100 bp to a number of kilobases, although this is dependent on their origin. Many are short enough to be cloned into plasmid vectors and are useful in that they may be manipulated easily and are relatively stable both in transit and in the laboratory. The DNA sequences representing the gene probe are usually excised from the cloning vector by digestion with restriction enzymes and purified. In this way vector sequences which may hybridise non-specifically and cause high background signals in hybridisation experiments are removed. There are various ways of labelling DNA probes and these are described in Section 5.9.4.

6.4.2 RNA gene probes

It is also possible to prepare cRNA probes or riboprobes by *in vitro* transcription of gene probes cloned into a suitable vector. A good example of such a vector is the phagemid pBluescript SK since at each end of the multiple cloning site where the cloned DNA fragment resides are promoters for T3 or T7 RNA polymerase (Section 6.3.3). The vector is then made linear with a restriction enzyme and T3 or T7 RNA polymerase is used to transcribe the cloned DNA fragment. Provided a labelled NTP is added in the reaction a riboprobe labelled to a high specific activity will be produced (Fig. 6.26). One advantage of riboprobes is that they are single stranded and their sensitivity is generally regarded as

superior to cloned double-stranded probes indicated in Section 6.4.1. They are used extensively in *in situ* hybridisation and for identifying and analysing mRNA and are described in more detail in Section 6.8.

6.5 SCREENING GENE LIBRARIES

6.5.1 Colony and plaque hybridisation

Once a cDNA or genomic library has been prepared the next task requires the identification of the specific fragment of interest. In many cases this may be more problematic than the library construction itself since many hundreds of thousands of clones may be in the library. One clone containing the desired fragment needs to be isolated from the library and therefore a number of techniques mainly based on hybridisation have been developed.

Colony hybridisation is one method used to identify a particular DNA fragment from a plasmid gene library (Fig. 6.27). A large number of clones are grown up to form colonies on one or more plates, and these are then replica plated onto nylon membranes placed on solid agar medium. Nutrients diffuse through the membranes and allow colonies to grow on them. The colonies are then lysed, and liberated DNA is denatured and bound to the membranes, so that the pattern of colonies is replaced by an identical pattern of bound DNA. The membranes are then incubated with a prehybridisation mix containing non-labelled non-specific DNA such as salmon sperm DNA to block non-specific sites. Following this denatured, labelled gene probe is added. Under hybridising conditions the probe will bind only to cloned fragments containing at least part of its corresponding gene (Section 5.9.3). The membranes are then washed to remove any unbound probe and the binding detected by autoradiography of the membranes. If non-radioactive labels have been used then alternative methods of detection must be employed (Section 5.9.4). By comparison of the patterns on the autoradiograph with the original plates of colonies, those that contain the desired gene (or part of it) can be identified and isolated for further analysis. A similar procedure is used to identify desired genes cloned into bacteriophage vectors. In this case the process is termed **plaque hybridisation**. It is the DNA contained in the bacteriophage particles found in each plaque that is immobilised on to the nylon membrane. This is then probed with an appropriately labelled complementary gene probe and the detection undertaken as for colony hybridisation.

6.5.2 PCR screening of gene libraries

In many cases it is now possible to use the PCR to screen cDNA or genomic libraries constructed in plasmids or bacteriophage vectors. This is usually undertaken with primers which anneal to the vector rather than the foreign DNA insert. The size of an amplified product may be used to characterise the cloned DNA and subsequent restriction mapping is then carried out (Fig. 6.28). The main advantage of the PCR over traditional hybridisation based screening is the rapidity of the technique, as PCR

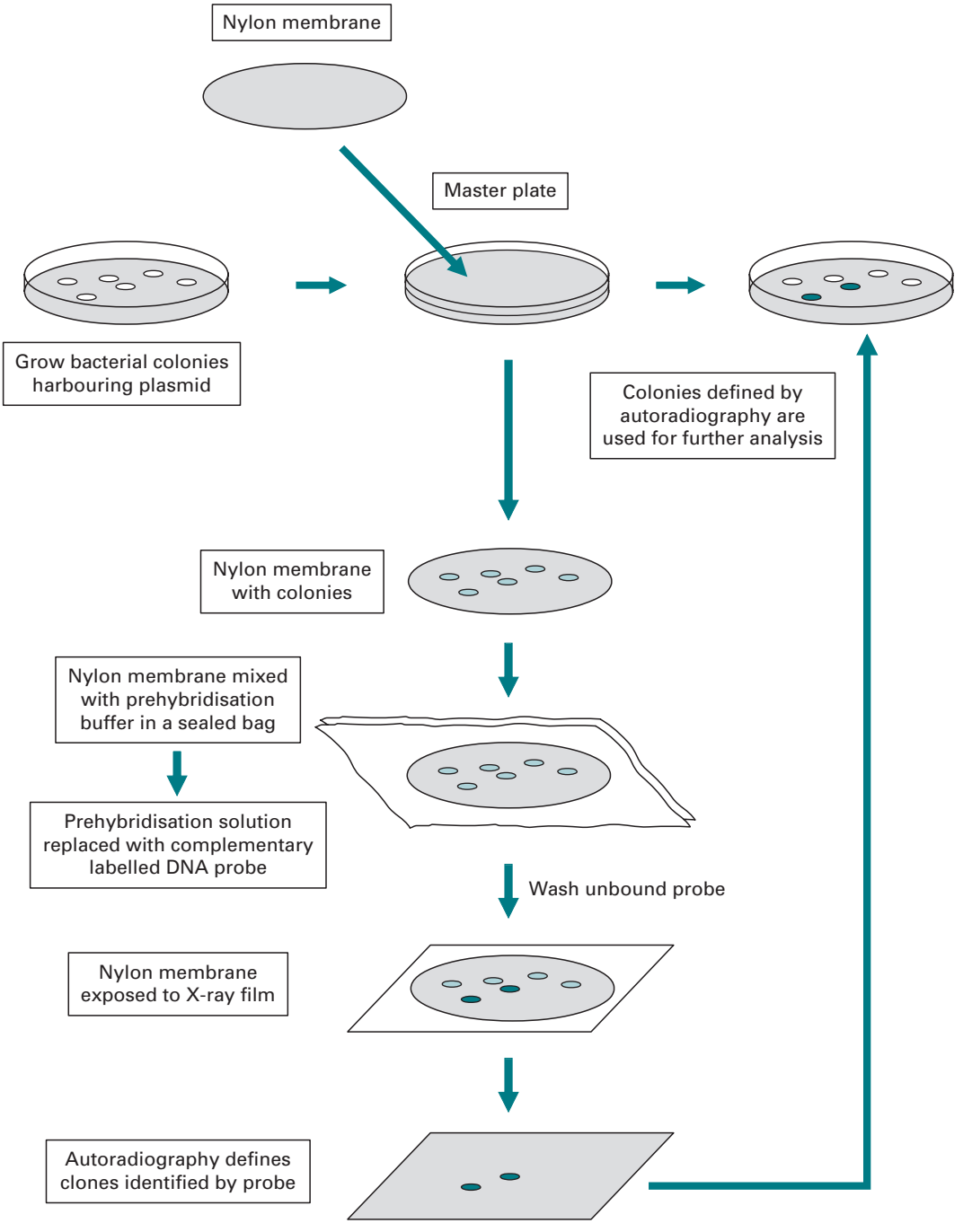

Fig. 6.27 Colony hybridisation technique for locating specific bacterial colonies harbouring recombinant plasmid vectors containing desired DNA fragments. This is achieved by hybridisation to a complementary labelled DNA probe and autoradiography.

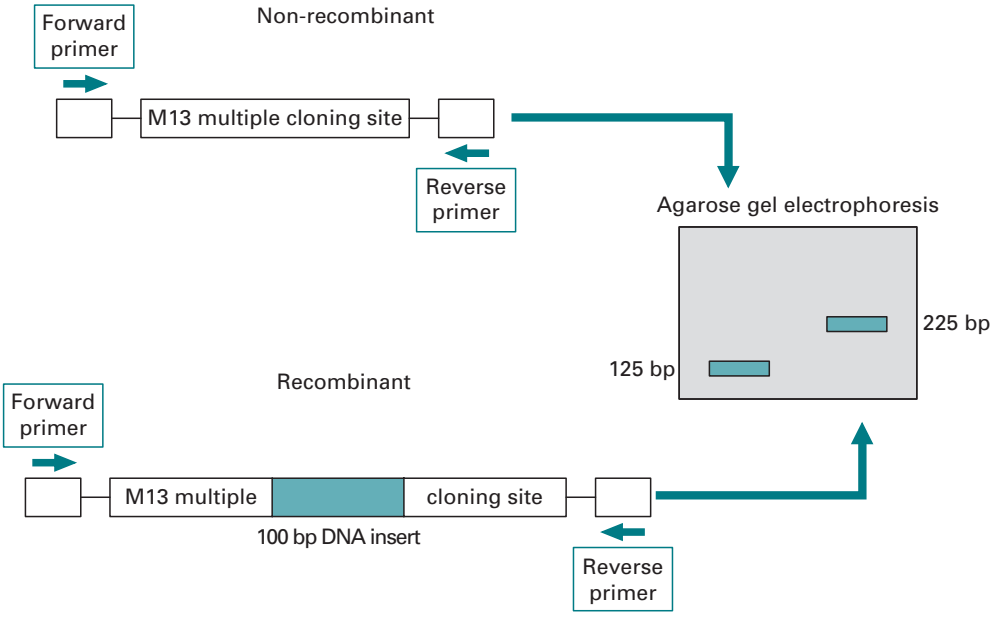

Fig. 6.28 PCR screening of recombinant vectors. In this figure, the M13 non-recombinant has no insert and so the PCR undertaken with forward and reverse sequencing primers gives rise to a product 125 bp in length. The M13 recombinant with an insert of 100 bp will give rise to a PCR product of 125 bp + 100 bp = 225 bp and thus may be distinguished from the non-recombinant by analysis on agarose gel electrophoresis.

screening may be undertaken in 3–4 h whereas it may be several days before detection by hybridisation is achieved. The PCR screening technique gives an indication of the size of the cloned insert rather than the sequence of the insert; however, PCR primers that are specific for a foreign DNA insert may also be used. This allows a more rigorous characterisation of clones from cDNA and genomic libraries.

6.5.3 Hybrid select/arrest translation

The difficulty of characterising clones and detecting a desired DNA fragment from a mixed cDNA library may be made simpler by two useful techniques termed **hybrid select (release) translation** or **hybrid arrest translation**. Following the preparation of a cDNA library in a plasmid vector the plasmid is extracted from part of each colony, and each preparation is then denatured and immobilised on a nylon membrane (Fig. 6.29). The membranes are soaked in total cellular mRNA, under stringent conditions (usually a temperature only a few degrees below T_m) in which hybridisation will occur only between complementary strands of nucleic acid. Hence each membrane will bind just one species of mRNA, since it has only one type of cDNA immobilised on it. Unbound mRNA is washed off the membranes, and then the bound mRNA is eluted and used to direct *in vitro* translation (Section 6.7). By **immunoprecipitation** or electrophoresis of the protein, the mRNA coding for a particular protein can be detected, and the clone containing its corresponding cDNA isolated. This technique is known as hybrid release translation. In a related method called

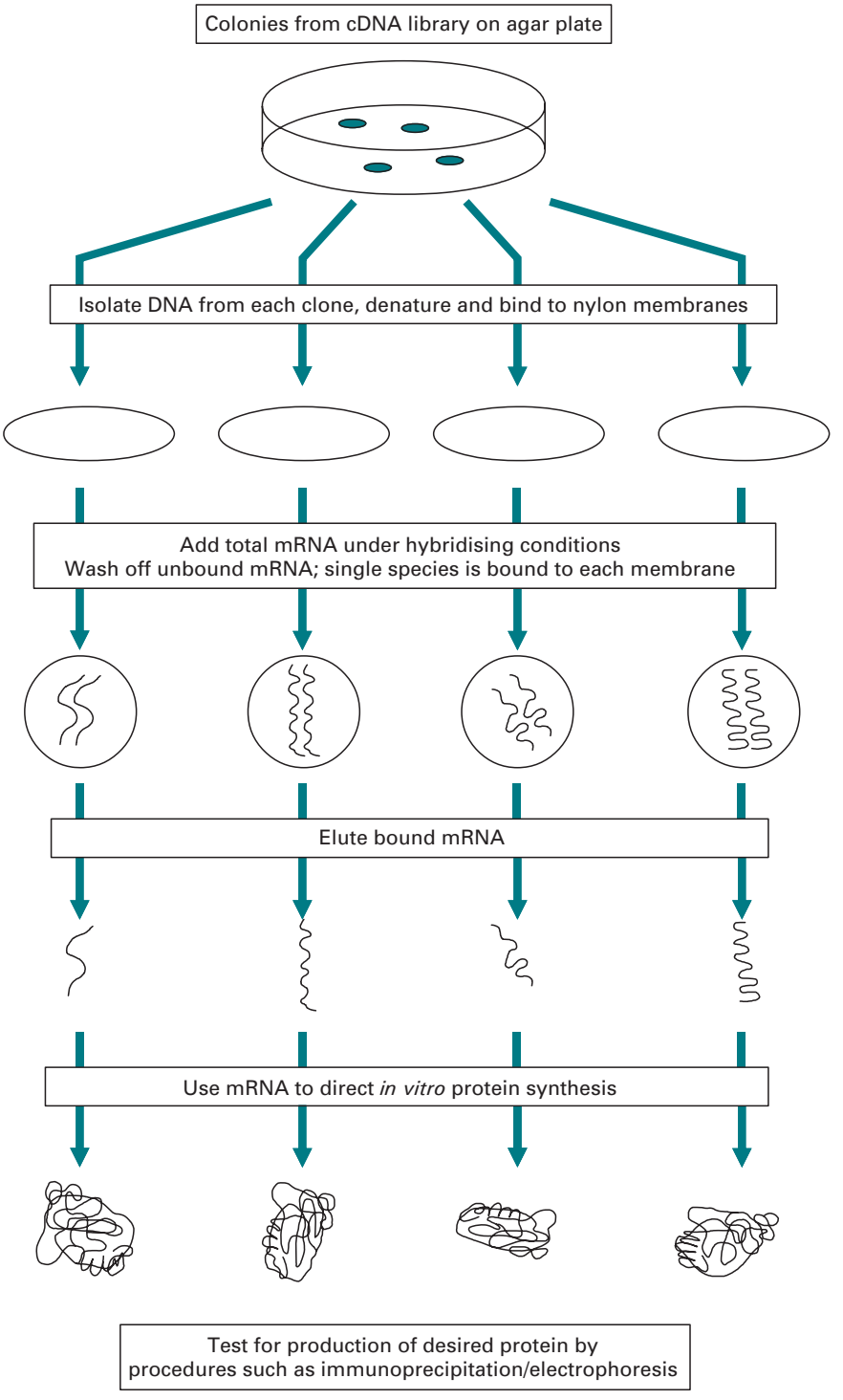

Fig. 6.29 General principles involved in the technique of a hybrid select translation.

hybrid arrest translation a positive result is indicated by the absence of a particular translation product when total mRNA is hybridised with excess cDNA. This is a consequence of the fact that mRNA cannot be translated when it is hybridised to another molecule.

6.5.4 Screening expression cDNA libraries

In some cases the protein for which the gene sequence is required is partially characterised and in these cases it may be possible to produce antibodies to that protein. This allows immunological screening to be undertaken rather than gene hybridisation. Such antibodies are useful since they may be used as the probe if little or no gene sequence is available. In these cases it is possible to prepare a cDNA library in a specially adapted vector termed an expression vector which transcribes and translates any cDNA inserted into it. The protein is usually synthesised as a fusion with another protein such as β-galactosidase. Common examples of expression vectors are those based on bacteriophage such as λgt11 and λZap or plasmids such as pEX. The precise requirements for such vectors are identical to vectors which are dedicated to producing proteins *in vitro* and are described in Section 6.7.1. In some cases expression vectors incorporate inducible promoters which may be activated by for example increasing the temperature allowing stringent control of expression of the cloned cDNA molecules (Fig. 6.30).

The cDNA library is plated out and nylon membrane filters prepared as for colony/plaque hybridisation. A solution containing the antibody to the desired protein is then added to the membrane. The membrane is then washed to remove any unbound protein and a further labelled antibody which is directed to the first antibody is applied. This allows visualisation of the plaque or colony that contains the cloned cDNA for that protein and this may then be picked from the agar plate and pure preparations grown for further analysis.

6.6 APPLICATIONS OF GENE CLONING

6.6.1 Sequencing cloned DNA

Most of the DNA sequencing now undertaken is based on the use of PCR products as the template; however, DNA fragments, including PCR products cloned into plasmid vectors, may be subjected to the chain termination sequencing (Section 5.9.5). However, due to the double-stranded nature of plasmids further manipulation needs to be undertaken before this may be attempted. In these cases the plasmids are denatured usually by alkali treatment. Although the plasmids containing the foreign DNA inserts may reanneal the kinetics of the reaction is such that the strands are single-stranded for a long enough period of time to allow the sequencing method to succeed. It is also possible to include denaturants such as formamide in the reaction to further prevent reannealing. In general, however, superior results may be gained with sequencing single-stranded DNA from M13 or single-stranded phagemids which means that the cloned DNA of interest is usually subcloned into these vectors. A further modification

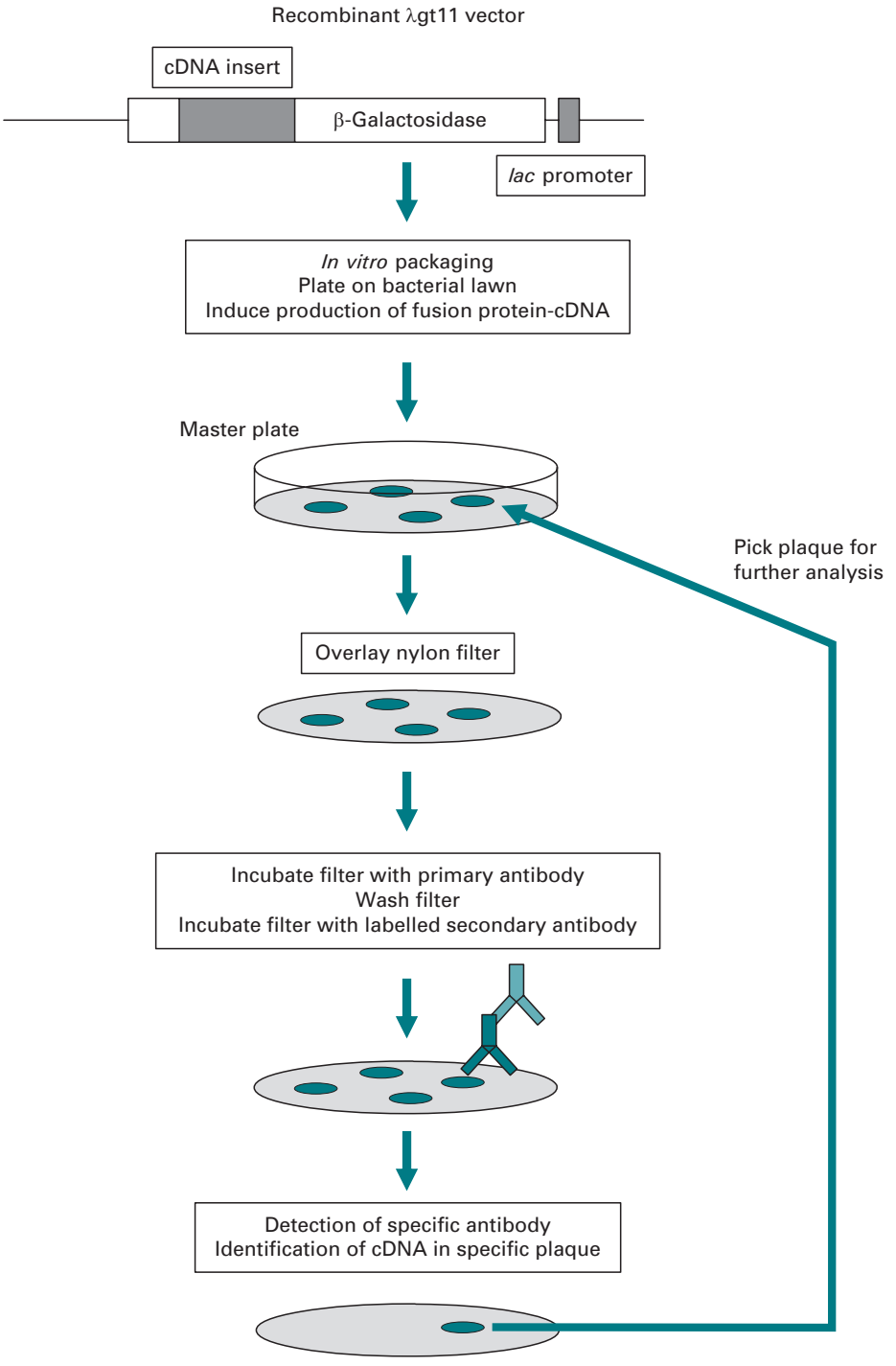

Fig. 6.30 Screening of cDNA libraries in expression vector λgt11. The cDNA inserted upstream from the gene for λβ-galactosidase will give rise to a fusion protein under induction (e.g. with IPTG). The plaques are then blotted onto a nylon membrane filter and probed with an antibody specific for the protein coded for by the cDNA. A secondary labelled antibody directed to the specific antibody can then be used to identify the location (plaque) of the cDNA.

that makes M13 useful in chain termination sequencing is the placement of universal priming sites at −20 or −40 bases from the start of the MCS. This allows any gene to be sequenced by using one universal primer since annealing of the primer prior to sequencing occurs outside the MCS and so is M13-specific rather than gene-specific. This obviates the need to synthesise new oligonucleotide primers for each new foreign DNA insert. A further, reverse priming site is also located at the opposite end of the polylinker allowing sequencing in the opposite orientation to be undertaken.

6.6.2 *In vitro* mutagenesis

One of the most powerful developments in molecular biology has been the ability to artificially create defined mutations in a gene and analyse the resulting protein following *in vitro* expression. Numerous methods are now available for producing **site-directed mutations** many of which now involve the PCR. Commonly termed **protein engineering**, this process involves a logical sequence of analytical and computational techniques centred around a design cycle. This includes the biochemical preparation and analysis of proteins, the subsequent identification of the gene encoding the protein and its modification. The production of the modified protein and its further biochemical analysis completes the concept of rational redesign to improve a protein's structure and function (Fig. 6.31).

The use of design cycles and rational design systems are exemplified by the study and manipulation of subtilisin. This is a serine protease of broad specificity and of considerable industrial importance being used in soap powder and in the food and leather industries. Protein engineering has been used to alter the specificity, pH profile and stability to oxidative, thermal and alkaline inactivation. Analysis of homologous thermophiles and their resistance to oxidation has also been improved. Engineered subtilisins of improved bleach resistance and wash performance are now used in many brands of washing powders. Furthermore mutagenesis has played an important role in the re-engineering of important therapeutic proteins such as the Herceptin antibody which has been used to successfully treat certain types of breast cancer.

6.6.3 Oligonucleotide-directed mutagenesis

The traditional method of site-directed mutagenesis demands that the gene be already cloned or subcloned into a single-stranded vector such as M13. Complete sequencing of the gene is essential to identify a potential region for mutation. Once the precise base change has been identified an oligonucleotide is designed that is complementary to part of the gene but has one base difference. This difference is designed to alter a particular codon, which, following translation, gives rise to a different amino acid and hence may alter the properties of the protein.

The oligonucleotide and the single-stranded DNA are annealed and DNA polymerase is added together with the dNTPs. The primer for the reaction is the 3′ end of the oligonucleotide. The DNA polymerase produces a new complementary DNA strand to the existing one but which incorporates the oligonucleotide with the base mutation. The subsequent cloning of the recombinant produces multiple copies, half of which

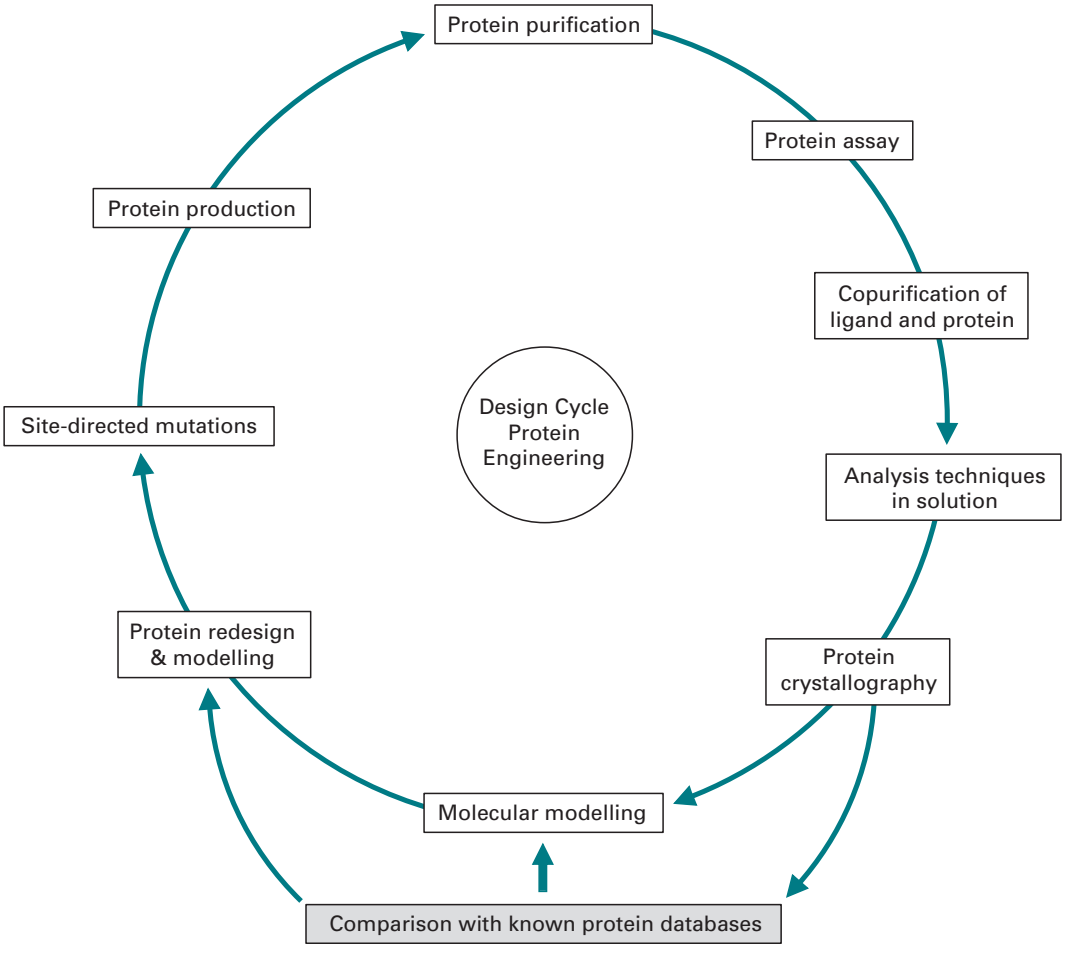

Fig. 6.31 Protein design cycle used in the rational redesign of proteins and enzymes.

contain a sequence with the mutation and half contain the wild-type sequence. Plaque hybridisation using the oligonucleotide as the probe is then used at a stringency that allows only those plaques containing a mutated sequence to be identified (Fig. 6.32). Further methods have also been developed which simplify the process of detecting the strands with the mutations.

6.6.4 PCR-based mutagenesis

The PCR has been adapted to allow mutagenesis to be undertaken and this relies on single bases mismatched between one of the PCR primers and the target DNA to become incorporated into the amplified product following thermal cycling.

The basic **PCR mutagenesis** system involves the use of two primary PCR reactions to produce two overlapping DNA fragments both bearing the same mutation in the overlap region. The technique is termed **overlap extension PCR**. The two separate PCR products are made single-stranded and the overlap in sequence allows the products

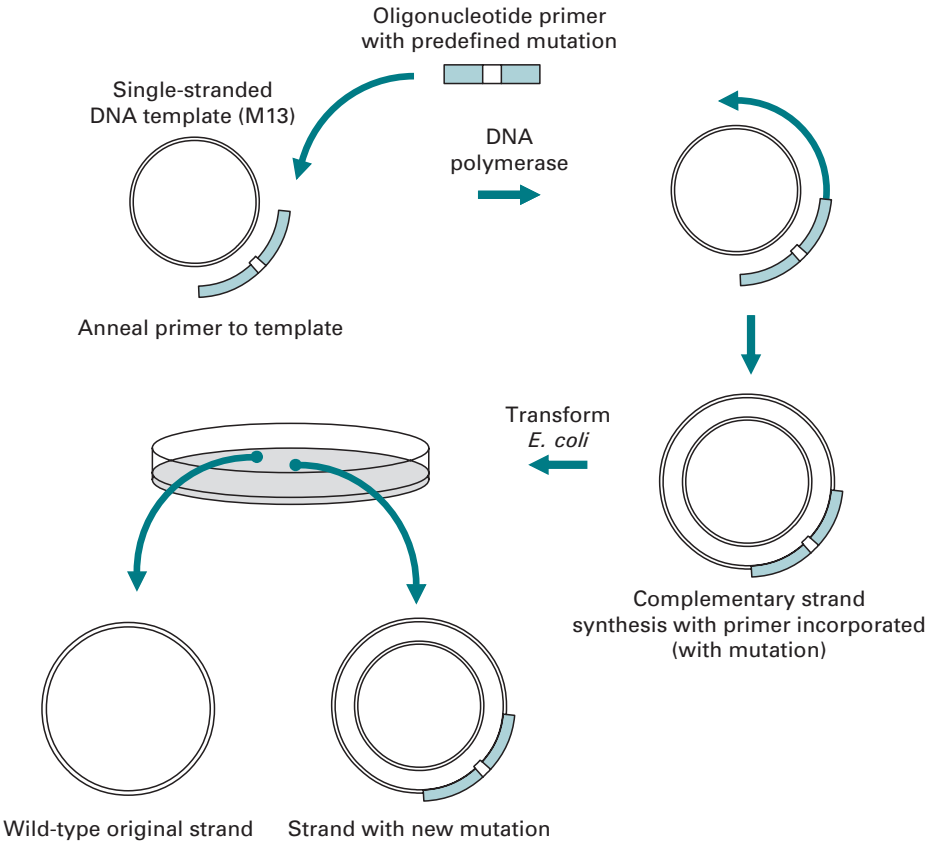

Fig. 6.32 Oligonucleotide-directed mutagenesis. This technique requires a knowledge of nucleotide sequence, since an oligonucleotide may then be synthesised with the base mutation. Annealing of the oligonucleotide to complementary (except for the mutation) single-stranded DNA provides a primer for DNA polymerase to produce a new strand and thus incorporates the primer with the mutation.

from each reaction to hybridise. Following this, one of the two hybrids bearing a free 3′ hydroxyl group is extended to produce a new duplex fragment. The other hybrid with a 5′ hydroxyl group cannot act as substrate in the reaction. Thus, the overlapped and extended product will now contain the directed mutation (Fig. 6.33). Deletions and insertions may also be created with this method although the requirements of four primers and three PCR reactions limits the general applicability of the technique. A modification of the overlap extension PCR may also be used to construct directed mutations; this is termed **megaprimer PCR**. This method utilises three oligonucleotide primers to perform two rounds of PCR. A complete PCR product, the megaprimer is made single-stranded and this is used as a large primer in a further PCR reaction with an additional primer.

The above are all methods for creating rational defined mutations as part of a design cycle system. However it is also possible to introduce random mutations into a gene and select for enhanced or new activities of the protein or enzyme it encodes. This accelerated form of artificial molecular evolution may be undertaken using

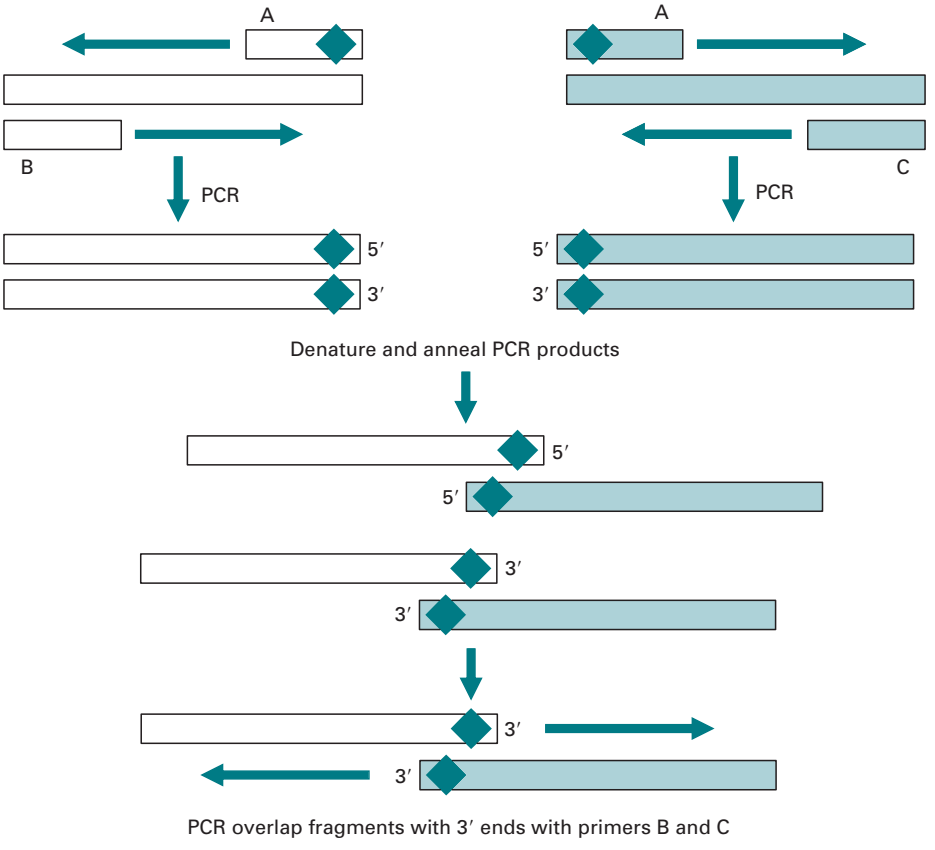

Fig. 6.33 Construction of a synthetic DNA fragment with a predefined mutation using overlap PCR mutagenesis.

error-prone PCR where deliberate and random mutations are introduced by a low-fidelity PCR amplification reaction. The resulting amplified gene is then translated and its activity assayed. This has already provided novel evolved enzymes such as a *p*-nitrobenzyl esterase which exhibits an unusual and surprising affinity for organic solvents. This accelerated evolutionary approach to protein engineering has been useful in the production of novel phage displayed antibodies and in the development of antibodies with enzymic activities (catalytic antibodies).

6.7 EXPRESSION OF FOREIGN GENES

One of the most useful applications of recombinant DNA technology is the ability to artificially synthesise large quantities of natural or modified proteins in a host cell such as bacteria or yeast. The benefits of these techniques have been enjoyed for many years since the first insulin molecules were cloned and expressed in 1982 (Table 6.3). Contamination of other proteins such as the blood product factor VIII with infectious agents has also increased the need to develop effective vectors for ***in vitro* expression**

Table 6.3 **A number of recombinant DNA-derived human therapeutic reagents**

Therapeutic area	Recombinant product
Drugs	Erythropoietin
	Insulin
	Growth hormone
	Coagulation factors (e.g. factor VIII)
	Plasminogen activator
Vaccines	Hepatitis B
Cytokines/growth factors	GM-CSF
	G-CSF
	Interleukins
	Interferons

Notes: GM-CSF, granulocyte–macrophage colony-stimulating factor; G-CSF, granulocyte colony-stimulating factor.

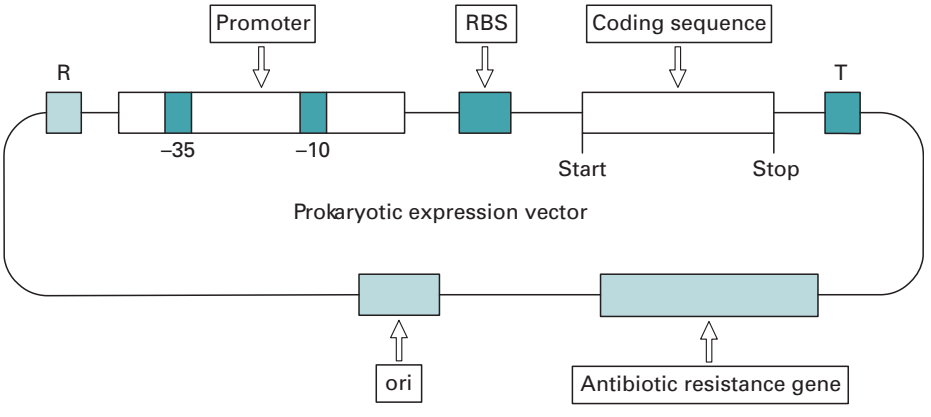

Fig. 6.34 Components of a typical prokaryotic expression vector. To produce a transcript (coding sequence) and translate it, a number of sequences in the vector are required. These include the promoter and ribosome-binding site (RBS). The activity of the promoter may be modulated by a regulatory gene (R), which acts in a way similar to that of the regulatory gene in the *lac* operon. T indicates a transcription terminator.

of foreign genes. In general the expression of foreign genes is carried out in specialised cloning vectors (Fig. 6.34). However it is possible to use cell-free transcription and translation systems that direct the synthesis of proteins without the need grow and maintain cells. *In vitro* translation is carried out with the appropriate amino acids, ribosomes, tRNA molecules and isolated mRNA fractions. Wheat germ extracts or rabbit reticulocyte lysates are usually the systems of choice for *in vitro* translation. The resulting

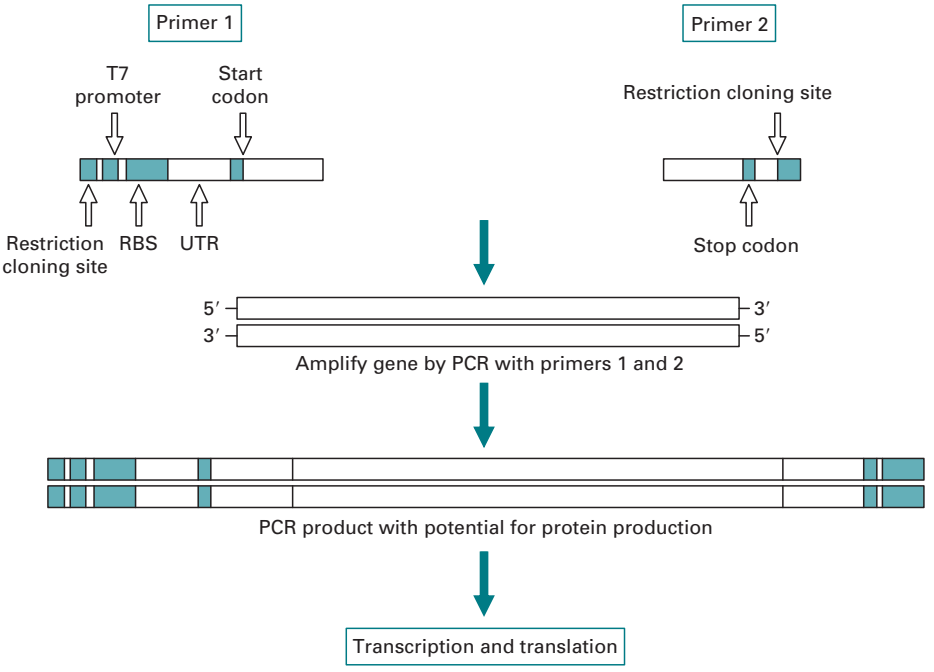

Fig. 6.35 Expression PCR (E-PCR). This technique amplifies a target sequence with one promoter that contains a transcriptional promoter, ribosome binding site (RBS), untranslated leader region (UTR) and start codon. The other primer contains a stop codon. The amplified PCR products may be used in transcription and translation to produce a protein.

proteins may be detected by polyacrylamide gel electrophoresis or by immunological detection using **western blotting**. Recently oligonucleotide PCR primers have been designed to incorporate a promoter for RNA polymerase and a ribosome-binding site. When the so-called **expression PCR** (E-PCR) is carried out the amplified products are denatured and transcribed by RNA polymerase after which they are translated *in vitro*. The advantage of this system is that large amounts of specific RNA are synthesised thus increasing the yield of specific proteins (Fig. 6.35).

6.7.1 Production of fusion proteins

For a foreign gene to be expressed in a bacterial cell, it must have particular promoter sequences upstream of the coding region, to which the RNA polymerase will bind prior to transcription of the gene. The choice of promoter is vital for correct and efficient transcription since the sequence and position of promoters are specific to a particular host such as *E. coli* (Section 5.5.4). It must also contain a ribosome-binding site, placed just before the coding region. Unless a cloned gene contains both of these sequences, it will not be expressed in a bacterial host cell. If the gene has been produced via cDNA from a eukaryotic cell, then it will certainly not have any such sequences. Consequently, expression vectors have been developed which contain promoter and ribosome-binding sites positioned just before one or more restriction

sites for the insertion of foreign DNA. These regulatory sequences, such as that from the *lac* operon of *E. coli*, are usually derived from genes that, when induced, are strongly expressed in bacteria. Since the mRNA produced from the gene is read as triplet codons, the inserted sequence must be placed so that its reading frame is in phase with the regulatory sequence. This can be ensured by the use of three vectors which differ only in the number of bases between promoter and insertion site, the second and third vectors being respectively one and two bases longer than the first. If an insert is cloned in all three vectors then in general it will subsequently be in the correct reading frame in one of them. The resulting clones can be screened for the production of a functional foreign protein (Section 6.5.4).

In some cases the protein is expressed as a fusion with a general protein such as β-galactosidase or glutathione-S-transferase (GST) to facilitate its recovery. It may also be tagged with a moiety such as a polyhistidine (6×His-Tag) which binds strongly to a nickel-chelate-nitrilotriacetate (Ni-NTA) chromatography column. The usefulness of this method is that the binding is independent of the three-dimensional structure of the 6×His-tag and so recovery is efficient even under strong denaturing conditions, often required for membrane proteins and inclusion bodies (Fig. 6.36). The tags are subsequently removed by cleavage with a reagent such as cyanogen bromide and the protein of interest purified by protein biochemical methods such as chromatography and polyacrylamide gel electrophoresis.

It is not only possible, but usually essential, to use cDNA instead of a eukaryotic genomic DNA to direct the production of a functional protein by bacteria. This is because bacteria are not capable of processing RNA to remove introns, and so any foreign genes must be pre-processed as cDNA if they contain introns. A further problem arises if the protein must be glycosylated, by the addition of oligosaccharides at specific sites, in order to become functional. Although the use of bacterial expression systems is somewhat limited for eukaryotic systems there are a number of eukaryotic expression systems based on plant, mammalian, insect and yeast cells. These types of cells can perform such post-translational modifications, producing a correct glycosylation pattern and in some cases the correct removal of introns. It is also possible to include a signal or address sequence at the 5′ end of the mRNA which directs the protein to a particular cellular compartment or even out of the cell altogether into the supernatant. This makes the recovery of expressed recombinant proteins much easier since the supernatant may be drawn off while the cells are still producing protein.

One useful eukaryotic expression system is based on the monkey COS cell line. These cells each contain a region derived from a mammalian monkey virus termed simian virus 40 (SV40). A defective region of the SV40 genome has been stably integrated into the COS cell genome. This allows the expression of a protein termed the large T antigen which is required for viral replication. When a recombinant vector having the SV40 origin of replication and carrying foreign DNA is inserted into the COS cells viral replication takes place. This results in a high level expression of foreign proteins. The disadvantage of this system is the ultimate lysis of the COS cells and limited insert capacity of the vector. Much interest is also currently focussed on other modified viruses, vaccinia virus and baculovirus. These have been developed for high-level expression in mammalian cells and insect cells respectively. The vaccinia virus

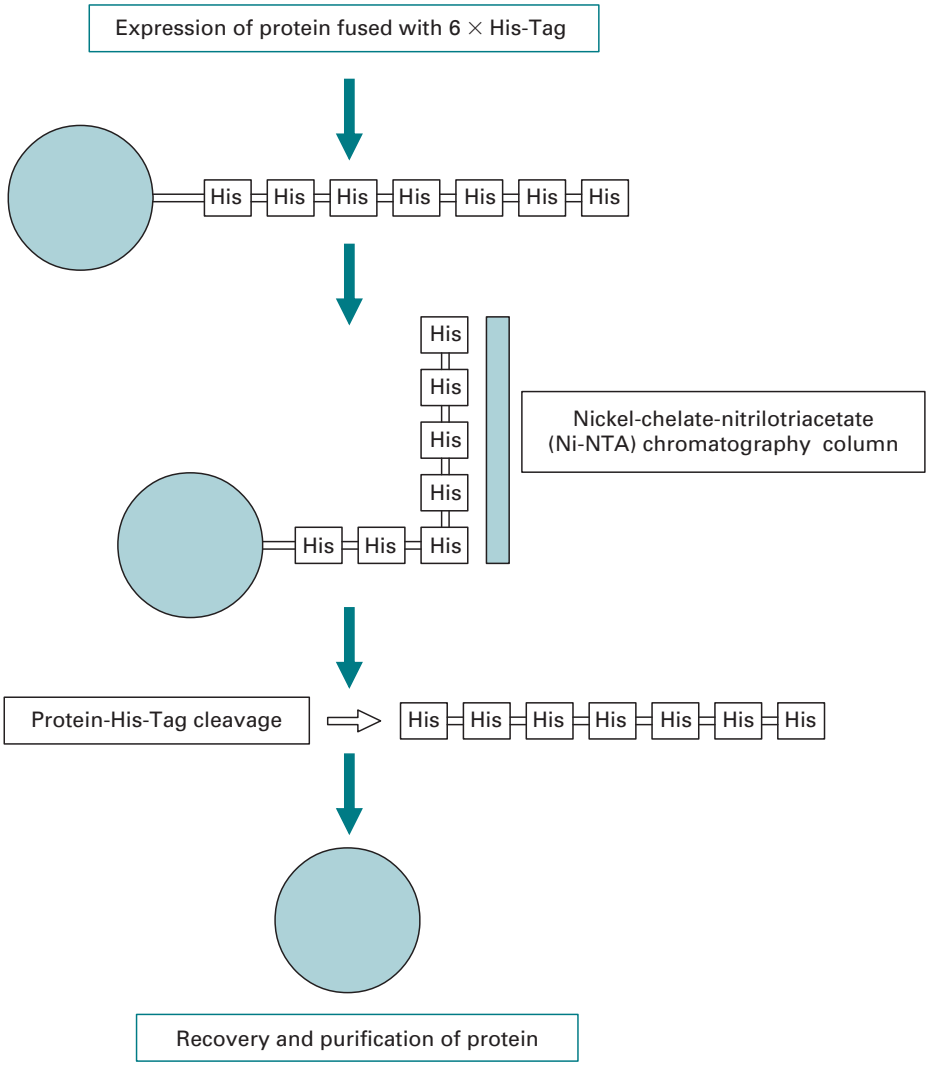

Fig. 6.36 Recovery of proteins using (6 × His-Tag) and (Ni-NTA) chromatography columns.

in particular has been used to correct the defective ion transport by introducing a wild-type cystic fibrosis gene into cells bearing a mutated cystic fibrosis (CFTR) gene. There is no doubt that the further development of these vector systems will enhance eukaryotic protein expression in the future.

6.7.2 Phage display techniques

As a result of the production of phagemid vectors and as a means of overcoming the problems of screening large numbers of clones generated from genomic libraries of antibody genes, a method for linking the phenotype or expressed protein with the genotype has been devised. This is termed **phage display**, since a functional protein is

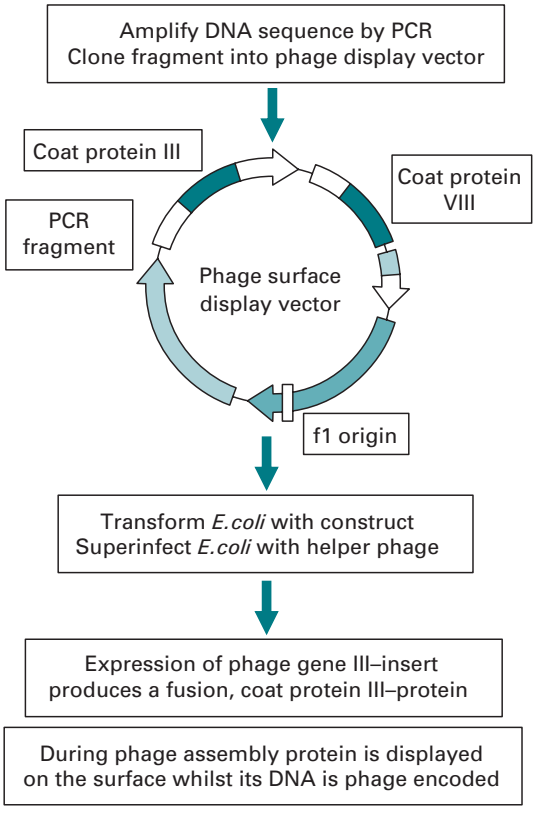

Fig. 6.37 Flow diagram indicating the main steps in the phage display technique.

linked to a major coat protein of a coliphage whilst the single-stranded gene encoding the protein is packaged within the virion. The initial steps of the method rely on the PCR to amplify gene fragments that represent functional domains or subunits of a protein such as an antibody. These are then cloned into a phage display vector which is an adapted phagemid vector (Section 6.3.3) and used to transform *E. coli*. A helper phage is then added to provide accessory proteins for new phage molecules to be constructed. The DNA fragments representing the protein or polypeptide of interest are also transcribed and translated, but linked to the major coat protein g III. Thus when the phage is assembled the protein or polypeptide of interest is incorporated into the coat of the phage and displayed, whilst the corresponding DNA is encapsulated (Fig. 6.37).

There are numerous applications for the display of proteins on the surface of bacteriophage viruses, bacteria and other organisms, and commercial organisations have been quick to exploit this technology. One major application is the analysis and production of engineered antibodies from which the technology was mainly developed. In general phage-based systems have a number of novel applications in terms of ease of selection rather than screening of antibody fragments, allowing analysis by methods such as affinity chromatography. In this way it is possible to generate large numbers of antibody heavy and light chain genes by PCR amplification and mix them in a random fashion. This **recombinatorial library** approach may allow new or novel partners to be formed

as well as naturally existing ones. This strategy is not restricted to antibodies and vast libraries of peptides may be used in this combinatorial chemistry approach to identify novel compounds of use in biotechnology and medicine.

Phage-based cloning methods also offer the advantage of allowing mutagenesis to be performed with relative ease. This may allow the production of antibodies with affinities approaching that derived from the human or mouse immune system. This may be brought about by using an error prone DNA polymerase in the initial steps of constructing a **phage display library**. It is possible that these types of libraries may provide a route to high affinity recombinant antibody fragments that are difficult to produce by more conventional hybridoma fusion techniques. Surface display libraries have also been prepared for the selection of ligands, hormones and other polypeptides in addition to allowing studies on protein–protein or protein–DNA interactions or determining the precise binding domains in these receptor–ligand interactions.

6.7.3 Alternative display systems

A number of display systems have been developed based on the original phage display technique. One interesting method is **ribosome display** where a sequence or even a library of sequences are transcribed and translated *in vitro*. However in the DNA library the sequences are fused to spacer sequences lacking a stop codon. During translation at the ribosome the protein protrudes from the ribosome and is locked in with the mRNA. The complex can be stabilised by adding salt. In this way it is possible to select the appropriate protein through binding to its ligand. Thus a high-affinity protein-ligand can be isolated which has the mRNA that originally encoded it. The mRNA may then be reverse transcribed into cDNA and amplified by PCR to allow further methods such as mutagenesis to be undertaken. A related technique, **mRNA display**, is similar except the association between the protein and mRNA is through a more stable covalent puromycin link rather than the salt-induced link as in ribosome display. Further display systems, based on yeast or bacteria, have also been developed and provide powerful *in vitro* selection methods.

6.8 ANALYSING GENES AND GENE EXPRESSION

6.8.1 Identifying and analysing mRNA

The levels and expression patterns of mRNA dictate many cellular processes and therefore there is much interest in the ability to analyse and determine levels of a particular mRNA. Technologies such as real-time or **quantitative PCR** and microchip expression arrays are currently being employed and refined for high throughput analysis. A number of other informative techniques have been developed that allow the fine structure of a particular mRNA to be analysed and the relative amounts of an RNA quantitated by non-PCR-based methods. This is important not only for gene regulation studies but may also be used as a marker for certain clinical disorders. Traditionally the Northern blot has been used for

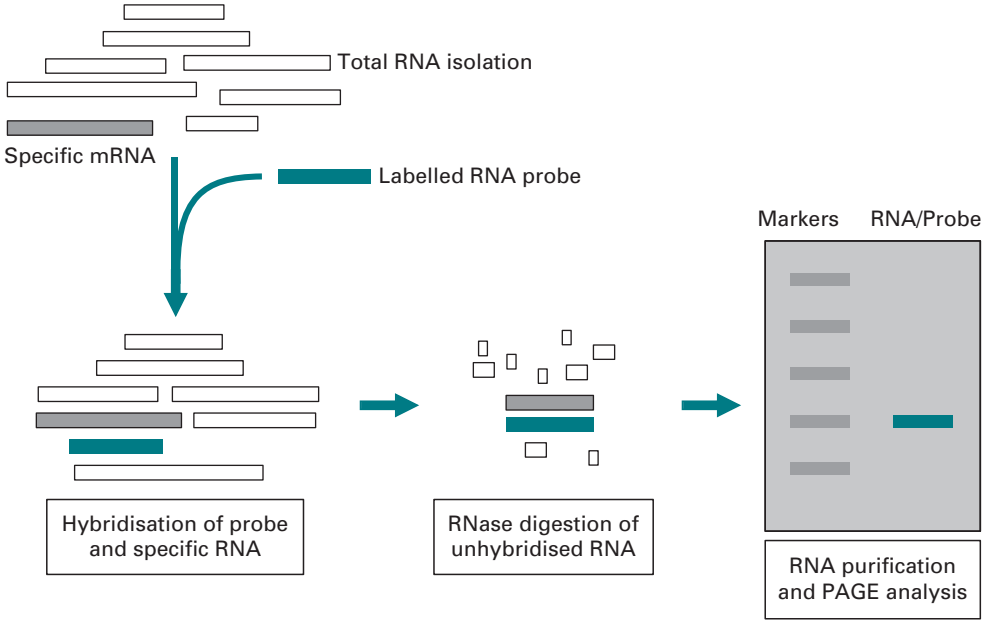

Fig. 6.38 Steps involved in the ribonuclease protection assay (RPA). PAGE, polyacrylamide gel electrophoresis.

detection of particular RNA transcripts by blotting extracted mRNA and immobilising it to a nylon membrane (Section 5.9.2). Subsequent hybridisation with labelled gene probes allows precise determination of the size and nature of a transcript. However, much use has been made of a number of nucleases that digest only single-stranded nucleic acids and not double-stranded molecules. In particular the **ribonuclease protection assay** (RPA) has allowed much information to be gained regarding the nature of mRNA transcripts (Fig. 6.38). In the RPA single-stranded mRNA is hybridised in solution to a labelled single-stranded RNA probe which is in excess. The hybridised part of the complex becomes protected whereas the unhybridised part of the probe made from RNA is digested with RNase A and RNase T1. The protected fragment may then be analysed on a high-resolution polyacrylamide gel. This method may give valuable information regarding the mRNA in terms of the precise structure of the transcript (transcription start site, intron/exon junctions, etc.). It is also quantitative and requires less RNA than a **Northern blot**. A related technique, **S1 nuclease mapping**, is similar although the unhybridised part of a DNA probe, rather than an RNA probe, is digested, this time with the enzyme S1 nuclease.

The PCR has also had an impact on the analysis of RNA via the development of a technique known as **reverse transcriptase–PCR** (RT–PCR). Here the RNA is isolated and a first strand cDNA synthesis undertaken with reverse transcriptase; the cDNA is then used in a conventional PCR (Section 6.2.5). Under certain circumstances a number of thermostable DNA polymerases have reverse transcriptase activity which obviates the need to separate the two reactions and allows the RT–PCR to be carried out in one tube. One of the main benefits of RT–PCR is the ability to identify rare or low levels of mRNA transcripts with great sensitivity. This is especially useful when

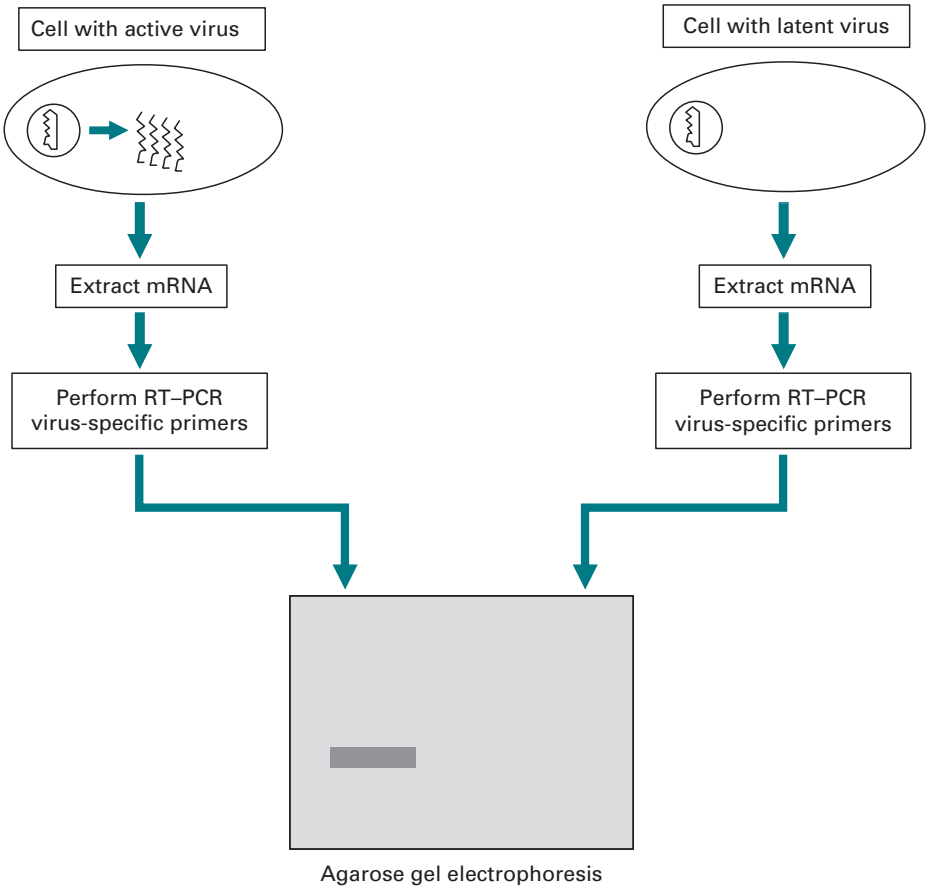

Fig. 6.39 Representation of the detection of active viruses using RT–PCR.

detecting, for example, viral gene expression and furthermore allows the means of differentiating between latent and active virus (Fig. 6.39). The level of mRNA production may also be determined by using a PCR-based method, termed quantitative PCR (Section 5.10.7).

In many cases the analysis of tissue-specific gene expression is required and again the PCR has been adapted provide a solution. This technique, termed **differential display**, is also an RT–PCR-based system requiring that isolated mRNA be first converted into cDNA. Following this, one of the PCR primers, designed to anneal to a general mRNA element such as the poly(A) tail in eukaryotic cells, is used in conjunction with a combination of arbitrary 6–7 bp primers which bind to the 5′ end of the transcripts. Consequently this results in the generation of multiple PCR products with reproducible patterns (Fig. 6.40). Comparative analysis by gel electrophoresis of PCR products generated from different cell types therefore allows the identification and isolation of those transcripts that are differentially expressed. As with many PCR-based techniques the time to identify such genes is dramatically reduced from the weeks that are required to construct and screen cDNA libraries to a few days.

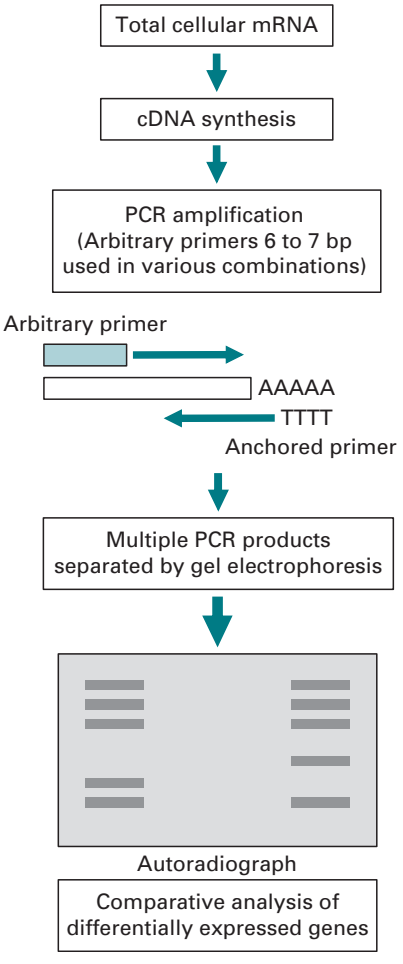

Fig. 6.40 Analysis of gene expression using differential display PCR.

6.8.2 Analysing genes *in situ*

Gross chromosomal changes are often detectable by microscopic examination of the chromosomes within a karyotype (Section 5.3). Single or restricted numbers of base substitutions, deletions, rearrangements or insertions are far less easily detectable but may induce similarly profound effects on normal cellular biochemistry. *In situ* hybridisation makes it possible to determine the chromosomal location of a particular gene fragment or gene mutation. This is carried out by preparing a radiolabelled DNA or RNA probe and applying this to a tissue or chromosomal preparation fixed to a microscope slide. Any probe that does not hybridise to complementary sequences is washed off and an image of the distribution or location of the bound probe is viewed by autoradiography (Fig. 6.41). Using tissue or cells fixed to slides it is also possible to carry out *in situ* PCR and qPCR. This is a highly sensitive technique where PCR is carried out directly on the tissue slide with the standard PCR reagents. Specially adapted thermal cycling machines are required to hold the slide preparations and allow the PCR to proceed.

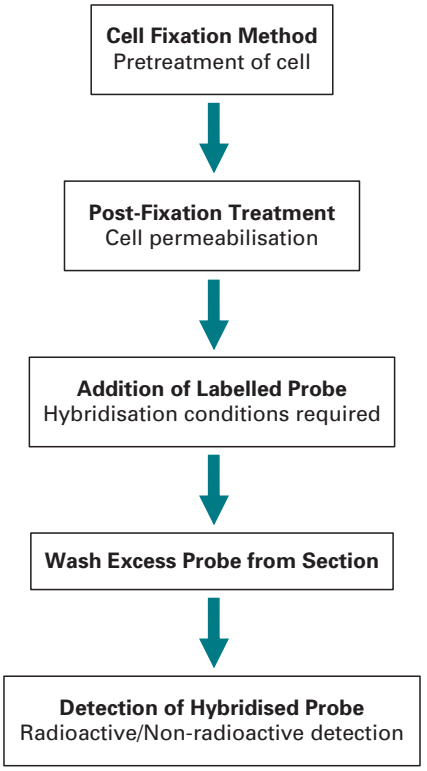

Fig. 6.41 General scheme for *in situ* hybridisation.

This allows the localisation and identification of, for example, single copies of intracellular viruses and in the case of qPCR the determination of initial concentrations of nucleic acid.

An alternative labelling strategy used in karyotyping and gene localisation is **fluorescent *in situ* hybridisation** (FISH). This method sometimes termed chromosome painting is based on *in situ* hybridisation but in which different gene probes are labelled with different fluorochromes, each specific for a particular chromosome. The advantage of this method is that separate gene regions may be identified and comparisons made within the same chromosome preparation. The technique is also likely to be highly useful in genome mapping for ordering DNA probes along a chromosomal segment (Section 6.9).

6.8.3 Analysing promoter–protein interactions

To determine potential transcriptional regulatory sequences genomic DNA fragments may be cloned into specially devised promoter probe vectors. These contain sites for insertion of foreign DNA which lies upstream of a reporter gene. A number of reporter genes are currently used, including the *lacZ* gene encoding β-galactosidase, the ***CAT* gene** encoding chloramphenicol acetyl transferase (CAT) and the *lux* gene which produces luciferase and is determined in a bioluminescent assay. Fragments of DNA potentially containing a promoter region are cloned into the vector and the constructs

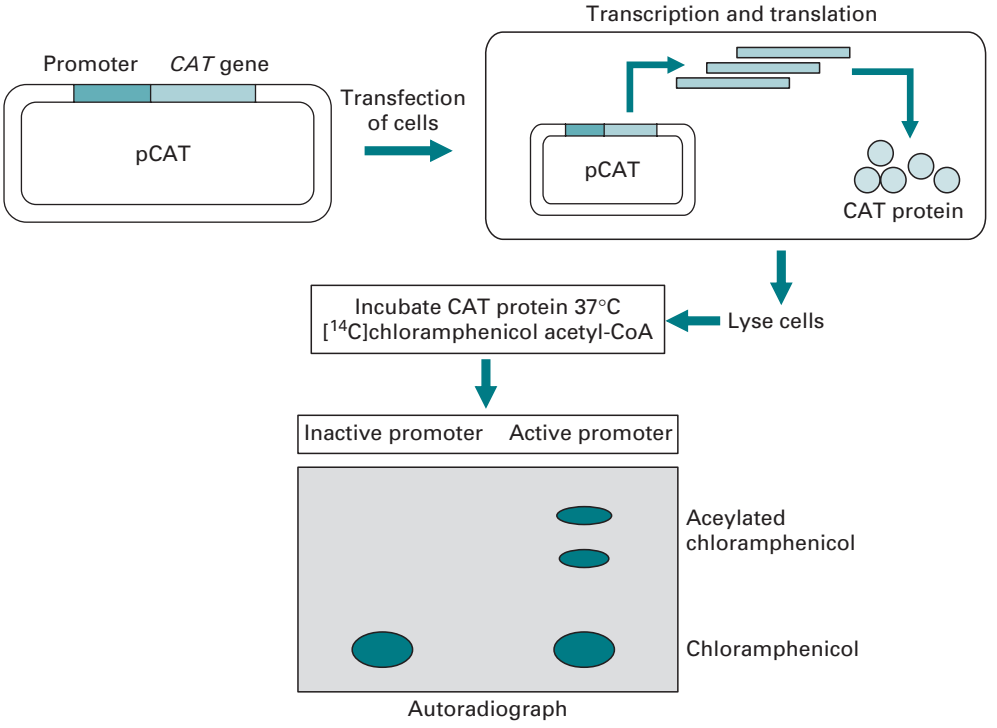

Fig. 6.42 Assay for promoters using the reporter gene for chloramphenicol acetyl transferase (CAT).

transfected into eukaryotic cells. Any expression of the reporter gene will be driven by the foreign DNA which must therefore contain promoter sequences (Fig. 6.42). These plasmids and other reporter genes such as those using **green fluorescent protein** (GFP) or the firefly luciferase gene allow quantitation of gene transcription in response to transcriptional activators.

The binding of a regulatory protein or transcription factor to a specific DNA site results in a complex that may be analysed by the technique termed **gel retardation**. Under gel electrophoresis the migration of a DNA fragment bound to a protein of a relatively large mass will be retarded in comparison to the DNA fragment alone. For gel retardation to be useful the region containing the promoter DNA element must be digested or mapped with a restriction endonuclease before it is complexed with the protein. The location of the promoter may then be defined by finding the position on the restriction map of the fragment that binds to the regulatory protein and therefore retards it during electrophoresis. One potential problem with gel retardation is the ability to define the precise nucleotide binding region of the protein, since this depends on the accuracy and detail of the restriction map and the convenience of the restriction sites. However it is a useful first step in determining the interaction of a regulatory protein with a DNA binding site.

DNA footprinting relies on the fact that the interaction of a DNA-binding protein with a regulatory DNA sequence will protect that DNA sequence from degradation by an enzyme such as DNase I. The DNA regulatory sequence is first labelled at one end with a radioactive label and then mixed with the DNA-binding protein

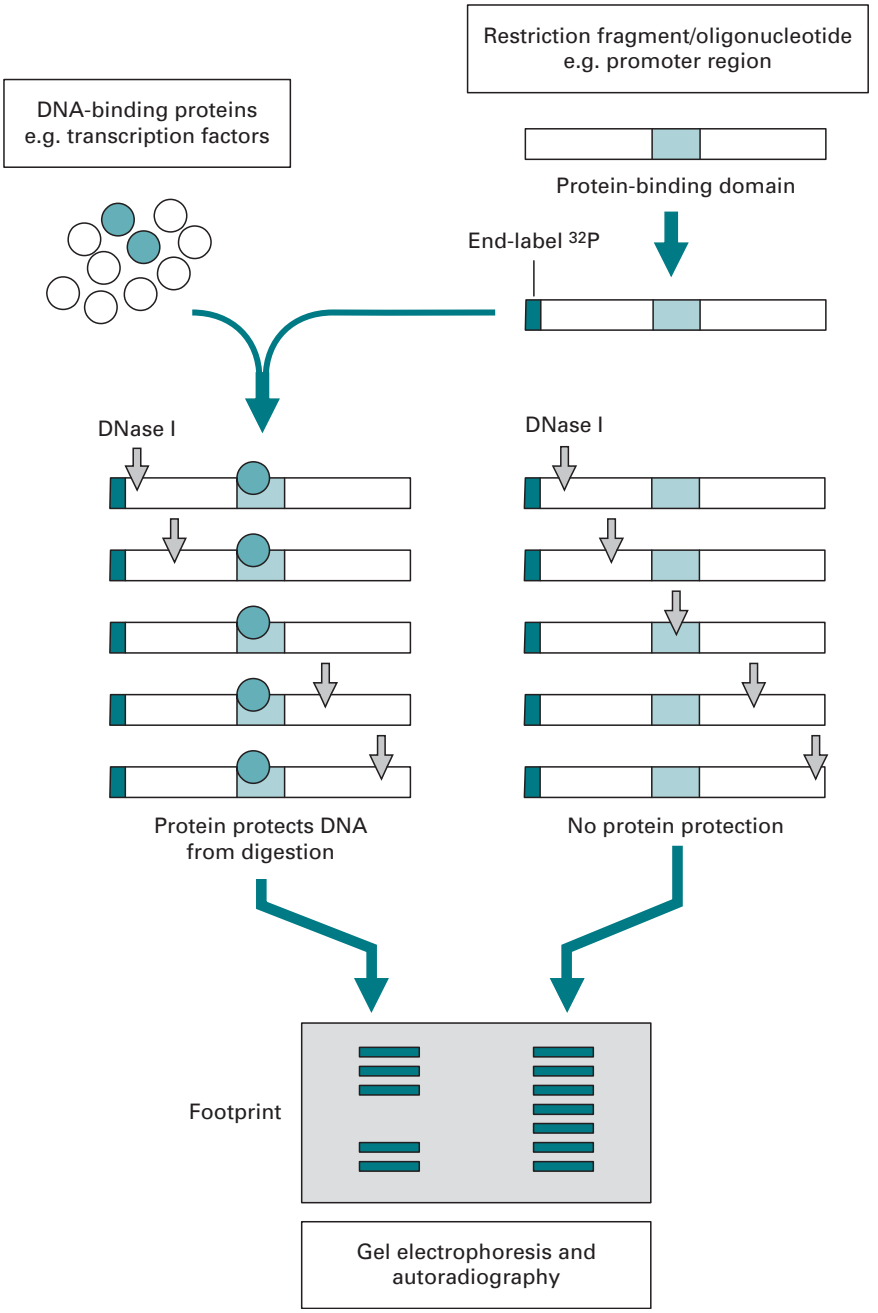

Fig. 6.43 Steps involved in DNA footprinting.

(Fig. 6.43). DNase I is added and conditions favouring a partial digestion are then carried out. This limited digestion ensures that a number of fragments are produced where the DNA is not protected by the DNA-binding protein. The region protected by the DNA-binding protein will remain undigested. All the fragments are then separated on a high-resolution polyacrylamide gel alongside a control digestion where no DNA-binding

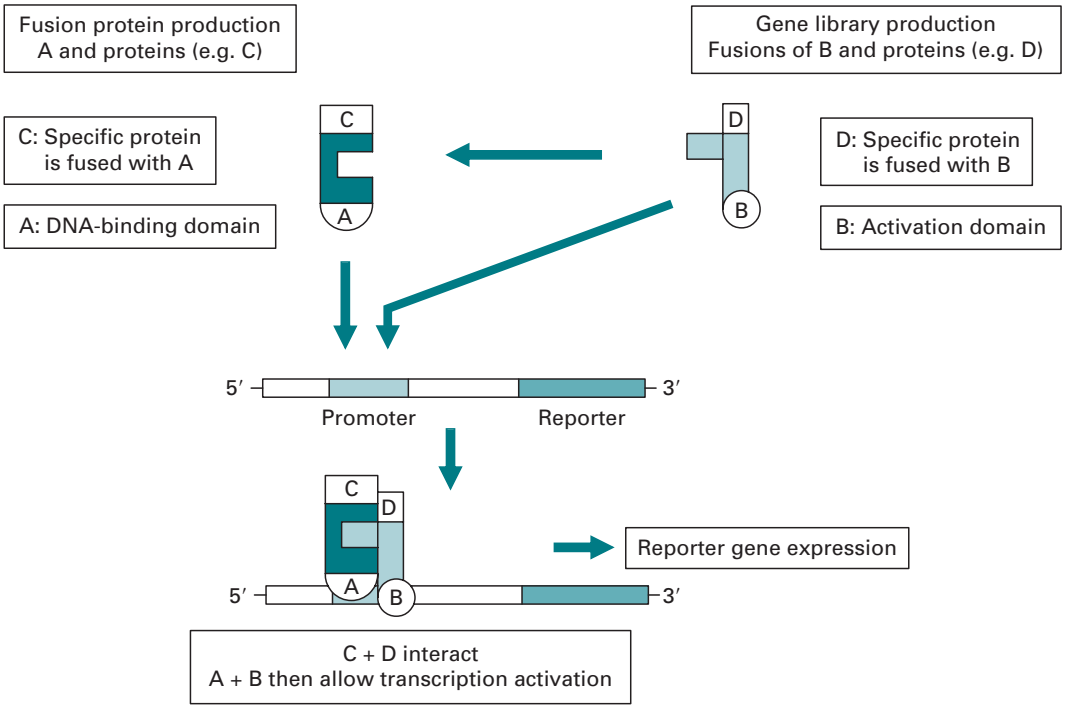

Fig. 6.44 Yeast two-hybrid system (interaction trapping technique). Transcription factors have two domains, one for DNA binding (A) and the other to allow binding to further proteins (B). Thus a recombinant molecule is formed from a protein (C) as a fusion with the DNA-binding domain. It cannot, however, activate transcription alone. Genes from a cDNA library (D) are expressed as a fusion with the activator domain (B) but also cannot initiate transcription alone. When the two fractions are mixed together, transcription is initiated if the domains are complementary and expression of a reporter gene takes place.

protein is present. The autoradiograph of a gel will contain a ladder of bands representing the partially digested fragments. Where DNA has been protected no bands appear; this region or hole is termed the DNA footprint. The position of the protein-binding sequence within the DNA may be elucidated from the size of the fragments either side of the footprint region. Footprinting is a more precise method of locating a DNA–protein interaction than gel retardation; however, it also is unable to give any information as to the precise interaction or the contribution of individual nucleotides.

In addition to the detection of DNA sequences that contribute to the regulation of gene expression an ingenious way of detecting the protein transcription factors has been developed. This is termed the **yeast two-hybrid system**. Transcription factors have two domains, one for DNA binding and the other to allow binding to further proteins (**activation domain**). These occur as part of the same molecule in natural transcription factors, for example TFIID (Section 5.5.4). However they may also be formed from two separate domains. Thus a recombinant molecule is formed encoding the protein under study as a fusion with the DNA-binding domain. It cannot however activate transcription. Genes from a cDNA library are expressed as a fusion with the activator domain; this also cannot initiate transcription. However, when the two fractions are mixed together transcription is initiated if the domains are complementary (Fig. 6.44). This is indicated

Table 6.4 **Use of transgenic mice for investigation of selected human disorders**

Gene/protein	Genetic lesion	Disorder in humans
Tyrosine kinase (TK)	Constitutive expression of gene	Cardiac hypertrophy
HIV transactivator	Expression of HIV *tat* gene	Kaposi's sarcoma
Angiotensinogen	Expression of rat angiotensinogen gene	Hypertension
Cholesterol ester transfer protein (CET protein)	Expression of *CET* gene	Atherosclerosis
Hypoxanthine-guanine phosphoribosyl transferase (HPRT)	Inactivation of *HPRT* gene	HPRT deficiency

by the transcription of a **reporter gene** such as the *CAT* gene. The technique is not just confined to transcription factors and may be applied to any protein system where interaction occurs.

6.8.4 Transgenics and gene targeting

In many cases it is desirable to analyse the effect of certain genes and proteins in an organism rather than in the laboratory. Furthermore the production of pharmaceutical products and therapeutic proteins is also desirable in a whole organism. This also has important consequences for the biotechnology and agricultural industry (Section 6.10) (Table 6.4). The introduction of foreign genes into germ line cells and the production of an altered organism is termed **transgenics**. There are two broad strategies for transgenesis. The first is **direct transgenesis** in mammals whereby recombinant DNA is injected directly into the male pronucleus of a recently fertilised egg. This is then raised in a foster mother animal resulting in an offspring that is all transgenic. **Selective transgenesis** is where the recombinant DNA is transferred into **embryo stem** (ES) cells. The cells are then cultured in the laboratory and those expressing the desired protein selected and incorporated into the inner cell mass of an early embryo. The resulting transgenic animal is raised in a foster mother but in this case the transgenic animal is a mosaic or chimeric since only a small proportion of the cells will be expressing the protein. The initial problem with both approaches is the random nature of the integration of the recombinant DNA into the genome of the egg or embryo stem cells. This may produce proteins in cells where it is not required or disrupt genes necessary for correct growth and development.

A refinement of this however is **gene targeting** which involves the production of an altered gene in an intact cell, a form of *in vivo* mutagenesis as opposed to *in vitro* mutagenesis (Section 6.6.2). The gene is inserted into the genome of, for example, an ES cell by specialised viral-based vectors. The insertion is non-random, however, since homologous sequences exist on the vector to the gene and on the gene to be targeted. Thus, **homologous recombination** may introduce a new genetic property to the cell, or inactivate an already existing one, termed **gene knockout**. Perhaps the most important aspect

of these techniques is that they allow animal models of human diseases to be created. This is useful since the physiological and biochemical consequences of a disease are often complex and difficult to study impeding the development of diagnostic and therapeutic strategies.

6.8.5 Modulating gene expression by RNAi

There are a number of ways of experimentally changing the expression of genes. Traditionally methods have focussed on altering the levels of mRNA by manipulation of promoter sequences or levels of accessory proteins involved in control of expression. In addition post-mRNA production methods have also been employed such as antisense RNA, where a nucleic acid sequence complementary to an expressed mRNA is delivered into the cell. This antisense sequence binds to the mRNA and prevents its translation. A development of this theme and a process that is found in a variety of normal cellular processes is termed **RNA interference** (RNAi) and uses microRNA. Here a number of techniques have been developed that allow the modulation of gene expression in certain cells. This type of cellular-based gene expression modulation will no doubt extend to many organisms in the next few years.

6.8.6 Analysing genetic mutations

There are several types of mutations that can occur in nucleic acids, either transiently or those that are stably incorporated into the genome. During evolution, mutations may be inherited in one or both copies of a chromosome, resulting in polymorphisms within the population (Section 5.3). Mutations may potentially occur at any site within the genome; however, there are several instances whereby mutations occur in limited regions. This is particularly obvious in prokaryotes, where elements of the genome (termed **hypervariable regions**) undergo extensive mutations to generate large numbers of variants, by virtue of the high rate of replication of the organisms. Similar hypervariable sequences are generated in the normal antibody immune response in eukaryotes. Mutations may have several effects upon the structure and function of the genome. Some mutations may lead to undetectable effects upon normal cellular functions, termed conservative mutations. An example of these are mutations that occur in intron sequences and therefore play no part in the final structure and function of the protein or its regulation. Alternatively, mutations may result in profound effects upon normal cell function such as altered transcription rates or on the sequence of mRNAs necessary for normal cellular processes.

Mutations occurring within exons may alter the amino acid composition of the encoded protein by causing amino acid substitution or by changing the reading frame used during translation. These point mutations were traditionally detected by Southern blotting or, if a convenient restriction site was available, by **restriction fragment length polymorphism** (RFLP) (Section 5.9). However, the PCR has been used to great effect in mutation detection since it is possible to use **allele-specific oligonucleotide PCR** (ASO–PCR) where two competing primers and one general primer are used in the reaction (Fig. 6.45). One of the primers is directly complementary to the known point mutation whereas the other is a wild-type primer; that is, the primers are identical

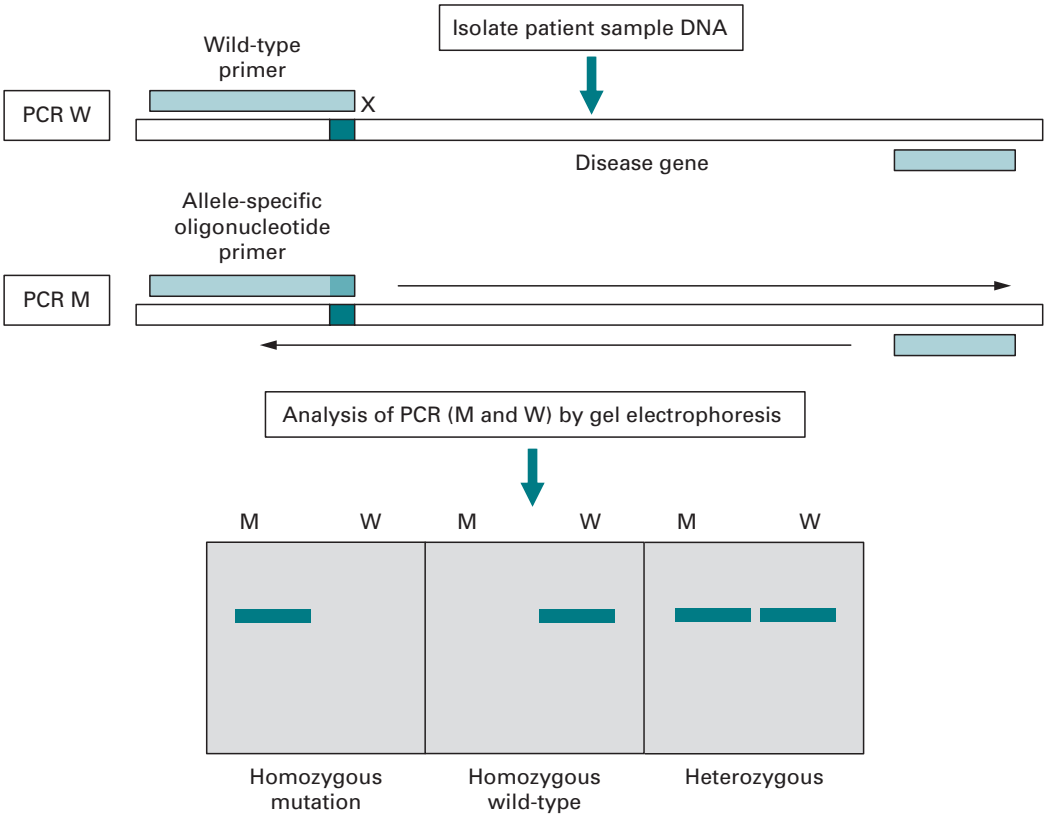

Fig. 6.45 Point mutation detection using allele-specific oligonucleotide PCR (ASO–PCR).

except for the terminal 3′ end base. Thus, if the DNA contains the point mutation only the primer with the complementary sequence will bind and be incorporated into the amplified DNA, whereas if the DNA is normal the wild-type primer is incorporated. The results of the PCR are analysed by agarose gel electrophoresis. A further modification of ASO–PCR has been developed where the primers are each labelled with a different fluorochrome. Since the primers are labelled differently a positive or negative result is produced directly without the need to examine the PCRs by gel electrophoresis.

Various modifications now allow more than one PCR to be carried out at a time (**multiplex PCR**), and hence the detection of more than one mutation is possible at the same time. Where the mutation is unknown it is also possible to use a PCR system with a gel-based detection method termed **denaturing gradient gel electrophoresis** (DGGE). In this technique a sample DNA heteroduplex containing a mutation is amplified by the PCR which is also used to attach a GC-rich sequence to one end of the heteroduplex. The mutated heteroduplex is identified by its altered melting properties through a polyacrylamide gel which contains a gradient of denaturant such as urea. At a certain point in the gradient the heteroduplex will denature relative to a perfectly matched homoduplex and thus may be identified. The **GC clamp** maintains the integrity of the end of the duplex on passage through the gel (Fig. 6.46). The sensitivity of this and other mutation detection methods has been substantially increased by the use of PCR, and further mutation

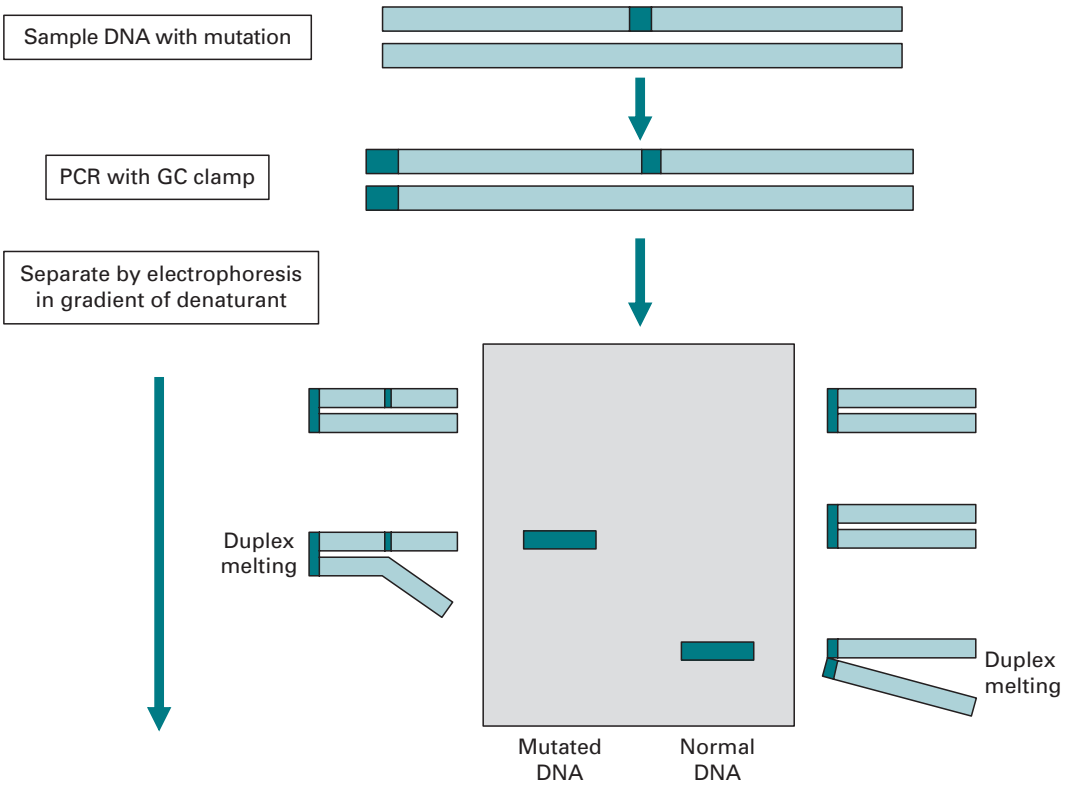

Fig. 6.46 Detection of mutations using denaturing gradient gel electrophoresis (DGGE).

techniques used to detect known or unknown mutations are indicated in Table 6.5. An extension of this principle is used in a number of detection methods employing denaturing high-performance liquid chromatography (dHPLC). Commonly known as **wave technology** the detection of denatured single strands containing mismatches is rapid allowing a high-throughput analysis of samples to be achieved.

6.8.7 Detecting DNA polymorphisms

Polymorphisms are particularly interesting elements of the human genome and as such may be used as the basis for differentiating between individuals. All humans carry repeats of sequences known as **minisatellite DNA** of which the number of repeats varies between unrelated individuals. Hybridisation of probes which anneal to these sequences using Southern blotting provides the means to type and identify those individuals (Section 5.3).

DNA fingerprinting is the collective term for two distinct genetic testing systems that use either 'multilocus' probes or 'single-locus' probes. Initially described DNA fingerprinting probes were multilocus probes and so termed because they detect **hypervariable minisatellites** throughout the genome, i.e. at multiple locations within the genome. In contrast, several single-locus probes were discovered which under

Table 6.5 **Main methods of detecting mutations in DNA samples**

Technique	Basis of method	Main characteristics of detection
Southern blotting	Gel based	Labelled probe hybridisation to DNA
Dot/slot blotting	Sample application	Labelled probe hybridisation to DNA
Allele-specific oligo-PCR (ASO–PCR)	PCR based	Oligonucleotide matching to DNA sample
Denaturing gradient gel electrophoresis (DGGE)	Gel/PCR based	Melting temperature of DNA strands
Single-stranded conformation polymorphism (SSCP)	Gel/PCR based	Conformation difference of DNA strands
Ligase chain reaction (LCR)	Gel/automated	Oligonucleotide matching to DNA sample
DNA sequencing	Gel based	Nucleotide sequence analysis of DNA
DNA microchips	Glass chip based	Sample DNA hybridisation to oligo arrays

specific conditions only detect the two alleles at a single locus and generate what have been termed DNA profiles because, unlike multilocus probes, the two-band pattern result is in itself insufficient to uniquely identify an individual.

Techniques based on the PCR have been coupled to the detection of minisatellite loci. The inherent larger size of such DNA regions was not best suited to PCR amplification; however, new PCR developments are beginning to allow this to take place. The discovery of polymorphisms within the repeating sequences of minisatellites has led to the development of a PCR-based method that distinguishes an individual on the basis of the random distribution of repeat types along the length of a person's two alleles for one such minisatellite. Known as **minisatellite variant repeat** (MVR) analysis or **digital DNA typing**, this technique can lead to a simple numerical coding of the repeat variation detected. Potentially this combines the advantages of PCR sensitivity and rapidity with the discriminating power of minisatellite alleles. Thus for the future there are a number of interesting identification systems under development and evaluation. Techniques for genetic detection of polymorphisms have been used in many cases of paternity testing and immigration control, and are becoming central factors in many criminal investigations. They are also valuable tools in plant biotechnology for cereal typing and in the field of pedigree analysis and animal breeding.

6.8.8 Microarrays and DNA microchips

One firmly established area under rapid development in molecular biology is the use of microarrays or **DNA microchips**. These provide a radically different approach to current laboratory molecular biology research strategies in that large-scale analysis and quantification of genes and gene expression is possible simultaneously. A microarray consists of an ordered arrangement of potentially hundreds of thousands of DNA sequences such

as oligonucleotides or cDNAs deposited onto a solid surface. The solid support may be either glass or silicon and currently the arrays are synthesised on or off the chip. They require complex fabrication methods similar to that used in producing computer microchips. Most commercial productions employ robotic ultrafine microarray deposition instruments which dispense volumes in the picolitre range. Alternatively on-chip fabrication as used by Affymetrix builds up layers of nucleotides using a process borrowed from the computer industry termed photolithography. Here wafer-thin masks with holes allow photoactivation of specific dNTPs which are linked together at specific regions on the chip. The whole process allows layers of oligonucleotides to be built up with each nucleotide at each position being defined by computer.

The arrays themselves may represent a variety of nucleic acid material. This may be mRNA produced in a particular cell type, termed **cDNA expression arrays**, or may alternatively represent coding and regulatory regions of a particular gene or group of genes. A number of arrays are now available that may determine mutations in DNA, mRNA transcript levels or other polymorphisms such as SNPs. Sample DNA is placed on the array and any unhybridised DNA washed off. The array is then analysed and scanned for patterns of hybridisation by detection of fluorescence signals. Any mutations or genetic polymorphisms in relevant genes may be rapidly analysed by computer interpretation of the resulting hybridisation pattern and mutation, transcript level or polymorphism defined. Indeed the collation and manipulation of data from microarrays presents as big a problem as fabricating the chips in the first place. The potential of microarrays appears to be limitless and a number of arrays have been developed for the detection of various genetic mutations including the cystic fibrosis CFTR gene (cystic fibrosis transmembrane regulator), the breast cancer gene BRCA1 and in the study of the human immunodeficiency virus (HIV).

At present microarrays require DNA to be highly purified, which limits their applicability. However as DNA purification becomes automated and microarray technology develops it is not difficult to envisage numerous laboratory tests on a single DNA microchip. This could not only be used for analysing single genes but large numbers of genes or DNA representing microorganisms, viruses, etc. Since the potential for quantitation of gene transcription exists expression arrays could also be used in defining a particular disease status. This technique may be very significant since it will allow large amounts of sequence information to be gathered very rapidly and assist in many fields of molecular biology, especially in large genome sequencing projects or in so-called **resequencing** projects where gene regions such as those containing potentially important polymorphisms require analysis in a number of samples.

One current application of microarray technology is the generation of a catalogue of SNPs across the human genome. Estimates indicate that there are approximately 10 million SNPs and importantly 200 000 coding or cSNPs that lie within genes and may point to the development of certain diseases. SNP analysis is therefore clearly a candidate for microarray analysis and developments such as Affymetrix Genome Wide SNP array enables the simultaneous analysis of nearly 1 million SNPs on one gene chip. In order to simplify the problem of the vast numbers of SNPs that need to be analysed the HapMap project currently analyses SNPs that are inherited as a block, and in theory as few as 500 000 SNPs will be required to genotype an individual.

An extension of microarray technology may also be used to analyse tissue sections. This process, termed tissue microarrays (TMA), uses tissue cores or biopsies from conventional paraffin-embedded tissues. Thousands of tissue cores are sliced and placed on a solid support such as glass where they may all be subjected to the same immunohistochemical staining process or analysis with gene probes using *in situ* hybridisation. As with DNA microarrays many samples may be analysed simultaneously, less tissue is required and greater standardisation is possible.

6.9 ANALYSING WHOLE GENOMES

Perhaps the most ambitious project in biosciences is the initiative to map and completely sequence a number of genomes from various organisms. The mapping and sequencing of a number of organisms indicated in Table 6.6. has been completed and many more are due for completion. A number have been completed already such as the bacterium *E. coli*. The demands of such large-scale mapping and sequencing have provided the impetus for the development and refinement of even the most standard of molecular biology techniques such as DNA sequencing. It has also led to new methods of identifying the important coding sequences that represent proteins and enzymes. The use of bioinformatics to collate, annotate and publish the information on the World Wide Web has also been an enormous undertaking. The availability of an informative map of the human genome that may be analysed and studied in detail chromosome by chromosome, such as the Map Viewer (NCBI), is just one of the rapid developments in the field of genome analysis and bioinformatics. Such is the power and ease of use of resources such as these that it is now inconceivable to work without these resources.

6.9.1 Physical genome mapping

In terms of genome mapping a physical map is the primary goal. Genetic linkage maps have also been produced by determining the recombination frequency between two particular loci. YAC-based vectors essential for large-scale cloning contain DNA inserts that are on average 300 000 bp in length, which is longer by a factor of ten than the longest inserts in the clones used in early mapping studies. The development of vectors with large insert capacity has enable the production of contigs. These are continuous overlapping cloned fragments that have been positioned relative to one another. Using these maps any cloned fragment may be identified and aligned to an area in one of the contig maps. In order to position cloned DNA fragments resulting from the construction of a library in a YAC or cosmid it is necessary to detect overlaps between the cloned DNA fragments. Overlaps are created because of the use of partial digestion conditions with a particular restriction endonuclease when constructing the libraries. This ensures that when each DNA fragment is cloned into a vector it has overlapping ends which theoretically may be identified and the clones positioned or ordered so that a physical map may be produced (Fig. 6.47).

In order to position the overlapping ends it is preferable to undertake DNA sequencing; however, due to the impracticality of this approach a fingerprint of each clone is

Table 6.6 **Current selected genome-sequencing projects**

Organism		Genome size (Mb)
Bacteria	*Escherichia coli*	4.6
Yeast	*Saccharomyces cerevisiae*	14
Roundworm	*Caenorhabditis elegans*	100
Fruit fly	*Drosophila melanogaster*	165
Puffer fish	*Fugu rubripes rubripes*	400
Mouse	*Mus musculus*	3000

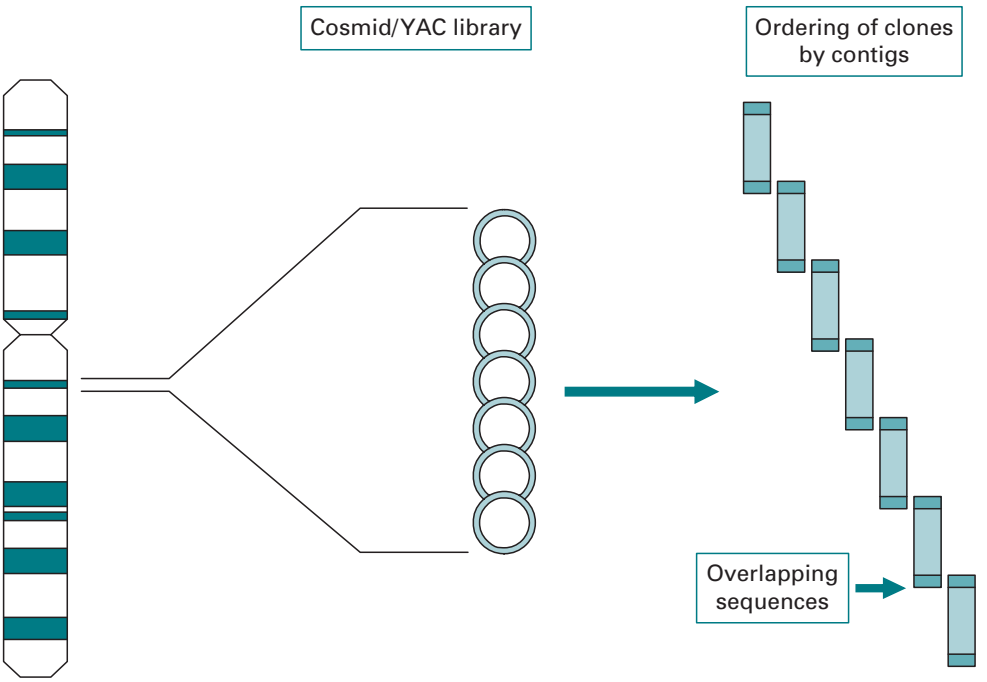

Fig. 6.47 Physical mapping using continuous overlapping cloned fragments (contigs). In order to assign the position of cloned DNA fragments resulting from the construction of a library in a YAC or cosmid vector, overlaps are detected between the clone fragments. These are created because of the use of partial digestion conditions when the libraries are constructed.

carried out by using restriction enzyme mapping. Although this is not an unambiguous method of ordering clones it is useful when also applying statistical probabilities of the overlap between clones. In order to link the contigs techniques such as *in situ* hybridisation may be used or a probe generated from one end of a contig in order to screen a different disconnected contig. This method of probe production and identification is termed **walking**, and has been used successfully in the production of physical maps

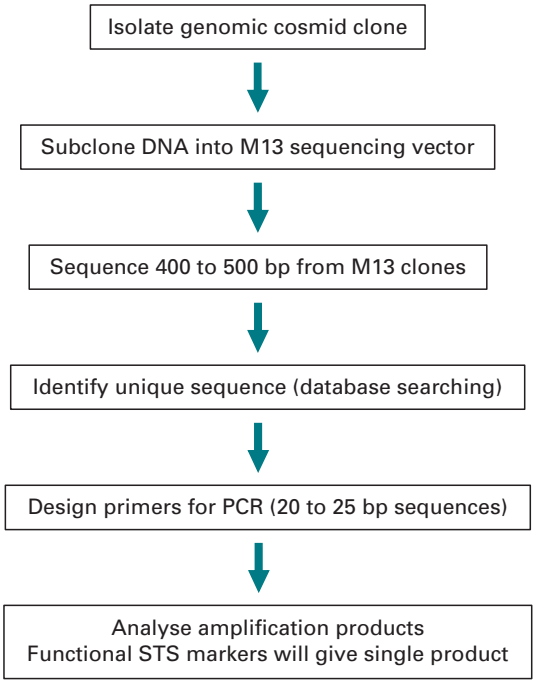

Fig. 6.48 General scheme of the production of a functional STS marker.

of *E. coli* and yeast genomes. This cycle of clone to fingerprint to contig is amenable to automation; however, the problem of closing the gaps between contigs remains very difficult.

In order to define a common way for all research laboratories to order clones and connect physical maps together an arbitrary molecular technique based on the polymerase chain reaction has been developed based on **sequence-tagged sites** (STS). This is a small unique sequence between 200–300 bp that is amplified by PCR (Fig. 6.48). The uniqueness of the STS is defined by the PCR primers that flank the STS. A PCR with those primers is performed and if the PCR results in selected amplification of target region it may be defined as a potential STS marker. In this way defining STS markers that lie approximately 100 000 bases apart along a contig map allows the ordering of those contigs. Thus, all groups working with clones have definable landmarks with which to order clones produced in their libraries.

An STS that occurs in two clones will overlap and thus may be used to order the clones in a contig. Clones containing the STS are usually detected by Southern blotting where the clones have been immobilised on a nylon membrane. Alternatively a library of clones may be divided into pools and and each pool PCR screened. This is usually a more rapid method of identifying an STS within a clone and further refinement of the PCR-based screening method allows the identification of a particular clone within a pool (Fig. 6.49). STS elements may also be generated from variable regions of the genome to produce a polymorphic marker that may be traced through families along with other DNA markers and located on a genetic linkage map. These

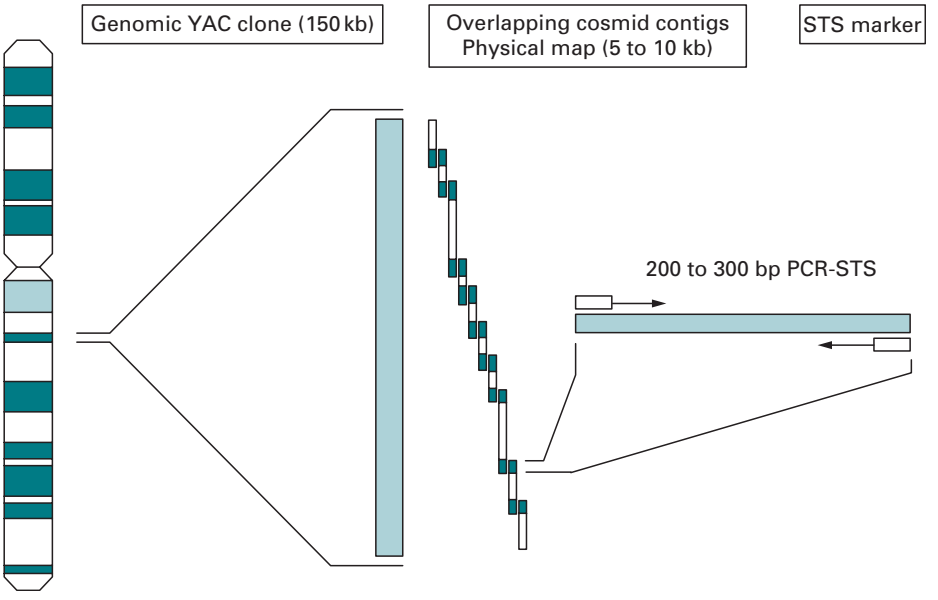

Fig. 6.49 The derivation of an STS marker. An STS is a small unique sequence of between 200 and 300 bp that is amplified by PCR and allows ordering along a contig map. Such sequences are definable landmarks with which to order clones produced in genome libraries and usually lie approximately 100 000 bp apart.

polymorphic STSs are useful since they may serve as markers on both a physical map and a genetic linkage map for each chromosome and therefore provide a useful marker for aligning the two types of map.

6.9.2 Gene discovery and localisation

A number of disease loci have been identified and located to certain chromosomes. This has been facilitated by the use of *in situ* mapping techniques such as FISH. In fact a number of genes have been identified and the protein determined where little was initially known about the gene except for its location. This method of gene discovery is known as positional cloning and was instrumental in the isolation of the *CFTR* gene responsible for the disorder cystic fibrosis (Fig. 6.50).

The genes that are actively expressed in a cell at any one time are estimated to be as little as 10% of the total. The remaining DNA is packaged and serves an as yet unknown function. Investigations have found that certain active genes may be identified by the presence of so-called **HTF** (*Hpa*II tiny fragments) islands often found at the 5′ end of genes. These are CpG-rich sequences that are not methylated and form tiny fragments on digestion with the restriction enzyme *Hpa*II. A further gene discovery method that has been used extensively in the past few years is a PCR-based technique giving rise to a product termed an **expressed sequence tag** (EST). This represents part of a putative gene for which a function has yet to be assigned. It is carried out on cDNA by using primers that bind to an anchor sequence such as a poly(A) tail and primers which bind to sequences at the 5′ end of the gene. Such PCRs may

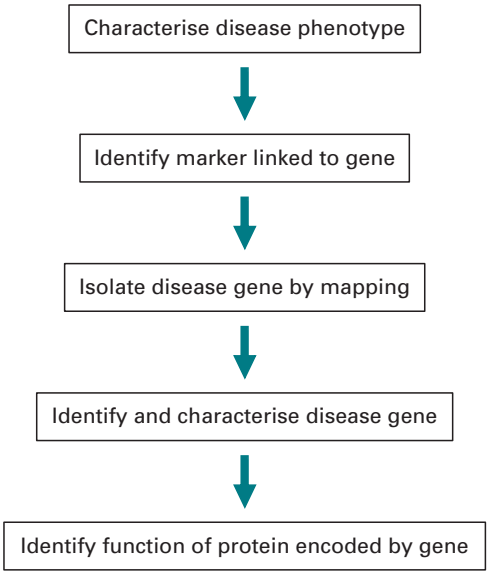

Fig. 6.50 The scheme of identification of a disease gene by positional cloning.

subsequently be used to map the putative gene to a chromosomal region or be used itself as a probe to search a genomic DNA library for the remaining parts of the gene. This type of information can be visualised using bioinformatics and useful information determined in a process termed data mining. Much interest currently lies in ESTs since they may represent a short cut to gene discovery.

A further gene isolation system that uses adapted vectors, termed **exon trapping** or **exon amplification**, may be used to identify exon sequences. Exon trapping requires the use of a specialised expression vector that will accept fragments of genomic DNA containing sequences for splicing reactions to take place. Following transfection of a eukaryotic cell line a transcript is produced that may be detected by using specific primers in a RT–PCR. This indicates the nature of the foreign DNA by virtue of the splicing sequences present. A list of further techniques that aid in the identification of a potential gene-encoding sequence is indicated in Table 6.7.

6.9.3 Genome mapping projects

As a result of the technological advances in large-scale DNA sequencing as indicated in Chapter 5 it is now possible not only to map genomes of various species but also to determine their sequence reliably and rapidly. The genomes of hundreds of species have been determined and this is increasing each month. Sequencing and mapping of the human genome was completed ahead of schedule and has provided many new insights into gene function and gene regulation. It was also a multi-collaboration effort that engaged many scientific research groups around the world and has given rise to many scientific, technical, financial and ethical debates. One interesting issue is the sequencing of the whole genome in relation to the coding sequences. Much of the human genome appears to be non-coding and composed of repetitive sequences.

Table 6.7 **Techniques used to determine putative gene-encoding sequences**

Identification method	Main details
Zoo blotting (cross-hybridisation)	Evolutionary conservation of DNA sequences that suggest functional significance
Homology searching	Gene database searching to gene family-related sequences
Identification of CpG islands	Regions of hypomethylated CpG frequently found 5′ to genes in vertebrate animals
Identification of open reading frames (ORF) promoters/splice sites/RBS	DNA sequences scanned for consensus sequences by computer
Northern blot hybridisation	mRNA detection by binding to labelled gene probes
Exon trapping technique	Artificial RNA splicing assay for exon identification
Expressed sequence tags (ESTs)	cDNAs amplified by PCR that represent part of a gene

Notes: RBS, ribosome binding site; cDNA, complementary DNA.

Estimates indicate that as little as 10% of the genome appears to encode enzymes and proteins. Current estimates equate this to approximately 20 000 genes which are important for human cellular development and maintenance. However it is the understanding of the complete function of many of the genes and their variants coupled with their interaction that now provides a major challenge. It also points to the fact that there is an extensive use of alternative splicing where exons are essentially mixed and matched to form different mRNA and thus different proteins. The study further aims to understand and possibly provide the eventual means of treating some of the 4000 genetic diseases in addition to other diseases whose inheritance is multifactorial. In this respect there are a number of specific genome projects such as the Cancer Genome Anatomy Project (CGAP) which aims to understand the part certain mutations play in the development of tumours.

6.10 PHARMACOGENOMICS

As a result of the developments in genomics new methods of providing targeted drug treatment are beginning to be developed. This area is linked to the proposal that it is possible to identify those people who react in a specific way to drug treatment by identifying their genetic make-up. In particular SNPs may provide a key marker of potential disease development and reaction to a particular treatment. A simple example that has been known for some time is the reaction to a drug used to treat a particular type of childhood leukaemia. Successful treatment of the majority of patients may be achieved with 6-mercaptopurine. A number of patients do not respond well, but in some cases it may be fatal to administer this drug. This is now known to be due to a mutation

in the gene encoding the enzyme that metabolises the drug. Thus, it is possible to analyse patient DNA prior to administration of a drug to determine what the likely response will be. The technology to deduce a patient's genotype is already developed and indicated in Section 6.8.7. It is also now possible to analyse SNPs which may also correlate with certain disease processes in a microarray type format. This opens up the possibility that it may be possible to assign a **pharmacogenetic** profile at birth, in much the same way as blood typing for later treatment. A further possibility is the determination of likely susceptibility to a disease based on genetic information. A number of companies including the Icelandic genetics company deCode are able to provide personal genetic information based on modelling and analysis of disease genes in large population studies for certain conditions such as diabetes.

6.11 MOLECULAR BIOTECHNOLOGY AND APPLICATIONS

It is a relatively short period of time since the early 1970s when the first recombinant DNA experiments were carried out. However, huge strides have been made not only in the development of molecular biology techniques but also in their practical application. The molecular basis of disease and the new areas of genetic analysis and gene therapy hold great promise. In the past medical science relied on the measurement of protein and enzyme markers which reflected disease states. It is possible now not only to detect such abnormalities at an earlier stage using mRNA techniques but also in some cases to predict such states using genome analysis. The complete mapping and sequencing of the human genome and the development of techniques such as DNA microchips will certainly accelerate such events. Perhaps even more difficult is

Table 6.8 **General classification of oncogenes and their cellular and biochemical functions**

Oncogene	Example	Main details
G-proteins	H-K- and N-*ras*	GTP-binding protein/GTPase
Growth factors	*sis*, *nt-2*, *hst*	β-chain of platelet-derived growth factor (PDGF)
Growth factor receptors	*erbB*	Epidermal growth factor receptor (EGFR)
	fms	Colony-stimulating factor-1 receptor
Protein kinases	*abl*, *src*	Protein tyrosine kinases
	mos, *ras*	Protein serine kinases
Nucleus-located transcription factors	*mye*	DNA-binding protein
	myb	DNA-binding protein
	jun, *fos*	DNA-binding protein

Table 6.9 **A number of selected examples of targets for gene therapy**

Disorder	Defect	Gene target	Target cell
Emphysema	Deficiency (α1-AT)	α1-Antitrypsin (α1-AT)	Liver cells
Gaucher disease (storage disorder)	GC deficiency	Glucocerebrosidase	GC fibroblasts
Haemoglobinopathies	Thalassaemia	β-Globin	Fibroblasts
Lesch–Nyhan syndrome	Metabolic deficiency	Hypoxanthine guanine phosphoribosyl transferase (HPRT)	HPRT cells
Immune system disorder	Adenosine deaminase deficiency	Adenosine deaminase (ADA)	T and B cells

Table 6.10 **Current selected plant/crops modified by genetic manipulation**

Crop or plant	Genetic modification
Canola (oil seed rape)	Insect resistance, seed oil modification
Maize	Herbicide tolerance, resistance to insects
Rice	Modified seed storage protein, insect resistance
Soya bean	Tolerance to herbicide, modified seed storage protein
Tomato	Modified ripening, resistance to insects and viruses
Sunflower	Modified seed storage protein

the elucidation of diseases that are multifactorial and involve a significant contribution from environmental factors. One of the best-studied examples of this type of disease is cancer. Molecular genetic analysis has allowed a discrete set of cellular genes, termed **oncogenes**, to be defined which play key roles in such events. These genes and their proteins are also major points in the cell cycle and are intimately invloved in cell regulation. A number of these are indicated in Table 6.8. In a number of cancers well-defined molecular events have been correlated with mutations in these oncogenes and therefore in the corresponding protein. It is already possible to screen and predict the fate of some disease processes at an early stage, a point which itself raises significant ethical dilemmas. In addition to understanding cellular processes both in normal and disease states great promise is also held in drug discovery and molecular gene therapy. A number of genetically engineered therapeutic proteins and enzymes have been developed and are already having an impact on disease management. In addition the correction of disorders at the gene level (**gene therapy**) is also under way and perhaps is one of the most startling applications of molecular biology to date. A number of these developments are indicated in Table 6.9.

The production of modified crops and animals for farming and as producers of important therapeutic proteins is also one of the most exciting developments of molecular biology. This has allowed the production of modified crops, improving their resistance to environmental factors and their stability (Table 6.10). The production of transgenic animals also holds great promise for improved livestock quality, low-cost production of pharmaceuticals and disease-free or disease-resistant strains. In the future this may overcome such factors as contamination with agents such as BSE. There is no doubt that improved methods of producing livestock by whole-animal cloning will also be a major benefit. All of these developments do however require debate and the many ethical considerations that arise from them require careful consideration.

6.12 SUGGESTIONS FOR FURTHER READING

Augen, J. (2005). *Bioinformatics in the Post-Genomic Era*. Reading, MA: Addison-Wesley.
Brooker, R. J. (2005). *Genetics Analysis and Principles*, 2nd edn. McGraw-Hill.
Brown, T. A. (2006). *Gene Cloning and DNA Analysis*. Oxford, UK: Wiley–Blackwell.
Primrose, S. B. and Twyman, R. (2006). *Principles of Gene Manipulation and Genomics*. Oxford, UK: Wiley–Blackwell.
Strachan, T. and Read, A. P. (2004). *Human Molecular Genetics*, 3rd edn. Oxford, UK: Bios.
Walker, J. M. and Rapley, R. (2008). *Molecular Biomethods Handbook*, 2nd edn. Totowa, NJ: Humana Press.
Watson, J. D., Caudy, A. A., Myers, R. M. and Witkowski, J. A. (2007). *Recombinant DNA: Genes and Genomes*. San Francisco, CA: W. H. Freeman.