

Molecular biology, bioinformatics and basic techniques

5.1 INTRODUCTION

The completion of the Human Genome Project has been heralded as one of the major landmark events in science. The human genome contains the blueprint for human development and maintenance and may ultimately provide the means to understand human cellular and molecular processes in both health and disease. The genome is the full complement of DNA from an organism and carries all the information needed to specify the structure of every protein the cell can produce. The realisation that DNA lies behind all of the cell's activities led to the development of what is termed molecular biology. Rather than a discrete area of biosciences, molecular biology is now accepted as a very important means of understanding and describing complex biological processes. The development of methods and techniques for studying processes at the molecular level has led to new and powerful ways of isolating, analysing, manipulating and exploiting nucleic acids. Moreover, to keep pace with the explosion in biological information a new area termed bioinformatics has evolved and provides a vital role in current biosciences. The completion of the Human Genome Project and numerous other genome projects has allowed the continued development of new exciting areas of biological sciences such as biotechnology, genome mapping, molecular medicine and gene therapy.

In considering the potential utility of molecular biological techniques it is important to understand the basic structure of nucleic acids and gain an appreciation of how this dictates the function *in vivo* and *in vitro*. Indeed many techniques used in molecular biology mimic in some way the natural functions of nucleic acids such as replication and transcription. This chapter is therefore intended to provide an overview of the general features of nucleic acid structure and function and describe some of the basic methods used in its isolation and analysis.

5.2 STRUCTURE OF NUCLEIC ACIDS

5.2.1 Primary structure of nucleic acids

DNA and RNA are macromolecular structures composed of regular repeating polymers formed from nucleotides. These are the basic building blocks of nucleic acids and are derived from **nucleosides** that are composed of two elements: a five-membered pentose carbon sugar (2-deoxyribose in DNA and ribose in RNA) and a nitrogenous base. The carbon atoms of the sugar are designated 'prime' (1', 2', 3', etc.) to distinguish them from the carbon atoms of nitrogenous bases, of which there are two types – purines and pyrimidines. A **nucleotide**, or nucleoside phosphate, is formed by the attachment of a phosphate to the 5' position of a nucleoside by an ester linkage (Fig. 5.1). Such nucleotides can be joined together by the formation of a second ester bond by reaction between the terminal phosphate group of one nucleotide and the 3' hydroxyl of another, thus generating a 5' to 3' phosphodiester bond between adjacent sugars; this process can be repeated indefinitely to give long polynucleotide molecules (Fig. 5.2). DNA has two such polynucleotide strands; however, since each strand has both a free 5' hydroxyl group at one end and a free 3' hydroxyl at the other, each strand has a polarity or directionality. The polarities of the two strands of the molecule are in opposite directions, and thus DNA is described as an '**antiparallel**' structure (Fig. 5.3).

The purine bases (composed of fused five- and six-membered rings), **adenine** (A) and **guanine** (G) are found in both RNA and DNA, as is the pyrimidine (a single six-membered ring) **cytosine** (C). The other pyrimidines are each restricted to one type of nucleic acid: **uracil** (U) occurs exclusively in RNA, whilst **thymine** (T) is limited to DNA. Thus it is possible to distinguish between RNA and DNA on the basis of the presence of ribose and uracil in RNA, and deoxyribose and thymine in DNA. However, it is the sequence of bases, which distinguishes one DNA (or RNA) molecule from another. It is conventional to write a nucleic acid sequence starting at the 5' end of the molecule, using single capital letters to represent each of the bases, for example CCGATCT. Note that there is usually no point in including the sugar or phosphate groups, since these are identical throughout the length of the molecule. Terminal phosphate groups can, when necessary, be indicated by use of a 'p'; thus 5' pCGGATCT 3' indicates the presence of a phosphate on the 5' end of the molecule.

5.2.2 Secondary structure of nucleic acids

The two polynucleotide chains in DNA are usually found in the shape of a **right-handed double helix**, in which the bases of the two strands lie in the centre of the molecule, with the sugar-phosphate backbones on the outside. A crucial feature of this double-stranded structure is that it depends on the sequence of bases in one strand being **complementary** to that in the other. A purine base attached to a sugar residue on one strand is always hydrogen bonded to a pyrimidine base attached to a sugar residue on the other strand. Moreover, adenine (A) always pairs with

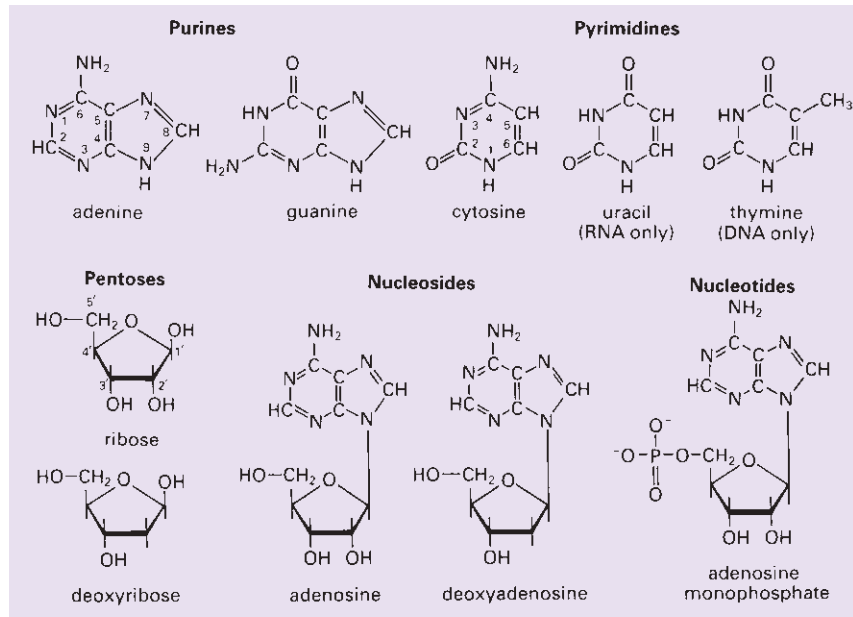


Fig. 5.1. Structure of bases, nucleosides and nucleotides.

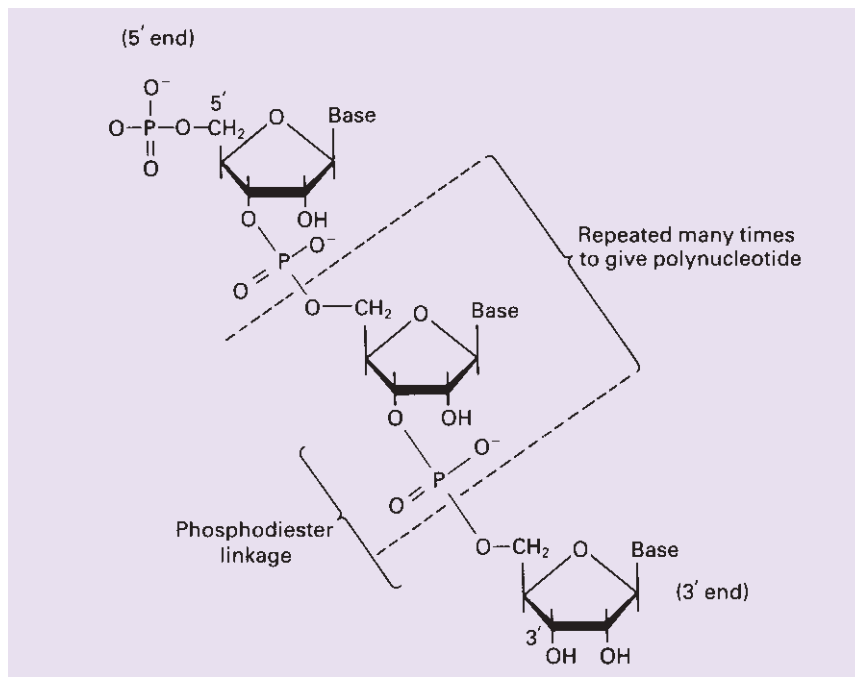


Fig. 5.2. Polynucleotide structure.

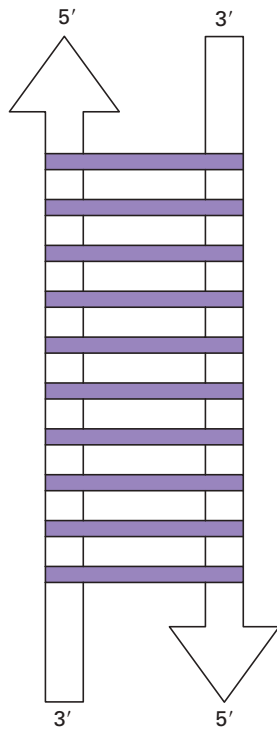


Fig. 5.3. The antiparallel nature of DNA. One strand in a double helix runs 5' to 3', whilst the other strand runs in the opposite direction 3' to 5'. The strands are held together by hydrogen bonds between the bases.

thymine (T) or uracil (U) in RNA, via two hydrogen bonds, and guanine (G) always pairs with cytosine (C) by three hydrogen bonds (Fig. 5.4). When these conditions are met a stable double-helical structure results in which the backbones of the two strands are, on average, a constant distance apart. Thus, if the sequence of one strand is known, that of the other strand can be deduced. The strands are designated as **plus** (+) and **minus** (−) and an RNA molecule complementary to the minus (−) strand is synthesised during transcription (Section 5.5.3). The base sequence may cause significant local variations in the shape of the DNA molecule and these variations are vital for specific interactions between the DNA and various proteins to take place. Although the three-dimensional structure of DNA may vary it generally adopts a double helical structure termed the B form or **B-DNA** *in vivo*. There are also other forms of right-handed DNA, such as A and C, that are formed when DNA fibres are subjected to different relative humidities (Table 5.1).

The major distinguishing feature of B-DNA is that it has approximately 10 bases for one turn of the double helix; furthermore distinctive major and minor grooves may be identified (Fig. 5.5). In certain circumstances, where repeated DNA sequences or motifs are found, the DNA may adopt a left-handed helical structure termed **Z-DNA**. This form of DNA was first synthesised in the laboratory and is

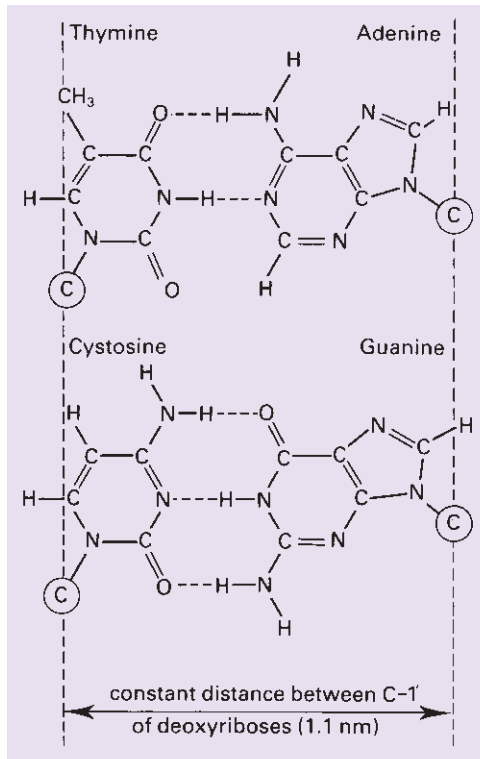


Fig. 5.4. Base-pairing in DNA. C in a circle represents carbon at the 1' position of deoxyribose.

Table 5.1 The various forms of DNA

DNA form	% humidity	Helix direction	Base/turn helix	Helix diameter (Å)
B	92%	RH	10	19
A	75%	RH	11	23
C	66%	RH	9.3	19
Z	(Pu-Py) _n	LH	12	18

RH, right-handed helix; LH, left-handed helix; Pu, Purine; Py, Pyrimidine.

Different forms of DNA may be obtained by subjecting DNA fibres to different relative humidities. The B form is the most common form of DNA whilst the A and C forms have been derived under laboratory conditions. The Z form may be produced with a DNA sequence made up from alternating purine and pyrimidine nucleotides.

thought is not to exist *in vivo*. The various forms of DNA serve to show that it is not a static molecule but dynamic and constantly in flux, and may be coiled, bent or distorted at certain times. Although RNA almost always exists as a single strand, it often contains sequences within the same strand which are **self-complementary**, and which can therefore base-pair if brought together by suitable folding of the

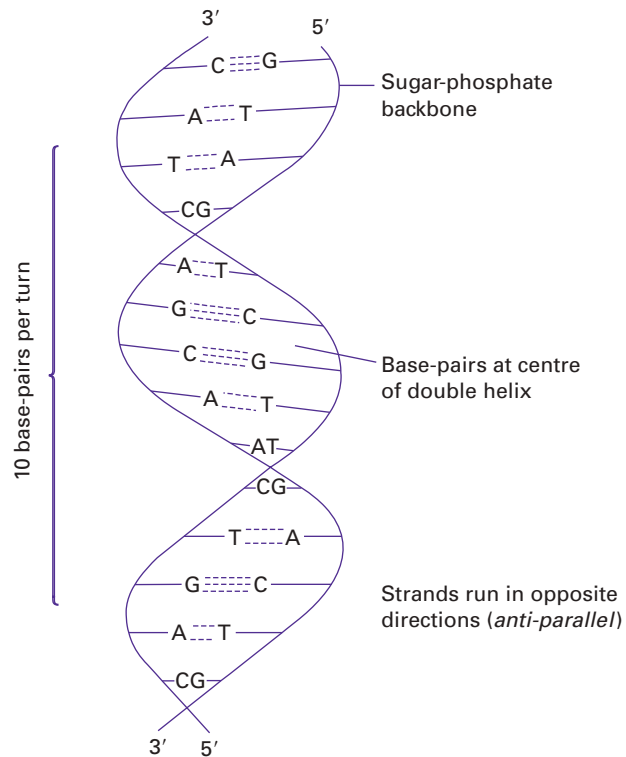


Fig. 5.5. The DNA double helix.

molecule. A notable example is transfer RNA (tRNA), which folds up to give a **clover leaf secondary structure** (Fig. 5.6).

5.2.3 Separation of double-stranded DNA

The two antiparallel strands of DNA are held together only by the weak forces of hydrogen bonding between complementary bases, and partly by hydrophobic interactions between adjacent, stacked base-pairs, termed **base-stacking**. Little energy is needed to separate a few base-pairs, and so, at any instant, a few short stretches of DNA will be opened up to the single-stranded conformation. However, such stretches immediately pair up again at room temperature, so the molecule as a whole remains predominantly double stranded.

If, however, a DNA solution is heated to approximately 90 °C or above there will be enough kinetic energy to **denature** the DNA completely, causing it to separate into single strands. This denaturation can be followed spectrophotometrically by monitoring the absorbance of light at 260 nm. The stacked bases of double-stranded DNA are less able to absorb light than the less constrained bases of single-stranded molecules, and so the absorbance of DNA at 260 nm increases as the DNA becomes denatured, a phenomenon known as the **hyperchromic effect**.

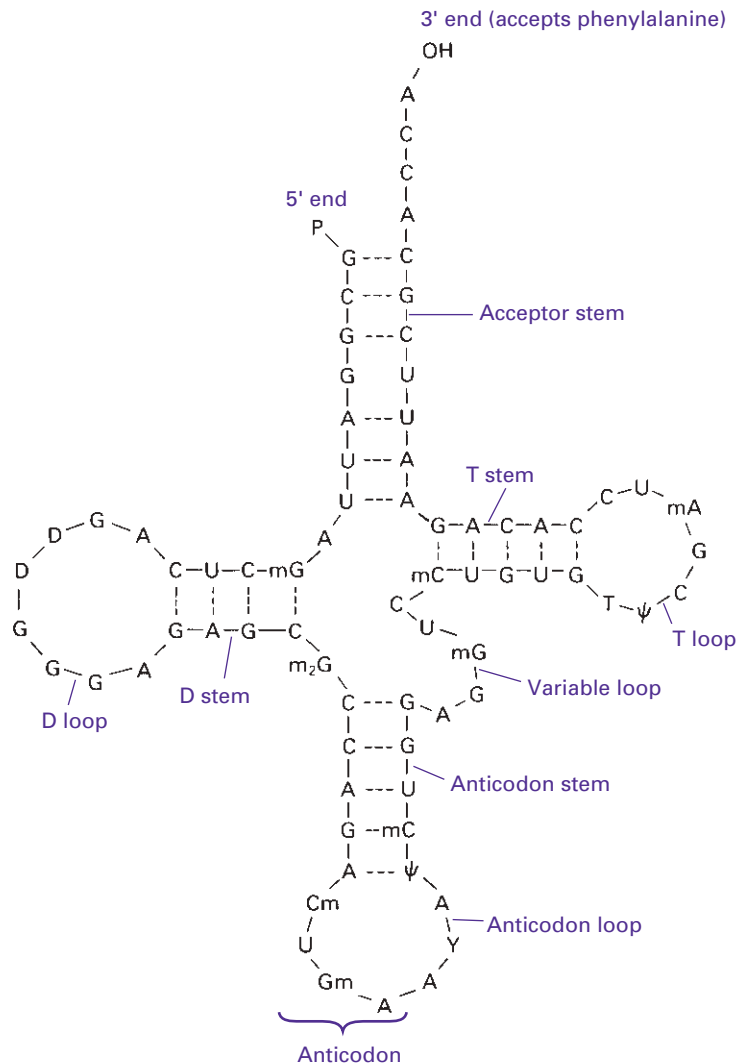


Fig. 5.6. Secondary structure of yeast tRNA^{Phe}. A single strand of 76 ribonucleotides forms four double-stranded 'stem' regions by base-pairing between complementary sequences. The anticodon will base-pair with UUU or UUC (both are codons for phenylalanine), phenylalanine is attached to the 3'-end by a specific aminoacyl tRNA synthetase. Several 'unusual' bases are present: D, dihydrouridine; T, ribothymidine; ψ , pseudouridine; Y, very highly modified, unlike any 'normal' base. mX indicates methylation of base X (m₂X shows dimethylation); Xm indicates methylation of ribose on the 2' position.

The absorbance at 260 nm may be plotted against the temperature of a DNA solution, which will indicate that little denaturation occurs below approximately 70°C, but further increases in temperature result in a marked increase in the extent of denaturation. Eventually a temperature is reached at which the sample is totally denatured, or melted. The temperature at which 50% of the DNA is melted is termed the **melting temperature** or T_m , and this depends on the nature of the

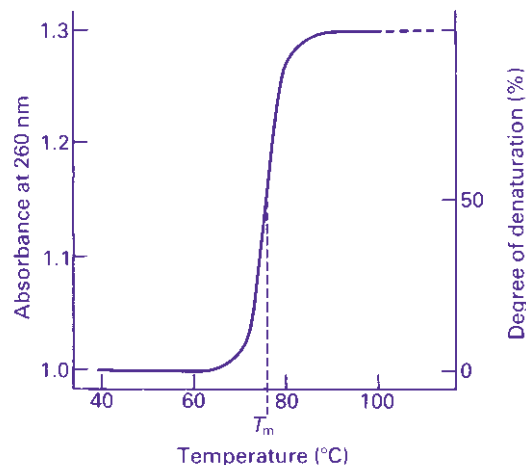


Fig. 5.7. Melting curve of DNA.

DNA (Fig. 5.7). If several different samples of DNA are melted, it is found that the T_m is highest for those DNA molecules that contain the highest proportion of cytosine and guanine, and T_m can actually be used to estimate the percentage (C + G) in a DNA sample. This relationship between T_m and (C + G) content arises because cytosine and guanine form three hydrogen bonds when base-paired, whereas thymine and adenine form only two. Because of the differential numbers of hydrogen bonds between A-T and C-G pairs those sequences with a predominance of C-G pairs will require greater energy to separate or denature them. The conditions required to separate a particular nucleotide sequence are also dependent on environmental conditions such as salt concentration.

If melted DNA is cooled it is possible for the separated strands to reassociate, a process known as **renaturation**. However, a stable double-stranded molecule will be formed only if the complementary strands collide in such a way that their bases are paired precisely, and this is an unlikely event if the DNA is very long and complex (i.e. if it contains a large number of different genes). Measurements of the rate of renaturation can give information about the complexity of a DNA preparation (Section 5.3).

Strands of RNA and DNA will associate with each other, if their sequences are complementary, to give double-stranded **hybrid molecules**. Similarly, strands of radioactively labelled RNA or DNA, when added to a denatured DNA preparation, will act as probes for DNA molecules to which they are complementary (Section 5.7). This **hybridisation** of complementary strands of nucleic acids is very useful for isolating a specific fragment of DNA from a complex mixture (Section 5.10). It is also possible for small single-stranded fragments of DNA (up to 40 bases in length) termed **oligonucleotides** to hybridise to a denatured sample of DNA. This type of hybridisation is termed **annealing** and again is dependent on the base sequence of the oligonucleotide and the salt concentration of the sample.

5.3 GENES AND GENOME COMPLEXITY

5.3.1 Gene complexity

Each region of DNA that codes for a single RNA or protein molecule is called a **gene**, and the entire set of genes in a cell, organelle or virus forms its **genome**. Cells and organelles may contain more than one copy of their genome. Genomic DNA from nearly all prokaryotic and eukaryotic organisms is also complexed with protein and termed **chromosomal DNA**. Each gene is located at a particular position along the chromosome, termed the locus, whilst the particular form of the gene is termed the **allele**. In mammalian DNA each gene is present in two allelic forms, which may be identical (**homozygous**) or may vary (**heterozygous**). Current estimates derived from the Human Genome Project indicate that there are approximately 35 000 genes present. However, various processing events may well increase the number of actual proteins found in the cell in relation to the number of genes. The occurrence of different alleles at the same site in the genome is termed **polymorphism**. In general the more complex an organism the larger its genome, although this is not always the case, since many higher organisms have non-coding sequences some of which are repeated numerous times and termed **repetitive DNA**. In mammalian DNA, repetitive sequences may be divided into low copy number and high copy number DNA. The latter is composed of repeat sequences that are dispersed throughout the genome and those that are clustered together. The repeat cluster DNA may be defined as so-called **classical satellite DNA**, **minisatellite** and **microsatellite DNA**, the last of these being composed mainly of dinucleotide repeats (Table 5.2). These sequences are termed polymorphic, collectively termed polymorphisms and vary between individuals; they also form the basis of **genetic fingerprinting** (Section 6.8.7).

Table 5.2 Repetitive satellite sequences found in DNA, and their characteristics

Types of repetitive DNA	Repeat unit size (bp)	Characteristics/motifs
Satellite DNA	5–200	Large repeat unit range (Mb) usually found at centromeres
Minisatellite DNA		
Telomere sequence	6	Found at the ends of chromosomes. Repeat unit may span up to 20 kb G-rich sequence
Hypervariable sequence	10–60	Repeat unit may span up to 20 kb
Microsatellite DNA	1–4	Mononucleotide repeat of adenine dinucleotide repeats common (CA). Usually known as VNTR (variable number tandem repeat)

bp, base-pairs; kb, kilobase-pairs.

5.3.2 Single nucleotide polymorphisms

In genomes, there is an additional source of polymorphic diversity termed **single nucleotide polymorphisms** (SNPs pronounced *snips*). SNPs are substitutions of one base at a precise location within the genome. Those that occur in coding regions are termed **cSNPs**. Estimates indicate that an SNP occurs once in every 300 bases and there are thought to be approximately 10 million in the human genome. Interest in SNPs lies in the fact that these differences may account for the differences in disease susceptibility, drug metabolism and response to environmental factors between individuals. There are now a number of initiatives to identify SNPs and produce a genome SNP map. A number of maps have been partially completed and a number of bioinformatics resources have been developed such as the SNP consortium.

5.3.3 Chromosomes and karyotypes

Higher organisms may be identified by using the size and shape of their genetic material at a particular point in the cell division cycle termed metaphase. At this point, DNA condenses to form a number of very distinct chromosome structures. Various morphological characteristics of chromosomes may be identified at this stage, including the centromere and the telomere. The array of chromosomes from a given organism may also be stained with dyes such as Giemsa stain and subsequently analysed by light microscopy. The complete array of chromosomes in an organism is termed the **karyotype**. In certain genetic disorders, aberrations in the size, shape and number of chromosomes may occur and thus the karyotype may be used as an indicator of the disorder (Section 6.8.2). Perhaps the best-known example of this is the correlation of Down's syndrome, where three copies of chromosome 21 (trisomy 21) exist rather than two as in the normal state.

5.3.4 Renaturation kinetics and genome complexity

When preparations of double-stranded DNA are denatured and allowed to renature, measurement of the rate of renaturation can give valuable information about the complexity of the DNA, i.e. how much information it contains (measured in base-pairs). The complexity of a molecule may be much less than its total length if some sequences are repetitive, but complexity will equal total length if all sequences are unique, appearing only once in the genome. In practice, the DNA is first cut randomly into fragments about 1 kb in length (Section 5.8), and is then completely denatured by heating above its T_m (Section 5.2.3). Renaturation at a temperature about 10°C below the T_m is monitored either by decrease in absorbance at 260 nm (the hypochromic effect), or by passing samples at intervals through a column of hydroxylapatite, which will adsorb only double-stranded DNA, and measuring how much of the sample is bound. The degree of renaturation after a given time will

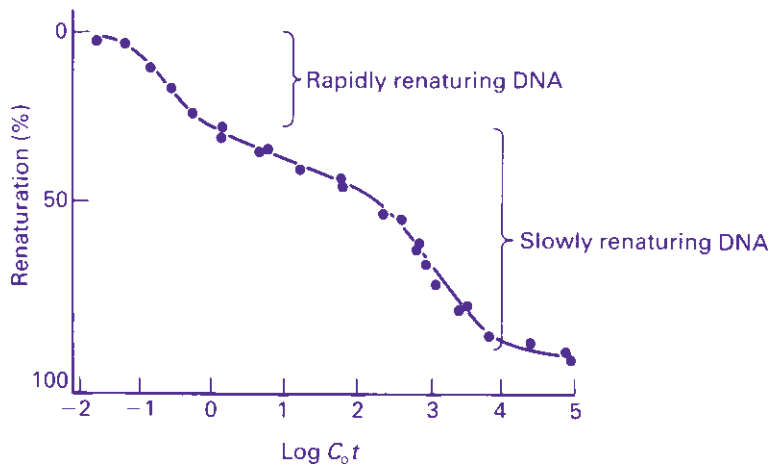


Fig. 5.8. Cot curve of human DNA. DNA was allowed to renature at 60°C after being completely dissociated by heat. Samples were taken at intervals and passed through a hydroxylapatite column to determine the percentage of double-stranded DNA present. This percentage was plotted against $\log C_0t$ (original concentration of DNA \times time of sampling).

depend on C_0 , the concentration (in nucleotides per unit volume) of double-stranded DNA prior to denaturation, and t , the duration of the renaturation in seconds.

For a given C_0 , it should be evident that a preparation of bacteriophage λ DNA (genome size 49 kb) will contain many more copies of the same sequence per unit volume than a preparation of human DNA (haploid genome size 3×10^6 kb), and will therefore renature far more rapidly, since there will be more molecules complementary to each other per unit volume in the case of λ DNA, and therefore more chance of two complementary strands colliding with each other. In order to compare the rates of renaturation of different DNA samples it is usual to measure C_0 and the time taken for renaturation to proceed half way to completion, $t_{1/2}$, and to multiply these values together to give a $C_0t_{1/2}$ value. The larger the $C_0t_{1/2}$, the greater the complexity of the DNA; hence λ DNA has a far lower $C_0t_{1/2}$ than does human DNA.

In fact, the human genome does not renature in a uniform fashion. If the extent of renaturation is plotted against $\log C_0t$ (this is known as a **Cot curve**), it is seen that part of the DNA renatures quite rapidly, whilst the remainder is very slow to renature (Fig. 5.8). This indicates that some sequences have a higher concentration than others; in other words, part of the genome consists of repetitive sequences. These repetitive sequences can be separated from the single-copy DNA by passing the renaturing sample through a hydroxylapatite column early in the renaturation process, at a time that gives a low value of C_0t . At this stage only the rapidly renaturing sequences will be double-stranded and they will therefore be the only ones able to bind to the column.

First position (5' end)	Second position				Third position (3' end)
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Fig. 5.9. The genetic code. Note that the codons in blue represent the start codon (ATG) and the three stop codons.

5.3.5 The nature of the genetic code

DNA encodes the primary sequence of a protein by utilising sets of three nucleotides, termed a **codon** or **triplet**, to encode a particular amino acid. The four bases (A, C, G and T) present in DNA allow a possible 64 triplet combinations; however, since there are only about 20 naturally occurring amino acids more than one codon may encode an amino acid. This phenomenon is termed the **degeneracy** of the genetic code. With the exception of a limited number of differences found in mitochondrial DNA and one or two other species, the genetic code appears to be universal. In addition to coding for amino acids, particular triplet sequences also indicate the beginning (**Start**) and the end (**Stop**) of a particular gene. Only one start codon exists (ATG) which also codes for the amino acid methionine, whereas three dedicated stop codons are available (TAT, TAG and TGA) (Fig. 5.9). A sequence flanked by a start and a stop codon containing a number of codons that may be read in-frame to represent a continuous protein sequence is termed an **open reading frame** (ORF).

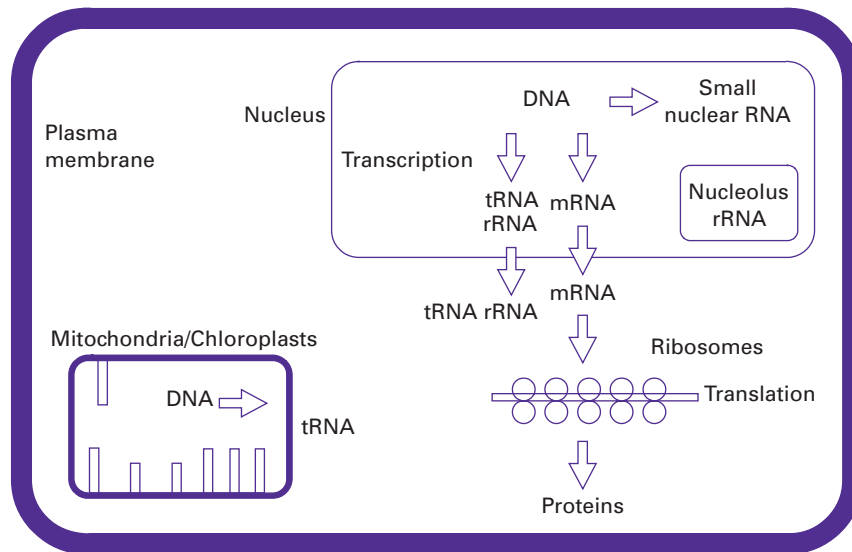


Fig. 5.10. Location of DNA and RNA molecules in eukaryotic cells and the flow of genetic information.

5.4 LOCATION AND PACKAGING OF NUCLEIC ACIDS

5.4.1 Cellular compartments

In general, DNA in eukaryotic cells is confined to the nucleus and organelles such as mitochondria or chloroplasts, which contain their own genome. The predominant RNA species are, however, normally located within the cytoplasm. The genetic information of cells and most viruses is stored in the form of DNA. This information is used to direct the synthesis of RNA molecules, which fall into three classes. Fig. 5.10 indicates the locations of nucleic acids in prokaryotic and eukaryotic cells.

- *Messenger RNA (mRNA)*: This contains sequences of ribonucleotides that code for the amino acid sequences of proteins. A single mRNA molecule codes for a single polypeptide chain in eukaryotes, but may code for several polypeptides in prokaryotes.
- *Ribosomal RNA (rRNA)*: This forms part of the structure of ribosomes, which are the sites of protein synthesis. Each ribosome contains only three or four different rRNA molecules, complexed with a total of between 55 and 75 proteins.
- *Transfer RNA (tRNA)*: These molecules carry amino acids to the ribosomes, and interact with the mRNA in such a way that their amino acids are joined together in the order specified by the mRNA. There is at least one type of tRNA for each amino acid.

In eukaryotic cells alone a further group of RNA molecules termed **small nuclear RNA** (snRNA) is present that function within the nucleus and promote the maturation of mRNA molecules. All RNA molecules are associated with their

respective binding proteins and are essential for their cellular functions. Nucleic acids from prokaryotic cells are less well compartmentalised, although they serve similar functions.

5.4.2 The packaging of DNA

The DNA in prokaryotic cells resides in the cytoplasm, although is associated with nucleoid proteins, where it is tightly coiled and supercoiled by topoisomerase enzymes to enable it to physically fit into the cell. By contrast, eukaryotic cells have many levels of packaging of the DNA within the nucleus, involving a variety of DNA-binding proteins.

First-order packaging involves the winding of the DNA around a core complex of four small proteins repeated twice and termed **histones** (H2A, H2B, H3 and H4). These are rich in the basic amino acids lysine and arginine and form a barrel-shaped core octamer structure. Approximately 180 bp of DNA is wound twice around the structure, which is termed a **nucleosome**. A further histone protein, H1, is found to associate with the outer surface of the nucleosome. The compacting effect of the nucleosome reduces the length of the DNA by a factor of 6.

Nucleosomes also associate to form a second order of packaging termed the 30 nm **chromatin fibre**, thus further reducing the length of the DNA by a factor of 7 (Fig. 5.11). These structure may be further folded and looped through the interaction with other non-histone proteins and ultimately form chromosome structures.

DNA is found closely associated with the nuclear lamina matrix, which forms a protein scaffold within the nucleus. The DNA is attached at certain positions within the scaffold, usually coinciding with origins of replication. Many other DNA-binding proteins are also present, such as high mobility group (HMG) proteins, which assist in promoting certain DNA conformations during processes such as replication or active gene expression.

5.5 FUNCTIONS OF NUCLEIC ACIDS

5.5.1 DNA replication

The double-stranded nature of DNA provides a means of replication during cell division, since the separation of two DNA strands allows complementary strands to be synthesised upon them. Many enzymes and accessory proteins are required for *in vivo* replication, which in prokaryotes begins at a region of the DNA termed the **origin of replication**.

DNA has to be unwound before any of the proteins and enzymes needed for replication can act, and this involves separating the double-helical DNA into single strands. This process is carried out by the enzyme DNA helicase. Furthermore, in order to prevent the single strands from re-annealing, small proteins termed **single-stranded DNA-binding proteins** (SSBs) attach to the single DNA strands (Fig. 5.12).

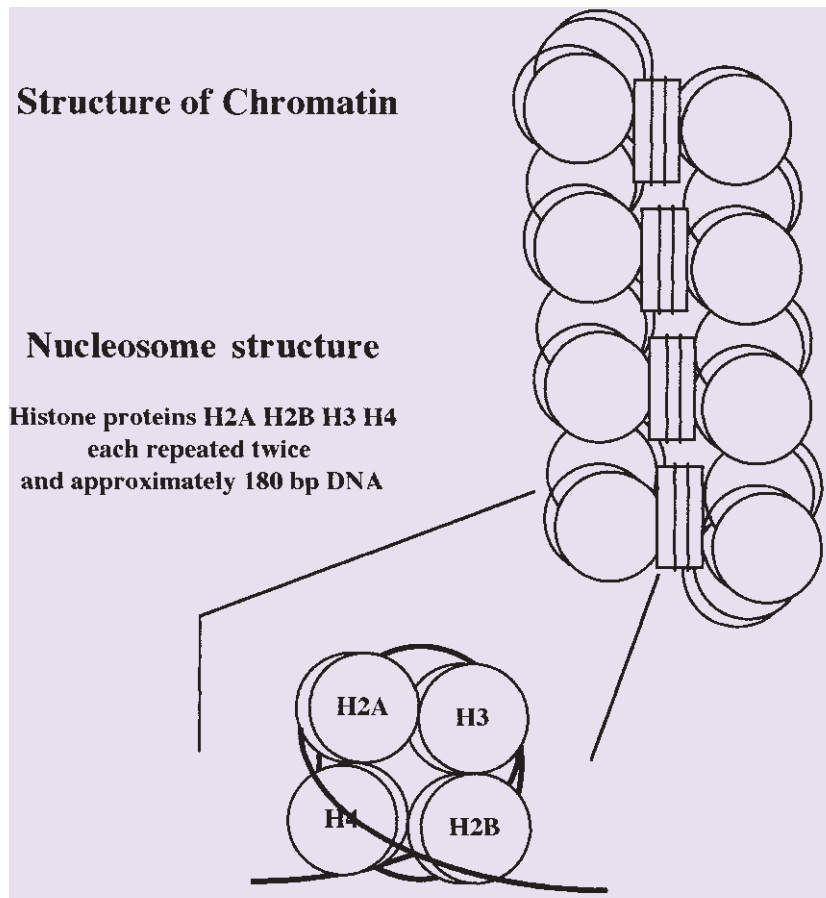


Fig. 5.11. Structure and composition of the nucleosome and chromatin.

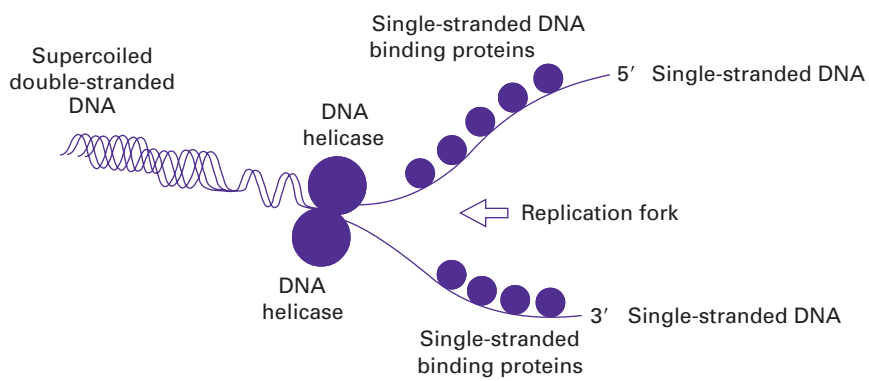


Fig. 5.12. Initial events at the replication fork involving DNA unwinding.

On each exposed single strand a short, complementary RNA chain termed a **primer** is first produced, using the DNA as a template. The primer is synthesised by an RNA polymerase enzyme known as a primase, which uses ribonucleoside triphosphates and itself requires no primer to function. Then DNA polymerase III (DNA pol III) also uses the original DNA as a template for synthesis of a DNA strand, using the RNA primer as a starting point. The primer is vital, since it leaves an exposed 3' hydroxyl group. This is necessary, since DNA polymerase III can add new nucleotides only to the 3' end and not the 5' end of a nucleic acid. Synthesis of the DNA strand therefore occurs only in a 5' to 3' direction from the RNA primer. This DNA strand is usually termed the **leading strand** and provides the means for continuous DNA synthesis.

Since the two strands of double helical DNA are antiparallel, only one can be synthesised in a continuous fashion. Synthesis of the other strand must take place in a more complex way. The precise mechanism was worked out by Reiji Okazaki in the 1960s. Here, the strand, usually termed the **lagging strand**, is produced in relatively short stretches of 1–2 kb termed **Okazaki fragments**. This is still in a 5' to 3' direction, using many RNA primers for each individual stretch. Thus discontinuous synthesis of DNA takes place and allows DNA polymerase III to work in the 5' to 3' direction. The RNA primers are then removed by DNA polymerase I, which has a 5' to 3' exonuclease and the gaps are filled by the same enzyme acting as a polymerase. The separate fragments are joined together by DNA ligase to give a newly formed strand of DNA on the lagging strand (Fig. 5.13).

The replication of eukaryotic DNA is less well characterised, involves multiple origins of replication and is certainly more complex than that of prokaryotes; however, in both cases the process involves 5' to 3' synthesis of new DNA strands. The net result of the replication is that the original DNA is replaced by two molecules, each containing one 'old' and one 'new' strand; the process is therefore known as **semi-conservative replication**. The ideas behind DNA synthesis, replication and the enzymes involved in them have been adopted in many molecular biological techniques and form the basis of many manipulations in genetic engineering.

5.5.2 DNA protection and repair systems

Cellular growth and division require the correct and coordinated replication of DNA. Mechanisms that proofread replicated DNA sequences and maintain the integrity of those sequences are, however, complex and are only beginning to be elucidated for prokaryotic systems. Bacterial protection is afforded by the use of a restriction modification system based on differential methylation of host DNA, so as to distinguish it from foreign DNA such as viruses. The most common is type II and consists of a host DNA methylase and restriction endonuclease that recognises short (4–6 bp) palindromic sequences and cleaves foreign unmethylated DNA at a particular target sequence. The enzymes involved in this process have been of enormous benefit for the manipulation and analysis of DNA, as indicated in Section 5.8.

Repair systems allow the recognition of altered, mispaired or missing bases in double-stranded DNA and invoke an excision repair process. Bacterial repair

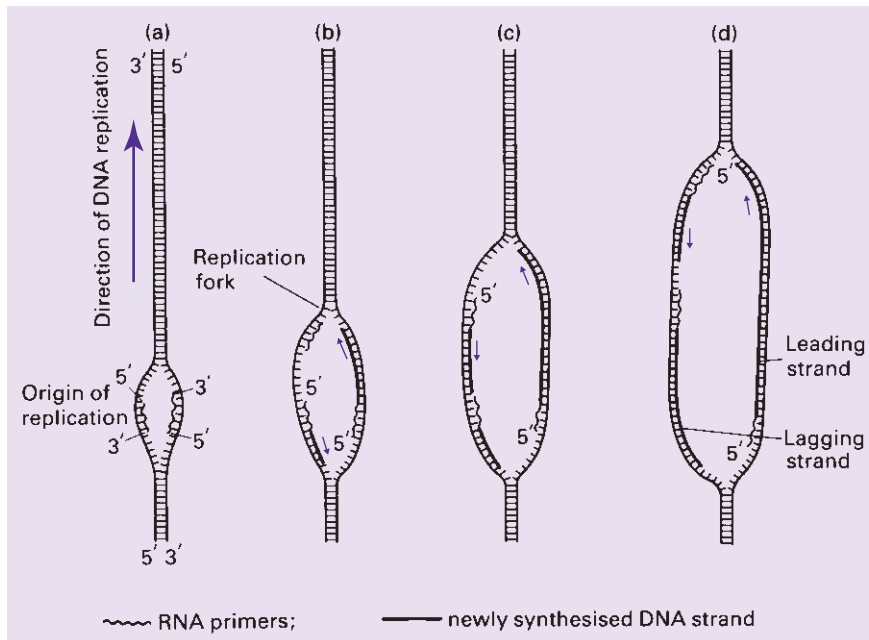


Fig. 5.13. DNA replication. (a) Double-stranded DNA separates at the origin of replication. RNA polymerase synthesises short DNA primer strands complementary to both DNA strands. (b) DNA polymerase III synthesises new DNA strands in a 5' to 3' direction, complementary to the exposed, old DNA strands, and continuing from the 3' end of each RNA primer. Consequently DNA synthesis is in the same direction as DNA replication for one strand (the leading strand) and in the opposite direction for the other (the lagging strand). RNA primer synthesis occurs repeatedly to allow the synthesis of fragments of the lagging strand. (c) As the replication fork moves away from the origin of replication, DNA polymerase III continues the synthesis of the leading strand, and synthesises DNA between RNA primers of the lagging strand. (d) DNA polymerase I removes RNA primers from the lagging strand and fills the resulting gaps with DNA. DNA ligase then joins the resulting fragments, producing a continuous DNA strand.

systems are based on the length of repairable DNA either during replication (**dam system**) or in general repair (**urr system**). In some cases damage to DNA activates a protein termed RecA to produce an **SOS response** that includes the activation of many enzymes and proteins; however, this has yet to be fully characterised. The recombination–repair systems in eukaryotic cells may share some common features with prokaryotes, although the precise mechanism has yet to be established. Defects in DNA repair may result in the stable incorporation of errors into genomic sequences that may underscore several genetic-based diseases.

5.5.3 Transcription of DNA

Expression of genes is carried out initially by the process of **transcription**, whereby a complementary RNA strand is synthesised by an enzyme termed

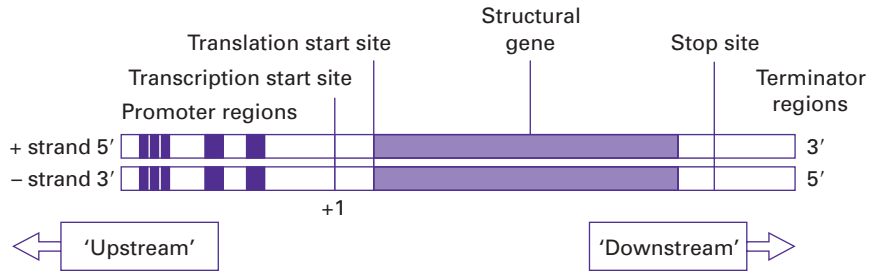


Fig. 5.14. Structure and nomenclature of a typical gene.

RNA polymerase from a DNA template encoding the gene. Most prokaryotic genes are made up of three regions. At the centre is the sequence that will be copied in the form of RNA, called the **structural gene**. To the 5' side (**upstream**) from the strand that will be copied (the plus (+) strand) lies a region called the **promoter**, and **downstream** from the transcription unit is the **terminator region**. Transcription begins when DNA-dependent RNA polymerase binds to the promoter region and moves along the DNA to the transcription unit. At the start of the transcription unit the polymerase begins to synthesise an RNA molecule complementary to the minus (–) strand of the DNA, moving along this strand in a 3' to 5' direction, and synthesising RNA in a 5' to 3' direction, using ribonucleoside triphosphates. The RNA will therefore have the same sequence as the plus strand of the DNA, apart from the substitution of uracil for thymine. On reaching the stop site in the terminator region, transcription is stopped, and the RNA molecule is released. The numbering of bases in genes is a useful way of identifying key elements. Point or base +1 is the residue located at the transcription start site, positive numbers denote 3' regions, whilst negative numbers denote 5' regions (Fig. 5.14).

In eukaryotes, three different RNA polymerases exist, designated I, II and III. Messenger RNA is synthesised by RNA polymerase II, while RNA polymerases I and III catalyse the synthesis of rRNA (I), tRNA and snRNA (III). Many non-expressed genes tend to have residues that are methylated, usually the C of a GC dinucleotide, and, in general, active genes tend to be hypomethylated. This is especially prevalent at the 5' flanking regions and is a useful means of discovering and identifying new genes.

5.5.4 Promoter and terminator sequences in DNA

Promoters are usually to the 5' end or upstream from the structural gene and have been best characterised in prokaryotes such as *Escherichia coli*. They comprise two highly conserved sequence elements: the **TATA box** (consensus sequence 'TATATT'), which is centred approximately 10 bp upstream from the transcription initiation site (–10 in the gene numbering system), and a 'G + C-rich' sequence which is centred about –25 bp upstream from the TATA box. The G + C element is thought to be important in the initial recognition and binding of RNA polymerase

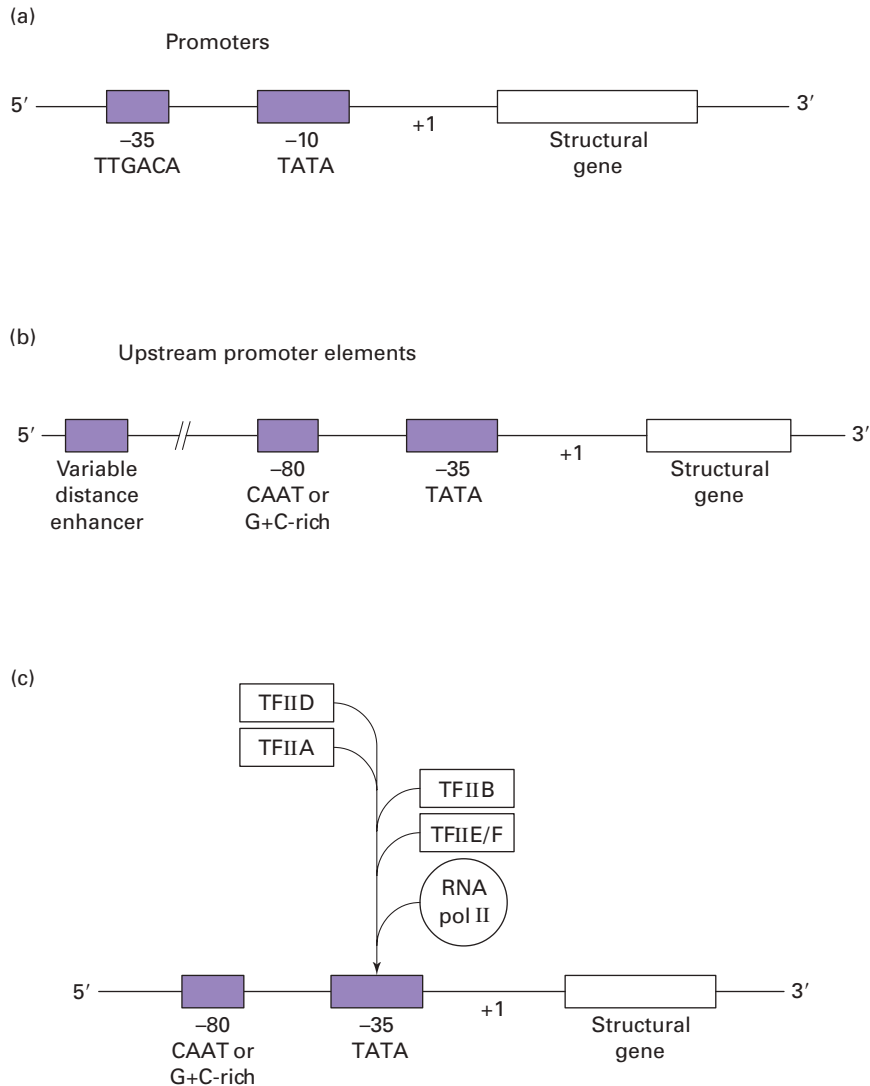


Fig. 5.15. (a) Typical promoter elements found in a prokaryotic cell (e.g. *E. coli*). (b) Typical promoter elements found in eukaryotic cells. (c) Generalised scheme of binding of transcription factors to the promoter regions of eukaryotic cells. Following the binding of the transcription factors IID, IIA, IIB, IIE and IIF a pre-initiation complex is formed. RNA polymerase II then binds to this complex and begins transcription from the start point +1.

to the DNA, while the -10 sequence is involved in the formation of a transcription initiation complex (Fig. 5.15a).

The promoter elements serve as recognition sites for DNA-binding proteins that control gene expression and these proteins are termed **transcription factors** or *trans-acting factors*. These proteins have a DNA-binding domain for interaction with promoters and an activation domain to allow interaction with other transcription factors. A well-studied example of a transcription factor (TF) is TFIIID,

which binds to the -35 promoter sequence in eukaryotic cells. Gene regulation occurs in most cases at the level of transcription, and primarily by the rate of transcription initiation, although control may also be by modulation of mRNA stability, or at other levels such as translation. Terminator sequences are less well characterised, but are thought to involve nucleotide sequences near the end of mRNA with the capacity to form a hairpin loop, followed by a run of U residues, which may constitute a termination signal for RNA polymerase.

In the case of eukaryotic genes, numerous short sequences spanning several hundred bases may be important for transcription, as compared with normally fewer than 100 bp for prokaryotic promoters. Particularly critical is the TATA box sequence, located approximately -35 bp upstream from the transcription initiation point in the majority of genes (Fig. 5.15b). This is analogous to the -10 sequence in prokaryotes. A number of other transcription factors also bind sequentially to form an initiation complex that includes RNA polymerase, subsequent to which transcription is initiated. In addition to the TATA box, a (CAT box (consensus GGCCAATCT) is often located at about -80 bp, which is an important determinant of promoter efficiency. Many upstream promoter elements (UPEs) have been described that are either general in their action or tissue (or gene) specific. GC elements that contain the sequence GGGCG may be present at multiple sites and in either orientation and are often associated with housekeeping genes such as those encoding enzymes involved in general metabolism. Some promoter sequence elements, such as the TATA box, are common to most genes, whilst others may be specific to particular genes or classes of genes.

Of particular interest is a class of promoter first investigated in simian virus 40 (SV40) and termed an enhancer. These sequences are distinguished from other promoter sequences by their unique ability to function over several kilobases either upstream or downstream from a particular gene in an orientation-independent manner. Even at such great distances from the transcription start point they may increase transcription by several hundred-fold. The precise interactions between transcription factors, RNA polymerase or other DNA-binding proteins and the DNA sequences to which they bind may be identified and characterised by the technique of DNA footprinting (Section 6.8.3). For transcription in eukaryotic cells to proceed, a number of transcription factors need to interact with the promoters and with each other. This cascade mechanism is indicated in Fig. 5.15c and is termed a pre-initiation complex. Once this has been formed around the -35 TATA sequence RNA polymerase II is able to transcribe the structural gene and form a complementary RNA copy (Section 5.5.6).

5.5.5 Transcription in prokaryotes

Prokaryotic gene organisation differs from that found in eukaryotes in a number of ways. Prokaryotic genes are generally found as continuous coding sequences. Moreover they are frequently found clustered into operons, which contain genes that relate to a particular function such as the metabolism of a substrate or synthesis of a product. This is particularly evident in the best-known operon

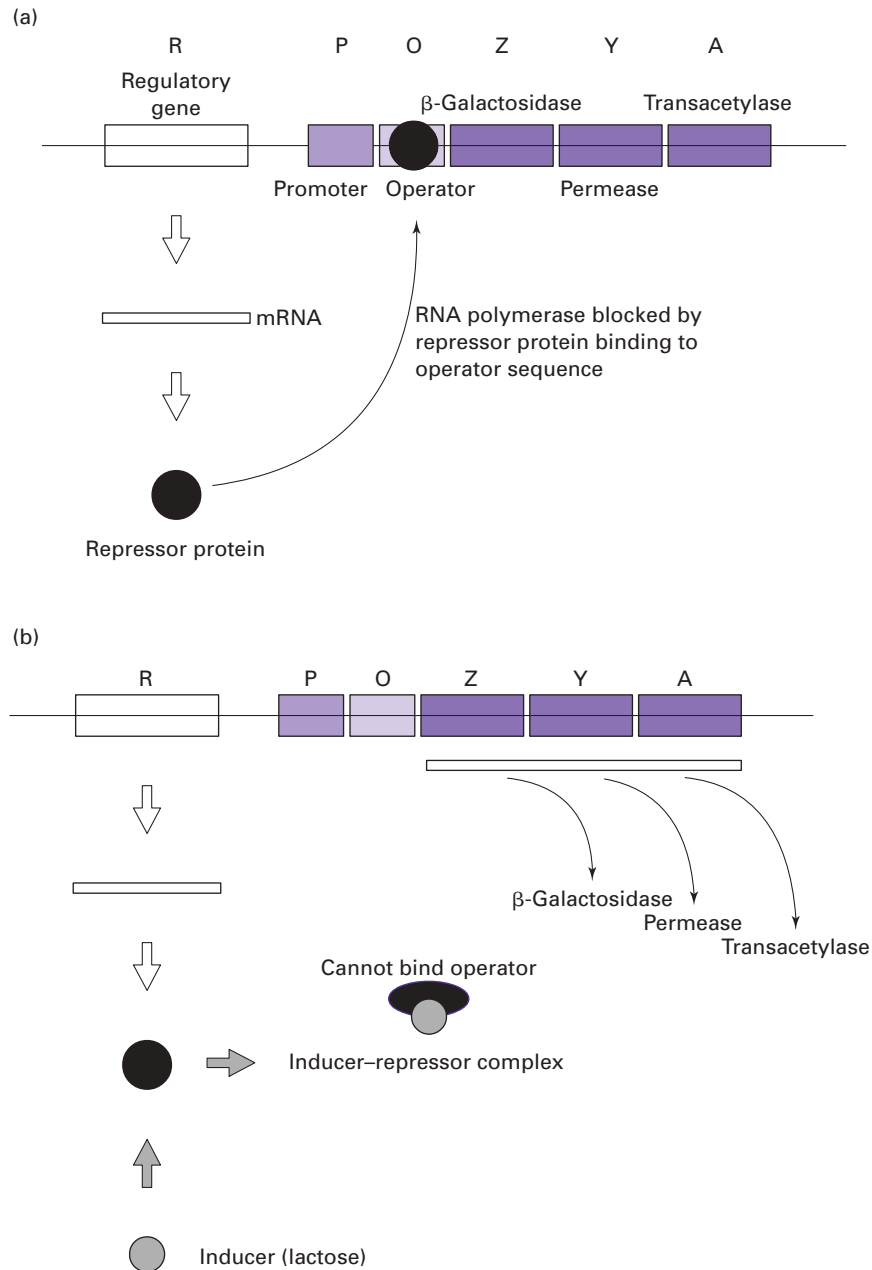


Fig. 5.16. Lactose operon (a) in a state of repression (no lactose present) and (b) following induction by lactose.

identified in *E. coli* termed the **lactose operon**, where three genes *lacZ*, *lacY* and *lacA* share the same promoter and are therefore switched on and off at the same time. In this model the absence of lactose results in a repressor protein binding to an operator region upstream of the Z, Y and A genes and prevents RNA polymerase from transcribing the genes (Fig. 5.16a). However, the presence of lactose

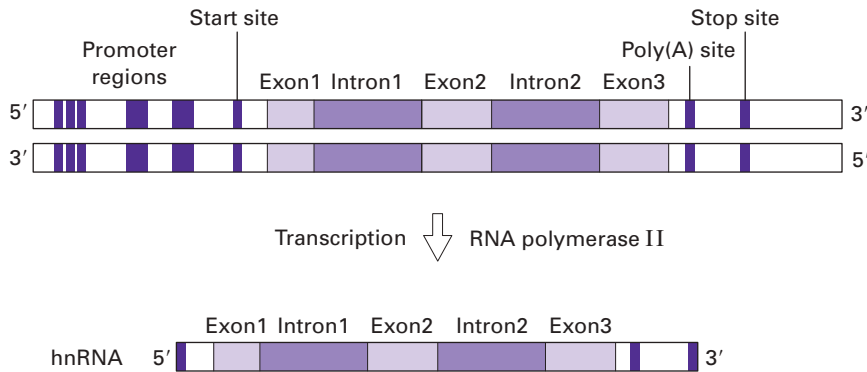


Fig. 5.17. Transcription of a typical eukaryotic gene to form heterogeneous nuclear RNA.

requires the genes to be transcribed to allow its metabolism. Lactose binds to the repressor protein and causes a conformational change in its structure. This prevents it binding to the operator and allows RNA polymerase to bind and transcribe the three genes (Fig. 5.16b). Transcription and translation in prokaryotes is also closely linked or coupled whereas in eukaryotic cells the two processes are distinct and take place in different cell compartments.

5.5.6 Post-transcriptional processing

Transcription of a eukaryotic gene results in the production of a **heterogeneous nuclear RNA** transcript (hnRNA) that faithfully represents the entire structural gene (Fig. 5.17). Three processing events then take place. The first processing step involves the addition of a methylated guanosine residue (m⁷Gppp) termed a cap to the 5' end of the hnRNA. This may be a signalling structure or aid in the stability of the molecule (Fig. 5.18). In addition, 150 to 300 adenosine residues termed a poly (A) tail are attached at the 3' end of the hnRNA by the enzyme poly (A) polymerase. The poly (A) tail allows the specific isolation of eukaryotic mRNA from total RNA by affinity chromatography (Section 5.7.2), its presence is thought to confer stability on the transcript.

Unlike prokaryotic transcripts, those from eukaryotes have their coding sequence (expressed regions or **exons**) interrupted by non-coding sequence (intervening regions or **introns**). Intron–exon boundaries are generally determined by the sequence GUAG and need to be removed or **spliced** before the mature mRNA is formed (Fig. 5.18). The process of intron splicing is mediated by small nuclear RNAs (snRNAs), which exist in the nucleus as ribonuclear protein particles. These are often found in a large nuclear structure complex termed the **spliceosome**, where splicing takes place. Introns are usually removed in a sequential manner from the 5' to the 3' end and their number varies between different genes. Some eukaryotic genes contain no introns, for example histone genes, whereas the gene for dystrophin, the gene responsible for muscular dystrophy, contains over 250 introns. In some cases, however, the same hnRNA transcript may be processed in

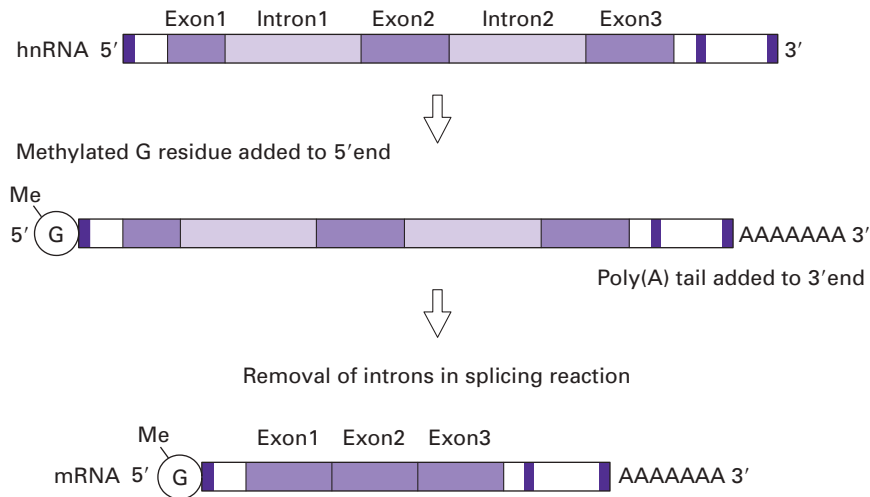


Fig. 5.18. Post-transcriptional modifications of heterogeneous nuclear RNA.

different ways to produce different mRNAs coding for different proteins in a process known as **alternative splicing**. Thus a sequence that constitutes an exon for one RNA species may be part of an excised intron in another. The particular type or amount of mRNA synthesised from a cell or cell type may be analysed by a variety of molecular biological techniques (Section 6.8.1).

5.5.7 Translation of mRNA

Messenger RNA molecules are read and translated into protein by complex RNA–protein particles termed ribosomes. The ribosomes are termed 70 S or 80 S depending on their sedimentation coefficient. Prokaryotic cells have 70 S ribosomes whilst those of the eukaryotic cytoplasm are 80 S. Ribosomes are composed of two subunits that are held apart by ribosomal binding proteins until translation proceeds. There are sites on the ribosome for the binding of one mRNA and two tRNA molecules and the translation process is in three stages.

- **Initiation:** This involves the assembly of the ribosome subunits and the binding of the mRNA.
- **Elongation:** This is where specific amino acids are used to form polypeptides, this being directed by the codon sequence in the mRNA.
- **Termination:** This involves the disassembly of the components of translation following the production of a polypeptide.

Transfer RNA molecules is also essential for translation. Each of these is covalently linked to a specific amino acid, forming an **aminoacyl tRNA**, and each has a triplet of bases exposed that is complementary to the codon for that amino acid. This exposed triplet is known as the **anticodon**, and allows the tRNA to act as an ‘adaptor’ molecule, bringing together a codon and its corresponding amino acid.

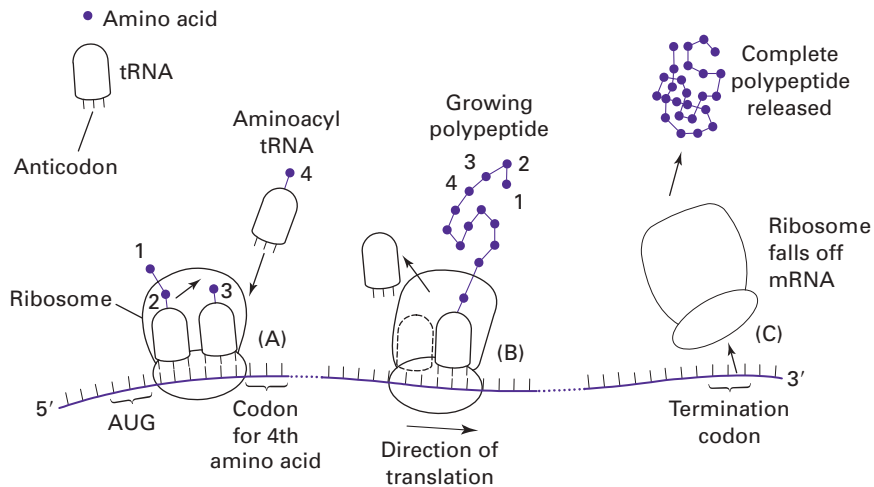


Fig. 5.19. Translation. Ribosome A has moved only a short way from the 5' end of the mRNA, and has built up a dipeptide (on one tRNA) that is about to be transferred onto the third amino acid (still attached to tRNA). Ribosome B has moved much further along the mRNA and has built up an oligopeptide that has just been transferred onto the most recent aminoacyl tRNA. The resulting free tRNA leaves the ribosome and will receive another amino acid. The ribosome moves towards the 3' end of the mRNA by a distance of three nucleotides, so that the next codon can be aligned with its corresponding aminoacyl tRNA on the ribosome. Ribosome C has reached a termination codon, has released the completed polypeptide, and has fallen off the mRNA.

The process of linking an amino acid to its specific tRNA is termed charging and is carried out by the enzyme aminoacyl tRNA synthetase.

In prokaryotic cells the ribosome binds to the 5' end of the mRNA at a sequence known as the **ribosome binding site** or sometimes termed the **Shine–Dalgarno sequence**, after the discoverers of the sequence. In eukaryotes the situation is similar but involves a Kozak sequence located around the initiation codon. Following translation initiation the ribosome moves towards the 3' end of the mRNA, allowing an aminoacyl tRNA molecule to base-pair with each successive codon, thereby carrying in amino acids in the correct order for protein synthesis. There are two sites for tRNA molecules in the ribosome: the A site and the P site. When these sites are occupied, directed by the sequence of codons in the mRNA, the ribosome allows the formation of a peptide bond between the amino acids. The process is also under the control of an enzyme, peptidyl transferase. When the ribosome encounters a **termination codon** (UAA, UGA or UAG) a release factor binds to the complex and translation stops, the polypeptide and its corresponding mRNA are released and the ribosome divides into its two subunits (Fig. 5.19). A myriad of accessory initiation and elongation protein factors are involved in this process. In eukaryotic cells the polypeptide may then be subjected to **post-translational modifications** such as glycosylation, and by virtue of specific amino acid signal sequences may be directed to specific cellular compartments or exported from the cell.

Since the mRNA base sequence is read in triplets, an error of one or two nucleotides in positioning of the ribosome will result in the synthesis of an incorrect polypeptide. Thus it is essential for the correct **reading frame** to be used during translation. This is ensured in prokaryotes by base-pairing between the Shine–Dalgarno sequence (Kozak sequence in eukaryotes) and a complementary sequence of one of the ribosome's rRNA molecules, thus establishing the correct starting point for movement of the ribosome along the mRNA. However, if a mutation such as a deletion/insertion takes place within the coding sequence, it will also cause a shift of the reading frame and result in an aberrant polypeptide. Genetic mutations and polymorphisms are considered in more detail in Section 6.8.5.

5.5.8 Control of protein production: RNA interference

There are a number of mechanisms by which protein production is controlled; however, the control may be at either the gene or the protein level. Typically this could include controlling levels of expression of mRNA, an increase or decrease in mRNA turnover, or controlling mRNA availability for translation. One recently discovered control mechanism that has also been adapted as a molecular biological technique to aid in the modulation of mRNA is termed **RNA interference** (RNAi). This involves the synthesis of short double-stranded RNA molecules that are cleaved into 21–23 nucleotide-long fragments to form an **RNA-induced silencing complex** (RISC). This complex potentially uses the short RNA molecules complementary to mRNA transcripts that, following hybridisation, allow an RNase to destroy the bound mRNA. The technique has important implications for medical conditions where, for example, increased levels of specific mRNA molecules in certain cancers and viral infections may be reduced using RNAi.

5.6 THE MANIPULATION OF NUCLEIC ACIDS: BASIC TOOLS AND TECHNIQUES

5.6.1 Enzymes used in molecular biology

The discovery and characterisation of a number of key enzymes has enabled the development of various techniques for the analysis and manipulation of DNA. In particular, the enzymes termed type II restriction endonucleases have come to play a key role in all aspects of molecular biology. These enzymes recognise certain DNA sequences, usually 4–6 bp in length, and cleave them in a defined manner. The sequences recognised are palindromic or of an inverted repeat nature; that is, they read the same in both directions on each strand. When cleaved they leave a flush-ended or staggered (also termed a **cohesive-ended**) fragment depending on the particular enzyme used (Fig. 5.20). An important property of staggered ends is that those produced from different molecules by the same enzyme are complementary (or 'sticky') and so will anneal to each other. The annealed strands are held together only by hydrogen bonding between complementary bases on opposite strands. Covalent joining of ends on each of the two

(a) Enzyme	Recognition sequence	Products	
<i>HpaII</i>	$\begin{array}{c} \downarrow \\ 5'-\text{CCGG}-3' \\ 3'-\text{GGCC}-5' \\ \uparrow \end{array}$	$\begin{array}{c} 5'-\text{C} \\ 3'-\text{GGC} \end{array}$	$\begin{array}{c} \text{CCG}-3' \\ \text{C}-5' \end{array}$
<i>HaeIII</i>	$\begin{array}{c} \downarrow \\ 5'-\text{GGCC}-3' \\ 3'-\text{CCGG}-5' \\ \uparrow \end{array}$	$\begin{array}{c} 5'-\text{GG} \\ 3'-\text{CC} \end{array}$	$\begin{array}{c} \text{CC}-3' \\ \text{GG}-5' \end{array}$
<i>BamHI</i>	$\begin{array}{c} \downarrow \\ 5'-\text{GGATCC}-3' \\ 3'-\text{CCTAGG}-5' \\ \uparrow \end{array}$	$\begin{array}{c} 5'-\text{G} \\ 3'-\text{CCTAG} \end{array}$	$\begin{array}{c} \text{GATCC}-3' \\ \text{G}-5' \end{array}$
<i>HpaI</i>	$\begin{array}{c} \downarrow \\ 5'-\text{GTTAAC}-3' \\ 3'-\text{CAATTG}-5' \\ \uparrow \end{array}$	$\begin{array}{c} 5'-\text{GTT} \\ 3'-\text{CAA} \end{array}$	$\begin{array}{c} \text{AAC}-3' \\ \text{TTG}-5' \end{array}$
(b) <i>EcoRI</i>	$\begin{array}{c} \downarrow \\ \text{GAATTC} \end{array}$		
<i>HindIII</i>	$\begin{array}{c} \downarrow \\ \text{AAGCTT} \end{array}$		
<i>PvuII</i>	$\begin{array}{c} \downarrow \\ \text{CAGCTG} \end{array}$		
<i>BamHI</i>	$\begin{array}{c} \downarrow \\ \text{GGATCC} \end{array}$		

Fig. 5.20. Recognition sequences of some restriction enzymes showing (a) full descriptions and (b) conventional representations. Arrows indicate positions of cleavage. Note that all the information in (a) can be derived from knowledge of a single strand of the DNA, whereas in (b) only one strand is shown, drawn 5' to 3'; this is the conventional way of representing restriction sites.

strands may be brought about by the enzyme DNA ligase (Section 6.2.2). This is widely exploited in molecular biology to enable the construction of **recombinant DNA**, i.e. the joining of DNA fragments from different sources. Approximately 500 restriction enzymes have been characterised that recognise over 100 different target sequences. A number of these, termed **isoschizomers**, recognise different target sequences but produce the same staggered ends or overhangs. A number of other enzymes have proved to be of value in the manipulation of DNA, as summarised in Table 5.3, and are indicated at appropriate points within the text.

5.7 ISOLATION AND SEPARATION OF NUCLEIC ACIDS

5.7.1 Isolation of DNA

The use of DNA for analysis or manipulation usually requires that it be isolated and purified to a certain extent. DNA is recovered from cells by the gentlest possible method of cell rupture to prevent the DNA from fragmenting by mechanical shearing. This is usually in the presence of EDTA, which chelates the Mg^{2+} needed for enzymes that degrade DNA (DNases). Ideally, cell walls, if present, should be digested enzymatically (e.g. lysozyme treatment of bacteria), and the cell

Table 5.3 Types and examples of typical enzymes used in the manipulation of nucleic acids

Enzyme	Specific example	Use in nucleic acid manipulation
	DNA pol I	DNA-dependent DNA polymerase 5'→3'/3'→5' exonuclease activity
DNA polymerases	Klenow	DNA pol I lacks 5'→3' exonuclease activity
	T4 DNA pol	Lacks 5'→3' exonuclease activity
	<i>Taq</i> DNA pol	Thermostable DNA polymerase used in PCR
	<i>Tth</i> DNA pol	Thermostable DNA polymerase with RT activity
	T7 DNA pol	Used in DNA sequencing
RNA polymerases	T7 RNA pol	DNA-dependent RNA polymerase
	T3 RNA pol	DNA-dependent RNA polymerase
	Q β replicase	RNA-dependent RNA polymerase, used in RNA-amplification
Nucleases	DNase I	Non-specific endonuclease that cleaves DNA
	Exonuclease III	DNA-dependent 3'→5' stepwise removal of nucleotides
	RNase A	RNases used in mapping studies
	RNase H	Used in second strand cDNA synthesis
	S1 nuclease	Single-strand-specific nuclease
Reverse transcriptase	AMV-RT	RNA-dependent DNA polymerase, used in cDNA synthesis
Transferases	Terminal transferase (TdT)	Adds homopolymer tails to the 3' end of DNA
Ligases	T4 DNA ligase	Links 5'-phosphate and 3'-hydroxyl ends via phosphodiester bond
Kinases	T4 polynucleotide kinase (PNK)	Transfers terminal phosphate groups from ATP to 5'-OH groups
Phosphatases	Alkaline phosphatase	Removes 5'-phosphates from DNA and RNA
Transferases	Terminal transferase	Adds homopolymer tails to the 3' end of DNA
Methylases	<i>Eco</i> RI methylase	Methylates specific residues and protects from cleavage by restriction enzymes

PCR, polymerase chain reaction; RT, reverse transcriptase; cDNA, complementary DNA; AMV, avian myeloblastosis virus.

membrane should be solubilised using detergent. If physical disruption is necessary, it should be kept to a minimum, and should involve cutting or squashing of cells, rather than the use of shear forces. Cell disruption (and most subsequent steps) should be performed at 4°C, using glassware and solutions that have been autoclaved to destroy DNase activity.

After release of nucleic acids from the cells, RNA can be removed by treatment with ribonuclease (RNase) that has been heat treated to inactivate any DNase contaminants; RNase is relatively stable to heat as a result of its disulphide bonds, which ensure rapid renaturation of the molecule on cooling. The other major contaminant, protein, is removed by shaking the solution gently with water-saturated phenol, or with a phenol/chloroform mixture, either of which

will denature proteins but not nucleic acids. Centrifugation of the emulsion formed by this mixing produces a lower, organic phase, separated from the upper, aqueous phase by an interface of denatured protein. The aqueous solution is recovered and deproteinised repeatedly, until no more material is seen at the interface. Finally, the deproteinised DNA preparation is mixed with two volumes of absolute ethanol, and the DNA allowed to precipitate out of solution in a freezer. After centrifugation, the DNA pellet is redissolved in a buffer containing EDTA to inactivate any DNases present. This solution can be stored at 4 °C for at least a month. DNA solutions can be stored frozen, although repeated freezing and thawing tends to damage long DNA molecules by shearing. The procedure described above is suitable for total cellular DNA. If the DNA from a specific organelle or viral particle is needed, it is best to isolate the organelle or virus before extracting its DNA, since the recovery of a particular type of DNA from a mixture is usually rather difficult. Where a high degree of purity is required, DNA may be subjected to density gradient ultracentrifugation through caesium chloride, which is particularly useful for the preparation of plasmid DNA. A flowchart of DNA extraction is indicated in Fig. 5.21. It is possible to check the integrity of the DNA by agarose gel electrophoresis and determine the concentration of the DNA by using the fact that 1 absorbance unit equates to 50 $\mu\text{g cm}^{-3}$ of DNA and so:

$$50 \times A_{260} = \text{concentration of DNA sample } (\mu\text{g cm}^{-3})$$

Contaminants may also be identified by scanning ultraviolet spectrophotometry from 200 nm to 300 nm. A ratio of 260 nm to 280 nm of approximately 1.8 indicates that the sample is free from protein contamination, which absorbs strongly at 280 nm.

5.7.2 Isolation of RNA

The methods used for RNA isolation are very similar to those described above for DNA; however, RNA molecules are relatively short, and therefore less easily damaged by shearing, so cell disruption can be rather more vigorous. RNA is, however, very vulnerable to digestion by RNases, which are present endogenously in various concentrations in certain cell types and exogenously on fingers. Gloves should therefore be worn, and a strong detergent should be included in the isolation medium to immediately denature any RNases. Subsequent deproteinisation should be particularly rigorous, since RNA is often tightly associated with proteins. DNase treatment can be used to remove DNA, and RNA can be precipitated by ethanol. One reagent in particular that is commonly used in RNA extraction is guanadinium thiocyanate, which is both a strong inhibitor of RNase and a protein denaturant. A flowchart of RNA extraction is indicated in Fig. 5.22. It is possible to check the integrity of an RNA extract by analysing it by agarose gel electrophoresis. The most abundant RNA species are the rRNA molecules, 23 S and 16 S for prokaryotes and 18 S and 28 S for eukaryotes. These appear as discrete bands on the agarose gel and indicate that the other RNA components are likely to be intact. This is usually carried out under denaturing conditions to prevent

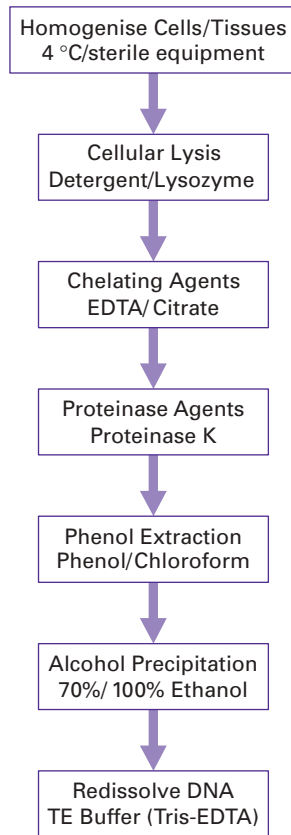


Fig. 5.21. General steps involved in extracting DNA from cells or tissues.

secondary structure formation in the RNA. The concentration of the RNA may be estimated by using ultraviolet spectrophotometry. At 260 nm 1 absorbance unit equates to $40 \mu\text{g cm}^{-3}$ of RNA and therefore:

$$40 \times A_{260} = \text{concentration of RNA sample } \mu\text{g cm}^{-3}$$

Contaminants may also be identified in the same way as for DNA by scanning ultraviolet spectrophotometry; however, in the case of RNA a 260 nm to 280 nm ratio of approximately 2 would be expected for a sample containing no protein (Section 5.7.1).

In many cases it is desirable to isolate eukaryotic mRNA, which constitutes only 2–5% of cellular RNA, from a mixture of total RNA molecules. This may be carried out by affinity chromatography on oligo(dT)-cellulose columns. At high salt concentrations, the mRNA containing poly(A) tails binds to the complementary oligo(dT) molecules of the affinity column, and so mRNA will be retained; all other RNA molecules can be washed through the column by further high salt solution. Finally, the bound mRNA can be eluted using a low concentration of salt (Fig 5.23). Nucleic acid species may also be subfractionated by more physical

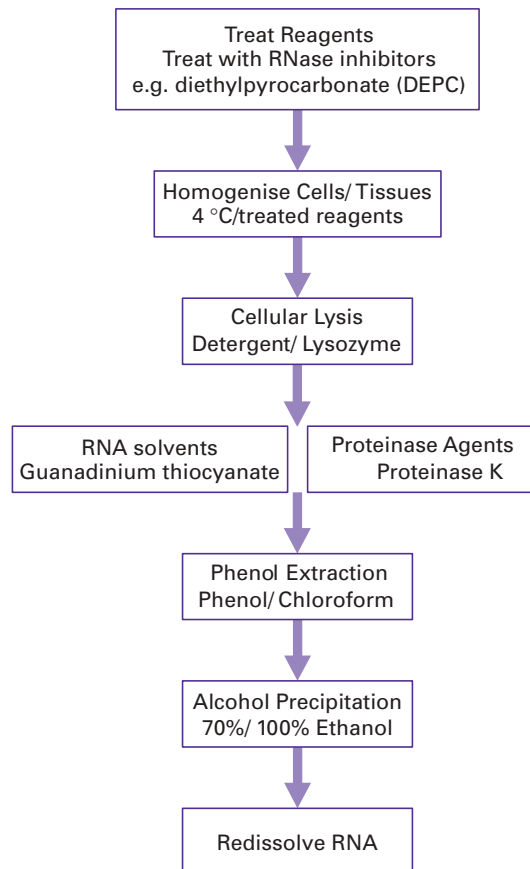


Fig. 5.22. General steps involved in extracting RNA from cells or tissues.

means such as electrophoretic or chromatographic separations based on differences in nucleic acid fragment sizes or physicochemical characteristics.

5.7.3 Automated and kit-based extraction of nucleic acids

Automation and kit-based manipulation in molecular biology is steadily increasing, and the extraction of nucleic acids by these means is no exception. There are many commercially available kits for nucleic acid extraction, although many rely on the methods described in Sections 5.8.1 and 5.8.2, their advantage lies in the fact that the reagents are standardised and quality control tested, providing a high degree of reliability. For example, the use of glass bead preparations for DNA purification has been used increasingly and with reliable results. Small compact column-type preparations are also used extensively in research and in routine DNA analysis such as QIAGEN columns. Essentially the same reagents for nucleic acid extraction may be used in a format that allows reliable and automated extraction. This is of particular use where a large number of DNA extractions are required.

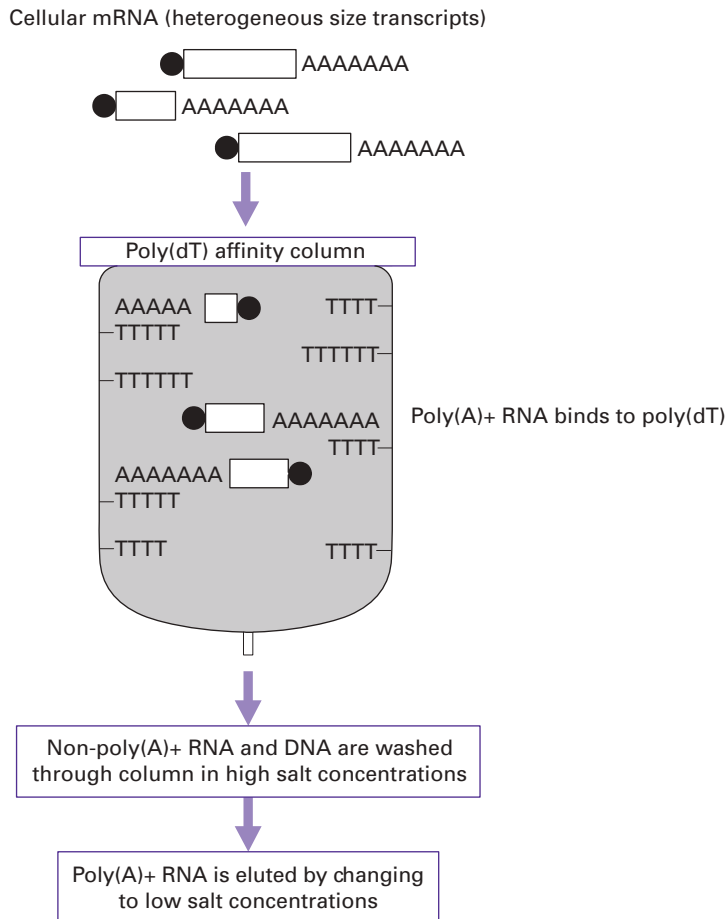
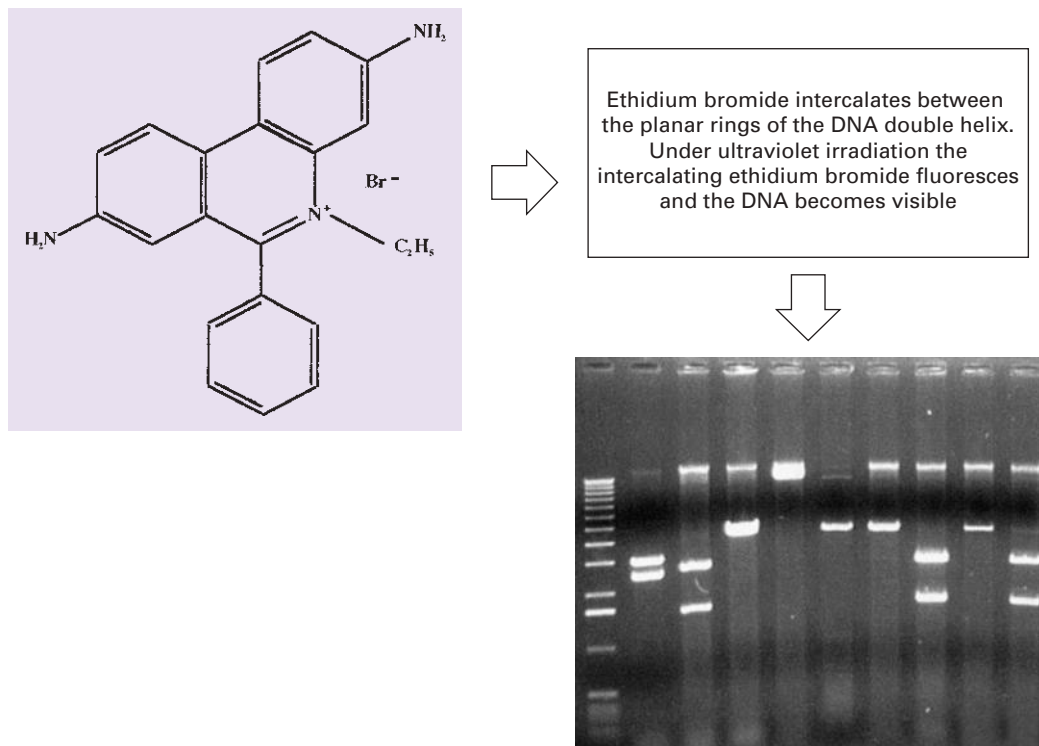


Fig. 5.23. Affinity chromatography of poly(A)+ RNA.

There are also many kit-based extraction methods for RNA, these in particular have overcome some of the problems of RNA extraction such as RNase contamination. A number of fully automated nucleic acid extraction machines are now employed in areas where high throughput is required, for example clinical diagnostic laboratories. Here, the raw samples such as blood specimens are placed in 96- or 384-well microtitre plates and these follow a set computer-controlled processing pattern carried out robotically. Thus the samples are rapidly manipulated and extracted in approximately 45 min without any manual operations being undertaken.

5.7.4 Electrophoresis of nucleic acids

Electrophoresis in agarose or polyacrylamide gels is the usual way to separate DNA molecules according to size. The technique can be used analytically or preparatively and can be qualitative or quantitative. Large fragments of DNA such as chromosomes may also be separated by a modification of electrophoresis termed



A photograph of an agarose gel stained with ethidium bromide and illuminated with UV irradiation showing discrete DNA bands

Fig. 5.24. The use of ethidium bromide to detect DNA.

pulsed field gel electrophoresis (PFGE). The easiest and most widely applicable method is electrophoresis in horizontal agarose gels, followed by staining with ethidium bromide. This dye binds to DNA by insertion between stacked base-pairs (**intercalation**), and it exhibits a strong orange/red fluorescence when illuminated with ultraviolet light (Fig 5.24). Very often electrophoresis is used to check the purity and intactness of a DNA preparation or to assess the extent of an enzymatic reaction during, for example, the steps involved in the cloning of DNA. For such checks minigels are particularly convenient, since they need little preparation, use small samples and quickly give results. Agarose gels can be used to separate molecules larger than about 100 bp. For higher resolution or for the effective separation of shorter DNA molecules, polyacrylamide gels are the preferred method.

When electrophoresis is used preparatively, the piece of gel containing the desired DNA fragment is physically removed with a scalpel. The DNA may be recovered from the gel fragment in various ways. This may include crushing with a glass rod in a small volume of buffer, using agarase to digest the agarose, thus leaving the DNA, or by the process of **electroelution**. In this method the piece of gel is sealed in a length of dialysis tubing containing buffer and is then placed

between two electrodes in a tank containing more buffer. Passage of an electrical current between the electrodes causes DNA to migrate out of the gel piece, but it remains trapped within the dialysis tubing and can therefore be recovered easily.

5.7.5 Automated analysis of nucleic acid fragments

Gel electrophoresis remains the established method for the separation and analysis of nucleic acids. However, a number of automated systems using pre-cast gels are available that are gaining popularity. This is especially useful in situations where a large number of samples or high throughput analysis is required. In addition, new technologies such as Agilent's *Lab-on-a-chip* have been developed that obviate the need to prepare electrophoretic gels. These systems employ microfluidic circuits where a small cassette unit that contains interconnected microreservoirs is used. The sample is applied in one area and driven through microchannels under computer-controlled electrophoresis. The channels lead to reservoirs allowing, for example, incubation with other reagents such as dyes for a specified time. Electrophoretic separation is thus carried out in a microscale format. The small sample size minimises sample and reagent consumption, and as such is useful for DNA and RNA sample analysis. In addition the units, being computer controlled, allow data to be captured within a very short time scale. More recently, alternative methods of analysis including high performance liquid chromatography-based approaches have gained in popularity, especially for mutation analysis (Section 6.8.6). Mass spectrometry is also becoming increasingly used for nucleic acid analysis (Section 9.2.4).

5.8 MOLECULAR BIOLOGY AND BIOINFORMATICS

5.8.1 Basic bioinformatics

Bioinformatics has become a vital resource for molecular biological research and is also increasingly used in the routine detection of DNA mutations. This growth of bioinformatics has been driven by the increase in genetic sequence information and the need to store, analyse and manipulate it. There are now a huge number of sequences stored in genetic databases from a variety of organisms, including the human genome. Indeed the genetic information from various organisms is an indispensable starting point for molecular biology research. The largest of the so-called *primary databases* include GenBank at the National Institutes of Health (NIH) in the USA, EMBL at the European Bioinformatics Institute (EBI) in Cambridge, UK, and the DNA database of Japan (DDBJ) at Mishima. These databases contain the nucleotide sequences that are annotated to allow easy identification. There are also many other databases such as *secondary databases* that contain information relating to sequence motifs, such as core sequences representing cytochrome P450 domains or DNA-binding domains. Importantly all of the databases may be accessed over the internet. A number of these important databases and internet resources are listed in Table 5.4.

Table 5.4 Nucleic acid and protein database resources available on the World Wide Web

Database or resource	URL (uniform resource locator)
General DNA sequence databases	
EMBL	European Bioinformatics Institute < http://www.ebi.ac.uk >
GenBank	US genetic database resource < http://www.ncbi.nlm.nih.gov >
DDBJ	Japanese genetic database < http://www.ddbj.nig.ac.jp >
Protein sequence databases	
Swiss-Prot	European protein sequence database < http://www.expasy.org >
UniProt TREMBL	European protein sequence database < http://www.ebi.ac.uk/trembl >
Protein structure databases	
PDB	Protein structure database < http://www.rcsb.org >
Genome project databases	
Human Genome Database, USA	< http://gdbwww.gdb.org >
dbEST (cDNA and partial sequences)	< http://www.ncbi.nih.gov/dbEST/index.html >
G�n�thon Genetic maps based on repeat markers	< http://www.genethon.fr >

DNA databases and other nucleic acid sequence and protein analysis software may all be accessed over the internet given the relevant software and authority. This is now relatively straightforward using web browsers that provide a user friendly graphical interface for data analysis and manipulation. Consequently the new expanding and exciting areas of bioscience research are those that analyse genome and cDNA sequence databases (**genomics**) and also their protein counterparts (**proteomics**). This is sometimes referred to as *in silico* research.

5.8.2 Analysing information using bioinformatics

One of the most useful bioinformatics resources is termed **BLAST** (basic local alignment search tool) located at the NCBI (<www.ncbi.nlm.nih.gov>). This allows a DNA sequence to be submitted via the World Wide Web in order to compare it to all the sequences contained within a DNA database. This is very useful, since it is possible once a nucleotide sequence has been deduced by, for example, Sanger sequencing, to identify sequences of similarity. Indeed if human sequences are used and have already been mapped it is possible to locate their position to a particular chromosome using NCBI, GenomeMap. A further resource such as an ORF (open reading frame) finder allows a search to be undertaken for ORFs, for example sequences beginning with a start codon (ATG) and continuing with a significant number of 'coding' triplets before a stop codon is reached. There are a number of other sequences that may be used to define coding sequences: these include ribosome binding sites, splice site junctions, poly(A) polymerase sequences and promoter sequences that lie outside the coding regions. These may

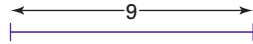

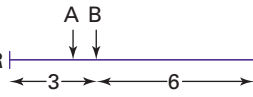
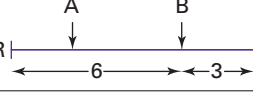
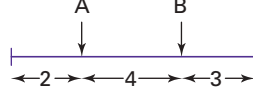
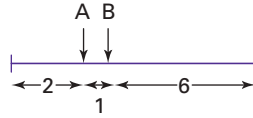
Treatment	Measured sizes of fragments (kb)	Interpretation
No digestion	9	
Enzyme A	2 + 7	
Enzyme B	3 + 6	EITHER  OR 
Enzymes A + B	2, 3 + 4	
	alternative result 1, 2 + 6	

Fig. 5.25. Restriction mapping of DNA. Note that each experimental result and its interpretation should be considered in sequence, thus building up an increasingly unambiguous map.

also be identified using resources such as the nucleotide identification system (NIX). Here, secondary databases are queried with the unknown sequence and an output is generated indicating any potential promoter, exon/intron sequences, etc.

5.9 MOLECULAR ANALYSIS OF NUCLEIC ACID SEQUENCES

5.9.1 Restriction mapping of DNA fragments

Restriction mapping involves the size analysis of restriction fragments produced by several restriction enzymes individually and in combination (Section 5.6.1). The principle of this mapping is illustrated in Fig. 5.25, in which the restriction sites of two enzymes, A and B, are being mapped. Cleavage with A gives fragments 2 and 7 kilobases (kb) from a 9 kb molecule; hence we can position the single A site 2 kb from one end. Similarly, B gives fragments 3 and 6 kb away, so it has a single site 3 kb from one end; but it is not possible at this stage to say whether it is near to A's site or at the opposite end of the DNA. This can be resolved by a double digestion. If the resultant fragments are 2, 3 and 4 kb away, then A and B cut at opposite ends of the molecule; if they are 1, 2 and 6 kb away, the sites are near to each other. Not surprisingly, the mapping of real molecules is rarely as simple as

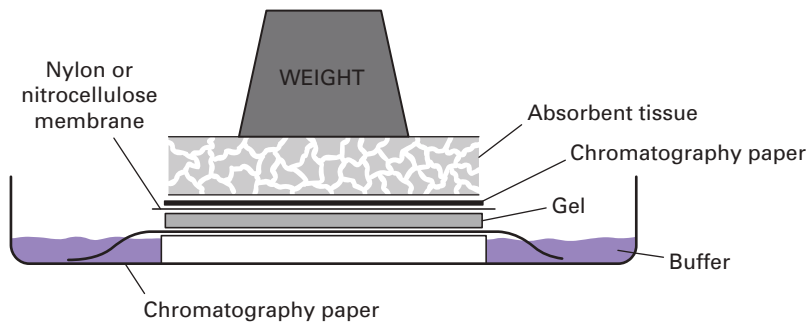


Fig. 5.26. Southern blot apparatus.

this, and bioinformatic analysis of the restriction fragment lengths is usually needed to construct a map.

5.9.2 Nucleic acid blotting methods

Electrophoresis of DNA restriction fragments allows separation based on size; however, it provides no indication as to the presence of a specific, desired fragment among the complex sample. This can be achieved by transferring the DNA from the intact gel onto a piece of nitrocellulose or nylon membrane placed in contact with it. This provides a more permanent record of the sample, since DNA begins to diffuse out of a gel that is left for a few hours. First the gel is soaked in alkali to render the DNA single stranded. It is then transferred to the membrane so that the DNA becomes bound to it in exactly the same pattern as that originally on the gel. This transfer, named a **Southern blot** after its inventor Ed Southern, can be performed electrophoretically or by drawing large volumes of buffer through both gel and membrane, thus transferring DNA from one to the other by capillary action (Fig. 5.26). The point of this operation is that the membrane can now be treated with a labelled DNA molecule, for example a **gene probe** (Section 5.9.3). This single-stranded DNA probe will hybridise under the right conditions to complementary fragments immobilised onto the membrane. The conditions of hybridisation, including the temperature and salt concentration, are critical for this process to take place effectively. This is usually referred to as the **stringency** of the hybridisation and it is particular for each individual gene probe and for each sample of DNA. A series of washing steps with buffer is then carried out to remove any unbound probe and the membrane is developed, like a photograph, after which the precise location of the probe and its target may be visualised. It is also possible to analyse DNA from different species or organisms by blotting the DNA and then using a gene probe representing a protein or enzyme from one of the organisms. In this way it is possible to search for related genes in different species. This technique is generally termed **zoo blotting**.

The same basic process of nucleic acid blotting can be used to transfer RNA from gels onto similar membranes. This allows the identification of specific mRNA

sequences of a defined length by hybridisation to a labelled gene probe and is known as **northern blotting**. It is possible with this technique not only to detect specific mRNA molecules but also to quantify the relative amounts of the specific mRNA. It is usual to separate the mRNA transcripts by gel electrophoresis under denaturing conditions, since this improves resolution and allows a more accurate estimation of the sizes of the transcripts (Section 5.7.2). The format of the blotting may be altered from transfer from a gel to direct application to slots on a specific blotting apparatus containing the nylon membrane. This is termed **slot** or **dot blotting** and provides a convenient means of measuring the abundance of specific mRNA transcripts without the need for gel electrophoresis; however, it does not provide information regarding the size of the fragments.

5.9.3 Design and production of gene probes

The availability of a gene probe is essential in many molecular biological techniques yet in many cases is one of the most difficult steps. The information needed to produce a gene probe may come from many sources. However, the availability of bioinformatics resources and **genetic databases** has ensured that this is the usual starting point for gene probe design.

In some cases it is possible to use related genes, i.e. from the same gene family, to gain information on the most useful DNA sequence to use as a probe. Similar proteins or DNA sequences but from different species may also provide a starting point with which to produce a so-called heterologous gene probe. Although in some cases probes are already produced and cloned it is possible, armed with a DNA sequence from a DNA database, to chemically synthesise a single-stranded **oligonucleotide probe**. This is usually undertaken by computer-controlled **gene synthesisers**, which link dNTPs (deoxyribonucleoside triphosphates) together based on a desired sequence. It is essential to carry out certain checks before probe production to determine that the probe is unique, is not able to self-anneal or that it is self-complementary – all of which may compromise its use.

Where little information on the DNA is available to prepare a gene probe, it is possible in some cases to use the knowledge gained from analysis of the corresponding protein. Thus it is possible to isolate and purify proteins and sequence part of the N-terminal end or an internal region of the protein. From our knowledge of the genetic code, it is possible to predict the various DNA sequences that could code for the protein, and then to synthesise appropriate oligonucleotide sequences chemically. Owing to the degeneracy of the genetic code, most amino acids are coded for by more than one codon, therefore there will be more than one possible nucleotide sequence that could code for a given polypeptide (Fig. 5.27). The longer the polypeptide, the greater is the number of possible oligonucleotides that must be synthesised. Fortunately, there is no need to synthesise a sequence longer than about 20 bases, since this should hybridise efficiently with any complementary sequences and should be specific for one gene. Ideally, a section of the protein should be chosen that contains as many tryptophan and methionine residues as possible, since these have unique codons and there will therefore be

Polypeptide		Phe	Met	Pro	Trp	His	
Corresponding nucleotide sequences	5'	TTC	ATC	CCC A G	TGG	T CAC	3'

Fig. 5.27. Oligonucleotide probes. Note that only methionine and tryptophan have unique codons. It is impossible to predict which of the indicated codons for phenylalanine, proline and histidine will be present in the gene to be probed, so all possible combinations must be synthesised (16 in the example shown).

fewer possible base sequences that could code for that part of the protein. The synthetic oligonucleotides can then be used as probes in a number of molecular biological methods.

5.9.4 Labelling DNA gene probe molecules

An essential feature of a gene probe is that it can be visualised or labelled by some means. This allows any complementary sequence that the probe binds to be flagged up or identified.

There are two main types of label used for gene probes. Traditionally labelling has been carried out using radioactive labels, but non-radioactive labels are gaining in popularity.

Perhaps the most common radioactive label is 32-phosphorus (^{32}P), although for certain techniques 35-sulphur (^{35}S) and tritium (^3H) are used. These may be detected by the process of autoradiography (Section 14.2.3), where the labelled probe molecule, bound to sample DNA, located for example on a nylon membrane, is placed in contact with an X-ray-sensitive film. Following exposure the film is developed and fixed just as a black-and-white negative. The exposed film reveals the precise location of the labelled probe and therefore the DNA to which it has hybridised.

Non-radioactive labels are increasingly being used to label DNA gene probes. Until recently, radioactive labels were more sensitive than their non-radioactive counterparts. However, recent developments have led to similar sensitivities, which, when combined with the improved safety of non-radioactive labels, have led to their greater acceptance.

The labelling systems are termed either direct or indirect. Direct labelling allows an enzyme reporter such as alkaline phosphatase to be coupled directly to the DNA. Although this may alter the characteristics of the DNA gene probe, it offers the advantage of rapid analysis, since no intermediate steps are needed. However, indirect labelling is at present more popular. This relies on the incorporation of a nucleotide that has a label attached. At present three of the main labels in use are biotin, fluorescein and digoxigenin. These molecules are linked covalently to nucleotides using a carbon spacer arm of 7, 14 or 21 atoms. Specific binding proteins may then be used as a bridge between the nucleotide and a

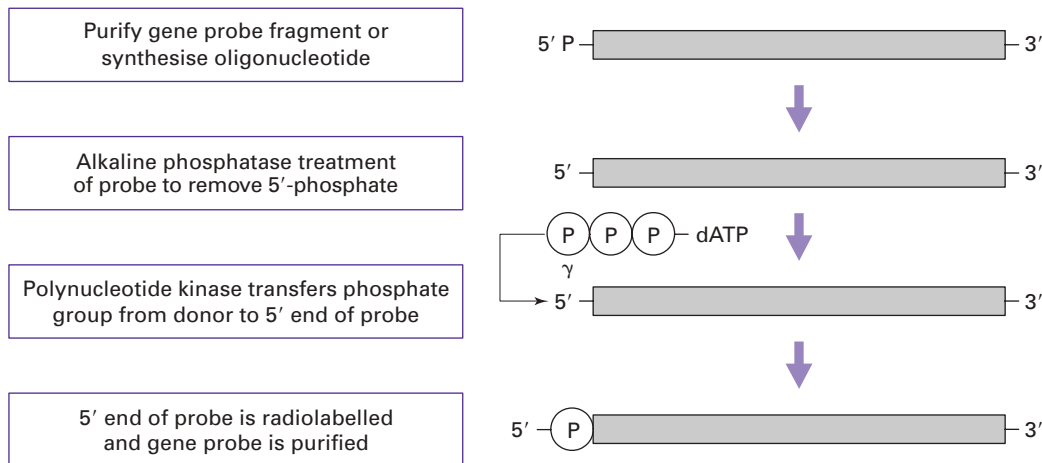


Fig. 5.28. End-labelling of a gene probe at the 5' end with alkaline phosphatase and polynucleotide kinase.

reporter protein such as an enzyme. For example, biotin incorporated into a DNA fragment is recognised with a very high affinity by the protein streptavidin. This may be either coupled or conjugated to a reporter enzyme molecule such as alkaline phosphatase. This is able to convert a colourless substrate *p*-nitrophenol phosphate (PNPP) into a yellow-coloured compound *p*-nitrophenol (PNP) and also offers a means of signal amplification. Alternatively, labels such as digoxigenin incorporated into DNA sequences may be detected by monoclonal antibodies, again conjugated to reporter molecules such as alkaline phosphatase. Thus, rather than the detection system relying on autoradiography, which is necessary for radiolabels, a series of reactions resulting in the production of a colour, light, or the product of a chemiluminescence reaction take place. This has important practical implications, since autoradiography may take 1–3 days whereas colour and chemiluminescent reactions take minutes.

5.9.5 End-labelling of DNA molecules

The simplest form of labelling DNA is by **5' or 3' end-labelling**. 5' End-labelling involves a phosphate transfer or exchange reaction where the 5' phosphate of the DNA to be used as the probe is removed and in its place a labelled phosphate, usually using ^{32}P , is added. This is carried out using two enzymes: the first, alkaline phosphatase, is used to remove the existing phosphate group from the DNA. After removal of the phosphate from the DNA, a second enzyme, polynucleotide kinase, is added, which catalyses the transfer of a phosphate group (^{32}P -labelled) to the 5' end of the DNA. The newly labelled probe is then purified, usually by chromatography through a Sephadex column and may be used directly (Fig. 5.28).

Using the other end of the DNA molecule, the 3' end, is slightly less complex. Here a new, labelled dNTP (e.g. ^{32}P - α dATP or biotin-labelled dNTP) is added to the

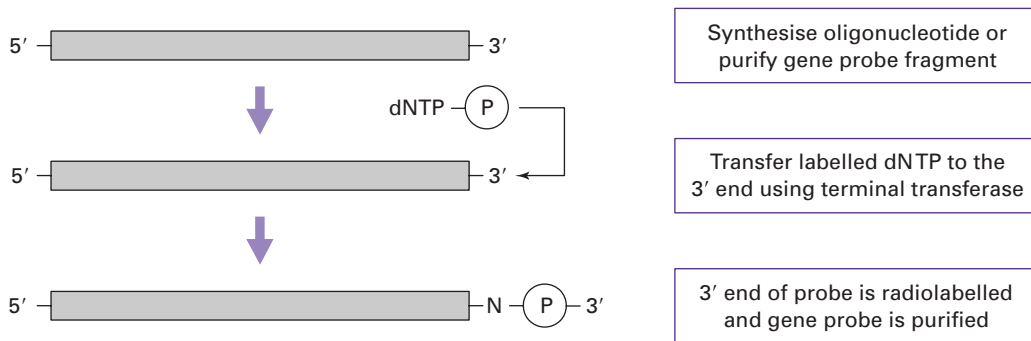


Fig. 5.29. End-labelling of a gene probe at the 3' end using terminal transferase. Note that the addition of a labelled dNTP at the 3' end alters the sequence of the gene probe.

3' end of the DNA by the enzyme terminal transferase. Although this is a simpler reaction, a potential problem exists because a new nucleotide is added to the existing sequence and so the complete sequence of the DNA is altered, which may affect its hybridisation to its target sequence. End-labelling methods also suffer from the fact that only one label is added to the DNA so they are of a lower activity in comparison with methods that incorporate label along the length of the DNA (Fig. 5.29).

5.9.6 Random primer labelling and nick translation

The DNA to be labelled is first denatured and then placed under renaturing conditions in the presence of a mixture of many different random sequences of hexamers or hexanucleotides. These hexamers will, by chance, bind to the DNA sample wherever they encounter a complementary sequence and so the DNA will rapidly acquire an approximately random sprinkling of hexanucleotides annealed to it. Each of the hexamers can act as a primer for the synthesis of a fresh strand of DNA catalysed by DNA polymerase, since it has an exposed 3'-hydroxyl group. The Klenow fragment of DNA polymerase is used for random primer labelling because it lacks a 5' to 3' exonuclease activity. This is prepared by cleavage of DNA polymerase with subtilisin, giving a large enzyme fragment that has no 5' to 3' exonuclease activity, but which still acts as a 5' to 3' polymerase. Thus, when the Klenow enzyme is mixed with the annealed DNA sample in the presence of dNTPs, including at least one that is labelled, many short stretches of labelled DNA will be generated (Fig. 5.30). In a similar way to random primer labelling, the polymerase chain reaction may also be used to incorporate radioactive or non-radioactive labels (Section 5.10.5).

A further traditional method of labelling DNA is by the process of **nick translation**. Low concentrations of DNase I are used to make occasional single-strand nicks in the double-stranded DNA that is to be used as the gene probe. DNA polymerase then fills in the nicks, using an appropriate dNTP, at the same time making a new nick to the 3' side of the previous one (Fig. 5.31). In this way the nick

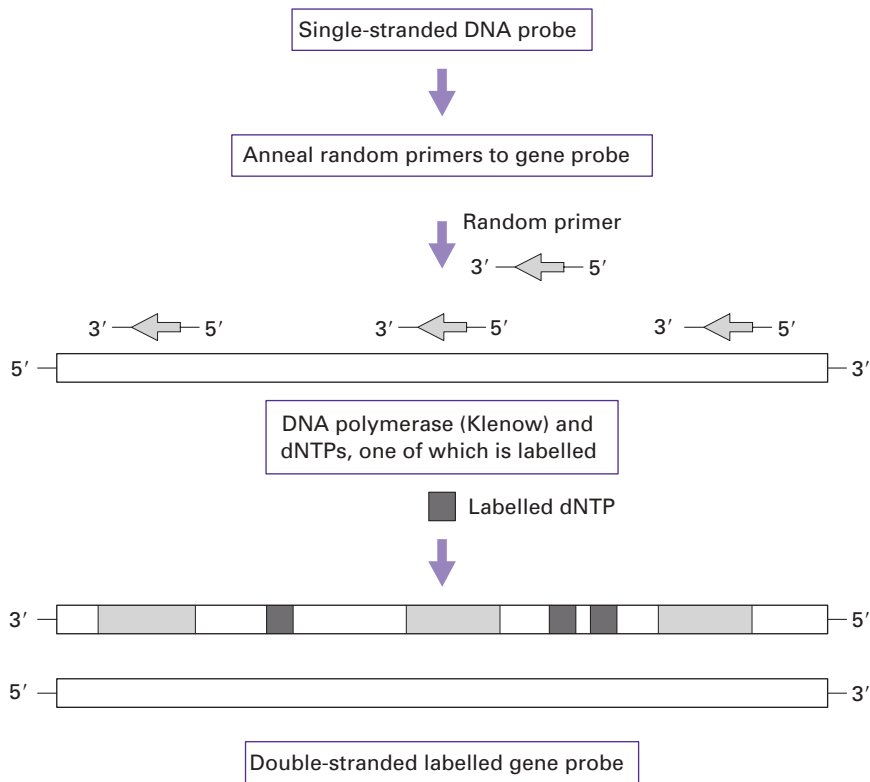


Fig. 5.30. Random primer gene probe labelling. Random primers are incorporated and used as a start point for Klenow DNA polymerase to synthesise a complementary strand of DNA whilst incorporating a labelled dNTP at complementary sites.

is translated along the DNA. If labelled dNTPs are added to the reaction mixture, they will be used to fill in the nicks, and so the DNA can be labelled to a very high specific activity.

5.9.7 Molecular beacon-based probes

A more recent development in the design of labelled oligonucleotide hybridisation probes is that of **molecular beacons**. These probes contain a fluorophore at one end of the probe and a quencher molecule at the other. The oligonucleotide has a stem-loop structure, where the stem places the fluorophore and quencher in close proximity. The loop structure is designed to be complementary to the target sequence. When the stem-loop structure is formed, the fluorophore is quenched by fluorescence resonance energy transfer (FRET), i.e. the energy is transferred from the fluorophore to the quencher and given off as heat. The elegance of these types of probe lies in the fact that, upon hybridisation to a target sequence, the stem and loop move apart, the quenching is then lost and emission of light occurs from the fluorophore upon excitation. These types of probe have also been used to

5.10 The polymerase chain reaction

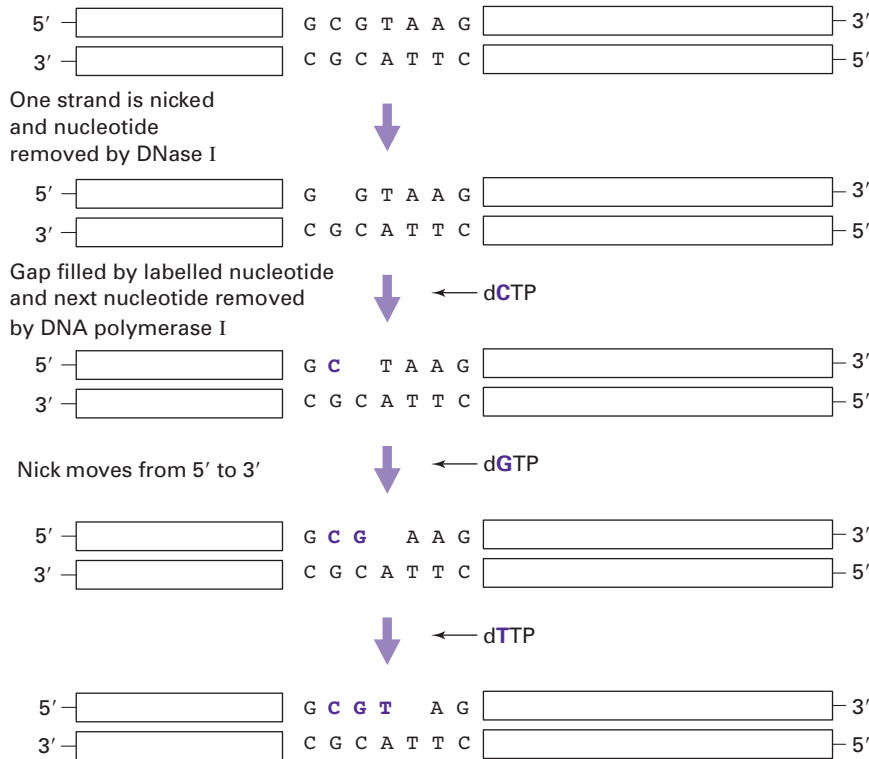


Fig. 5.31. Nick translation. The removal of nucleotides and their subsequent replacement with labelled nucleotides by DNA polymerase I increase the label in the gene probe as nick translation proceeds.

detect nucleic acid amplification systems products such as the polymerase chain reaction (PCR) and have the advantage that it is unnecessary to remove the unhybridised probes.

5.10 THE POLYMERASE CHAIN REACTION

5.10.1 Basic concept of the PCR

The [polymerase chain reaction](#), or PCR, is one of the mainstays of molecular biology. One of the reasons for the wide adoption of the PCR is the elegant simplicity of the reaction and relative ease of the practical manipulation steps. Indeed, combined with the relevant bioinformatics resources for its design and for determination of the required experimental conditions, it provides a rapid means for DNA identification and analysis. It has opened up the investigation of cellular and molecular processes to those outside the field of molecular biology.

The PCR is used to amplify a precise fragment of DNA from a complex mixture of starting material usually termed the [template DNA](#) and in many cases requires little purification of the DNA. It does require the knowledge of some DNA

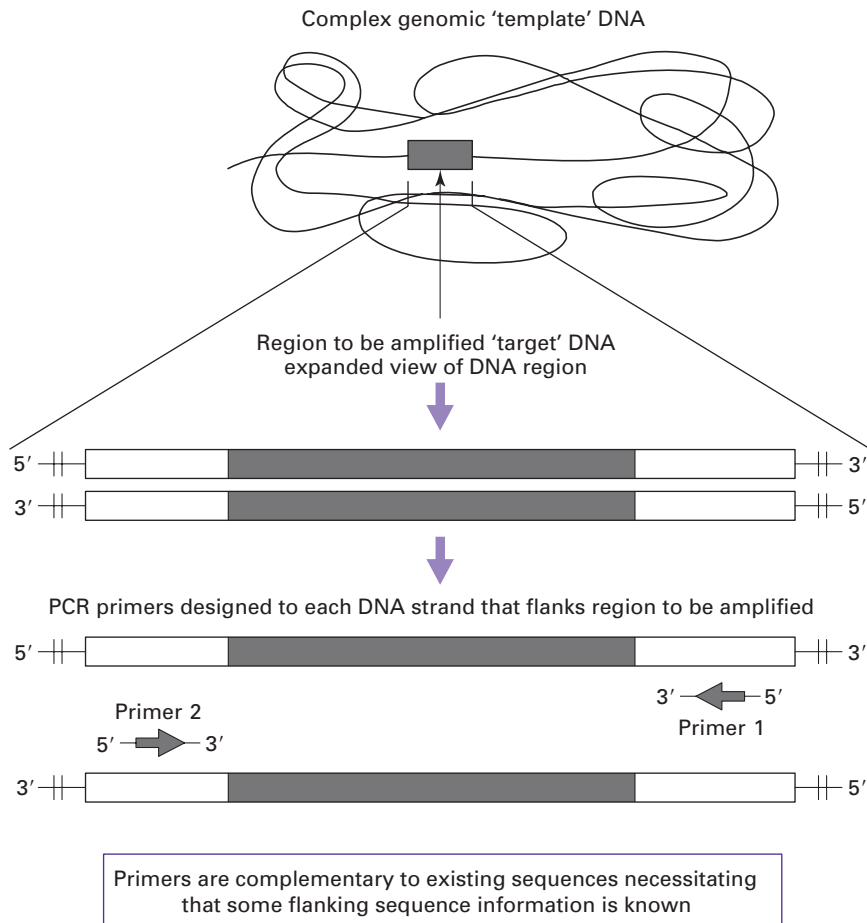


Fig. 5.32. The location of polymerase chain reaction (PCR) primers. PCR primers designed for sequences adjacent to the region to be amplified allowing a region of DNA (e.g. a gene) to be amplified from a complex starting material of genomic template DNA.

sequences which flank the fragment of DNA to be amplified (**target DNA**). From this information two **oligonucleotide primers** may be chemically synthesised, each complementary to a stretch of DNA to the 3' side of the target DNA, one oligonucleotide for each of the two DNA strands (Fig. 5.32). It may be thought of as a technique analogous to the DNA replication process that takes place in cells, since the outcome is the same – the generation of new complementary DNA stretches based upon the existing ones. It is also a technique that has replaced, in many cases, the traditional DNA cloning methods, since it fulfils the same function, the production of large amounts of DNA from limited starting material. However, this is achieved in a fraction of the time needed to clone a DNA fragment (Chapter 6). Although not without its drawbacks, the PCR is a remarkable development that is changing the approach of many scientists to the analysis of nucleic acids and continues to have a profound impact on core biosciences and biotechnology.

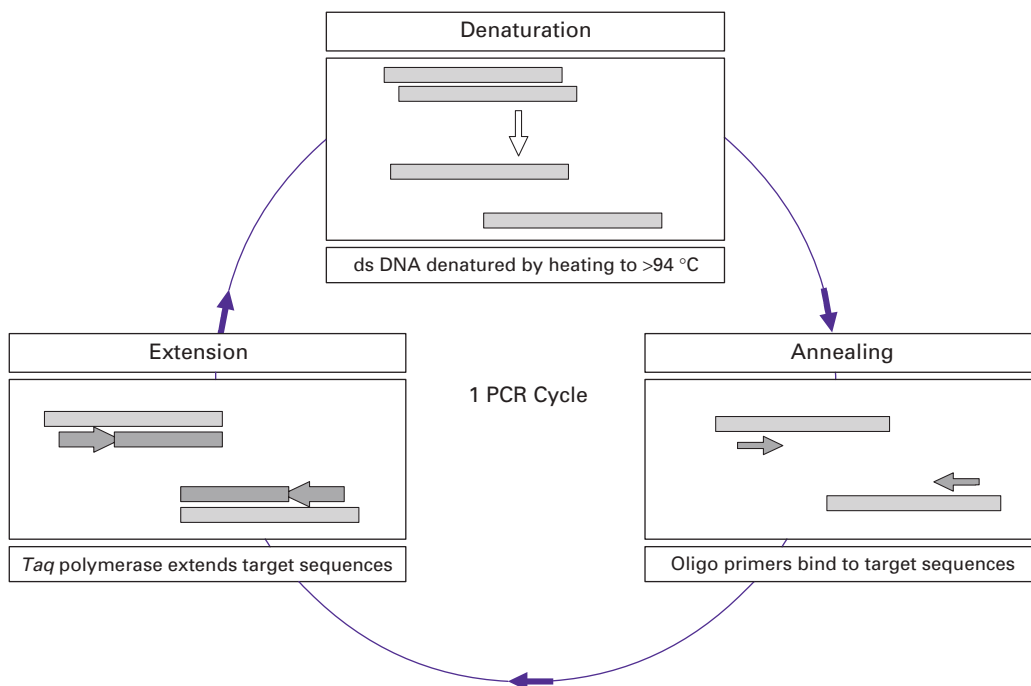


Fig. 5.33. A simplified scheme of one PCR cycle that involves denaturation, annealing and extension. ds, double-stranded.

5.10.2 Stages in the PCR

The PCR consists of three defined sets of times and temperatures termed steps: (i) **denaturation**, (ii) **annealing** and (iii) **extension**. Each of these steps is repeated 30–40 times, termed **cycles** (Fig. 5.33). In the first cycle the double-stranded template DNA is (i) denatured by heating the reaction to above 90°C. Within the complex DNA the region to be specifically amplified (target) is made accessible. The temperature is then cooled to between 40°C and 60°C. The precise temperature is critical and each PCR system has to be defined and optimised. One useful technique for optimisation is **touchdown PCR** where a programmable cycler is used to incrementally decrease the annealing temperature until the optimum is derived. Reactions that are not optimised may give rise to other DNA products in addition to the specific target or may not produce any amplified products at all. The annealing step allows the hybridisation of the two oligonucleotide primers, which are present in excess, to bind to their complementary sites that flank the target DNA. The annealed oligonucleotides act as primers for DNA synthesis, since they provide a free 3'-hydroxyl group for DNA polymerase. The DNA synthesis step is termed **extension** and is carried out by a thermostable DNA polymerase, most commonly *Taq* DNA polymerase.

DNA synthesis proceeds from both of the primers until the new strands have been extended along and beyond the target DNA to be amplified. It is important to note that, since the new strands extend beyond the target DNA they will contain a region near their 3' ends that is complementary to the other primer. Thus, if another round of DNA synthesis is allowed to take place, not only will the original strands be used as templates but also the new strands. Most interestingly, the products obtained from the new strands will have a precise length, delimited exactly by the two regions complementary to the primers. As the system is taken through successive cycles of denaturation, annealing and extension, all the new strands will act as templates and so there will be an exponential increase in the amount of DNA produced. The net effect is to selectively amplify the target DNA and the primer regions flanking it (Fig. 5.34).

One problem with early PCR reactions was that the temperature needed to denature the DNA also denatured the DNA polymerase. However, the availability of a thermostable DNA polymerase enzyme isolated from the thermophilic bacterium *Thermus aquaticus* found in hot springs provided the means to automate the reaction. *Taq* DNA polymerase has a temperature optimum of 72°C and survives prolonged exposure to temperatures as high as 96°C and so is still active after each of the denaturation steps. The widespread utility of the technique is also due to the ability to automate the reaction and, as such, many thermal cyclers have been produced in which it is possible to program in the temperatures and times for a particular PCR reaction.

5.10.3 PCR primer design and bioinformatics

The specificity of the PCR lies in the design of the two oligonucleotide primers. These must not only be complementary to sequences flanking the target DNA but not be self-complementary or bind to each other to form dimers, since both actions prevent DNA amplification. They also have to be matched in their G + C content and have similar annealing temperatures. The increasing use of bioinformatics resources such as Oligo, Genrunner and Genefisher in the design of primers makes the design and the selection of reaction conditions much more straightforward. These resources allow the sequences to be specified; primer length, product size, G + C content etc. to be input; and following analysis provide a choice of matched primer sequences. Indeed the initial selection and design of primers without the aid of bioinformatics would now be unnecessarily time-consuming.

It is also possible to design primers with additional sequences at their 5' end, such as restriction endonuclease target sites or promoter sequences. However, modifications such as these require that the annealing conditions be altered to compensate for the areas of non-homology in the primers. A number of PCR methods have been developed where either one primer or both are random. This gives rise to arbitrary priming in genomic templates but interestingly may give rise to discrete banding patterns when analysed by gel electrophoresis. In many cases this technique may be used reproducibly to identify a particular organism

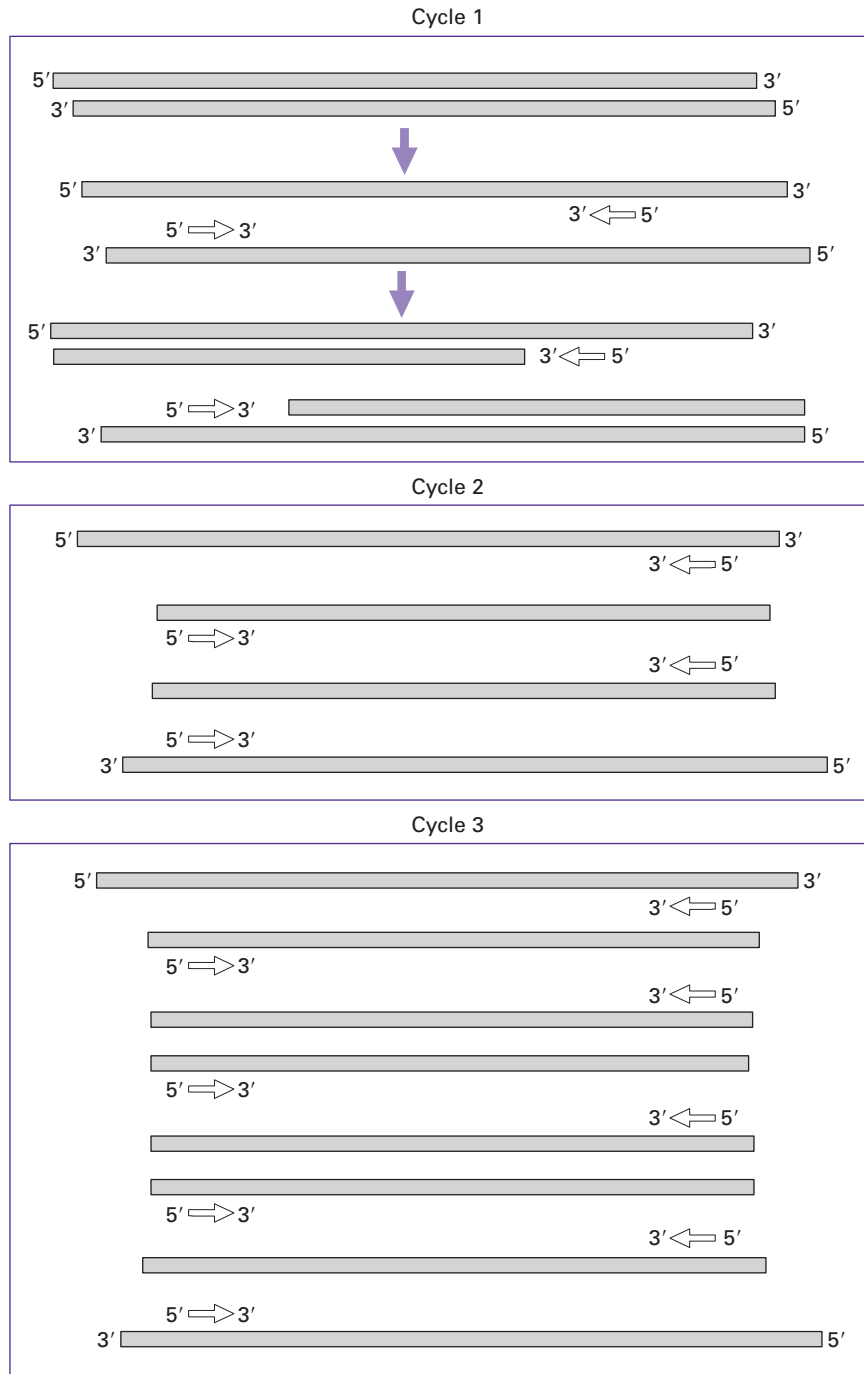


Fig. 5.34. Three cycles in the PCR. As the number of cycles in the PCR increases, the DNA strands that are synthesised and become available as templates are delimited by the ends of the primers. Thus specific amplification of the desired target sequence flanked by the primers is achieved. Primers are denoted as 5' to 3'.

or species. This is sometimes referred to as **rapid amplification of polymorphic DNA** (RAPD) and has been used successfully in the detection and differentiation of a number of pathogenic strains of bacteria. In addition, primers can now be synthesised with a variety of labels such as fluorophores bound to them allowing easier detection and quantitation (Section 5.9.4).

5.10.4 PCR amplification templates

DNA from a variety of sources may be used as the initial source of amplification templates. It is also a highly sensitive technique and requires only one or two molecules for successful amplification. Unlike many manipulation methods used in molecular biology, the PCR technique is sensitive enough to require very little template preparation. The extraction from many prokaryotic and eukaryotic cells may involve a simple boiling step. Indeed the components of many extraction techniques such as SDS and proteinase K may adversely affect the PCR. The PCR may also be used to amplify RNA, a process termed RT-PCR (**reverse transcriptase-PCR**). Initially a reverse transcription reaction that converts the RNA to complementary DNA is carried out (Section 6.2.5). This reaction normally involves the use of the enzyme reverse transcriptase, although some thermostable DNA polymerases used in the PCR such as *Tth* from *Thermus thermophilus*, have a reverse transcriptase activity under certain buffer conditions. This allows mRNA transcription products to be effectively analysed. It may also be used to differentiate latent viruses (detected by standard PCR) or active viruses that replicate and produce transcription products and are thus detectable by RT-PCR (Fig. 5.35). In addition the PCR may be extended to determine relative amounts of a transcription product.

5.10.5 Sensitivity of the PCR

The enormous sensitivity of the PCR system is also one of its main drawbacks, since the very large degree of amplification makes the system vulnerable to contamination. Even a trace of foreign DNA, such as that contained in dust particles, may be amplified to significant levels and may give misleading results. Hence cleanliness is paramount when carrying out PCR, and dedicated equipment, and in some cases dedicated laboratories, are used. It is possible that amplified products may also contaminate the PCR, although this may be overcome by ultraviolet irradiation to damage already amplified products so that they cannot be used as templates. A further interesting solution is to incorporate uracil into the PCR and then treat the products with the enzyme uracil *N*-glycosylase (UNG), which degrades any PCR amplified DNA products or **amplicons** with incorporated uracil, rendering them useless as templates. In addition, most PCRs are now undertaken using **hotstart**. Here, the reaction mixture is physically separated from the template or the enzyme. When the reaction begins, mixing occurs and thus avoids any mispriming that may have arisen.

5.10 The polymerase chain reaction

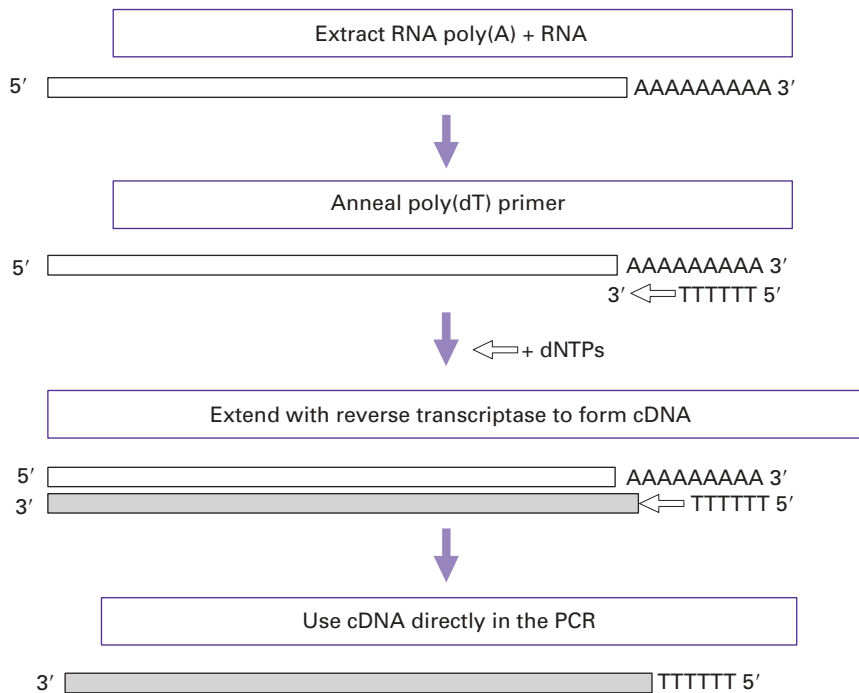


Fig. 5.35. Reverse transcriptase–PCR (RT–PCR), mRNA is converted to complementary DNA (cDNA) using the enzyme reverse transcriptase. The cDNA is then used directly in the PCR.

5.10.6 Applications of the PCR

Many traditional methods in molecular biology have now been superseded by the PCR and the applications for the technique appear to be unlimited. Some of the main techniques derived from the PCR are introduced in Chapter 6, whilst some of the main areas to which the PCR has been put to use are summarised in Table 5.5. The success of the PCR process has given impetus to the development of other amplification techniques that are based on either thermal cycling or non-thermal cycling (isothermal) methods. The most popular alternative to the PCR is termed the **ligase chain reaction** or LCR. This operates in a similar fashion to the PCR but a thermostable DNA ligase joins sets of primers together that are complementary to the target DNA. Following this, a similar exponential amplification reaction takes place producing amounts of DNA that are similar to the PCR. A number of alternative amplification techniques are listed in Table 5.6.

5.10.7 Quantitative and real time PCR

One of the most useful PCR applications is **quantitative PCR** or Q-PCR. This allows the PCR to be used as a means of identifying the initial concentrations of template

Table 5.5 Selected applications of the PCR. A number of the techniques are described in the text of Chapters 5 and 6

Field or area of study	Application	Specific examples or uses
General molecular biology	DNA amplification	Screening gene libraries
Gene probe production	Production/labelling	Use with blots/hybridisations
RNA analysis	RT-PCR	Active latent viral infections
Forensic science	Scenes of crime	Analysis of DNA from blood
Infection/disease monitoring	Microbial detection	Strain typing/analysis RAPDs
Sequence analysis	DNA sequencing	Rapid sequencing possible
Genome mapping studies	Referencing points in genome	Sequence-tagged sites (STS)
Gene discovery	mRNA analysis	Expressed sequence tags (EST)
Genetic mutation analysis	Detection of known mutations	Screening for cystic fibrosis
Quantification analysis	Quantitative PCR	5' Nuclease (TaqMan assay)
Genetic mutation analysis	Detection of unknown mutations	Gel-based PCR methods (DGGE)
Protein engineering	Production of novel proteins	PCR mutagenesis
Molecular archaeology	Retrospective studies	Dinosaur DNA analysis
Single-cell analysis	Sexing or cell mutation sites	Sex determination of unborn
<i>In situ</i> analysis	Studies on frozen sections	Localisation of DNA/RNA

RT, reverse transcriptase; RAPDs, rapid amplification polymorphic DNA; DDGE, denaturing gradient gel electrophoresis.

Table 5.6 Selected alternative amplification techniques to the PCR. Two broad methodologies exist that either amplify the target molecules such as DNA and RNA or detect the target and amplify a signal molecule bound to it

Technique	Type of assay	Specific examples or uses
Target amplification methods		
Ligase chain reaction (LCR)	Non-isothermal, employs thermostable DNA ligase	Mutation detection
Nucleic acid sequence based amplification (NASBA)	Isothermal, involving use of RNA, RNase H/reverse transcriptase, and T7 DNA polymerase	Viral detection, e.g. HIV
Signal amplification methods		
Branched DNA amplification (b-DNA)	Isothermal microwell format using hybridisation or target/capture probe and signal amplification	Mutation detection

HIV, human immunodeficiency virus.

DNA and is very useful for the measurement of, for example, a virus or an mRNA representing a protein expressed in abnormal amounts in a disease process. Early quantitative PCR methods involved the comparison of a standard or control DNA template amplified with separate primers at the same time as the specific target DNA. These types of quantification rely on the reaction being exponential and

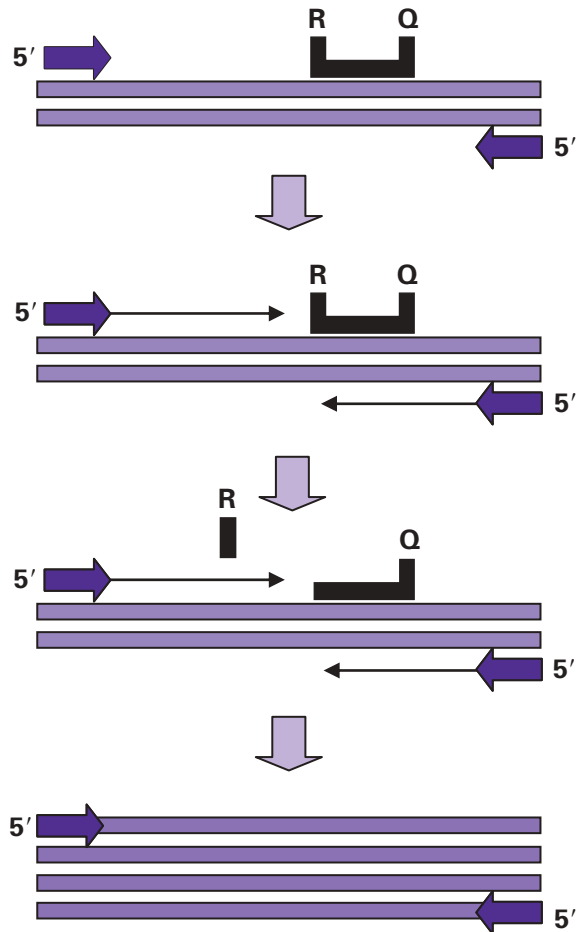


Fig. 5.36. 5' Nuclease assay (TaqMan assay). PCR is undertaken with RQ probe (reporter/quencher dye). As R-Q are in close proximity, fluorescence is quenched. During extension by *Taq* polymerase the probe is cleaved as a result of *Taq* having 5' nuclease activity. This cleaves R-Q probe and the reporter is released. This results in detectable increase in fluorescence and allows real time PCR detection.

so any factors affecting this may also affect the result. Other methods involve the incorporation of a radiolabel through the primers or nucleotides and their subsequent detection following purification of the amplicon. An alternative automated real time PCR method is the [5' fluorogenic exonuclease detection system](#) or [TaqMan assay](#) (Fig. 5.36). In its simplest form, a DNA binding dye such as SYBR Green is included in the reaction. As amplicons accumulate, SYBR green binds the double-stranded DNA proportionally. Fluorescence emission of the dye is detected after excitation. The binding of SYBR Green is non-specific. Therefore in order to detect specific amplicons an oligonucleotide probe labelled with a fluorescent reporter and quencher molecule, respectively, at either end is included in the reaction in place of SYBR Green. When the oligonucleotide probe binds to the target

sequence the 5' exonuclease activity of *Taq* polymerase degrades and releases the reporter from the quencher. A signal is thus generated that increases in direct proportion to the number of starting molecules. Thus a detection system is able to induce and detect fluorescence in real time as the PCR proceeds. Part of the system relies on the capillary-based thermal cycling using a specialised thermal cycler, for example the Roche [light cycler](#). Although relatively expensive in comparison with other methods for determining expression levels, it is simple, rapid and reliable. In addition to quantification, real time PCR systems may also be used for genotyping and for accurate determination of amplicon melting temperature using melting curve analysis.

This allows accurate amplicon identification and also offers the potential to detect mutations and SNPs. Further developments in probe-based PCR systems have also been used and include scorpion probe systems, amplifluor and real time LUX probes.

5.11 NUCLEOTIDE SEQUENCING OF DNA

5.11.1 Concepts of nucleic acid sequencing

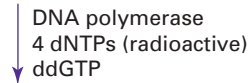
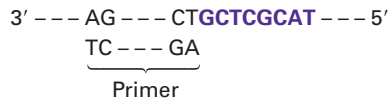
The determination of the order or sequence of bases along a length of DNA is one of the central techniques in molecular biology. Although it is now possible to derive amino acid sequence information with a degree of reliability, it is frequently more convenient and rapid to analyse the DNA coding information. Knowledge of the precise usage of codons, information regarding mutations and polymorphisms and the identification of gene regulatory control sequences are also possible only by analysing DNA sequences. Two techniques have been developed for this, one based on an enzymatic method frequently termed [Sanger sequencing](#), after its developer, and a chemical method called Maxam and Gilbert, named for the same reason. At present, Sanger sequencing is by far the most popular method and many commercial kits are available for its use. However, there are certain occasions such as the sequencing of short oligonucleotides where the Maxam–Gilbert method is more appropriate.

One absolute requirement for Sanger sequencing is that the DNA to be sequenced is in a single-stranded form. Traditionally this demanded that the DNA fragment of interest be inserted and cloned into a specialised bacteriophage vector termed M13, which is naturally single stranded (Section 6.3.3). Although M13 is still universally used, the advent of the PCR has provided the means not only to amplify a region of any genome or cDNA but also very quickly to generate the corresponding nucleotide sequence. This has led to an explosion in the accumulation of DNA sequence information and has provided much impetus for gene discovery and genome mapping (Section 6.9).

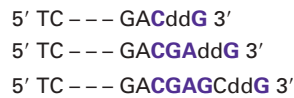
The Sanger method is simple and elegant and mimics in many ways the natural ability of DNA polymerase to extend a growing nucleotide chain from an existing template. Initially the DNA to be sequenced is allowed to hybridise with an oligonucleotide primer that is complementary to a sequence adjacent to the 3'

5.11 Nucleotide sequencing of DNA

Fragment to be sequenced, cloned in M13 phage



Synthesis of complementary second strands:



Denature to give single strands

Run on sequencing gel alongside products of ddCTP, ddATP and ddTTP reactions

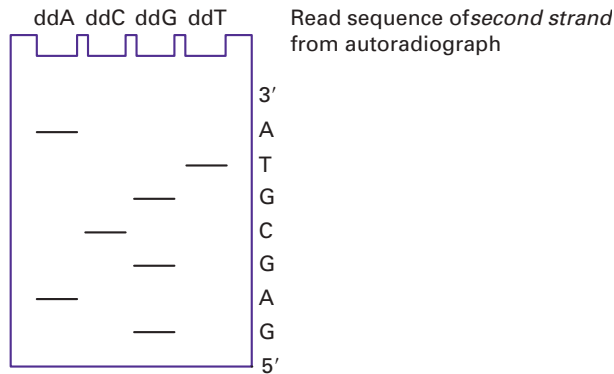


Fig. 5.37. Sanger sequencing of DNA.

side of the DNA within a vector such as M13 or in an amplicon. The oligonucleotide will then act as a primer for synthesis of a second strand of DNA, catalysed by DNA polymerase (Fig. 5.37). Since the new strand is synthesised from its 5' end, virtually the first DNA to be made will be complementary to the DNA to be sequenced. One of the dNTPs that must be provided for DNA synthesis is labelled with ^{32}P or ^{35}S , and so the newly synthesised strand will be radioactively labelled. This reaction mixture is left to incubate at room temperature for a few minutes.

5.11.2 Dideoxynucleotide chain terminators

The reaction mixture is then divided into four aliquots, representing the four dNTPs, A, C, G and T. In addition to all of the dNTPs being present in the A tube an analogue of dATP is added (2',3'-dideoxyadenosine triphosphate (ddATP)) that is similar to A but has no 3' hydroxyl group and so will terminate the growing chain

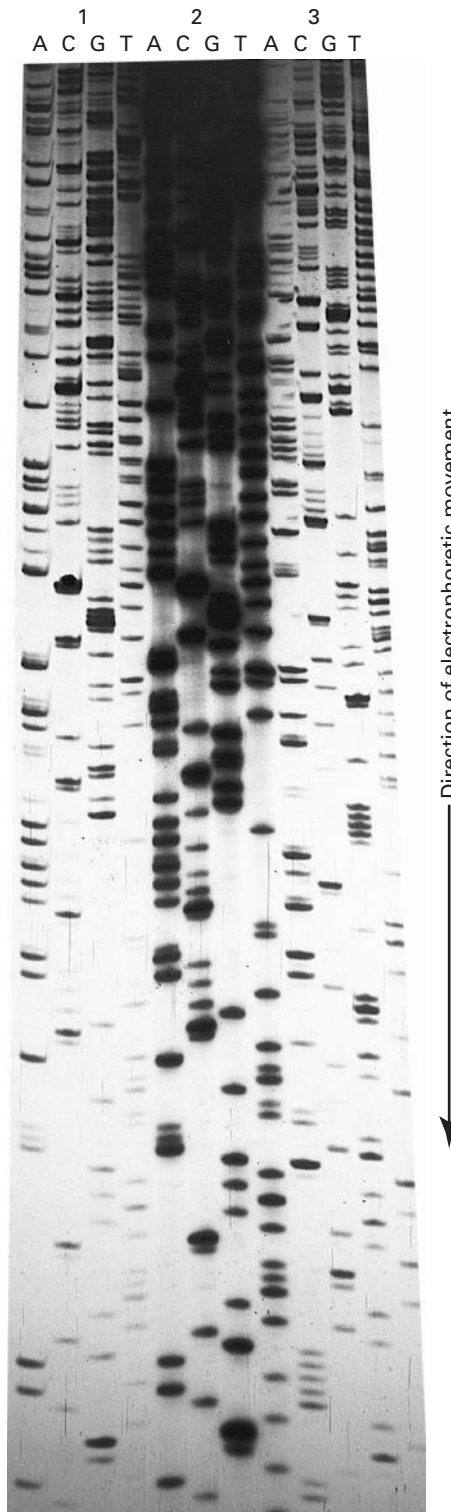


Fig. 5.38. Autoradiograph of a DNA sequencing gel. Samples were prepared using the Sanger dideoxy method of DNA sequencing. Each set of four samples was loaded into adjacent tracks, indicated by A, C, G and T, depending on the identity of the dideoxynucleotide used for that sample. Two sets of samples were labelled with ^{35}S (1 and 3) and one was labelled with ^{32}P (2). It is evident that ^{32}P generates darker but more diffuse bands than does ^{35}S , making the bands nearer the bottom of the autoradiograph easy to see. However, the broad bands produced by ^{32}P cannot be resolved near the top of the autoradiograph, making it impossible to read a sequence from this region. The much sharper bands produced by ^{35}S allow sequences to be read with confidence along most of the autoradiograph and so a longer sequence of DNA can be obtained from a single gel.

since a 5' to 3' phosphodiester linkage cannot be formed without a 3'-hydroxyl group. The situation for tube C is identical, except that ddCTP is added; similarly the G and T tubes contain ddGTP and ddTTP, respectively.

Since the incorporation of ddNTP rather than dNTP is a random event, the reaction will produce new molecules varying widely in length, but all terminating in the same type of base. Thus four sets of DNA sequences are generated, each terminating in a different type of base, but all having a common 5' end (the primer). The four labelled and chain-terminated samples are then denatured by heating and loaded next to each other on a polyacrylamide gel for electrophoresis. Electrophoresis is performed at approximately 70 °C in the presence of urea, to prevent renaturation of the DNA, since even partial renaturation alters the rates of migration of DNA fragments. Very thin, long gels are used for maximum resolution over a wide range of fragment lengths. After electrophoresis, the positions of radioactive DNA bands on the gel are determined by autoradiography. Since every band in the track from the ddATP sample must contain molecules that terminate at adenine, and those in the ddCTP terminate in cytosine, etc., it is possible to read the sequence of the newly synthesised strand from the autoradiograph, provided that the gel can resolve differences in length equal to a single nucleotide (Fig. 5.38). Under ideal conditions, sequences up to about 300 bases in length can be read from one gel.

5.11.3 Direct PCR pyrosequencing

Rapid PCR sequencing has also been made possible by the use of **pyrosequencing**. This is a sequencing by synthesis whereby a PCR template is hybridised to an oligonucleotide and incubated with DNA polymerase, ATP sulphurylase, luciferase and apyrase. During the reaction, the first of the four dNTPs are added and, if incorporated, release pyrophosphate (PP_i). The ATP sulphurylase converts the PP_i to ATP, which drives the luciferase-mediated conversion of luciferin to oxyluciferin to generate light. Apyrase degrades the resulting component dNTPs and ATP. This is followed by another round of dNTP addition. A resulting pyrogram provides an output of the sequence. The method provides short reads very quickly and is especially useful for the determination of mutations or SNPs.

It is also possible to undertake nucleotide sequencing directly from double-stranded molecules such as plasmid cloning vectors and PCR amplicons. The double-stranded DNA must be denatured prior to annealing with primer. In the case of plasmids an alkaline denaturation step is sufficient; however, for amplicons this is more problematic and a focus of much research. Unlike plasmids, amplicons are short and reanneal rapidly, therefore preventing the reannealing process or biasing the amplification towards one strand by using a primer ratio of 100:1 overcomes this problem to a certain extent. Denaturants such as formamide or DMSO have also been used with some success in preventing the reannealing of PCR strands following their separation.

It is possible to physically separate and retain one PCR strand by incorporating a molecule such as biotin into one of the primers. Following PCR, one strand with an affinity molecule may be removed by affinity chromatography with streptavidin,

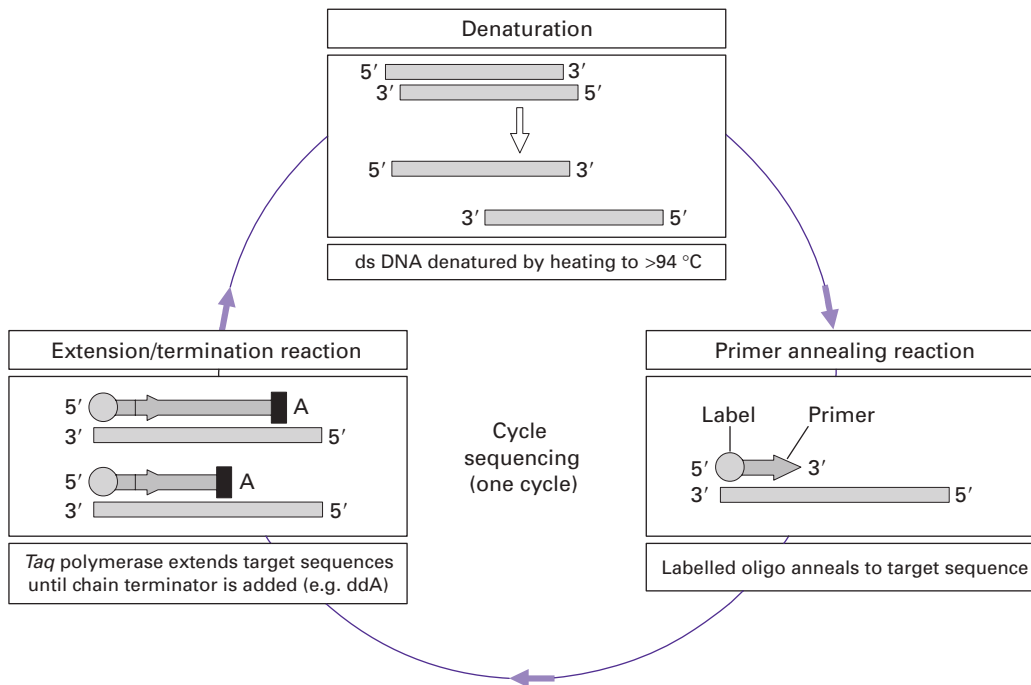


Fig. 5.39. Simplified scheme of cycle sequencing. Linear amplification takes place with the use of labelled primers. During the extension and termination reaction, the chain terminator dideoxynucleotides are incorporated into the growing chain. This takes place in four separate reactions (A, C, G and T). The products are then run on a polyacrylamide gel and the sequence analysed. The scheme indicates the events that take place in the A reaction only. ds, double-stranded.

leaving the complementary PCR strand. This affinity purification provides single-stranded DNA derived from the PCR amplicon and, although it is somewhat time consuming, does provide high quality single-stranded DNA for sequencing.

5.11.4 PCR cycle sequencing

One of the most useful methods of sequencing PCR amplicons is termed **PCR cycle sequencing**. This is not strictly a PCR, since it involves linear amplification with a single primer. Approximately 20 cycles of denaturation, annealing and extension take place. Radiolabelled or fluorescently labelled dideoxynucleotides are then introduced into the final stages of the reaction to generate the chain terminated extension products (Fig. 5.39). Automated direct PCR sequencing is increasingly being refined, allowing greater lengths of DNA to be analysed in one sequencing run, and provides a very rapid means of analysing DNA sequences.

5.11.5 Automated fluorescent DNA sequencing

Advances in fluorescent dye terminator and labelling chemistry has led to the development of high throughput automated sequencing techniques. Essentially most systems involve the use of dideoxynucleotides labelled with different fluorochromes. Thus the label is incorporated into the ddNTP and this is used to carry out chain termination as in the standard reaction indicated in Section 5.11.1. The advantage of this modification is that, since a different label is incorporated with each ddNTP, it is unnecessary to perform four separate reactions. Therefore the four chain-terminated products are run on the same track of a denaturing electrophoresis gel. Each product with their base-specific dye is excited by a laser and the dye then emits light at its characteristic wavelength. A diffraction grating separates the emissions, which are detected by a [charge-coupled device](#) and the sequence interpreted by a computer. The advantages of the technique include real time detection of the sequence. In addition the lengths of sequence that may be analysed are in excess of 500 bp (Fig. 5.40). Capillary electrophoresis is increasingly being used for the detection of sequencing products. This is where liquid polymers in thin capillary tubes are used, obviating the need to pour sequencing gels and requiring little manual operation. This substantially reduces the electrophoresis run times and allows high throughput to be achieved. A number of large-scale sequence facilities are now fully automated using 96-well microtitre-based formats. The derived sequences can be downloaded automatically to databases and manipulated using a variety of bioinformatics resources. Developments in the technology of DNA sequencing have made whole-genome sequencing projects a realistic proposition within achievable time scales, and a number of these have been completed or are nearing completion.

5.11.6 Maxam and Gilbert sequencing

Sanger sequencing is by far the most popular technique for DNA sequencing; however, an alternative technique developed at the same time may also be used. The chemical cleavage method of DNA sequencing developed by A. M. Maxam and W. Gilbert is often used for sequencing small fragments of DNA such as oligonucleotides, where Sanger sequencing is problematic. A radioactive label is added to either the 3' or the 5' end of a double-stranded DNA sample (Fig 5.41). The strands are then separated by electrophoresis under denaturing conditions, and analysed separately. DNA labelled at one end is divided into four aliquots and each is treated with chemicals that act on specific bases by methylation or removal of the base. Conditions are chosen so that, on average, each molecule is modified at only one position along its length; every base in the DNA strand has an equal chance of being modified. After the modification reactions, the separate samples are cleaved by piperidine, which breaks phosphodiester bonds exclusively at the 5' side of nucleotides whose base has been modified. The result is similar to that produced by the Sanger method, since each sample now contains radioactively labelled molecules of various lengths, all with one end in common (the labelled

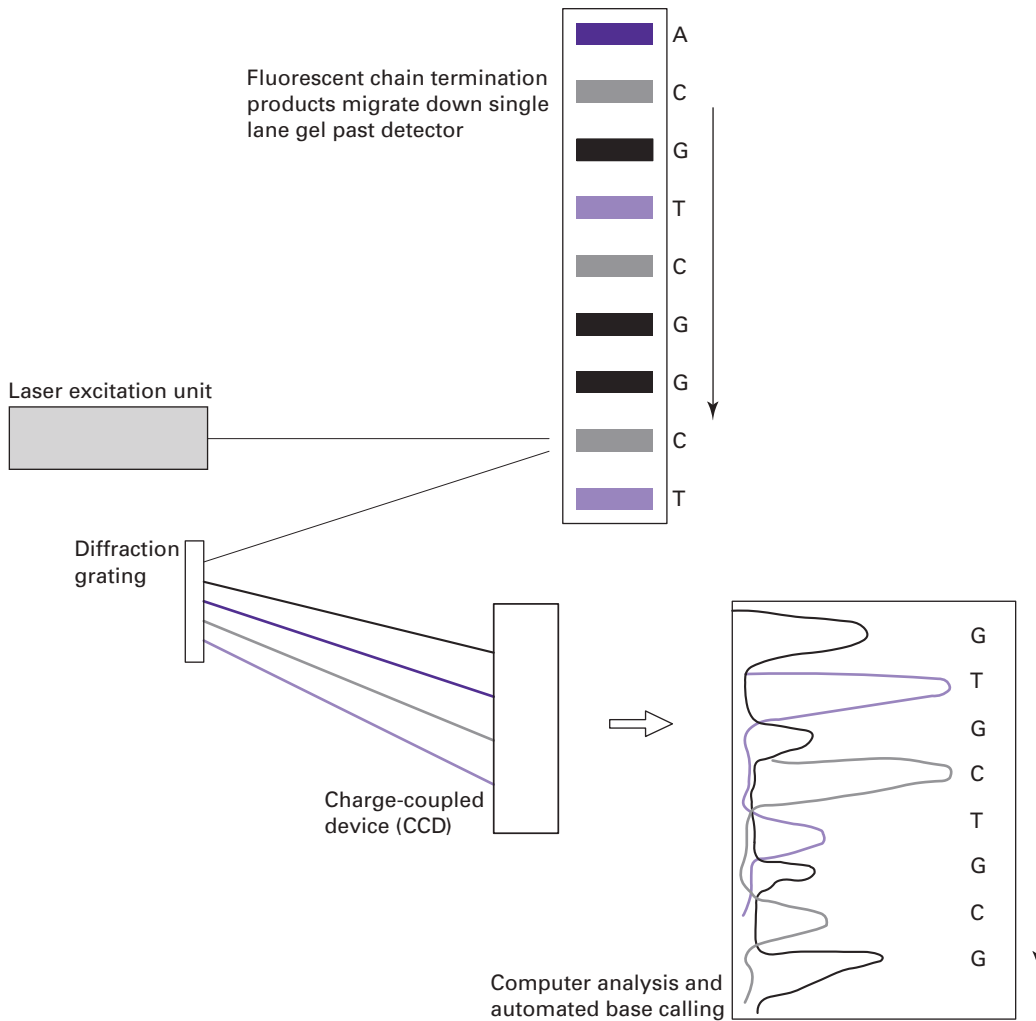


Fig. 5.40. Automated fluorescent sequencing detection using single-lane gel and charge-coupled device.

end), and with the other end cut at the same type of base. Analysis of the reaction products by electrophoresis is as described for the Sanger method.

The developments in DNA sequencing and techniques such as the PCR have allowed a means of rapidly identifying and analysing biological molecules. This, coupled with the rapid advancements in bioinformatics and the increasingly sophisticated methods of protein structure prediction, has led to the generalised scheme of work flow indicated in Fig. 5.42. These new methods of *in silico* methods will no doubt accelerate the understanding of molecular structure–function interactions in the coming years and be a central focus of bioscience research.

5.11 Nucleotide sequencing of DNA

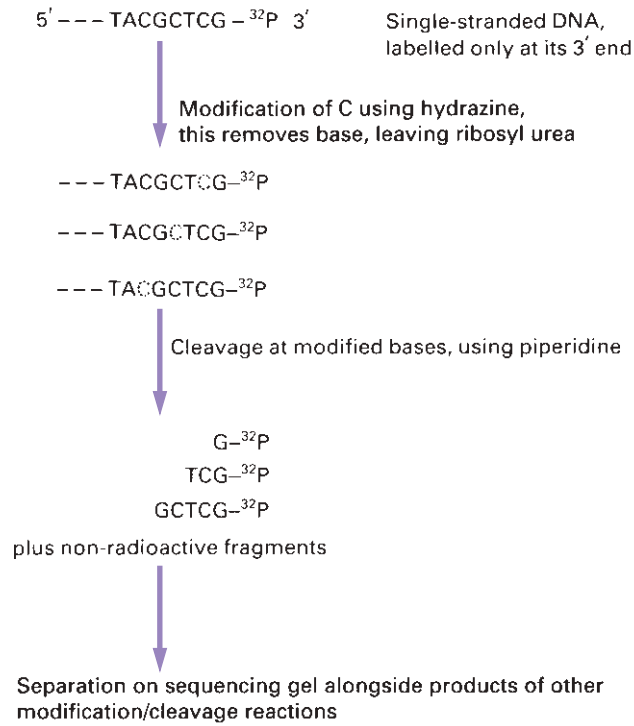


Fig. 5.41. Maxam and Gilbert sequencing of DNA. Only modification and cleavage of deoxycytidine is shown, but three more portions of the end-labelled DNA would be modified and cleaved at G, G + A, and T + C, and the products would be separated on the sequencing gel alongside those from the C reactions.

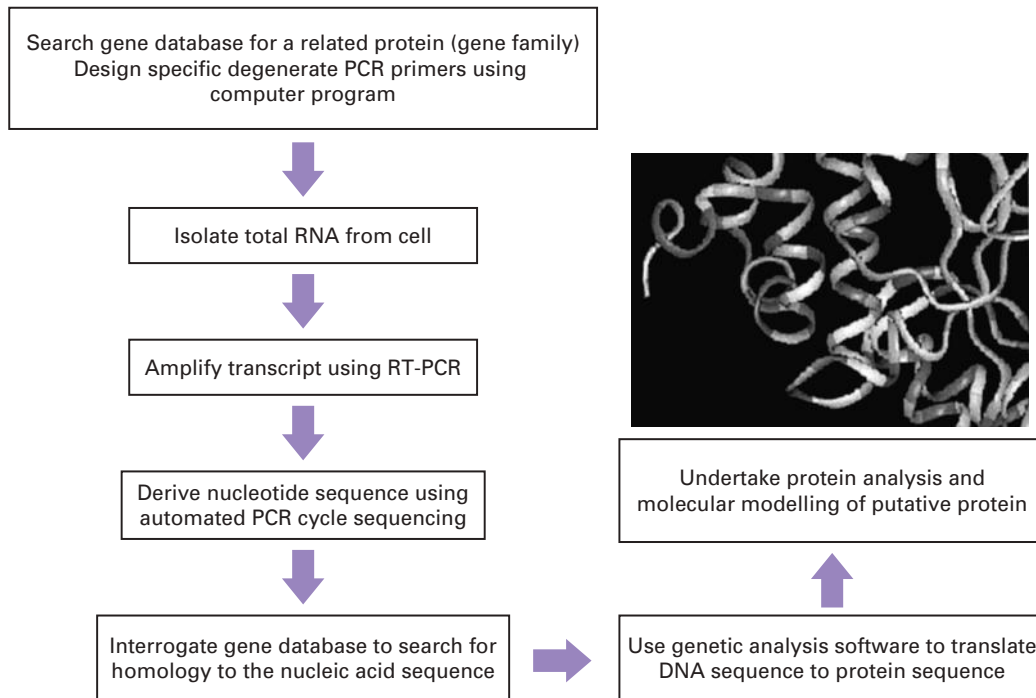


Fig. 5.42. Possible generalised scheme of work using bioinformatics to generate protein information.

5.12 SUGGESTIONS FOR FURTHER READING

- ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K. and WALTER, P. (2002). *Molecular Biology of the Cell*, 4th edn. Garland Publishing, New York (A comprehensive all round text.)
- BROWN, T. A. (2001). *Gene Cloning and DNA Analysis: An Introduction*. Blackwell Science, Oxford. (A very good introduction to genetics and molecular biology.)
- CAMPBELL, A. M. and HEYER, L. J. (2002). *Discovering Genomics, Proteomics and Bioinformatics*. Addison Wesley, Boston, MA. (A very useful resource for genomics and bioinformatics.)
- NEWTON, C. R. and GRAHAM, A. (1997). *PCR*, 2nd edn. Bios Scientific Publishers Ltd, Oxford. (An excellent introduction to the methods and application of the PCR.)
- RAPLEY, R. and HARBRON, S. (2004). *Molecular Analysis and Genome Discovery*. John Wiley & Sons, Chichester. (An up-to-date collection of key nucleic acid and proteins techniques in analysis and drug discovery.)
- READ, A. P. and STRACHEN, T. (2004). *Human Molecular Genetics*. Garland Science, London. (An excellent and very comprehensive textbook with excellent illustrations.)

Recombinant DNA and genetic analysis

6.1 INTRODUCTION

The genomics era has provided a new approach to understanding and discovering biological processes. Indeed the many genome mapping and sequencing projects completed or under way now require new methods of analysis such as automated microarray technology and bioinformatics. New areas have recently been developed, such as pharmacogenomics, metabolomics and systems biology, all of which aim to analyse large numbers of samples simultaneously. This type of massive parallel analysis is set to be the main driving force of discovery and analysis in the coming years. However, developing techniques of molecular biology and genetic analysis have their foundations in methods developed 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 led not only to the possibility of analysing DNA more effectively but also to 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 led to numerous discoveries and insights into gene structure, function and regulation. Since their initial use, methods for the production of gene libraries have been steadily refined and developed. Although the polymerase chain reaction (PCR; Section 5.10) has provided shortcuts to gene analysis, there are still many cases where gene cloning methods are not only useful but 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

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 (<i>E. coli</i>)	4000	700	340
Yeast	20 000	3500	1700
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

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 (4^6) 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 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, called DNA ligase, in a process termed **ligation**. This enzyme, usually isolated from bacteriophage T4 and called T4 DNA ligase, forms a covalent bond between the 5'-phosphate group at

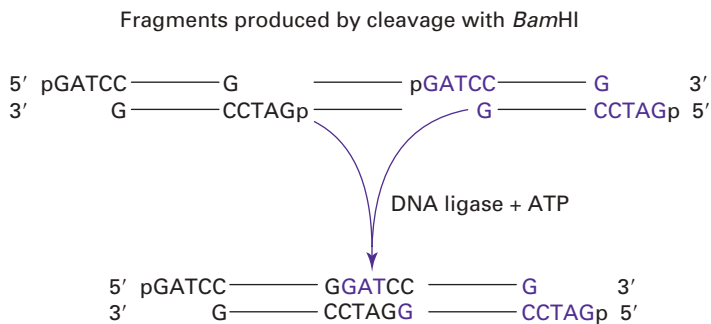


Fig. 6.1. Ligation molecules with cohesive ends. Complementary cohesive ends base-pair, forming a temporary link between two DNA fragments. This association of fragments is stabilised by the formation of 3' to 5' phosphodiester linkages between cohesive ends, a reaction catalysed by DNA ligase.

the end of one strand and the 3'-hydroxyl group 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 that of 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 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 then allows the whole recombined DNA to 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 **genome library**, which consists of the total chromosomal DNA of an organism, and a **cDNA library**, which represents the mRNA from a 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 its architecture, then genome 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

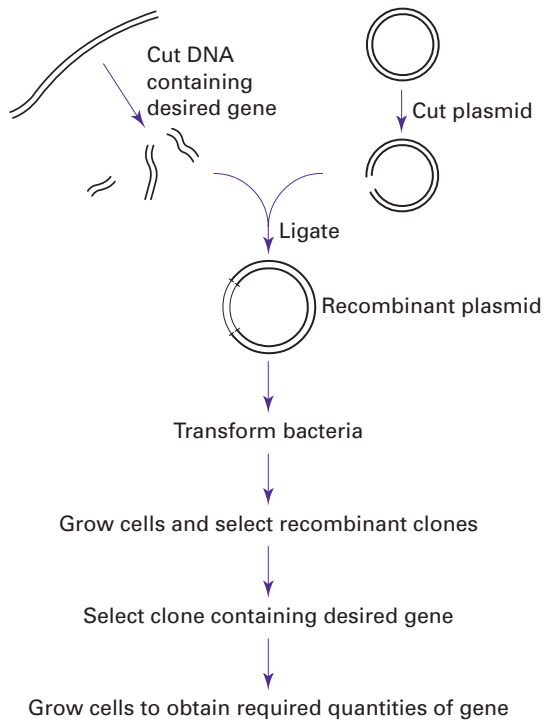


Fig. 6.2. Outline of gene cloning.

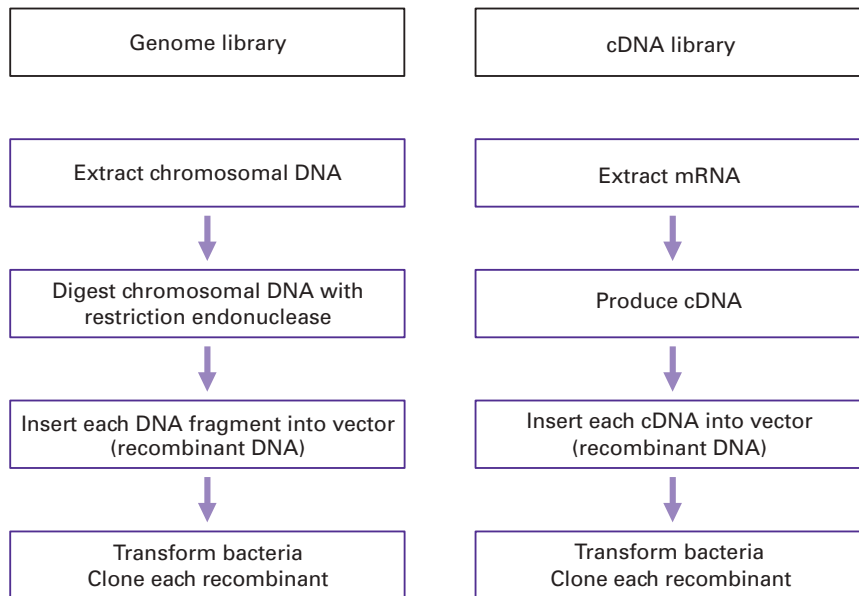


Fig. 6.3. Comparison of the general steps involved in the construction of genomic and complementary DNA (cDNA) libraries.

6.2 Constructing gene libraries

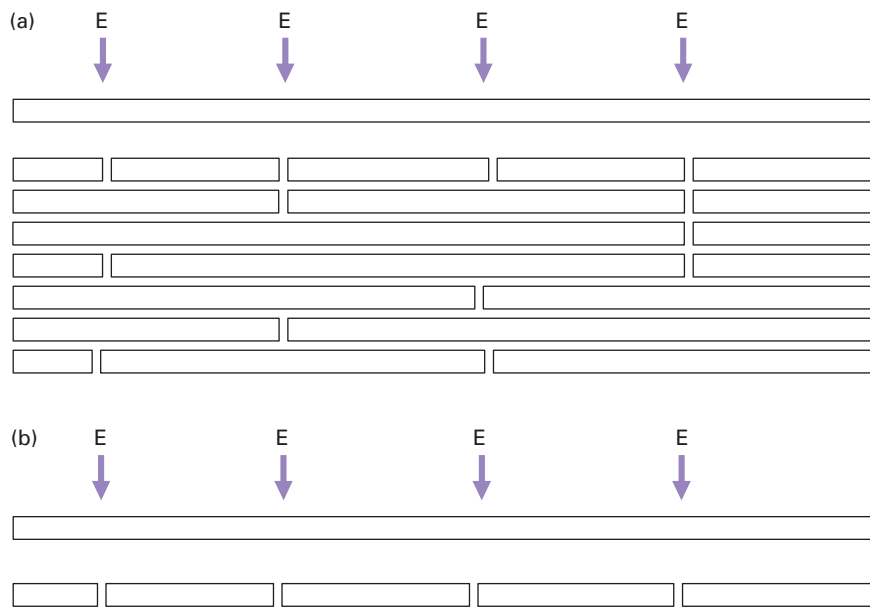


Fig. 6.4. Comparison of (a) partial and (b) complete digestion of DNA molecules at restriction enzymes sites (E).

consideration in the construction of genome or cDNA libraries is therefore 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 a genome library. In contrast, however, cDNA libraries represent only the mRNA being produced from a specific cell type at a particular time in the cell's development. 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 also 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 Genome DNA libraries

Genomic libraries are constructed by isolating the complete chromosomal DNA from a cell and digesting it into fragments of the desired average length with restriction endonucleases. This can be achieved by [partial restriction digestion](#) with 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 may need to be repaired to make them flush ended. This is achieved by using a modified DNA polymerase termed Klenow polymerase. The enzyme is prepared by cleavage of DNA polymerase with subtilisin, giving a large fragment that 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 deoxyribonucleoside triphosphates (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 *Escherichia 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 mRNA molecules, 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 cDNA molecules to all the mRNAs from the selected tissue. The cDNA may be inserted into vectors and then cloned. The production of cDNA 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 anneal randomly 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 of 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

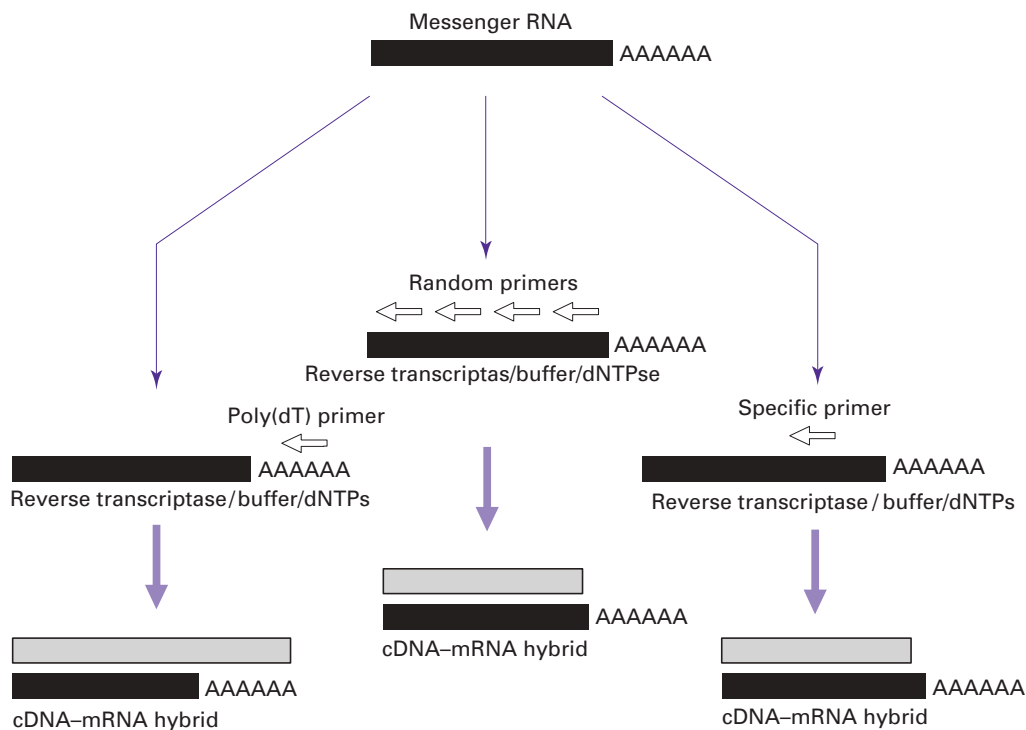


Fig. 6.5. Strategies for producing first-strand cDNA from mRNA.

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 the poly(A) of mRNA. Alkaline hydrolysis is then used to remove the RNA strand, leaving single-stranded DNA that 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 prepared 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 that 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

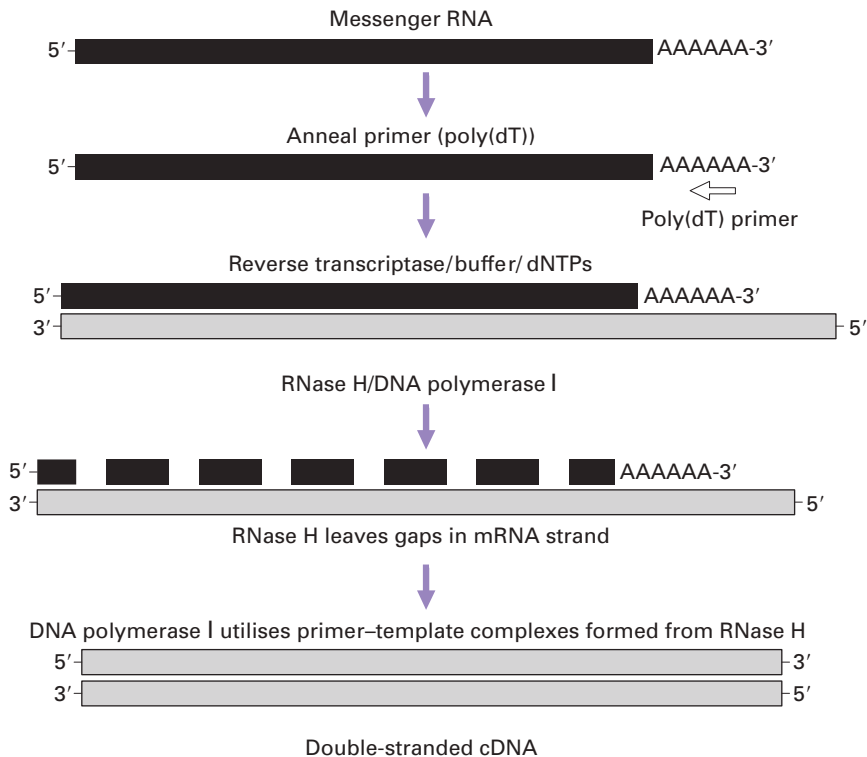


Fig. 6.6. Second-strand cDNA synthesis using the RNase H method.

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 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 pre-formed 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 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 is 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

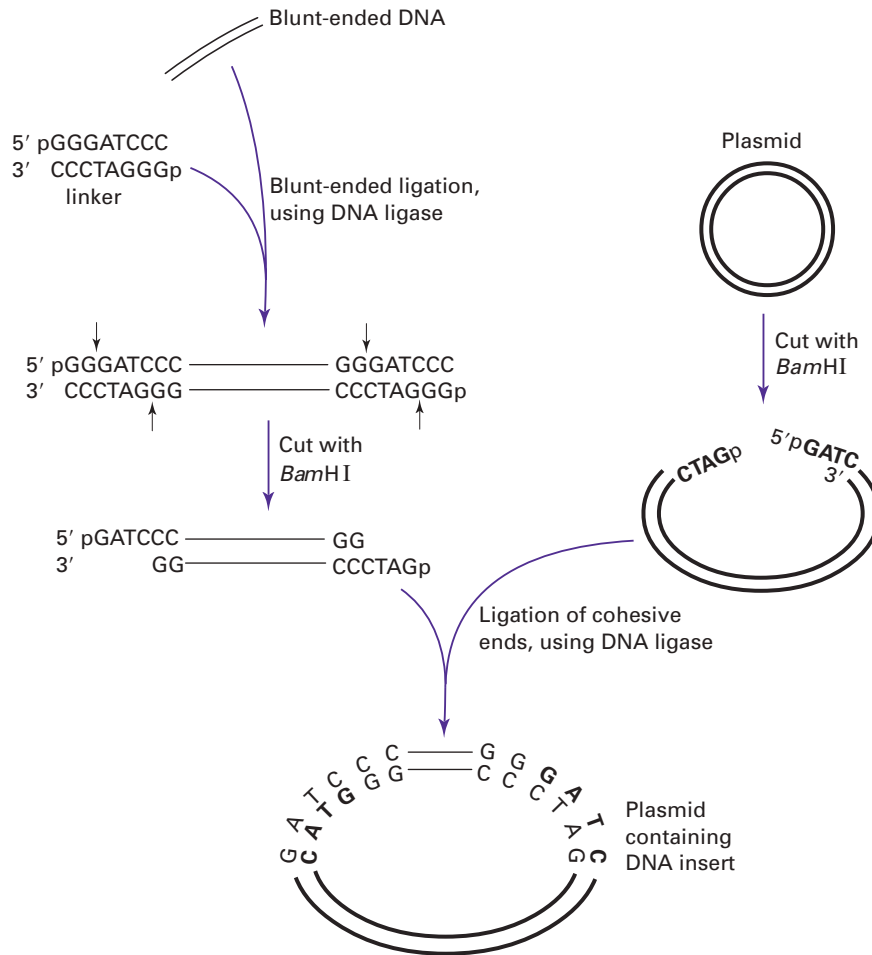


Fig. 6.7. Use of linkers. In this example, blunt-ended DNA is inserted into a specific restriction site on a plasmid, after ligation to a linker containing the same restriction site.

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 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.

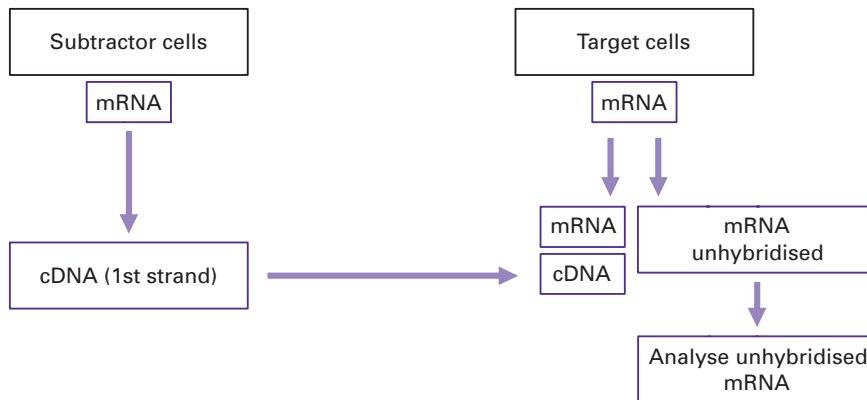


Fig. 6.8. Scheme of analysing specific mRNA molecules by subtractive hybridisation.

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 to analysing the differential expression of genes has been developed using the polymerase chain reaction (PCR). This technique, termed differential display, is explained in greater detail in Section 6.8.1.

6.2.9 Cloning PCR products

Whilst PCR has to some extent replaced cloning as a method for the generation of large quantities of a desired DNA fragment, there is, under certain circumstances, still 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 by which this may be achieved are blunt-ended and 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 endonucleases 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 (Sections 6.2.1 and 6.2.2).

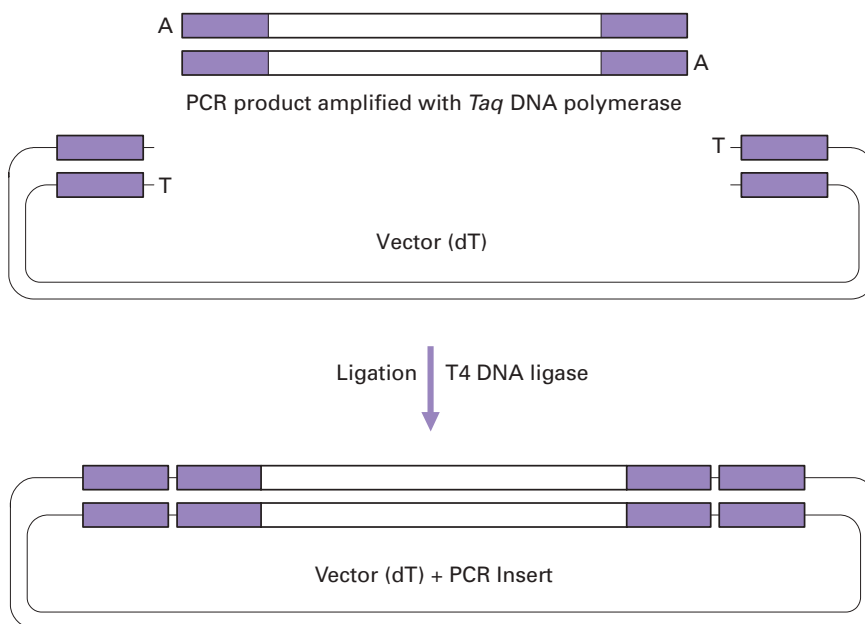


Fig. 6.9. Cloning of PCR products using dA:dT cloning.

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. Also any DNA cloned into a vector will permit 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 not only to store a copy of any particular fragment of DNA but also to produce unlimited amounts of it (Fig. 6.10).

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

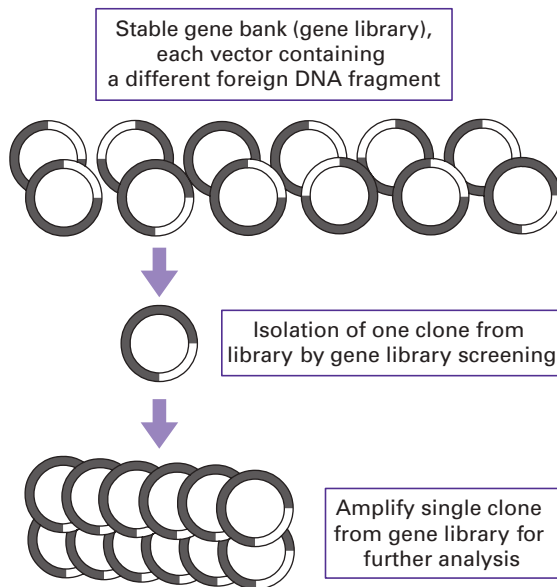


Fig. 6.10. Production of multiple copies of a single clone from a stable gene bank or library.

Table 6.2 Comparison of vectors generally available for cloning DNA fragments

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

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

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 F. Bolivar and R. 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**.

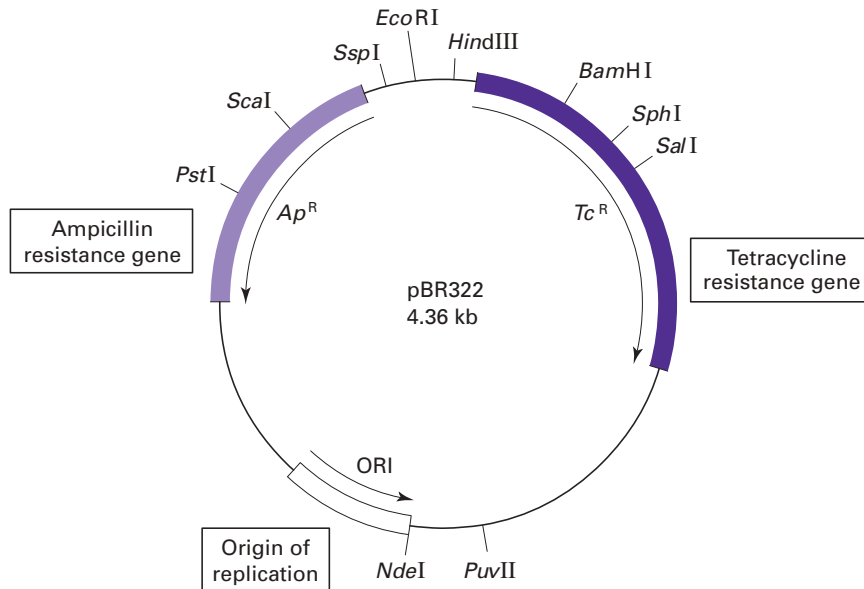


Fig. 6.11. Map and important features of pBR322.

- 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 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 that contain plasmid: if cells are plated on a 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 that can be used to open or linearise the circular plasmid. Linearising a plasmid allows a fragment of DNA to be inserted and the circle then closed again. The variety of sites not only makes it easier to find a restriction enzyme 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*HI would be isolated and purified. The plasmid pBR322 would also be digested at a single site, using *Bam*HI, and both samples would then be deproteinised to inactivate the restriction enzyme. *Bam*HI 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 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 an antibiotic, for example 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 contain simply 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 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, which 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

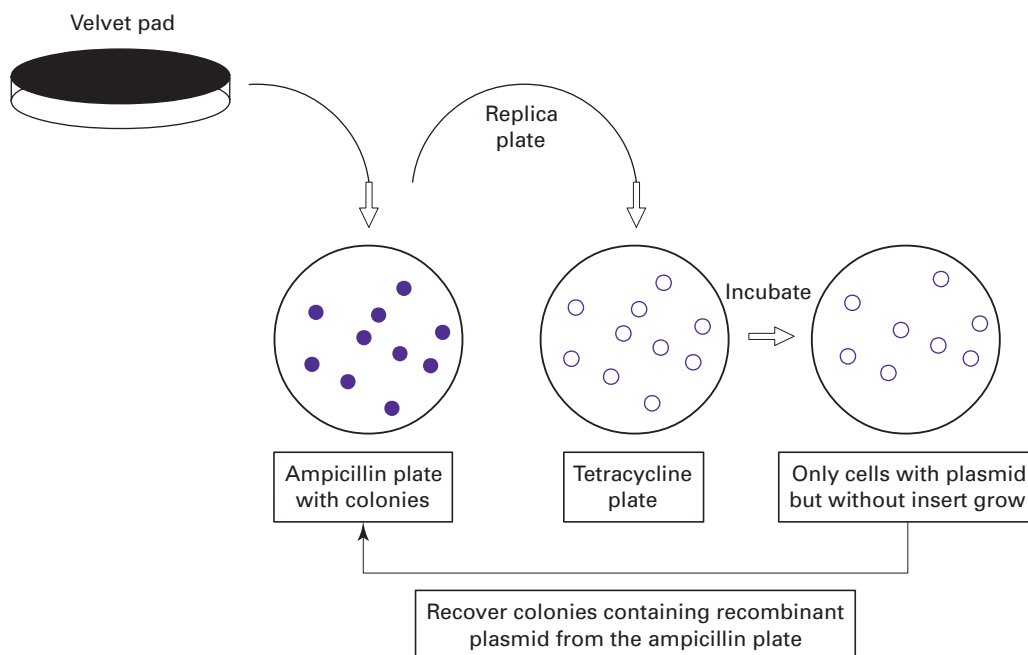


Fig. 6.12. Replica plating to detect recombinant plasmids. A sterile velvet pad is pressed onto the surface of an agar plate, picking up some cells from each colony growing on that plate. The pad is then pressed on to a fresh agar plate, thus inoculating it with cells in a pattern identical with that 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.

into a region termed the **multiple cloning site** or 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. 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 (Fig. 6.14). However, if the gene is disrupted by the insertion of a foreign fragment of DNA, a non-functional enzyme results that 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, merely 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.

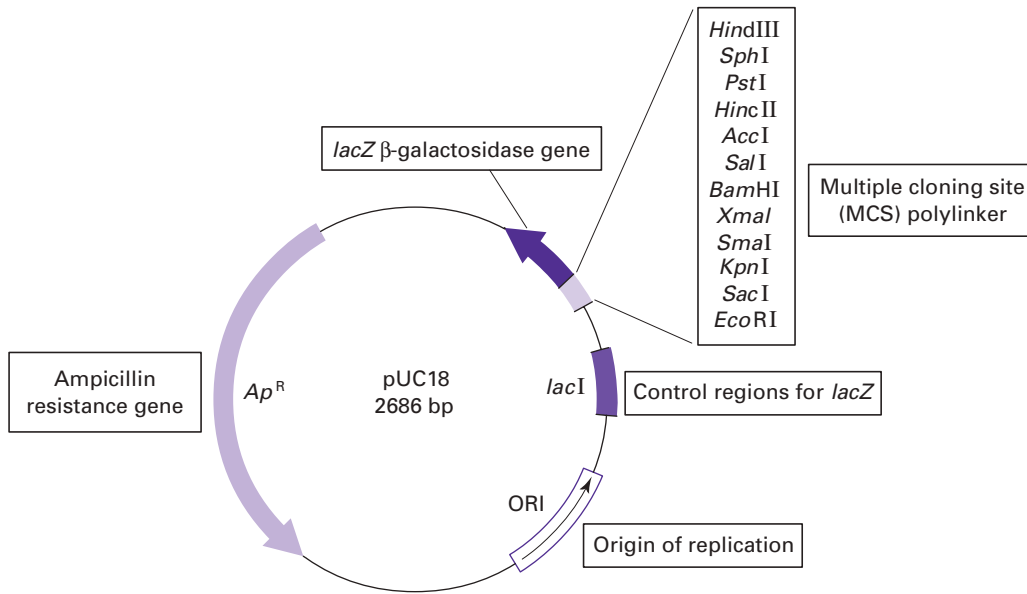


Fig. 6.13. Map and important features of pUC18.

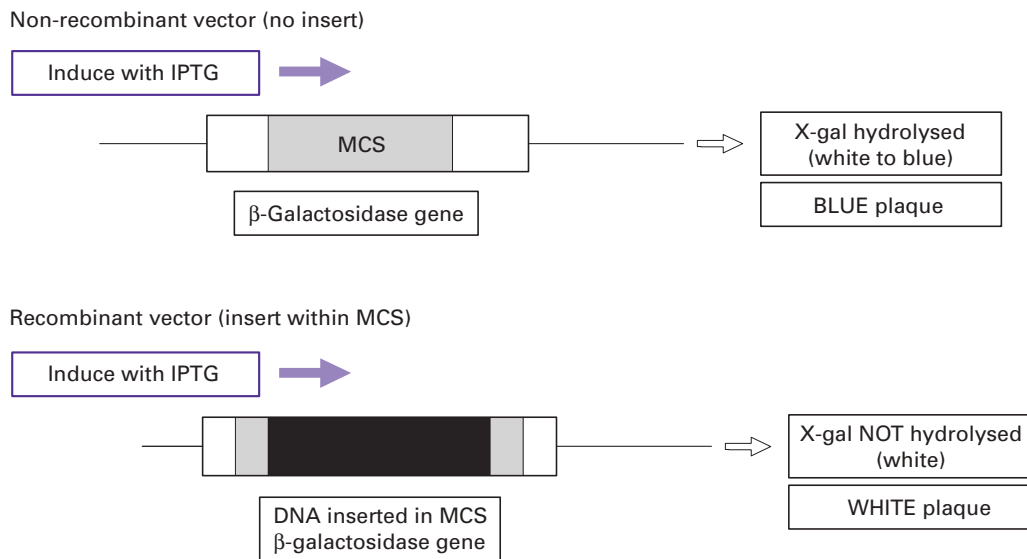


Fig. 6.14. Principle of blue/white selection for the detection of recombinant vectors.

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 (phages or bacterial viruses) have been adapted as vectors in order to propagate larger fragments of DNA in bacterial cells. Cloning vectors derived from phage λ are commonly used since they offer an approximately 16-fold advantage in cloning efficiency by 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. First, 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 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 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 is 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 phage λ 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 being λ gt10 and λ charon16A. 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, *EcoRI* and *BamHI*. 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 phage λ . The most notable examples of λ replacement vectors are λ EMBL and λ Zap.

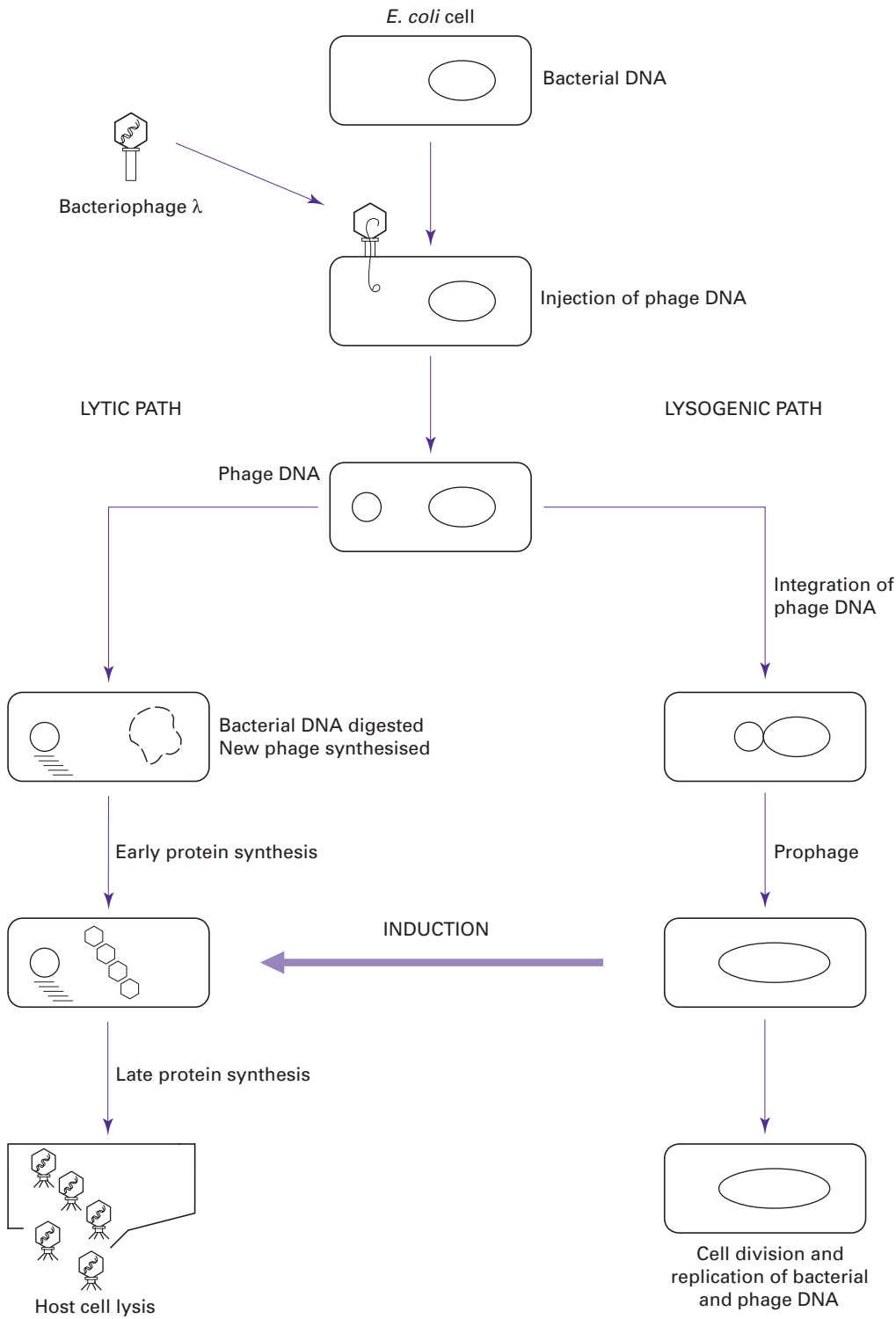


Fig. 6.15. The lysogenic and lytic cycles of bacteriophage λ .

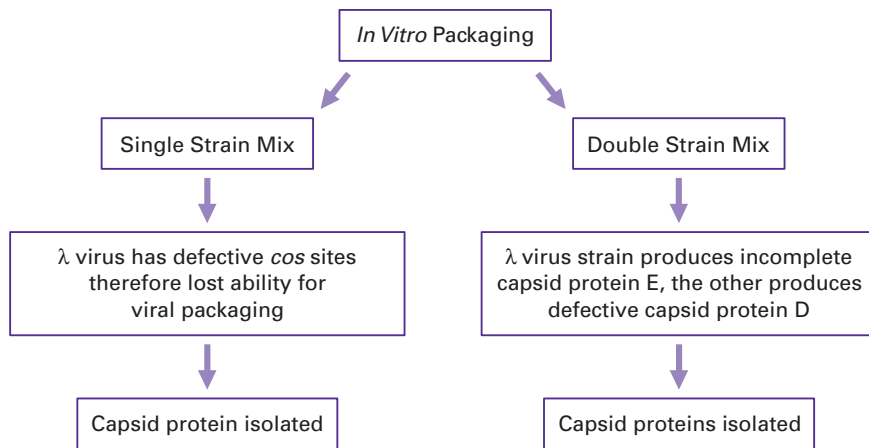


Fig. 6.16. Two strategies for producing *in vitro* packaging extracts for bacteriophage λ .

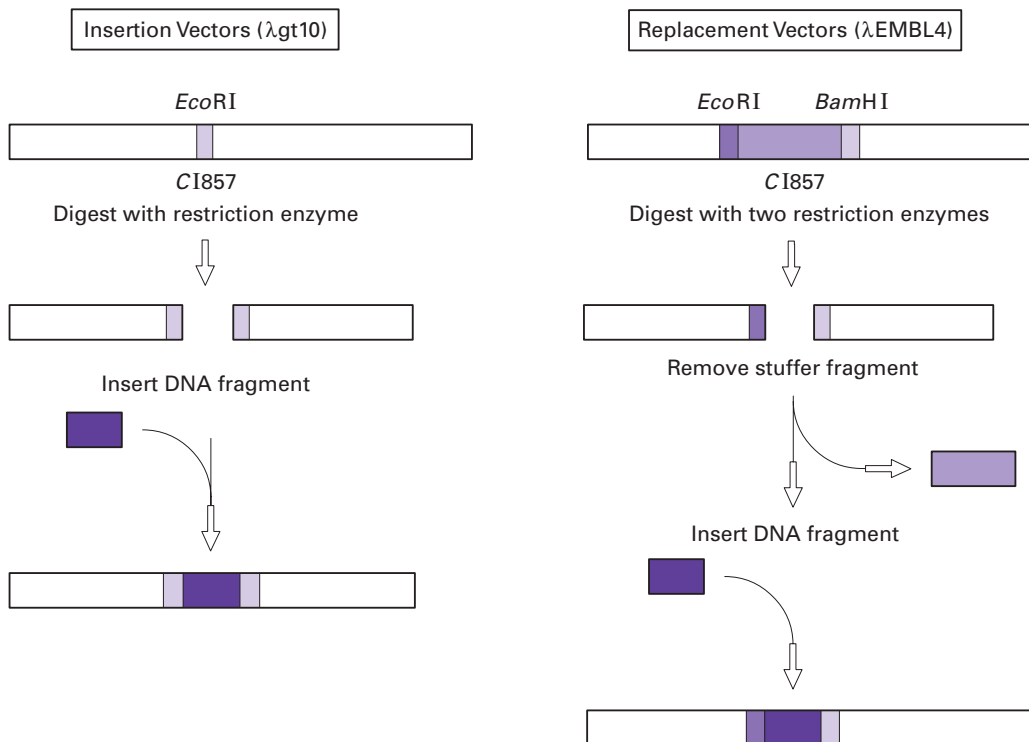


Fig. 6.17. General schemes used for cloning in λ insertion and λ replacement vectors. *CI857* is a temperature-sensitive mutation that promotes lysis at 42 °C after incubation at 37 °C.

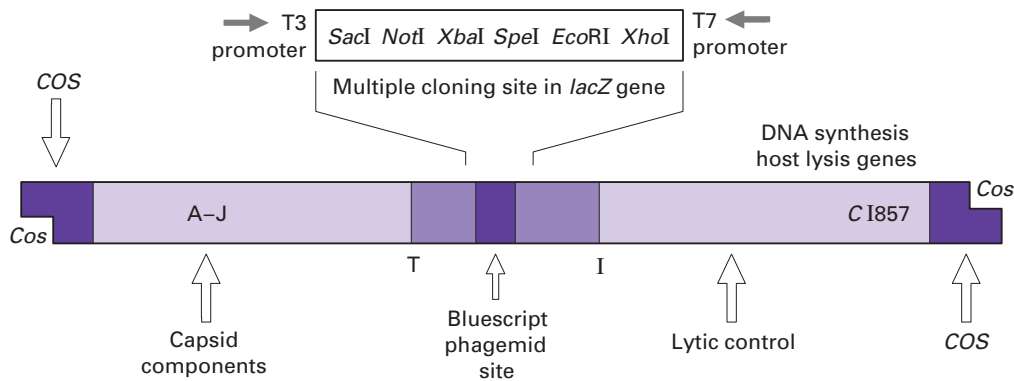


Fig. 6.18. General map of λ Zap cloning vector, indicating important areas of the vector. The multiple cloning site is based on the *lacZ* gene, providing blue/white selection based on the β -galactosidase gene. In between the initiator (I) site and terminator (T) site lie sequences encoding the phagemid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

λ Zap is a commercially produced cloning vector that includes unique cloning sites clustered into a MCS (Fig. 6.18). Furthermore the MCS is located within a *lacZ* region providing a blue/white selection system based on insertional inactivation (Fig. 6.14). 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 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). Another feature that makes this a useful cloning vector is the ability to produce RNA transcripts, termed **crRNA** 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 λ Zap 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).

The helper phage displace a strand within the λ Zap that 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 λ Zap 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 gene fragment that has been cloned into 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).

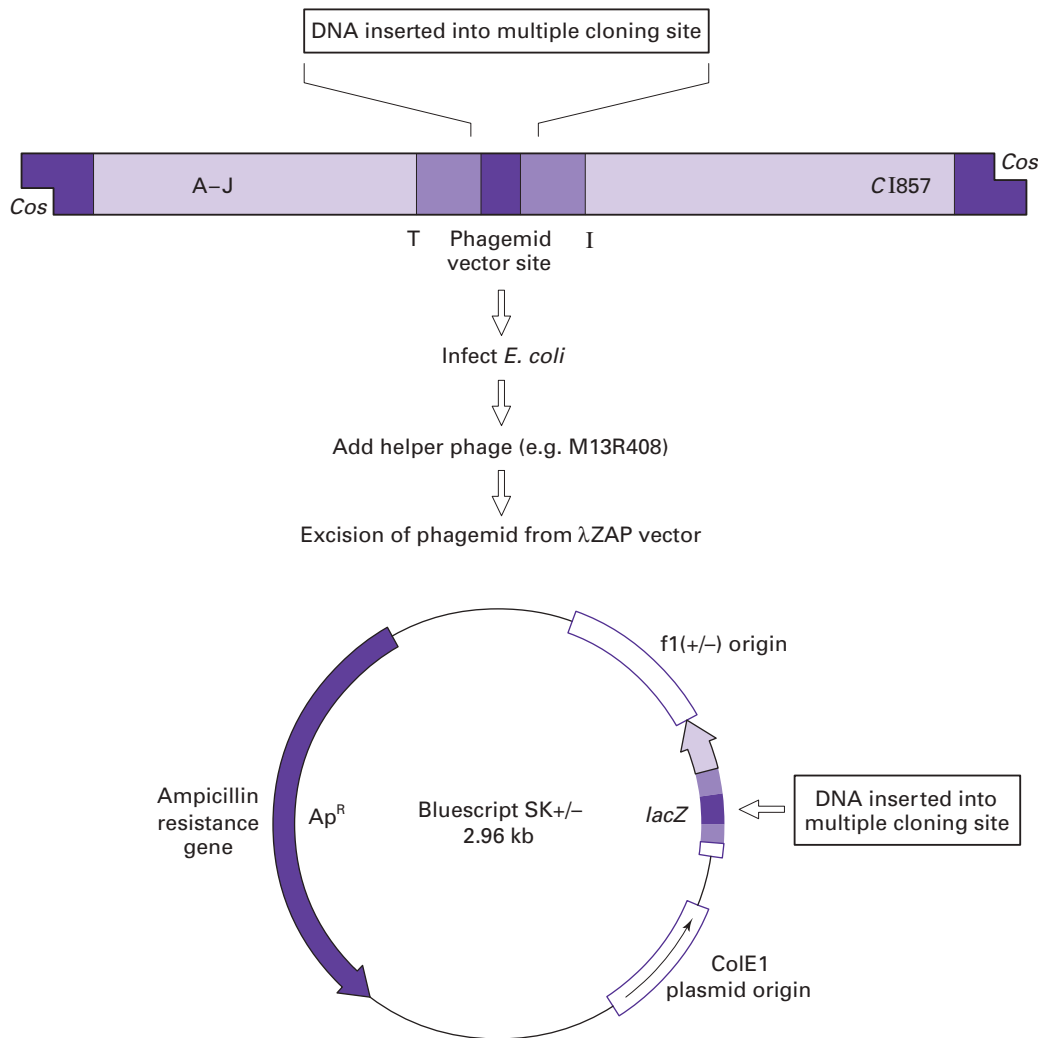


Fig. 6.19. Single-stranded DNA rescue of phagemid from λ Zap. The single-stranded phagemid pBluescript SK may be excised from λ Zap by addition of helper phage. This provides the necessary proteins and factors for transcription between the I and T sites in the parent phage to produce the phagemid with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

6.3.3 M13 and phagemid-based vectors

Much use has been made of single-stranded bacteriophage vectors such as M13 and vectors that 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 virus particles or virions for infection of further bacterial cells (lytic growth). The nature of these vectors makes them ideal for techniques such as chain termination sequencing

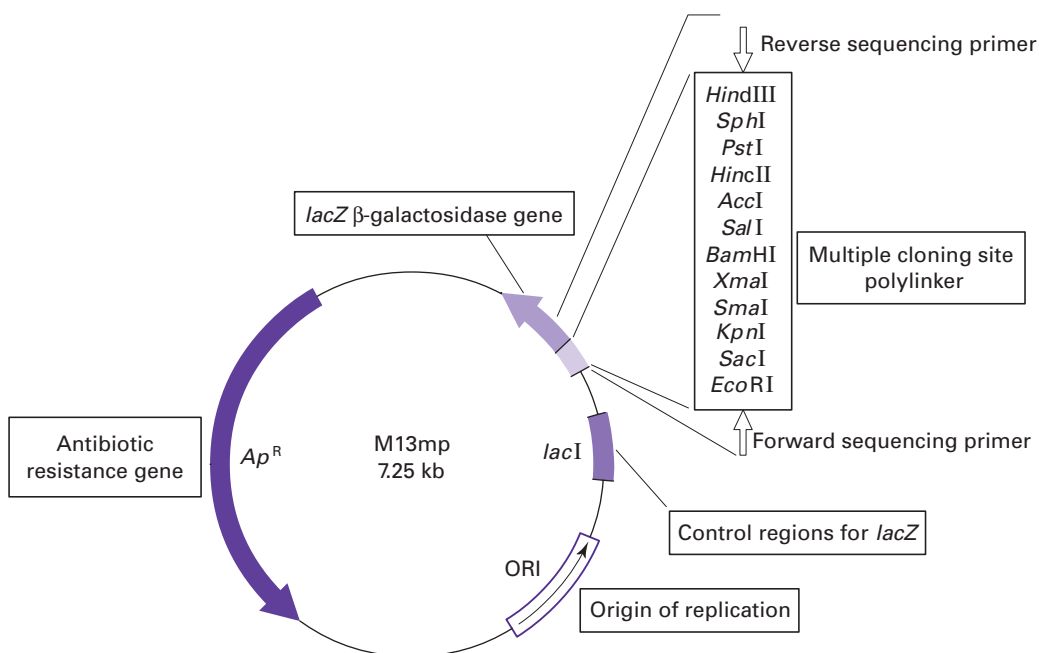


Fig. 6.20. Genetic map and important features of bacteriophage vector M13.

(Section 6.6.1) and *in vitro* mutagenesis (Section 6.6.2), 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 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 supercoiled 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

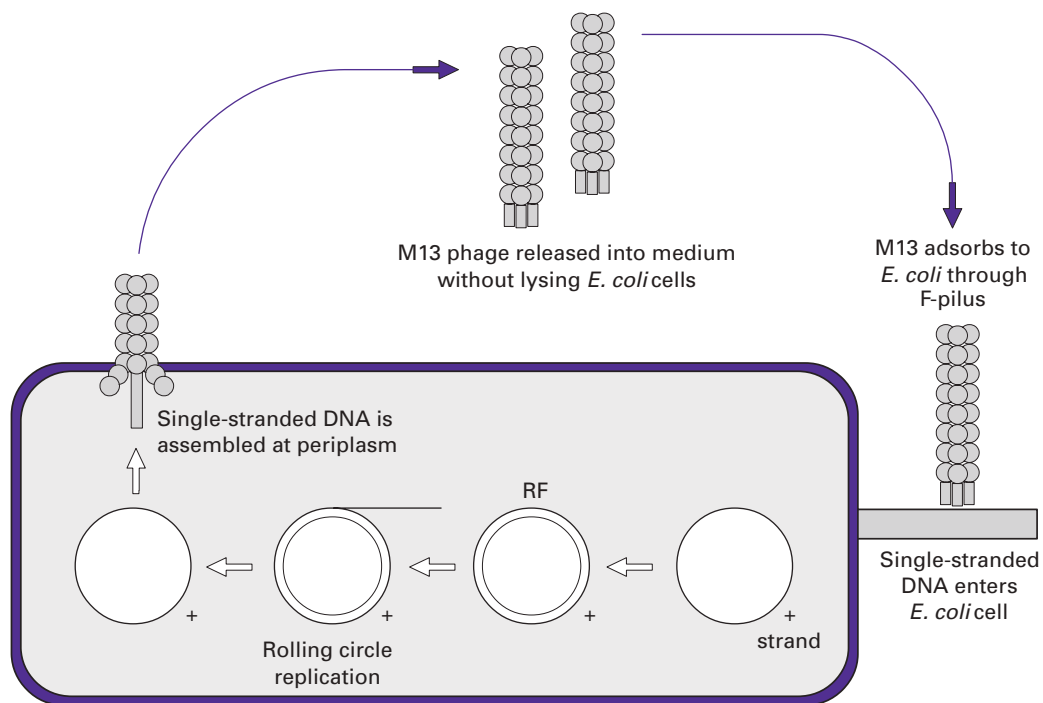


Fig. 6.21. Life cycle of bacteriophage M13. The bacteriophage virus enters the *E. coli* cell through the F-pilus. It then enters a stage where the circular single strands are converted to double strands. Rolling-circle replication then produces single strands, which are packaged and extruded through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

such as polyethylene glycol, and the DNA purified with phenol/chloroform (Section 5.7.1). Thus the phage 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, termed M13mp8/9, mp18/19 etc., are currently widely used 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 on the basis of the technique of blue/white selection (Section 6.3.1). As the series of vectors was developed, the number of restriction sites was increased in an asymmetric fashion. Thus M13mp8, mp12, mp18 and sister vectors that have the same MCS but in reverse orientation (M13mp9, mp13 and mp19, respectively) and have more restriction sites in the MCS, making the vector more useful as a 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. After superinfection with helper phage,

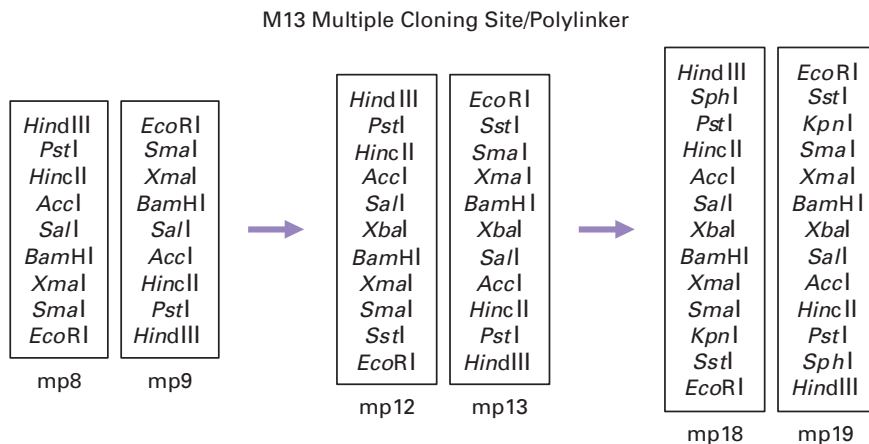


Fig. 6.22. Design and orientation of polylinkers in M13 series. Only the main restriction enzymes are indicated.

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 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 complementary RNA 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 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

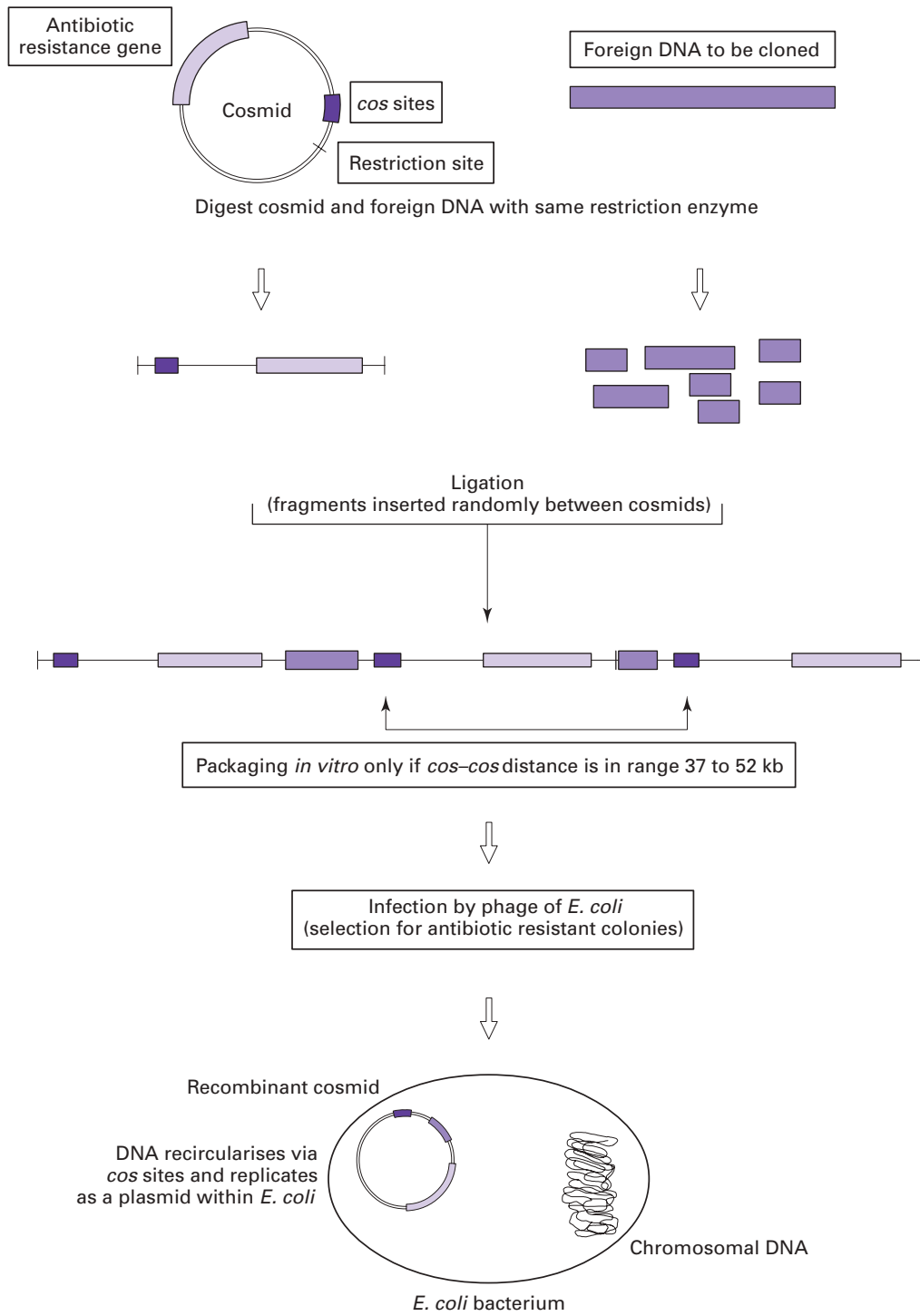


Fig. 6.23. Scheme for cloning foreign DNA fragments in cosmid vectors.

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 recircularizes through its *cos* sites and from then on 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 one is searching for the foreign DNA of interest. They have also had an enormous impact on 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 bacterial artificial chromosomes (BACs), mammalian artificial chromosomes (MACs) and on the virus P1 artificial chromosomes (PACs). However, perhaps the most significant development is vectors based on yeast artificial chromosomes.

6.3.6 Yeast artificial chromosome vectors

Yeast artificial chromosomes (YACs) are linear molecules composed of a centromere, telomere and a replication origin termed an ARS (autonomous replicating sequence) element. 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 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 repeated sequences that are sometimes difficult to clone in bacterial-based vectors may also be 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).

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 (YE_p) have been

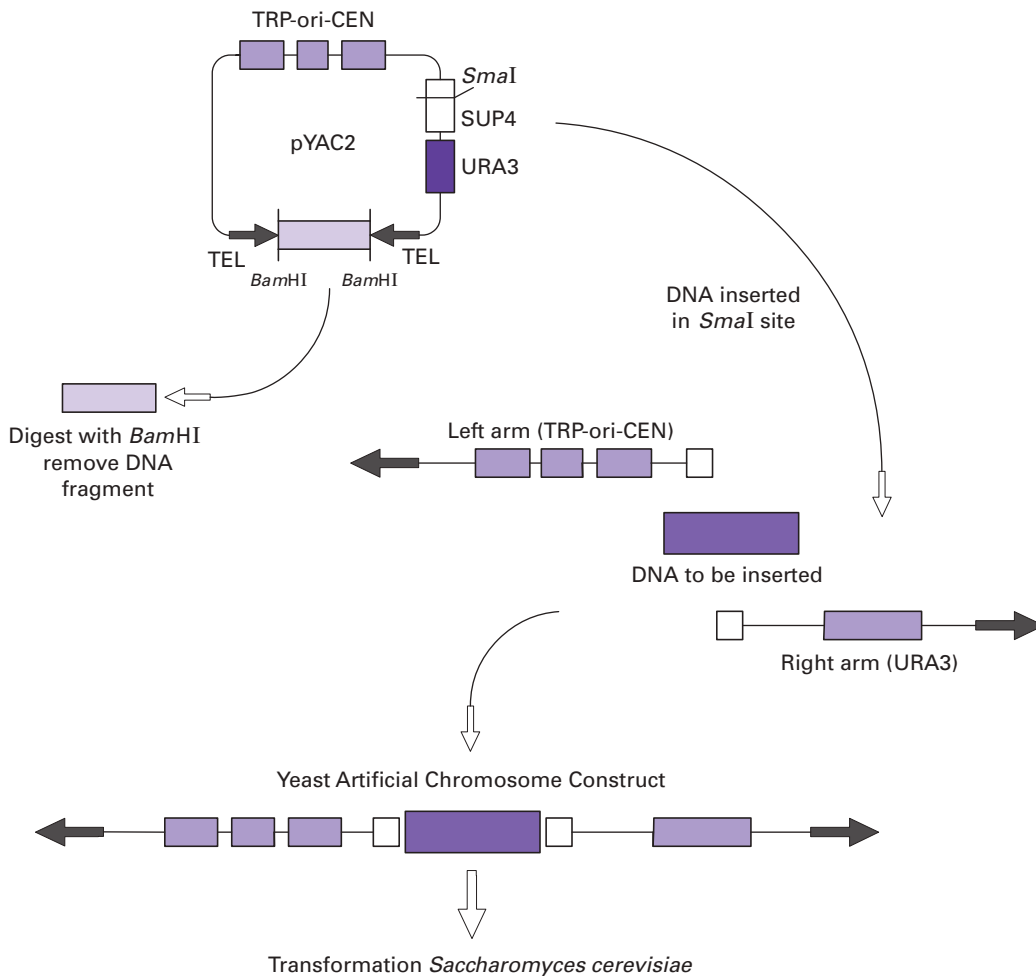


Fig. 6.24. Scheme for cloning large fragments of DNA into YAC vectors.

created by genetic manipulation using replication origins from the 2μ circle, and by incorporating a gene that will complement a defective gene in the host yeast cell. If, for example, a strain of yeast is used that 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 is therefore well suited for use in cloning. Of particular use has been the creation of **shuttle vectors** that have origins of replication for yeast and bacteria such as *E. coli*. 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 **Ti**, part of which is transferred into the nuclei of plant

cells 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 a 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 construct is 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 phage to introduce DNA into *E. coli* has also been well exploited and results in 10- to 100-fold higher efficiency for the introduction of recombinant DNA as compared with 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 *E. coli*.

It is possible to deliver recombinant molecules into animal cells by **transfection**, the efficiency of which can be increased by first precipitating the DNA with Ca^{2+} or making the membrane permeable with divalent cations or high molecular weight polymers such as DEAE-dextran or polyethylene glycol (PEG). 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 that fuse with the lipid membrane of cells and 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

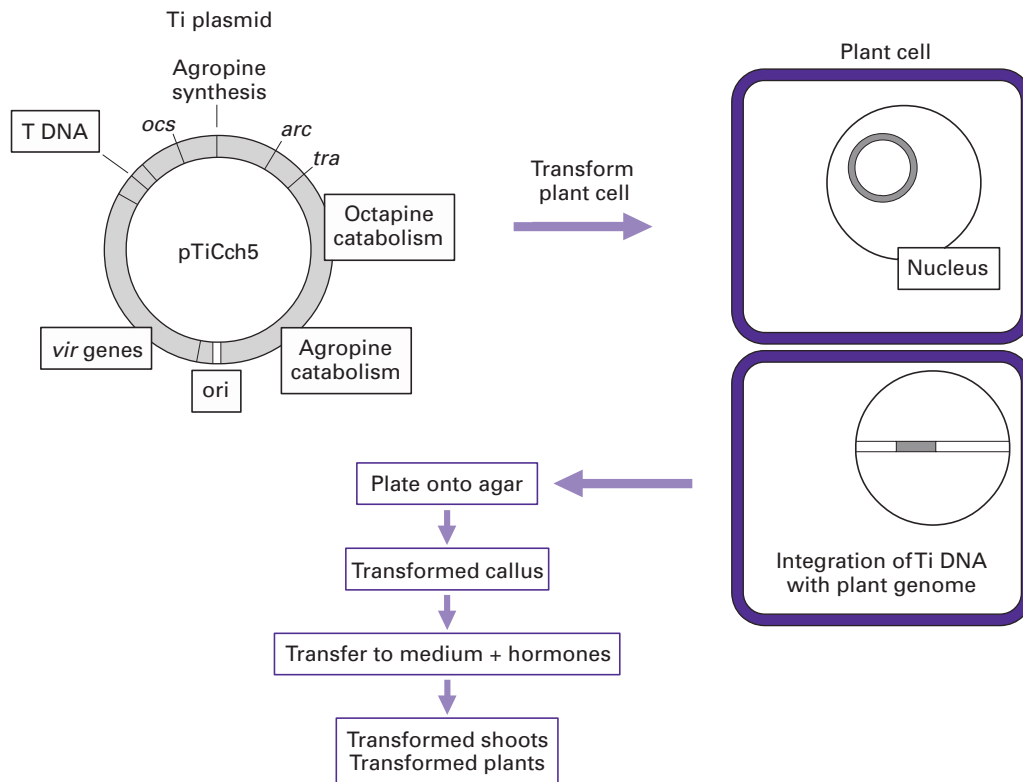


Fig. 6.25. Scheme for cloning in plant cells using the Ti plasmid.

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 nucleic acid database entries and availability of custom synthesis of oligonucleotides has provided a relatively straightforward means for designing and producing **gene probes** and primers for PCR. Gene probes and primers are usually designed using bioinformatics software and nucleic acid databases or gene family-related sequences as indicated in Section 5.9.3. 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

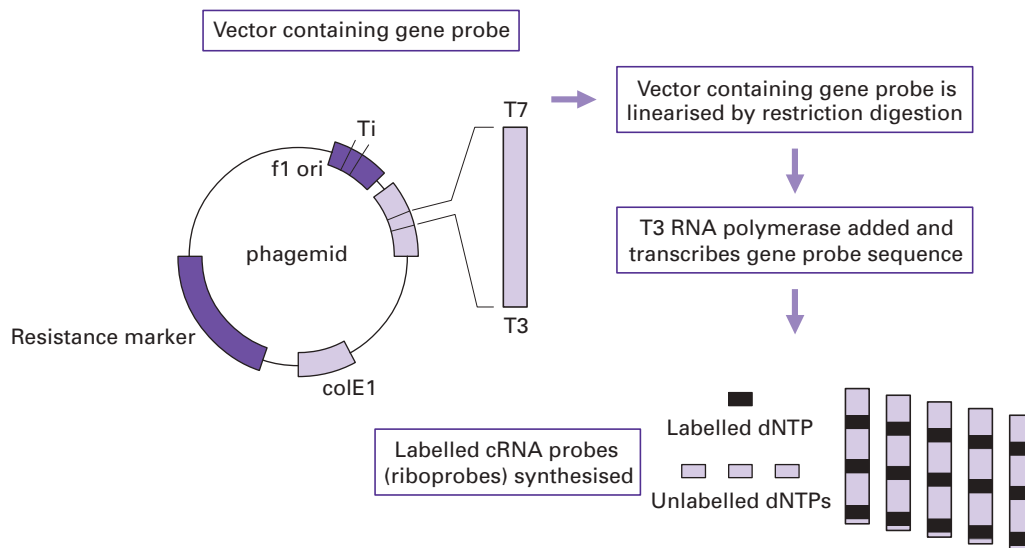


Fig. 6.26. Production of cDNA (riboprobes) using T3 RNA polymerase and phagemid vectors.

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 that 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 the cloned double-stranded probes mentioned 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 is 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 based mainly 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 a nylon membrane 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 5.9.2). 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 onto 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 that 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: PCR 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 inserts 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 genome libraries.

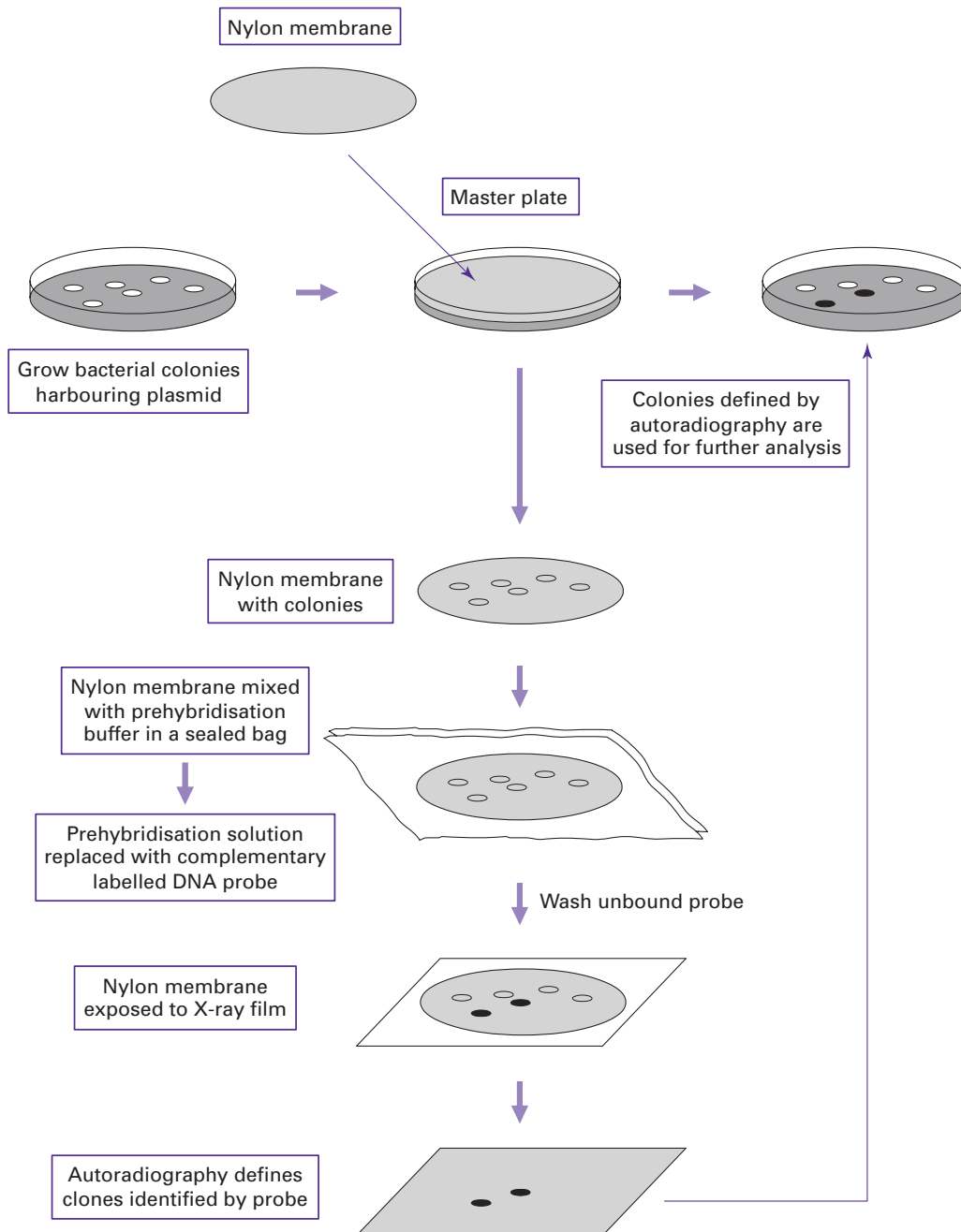


Fig. 6.27. Colony hybridisation technique for locating specific bacterial colonies harbouring recombinant plasmid vectors containing desired DNA fragments. This is achieved by hybridisation to a complementary labelled DNA probe and autoradiography.

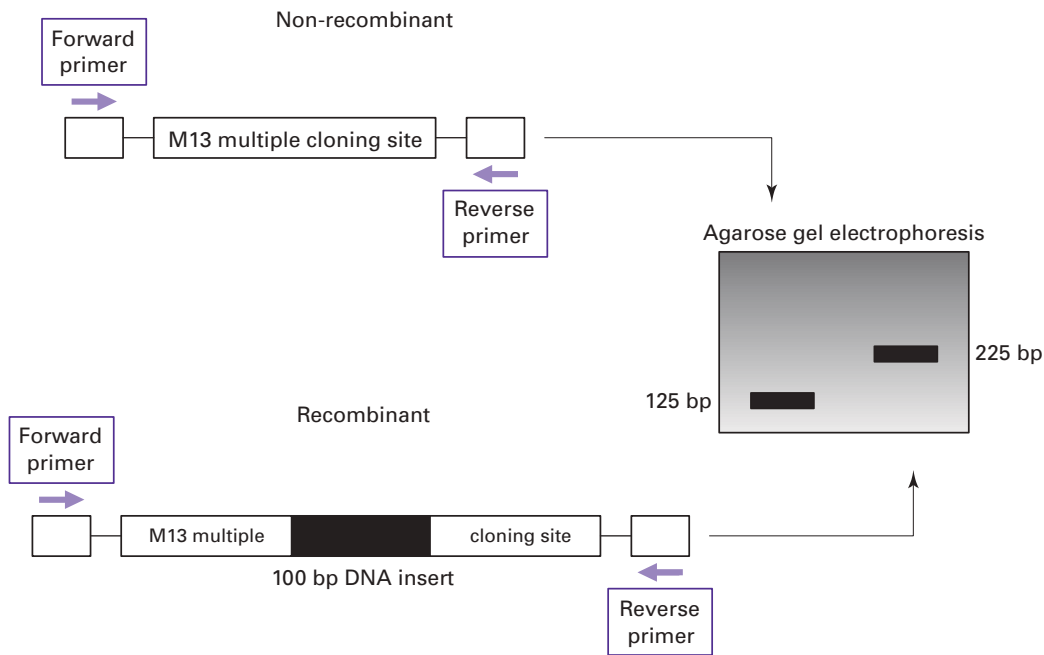


Fig. 6.28. PCR screening of recombinant vectors. In this figure, the M13 non-recombinant has no insert and so the PCR undertaken with forward and reverse sequencing primers gives rise to a product 125 bp in length. The M13 recombinant with an insert of 100 bp will give rise to a PCR product of 125 bp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

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 onto a nylon membrane (Fig. 6.29). The membranes are soaked in total cellular mRNA, under stringency conditions, i.e. usually at a temperature only a few degrees below the T_m at 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 onto 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 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.

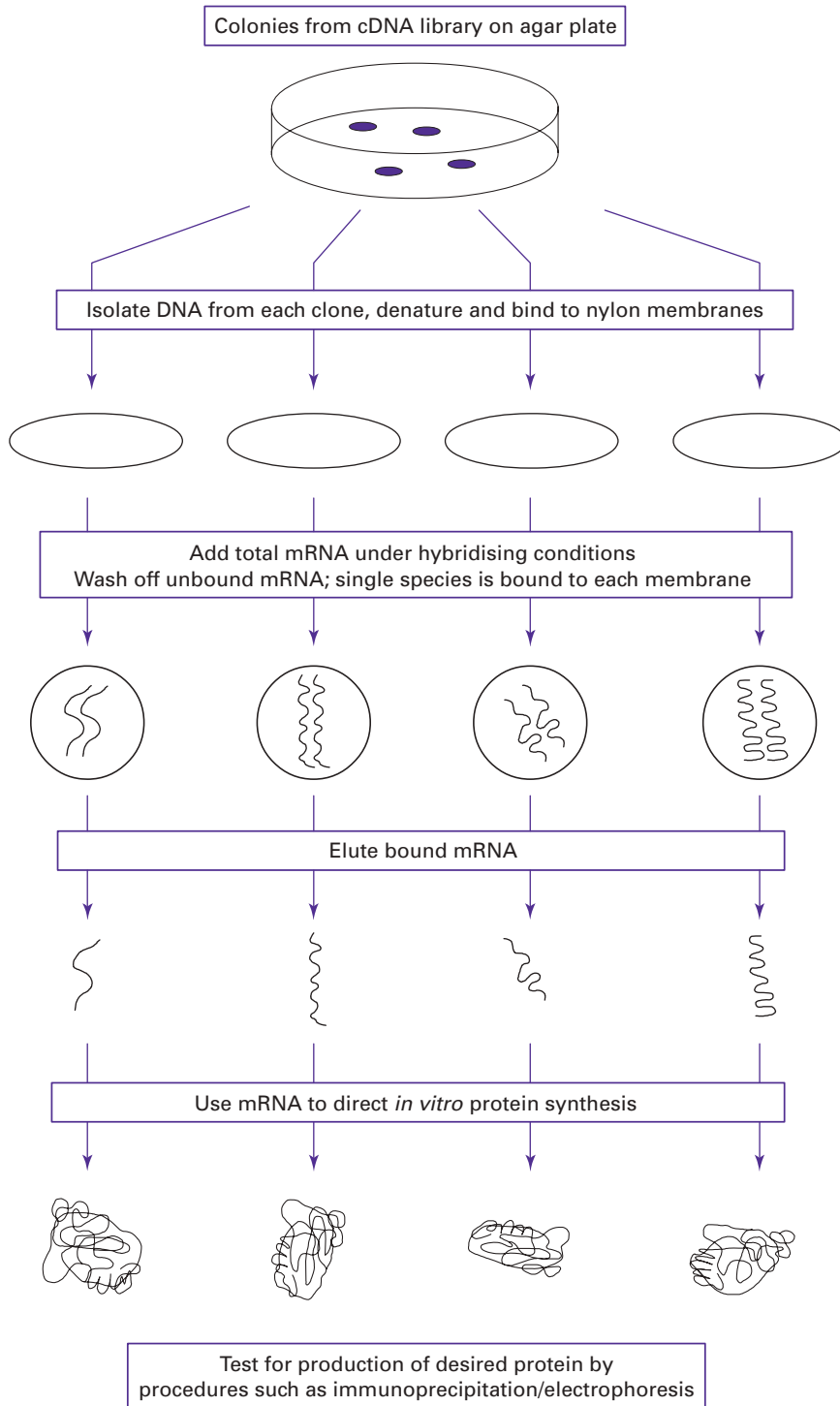


Fig. 6.29. General principles involved in the technique of a hybrid select translation.

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 λ gt11 and λ Zap or plasmids such as pEX. The precise requirements for such vectors are identical with vectors that are dedicated to producing proteins *in vitro* and are described in Section 6.7.1. In some cases, expression vectors incorporate inducible promoters that may be activated by, for example, increasing the temperature, thus 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 that 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

DNA fragments cloned into plasmid vectors may be subjected to the Sanger **chain termination sequencing method** detailed in Section 5.11.1. However, since plasmids are double stranded, 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 are such that the strands are single stranded for a long enough period to allow the sequencing method to succeed. It is also possible to include denaturants such as formamide to the reaction to further prevent reannealing. In general, however, the superior results gained with sequencing single-stranded DNA from M13 or single-stranded phagemids means that cloned DNA of interest is usually subcloned into such vectors.

M13 vectors are the traditional choice for chain termination sequencing because of the single-stranded nature of their DNA. A further modification 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

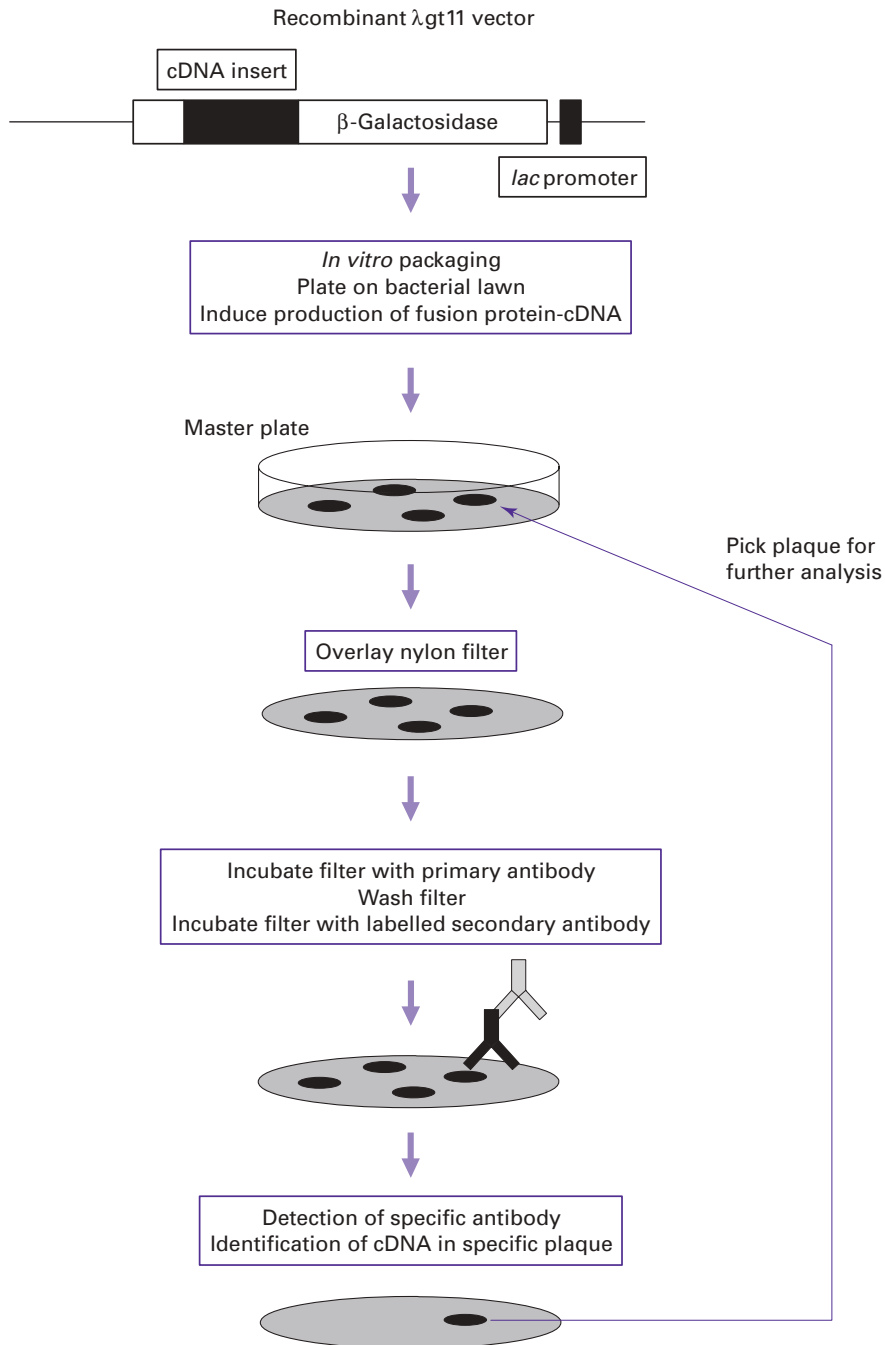


Fig. 6.30. Screening of cDNA libraries in expression vector λ gt11. The cDNA inserted upstream from the gene for β -galactosidase will give rise to a fusion protein under induction (e.g. with IPTG). The plaques are then blotted onto a nylon membrane filter and probed with an antibody specific for the 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.

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 and rational design

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 undertakes a logical sequence of analytical and computational techniques centred around a design cycle. This involves 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 as it is 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.

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, after 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 DNA strand which is complementary to the existing one but incorporates the oligonucleotide with the base mutation. The subsequent cloning of the recombinant produces multiple copies, half of which contain a sequence with the mutation and the other 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 that simplify the process of detecting the strands with the mutations.

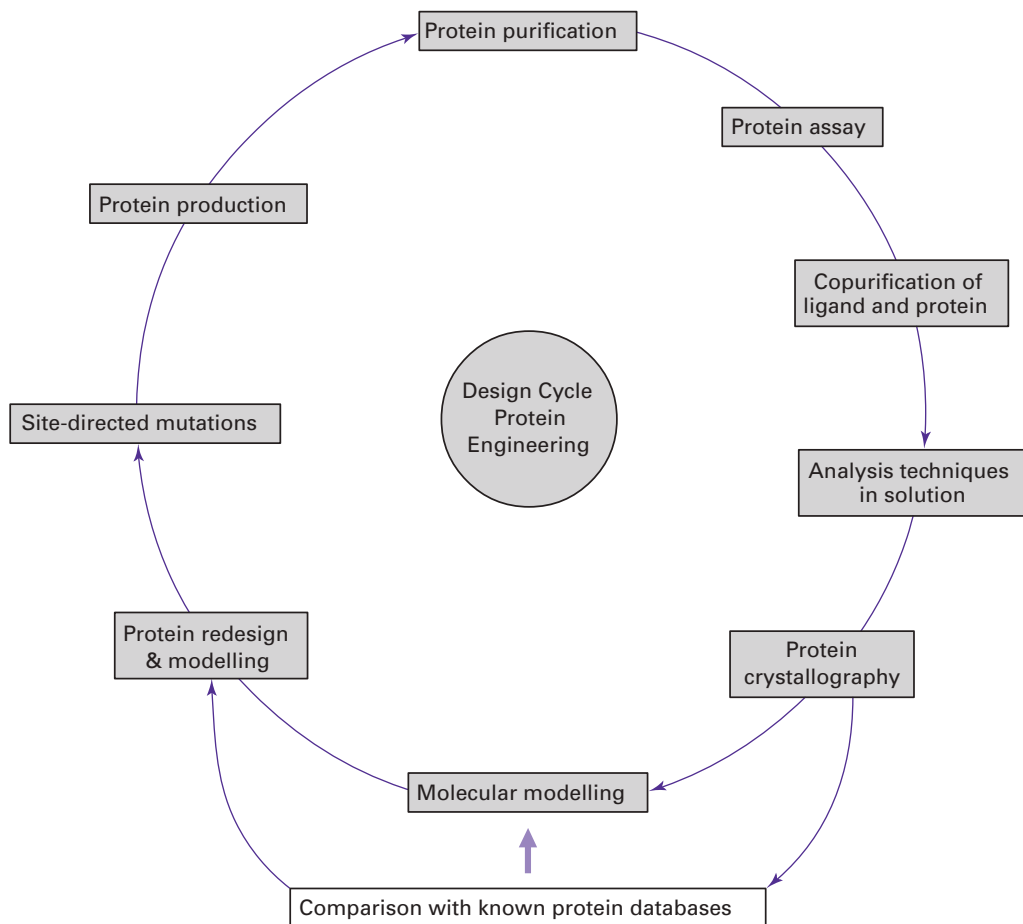


Fig. 6.31. Protein design cycle used in the rational redesign of proteins and enzymes.

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 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

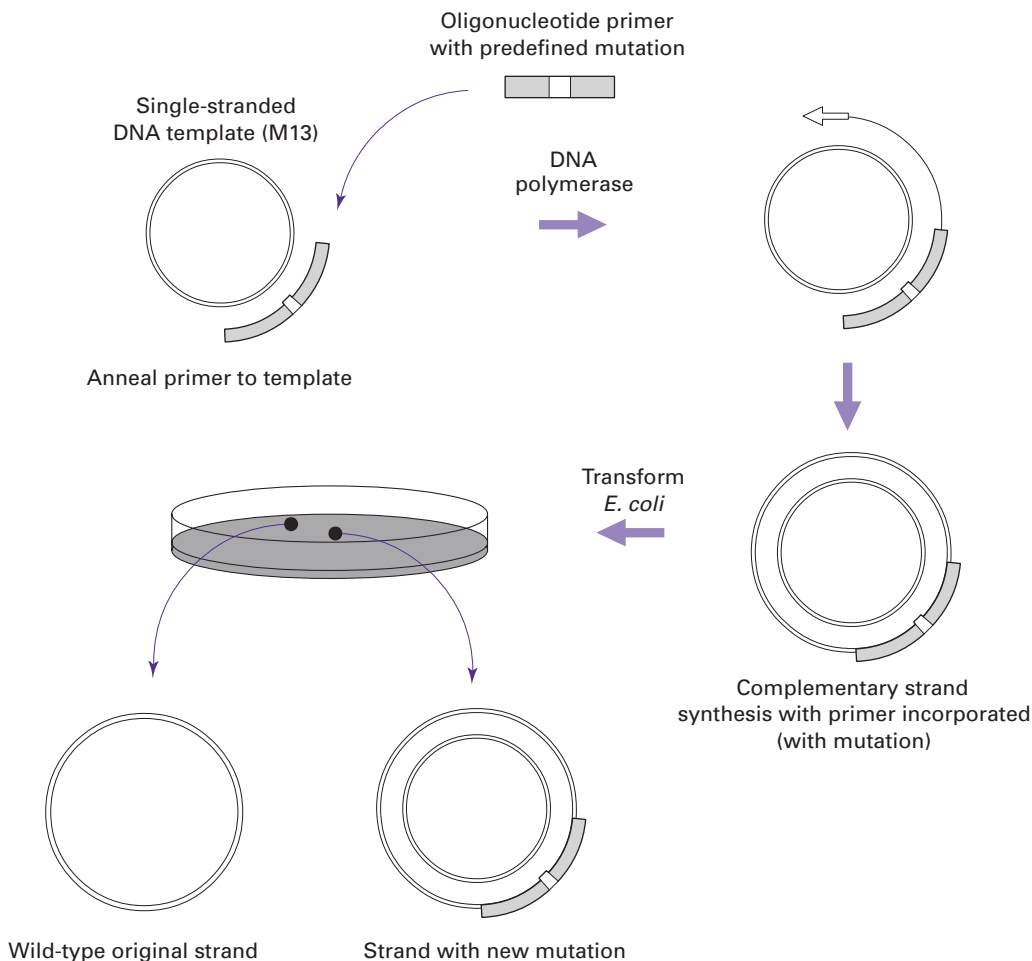


Fig. 6.32. Oligonucleotide-directed mutagenesis. This technique requires a knowledge of nucleotide sequence, since an oligonucleotide may then be synthesised with the base mutation. Annealing of the oligonucleotide to complementary (except for the mutation) single-stranded DNA provides a primer for DNA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

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 that it encodes. This accelerated form of artificial molecular evolution may be undertaken using **error-prone PCR**, where deliberate and random mutations are introduced by a low fidelity PCR amplification reaction. The resulting amplified gene is

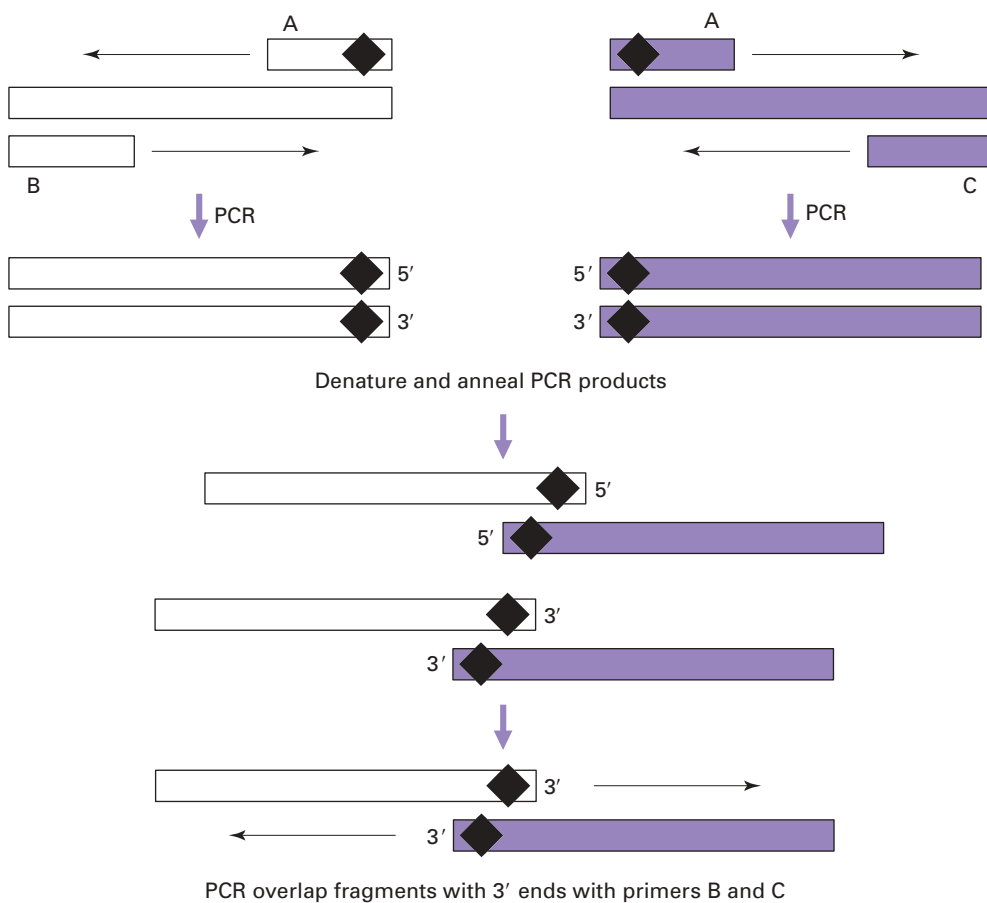


Fig. 6.33. Construction of a synthetic DNA fragment with a predefined mutation using overlap PCR mutagenesis.

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 (Section 6.7.2) and in the development of antibodies with enzymatic activities (catalytic antibodies) (Section 15.3.5).

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

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

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

in vitro expression 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 to 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 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 from 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 bacteria (Section 5.5.5). 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 that contain promoter and ribosome binding sites positioned just before one or more restriction sites for the insertion of foreign DNA. These

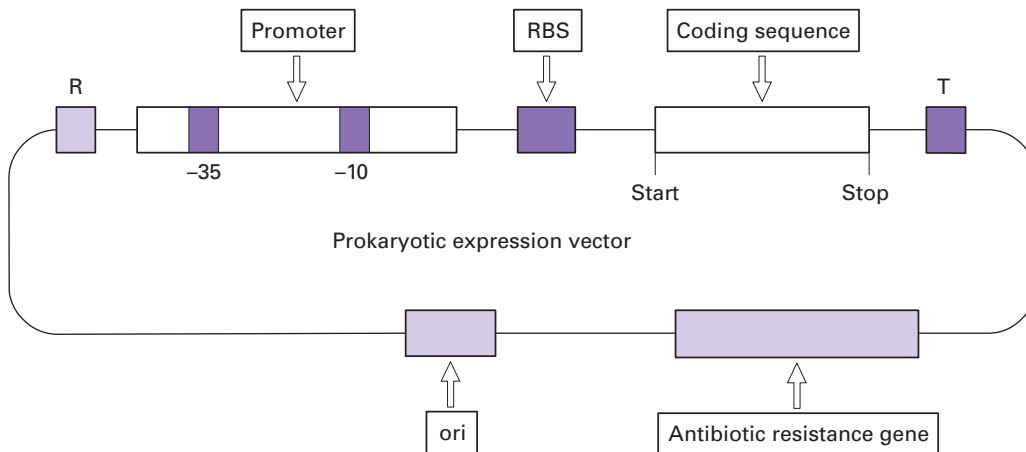


Fig. 6.34. Components of a typical prokaryotic expression vector. To produce a transcript (coding sequence) and translate it, a number of sequences in the vector are required. These include the promoter and ribosome-binding site (RBS). The activity of the promoter may be modulated by a regulatory gene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

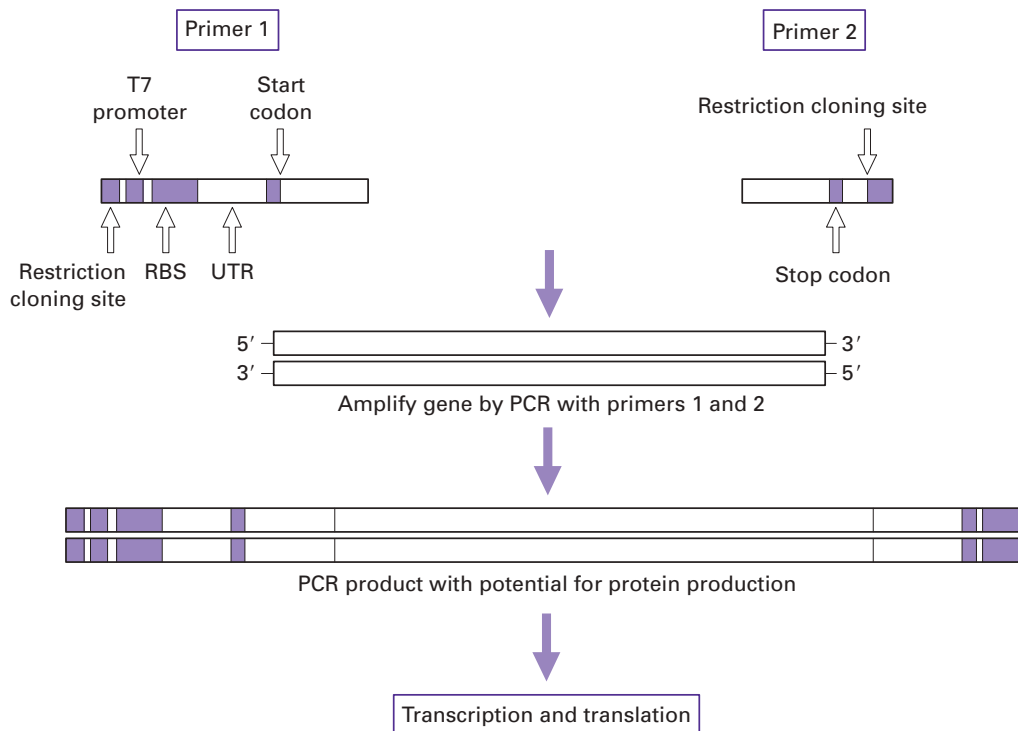


Fig. 6.35. Expression PCR (E-PCR). This technique amplifies a target sequence with one promoter that contains a transcriptional promoter, ribosome binding site (RBS), untranslated leader region (UTR) and start codon. The other primer contains a stop codon. The amplified PCR products may be used in transcription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

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 that 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 \times 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 \times His-Tag and so recovery is efficient even under the 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 cell 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 that 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 whilst 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, that 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 of expression of foreign proteins. The disadvantages of this system are the ultimate lysis of the COS cells and the limited insert capacity of the vector. Much interest is also currently focused on other modified viruses: vaccinia virus and baculovirus.

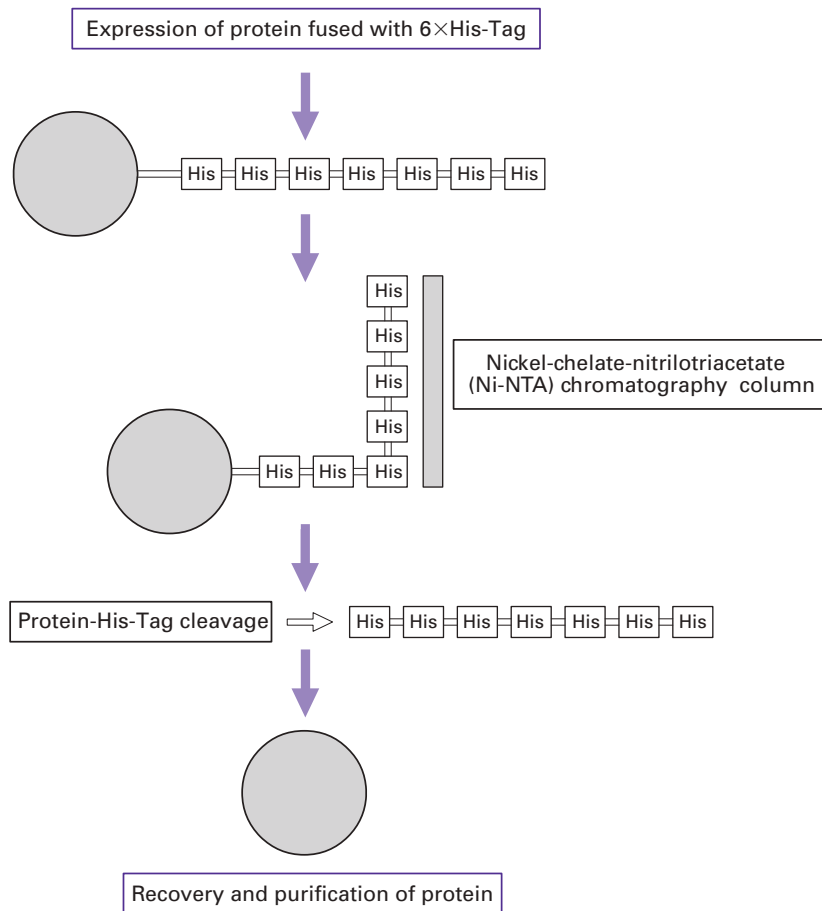


Fig. 6.36. Recovery of proteins using (6xHis-Tag) and (Ni-NTA) chromatography columns.

These have been developed for high level expression in mammalian cells and insect cells, respectively. The vaccinia virus 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 transmembrane regulator (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 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

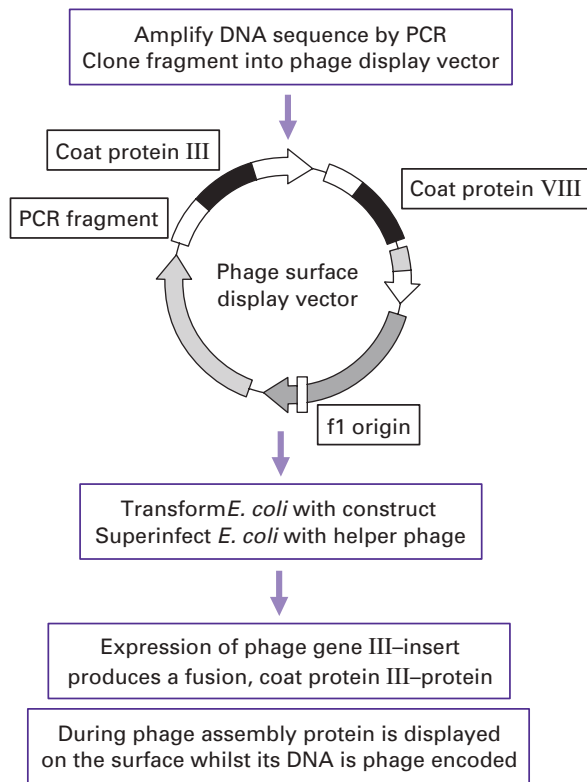


Fig. 6.37. Flow diagram indicating the main steps in the phage display technique.

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 that 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 gIII. 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 phage, viruses, bacteria and other organisms, and commercial organisations have been quick to exploit this technology. One major application is the analysis and production of the engineered antibodies from which the technology was mainly developed. Screening of novel recombinants may be carried out by techniques 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 provide 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 for 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 library may provide a route to high-affinity recombinant antibody fragments that are difficult to produce by more conventional hybridoma fusion techniques (Section 7.2.4). 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.

6.8 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 quantified by non-PCR based methods. This is not only important for gene regulation studies but may also be used as a marker for certain clinical disorders. Traditionally the northern blot has been used for detection of particular RNA transcripts by blotting extracted mRNA and immobilising it onto 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, recently 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 that 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.

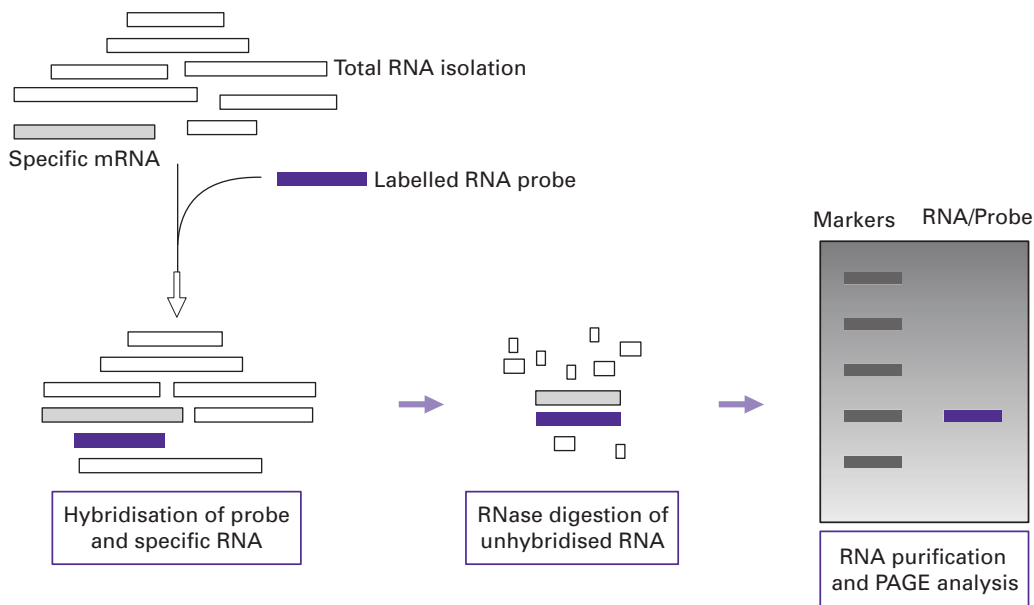


Fig. 6.38. Steps involved in the ribonuclease protection assay (RPA).

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 that 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 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 to 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 that 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 gelelectrophoresis 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.

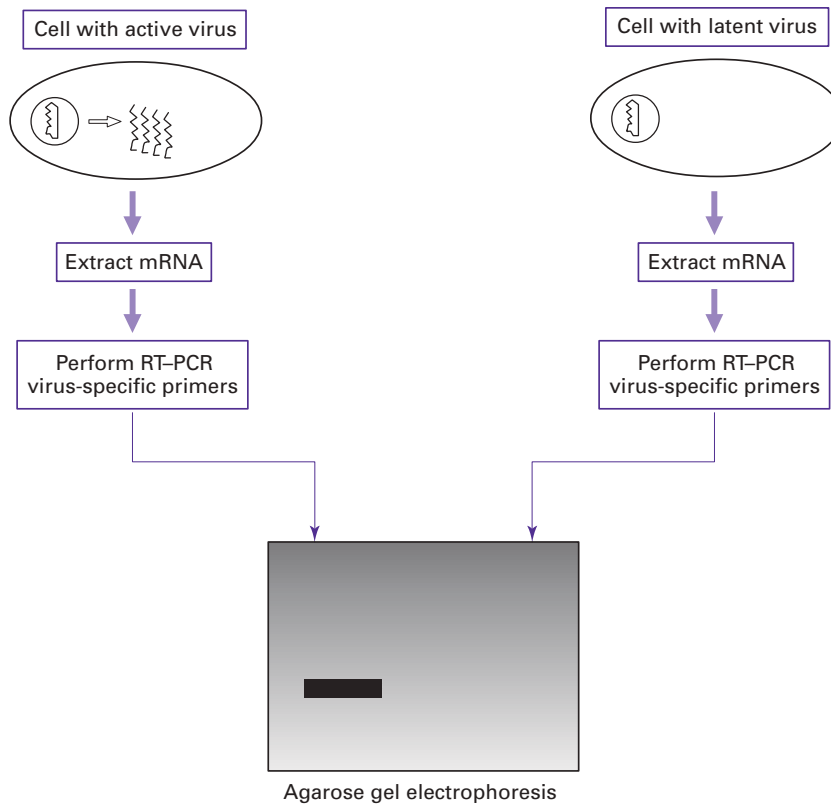


Fig. 6.39. Representation of the detection of active viruses using RT-PCR.

6.8.2 Analysing genes *in situ*

Gross chromosomal changes are often detectable by microscopic examination of the chromosomes within a karyotype (Section 5.3.3). 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. 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. This allows the localisation and identification of, for example, single copies of intracellular viruses.

An alternative labelling strategy used in karyotyping and gene localisation is *fluorescence in situ hybridisation* (FISH). This method, sometimes termed

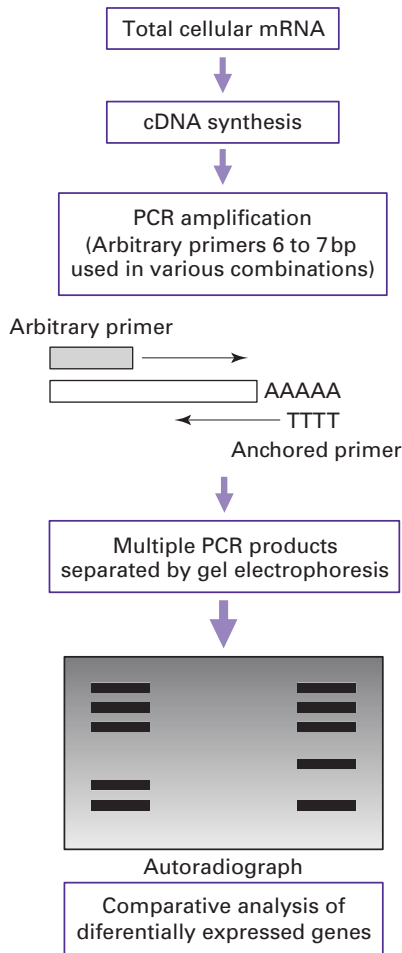


Fig. 6.40. Analysis of gene expression using differential display PCR.

chromosome painting, is based on *in situ* hybridisation but 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 that lies upstream from 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)

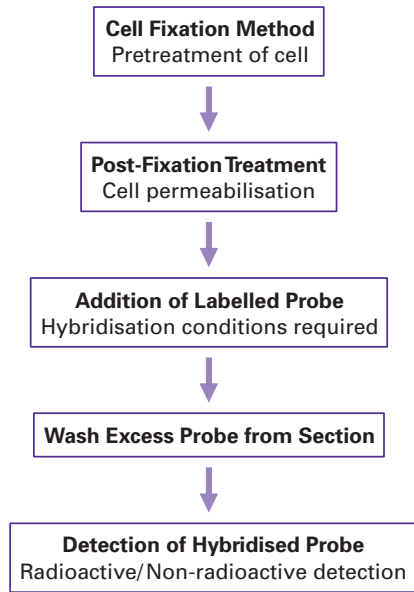


Fig. 6.41. General scheme for *in situ* hybridisation.

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 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 quantification 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 [gel retardation](#). Under gel electrophoresis, the migration of a DNA fragment bound to a protein of a relatively large mass will be retarded by comparison with 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, on the restriction map, the position 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

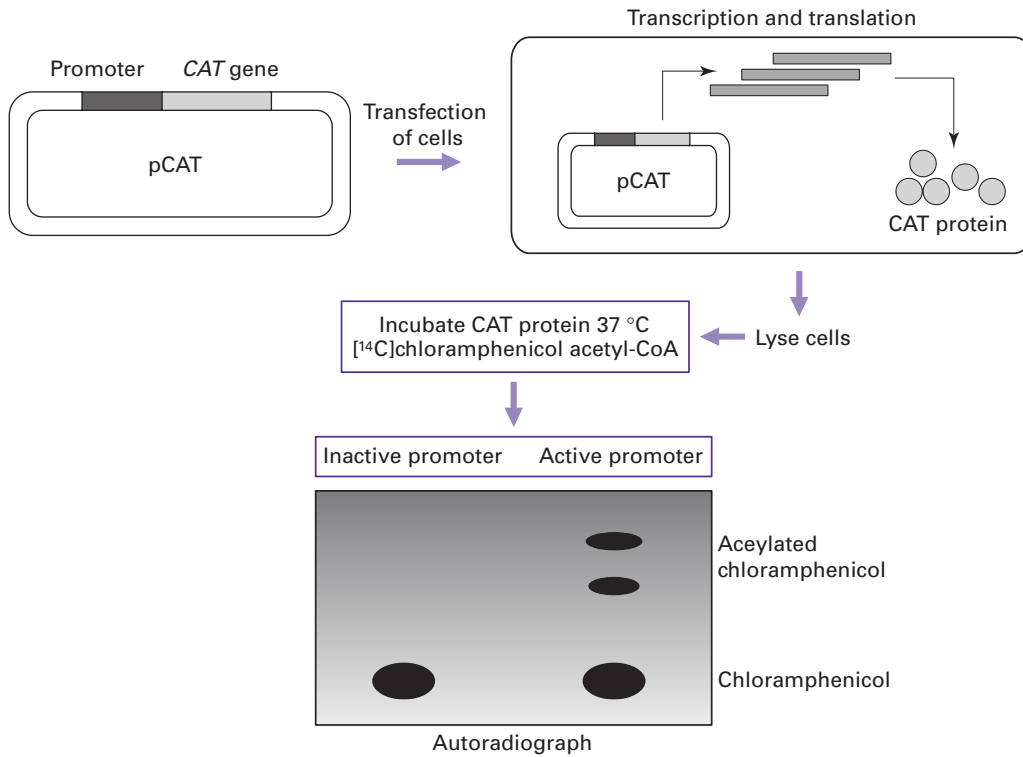


Fig. 6.42. Assay for promoters using the reporter gene for chloramphenicol acetyl transferase (CAT).

labelled at one end with a radioactive label and then mixed with the DNA-binding protein (Fig. 6.43). DNase I is added and partial digestion is 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 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 with or the contribution from 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

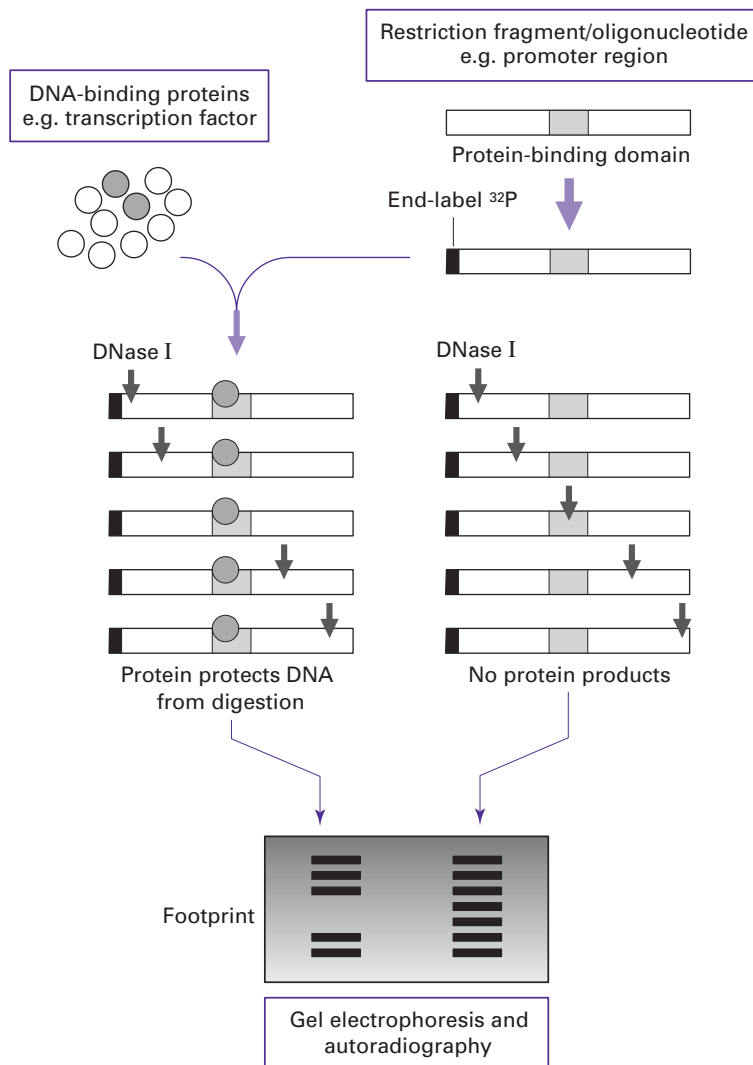


Fig. 6.43. Steps involved in DNA footprinting.

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. But when the two fractions are mixed together transcription is initiated if the domains are complementary (Fig. 6.44). This is indicated by the transcription of a reporter gene such as the *CAT* gene. The technique is not confined only to transcription factors and may be applied to any protein system where interaction occurs.

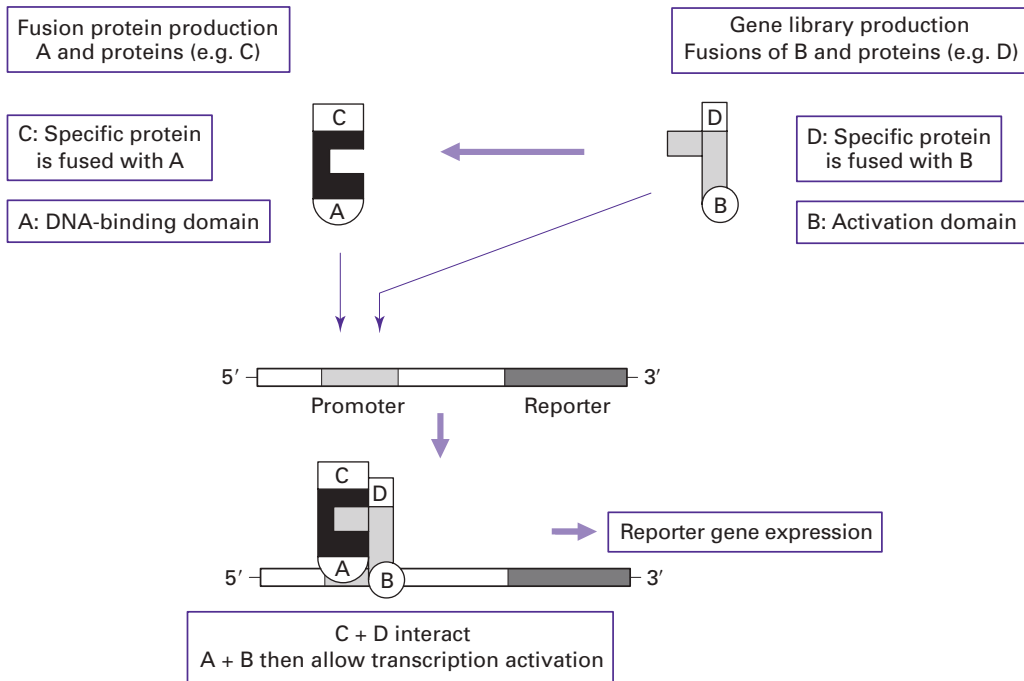


Fig. 6.44. Yeast two-hybrid system (interaction trapping technique). Transcription factors have two domains, one for DNA binding (A) and the other to allow binding to further proteins (B). Thus a recombinant molecule is formed from a protein (C) as a fusion with the DNA-binding domain. It cannot, however, 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.

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 (Table 6.4). This also has important consequences for the biotechnology and agricultural industries (see Table 6.10) (Section 6.11). 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 zygote 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

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 <i>tat</i> gene	Kaposi's sarcoma
Angiotensinogen	Expression of rat angiotensinogen gene	Hypertension
Cholesterol ester transfer protein (CET protein)	Expression of <i>CET</i> gene	Atherosclerosis
Hypoxanthine-guanine phosphoribosyl transferase (HPRT)	Inactivation of <i>HPRT</i> gene	HPRT deficiency

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 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 virus-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 on the cell, or inactivate an 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 focused on altering the levels of mRNA by manipulation of promoter sequences or levels of accessory proteins involved in the 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** or RNAi and uses micro RNA. Here, a number of techniques have been developed that allow the modulation of gene expression in certain cells. This type of cell-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 mutation that can occur in nucleic acids, either transiently or by being 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 occur potentially 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**. Examples 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 have traditionally been detected by Southern blotting or, if a convenient restriction site is available, by restriction fragment length polymorphism (RFLP) (Section 5.9.1). 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 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 simultaneous detection of more than one mutation is possible. 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 G + C rich sequence to one end of the heteroduplex. The mutated heteroduplex is identified by its altered melting properties through a polyacrylamide gel that contains a gradient of denaturant

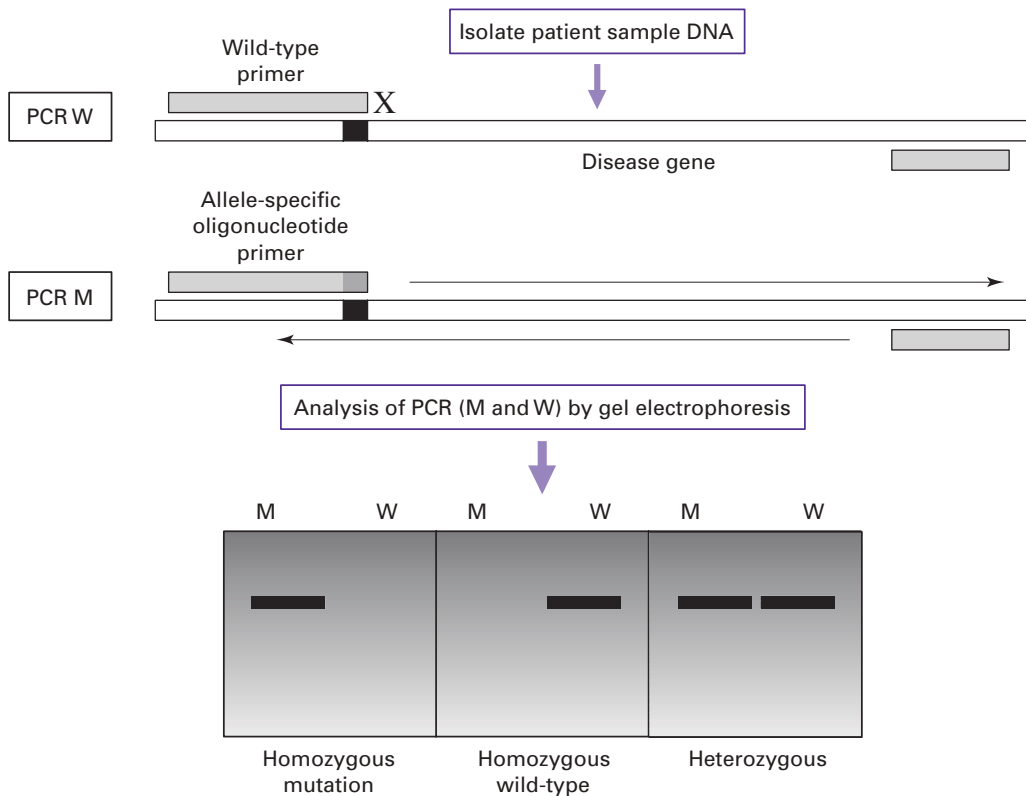


Fig. 6.45. Point mutation detection using allele-specific oligonucleotide PCR (ASO-PCR).

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 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 that anneal to these sequences using Southern blotting provides the means to type and identify those individuals (Section 5.3.1).

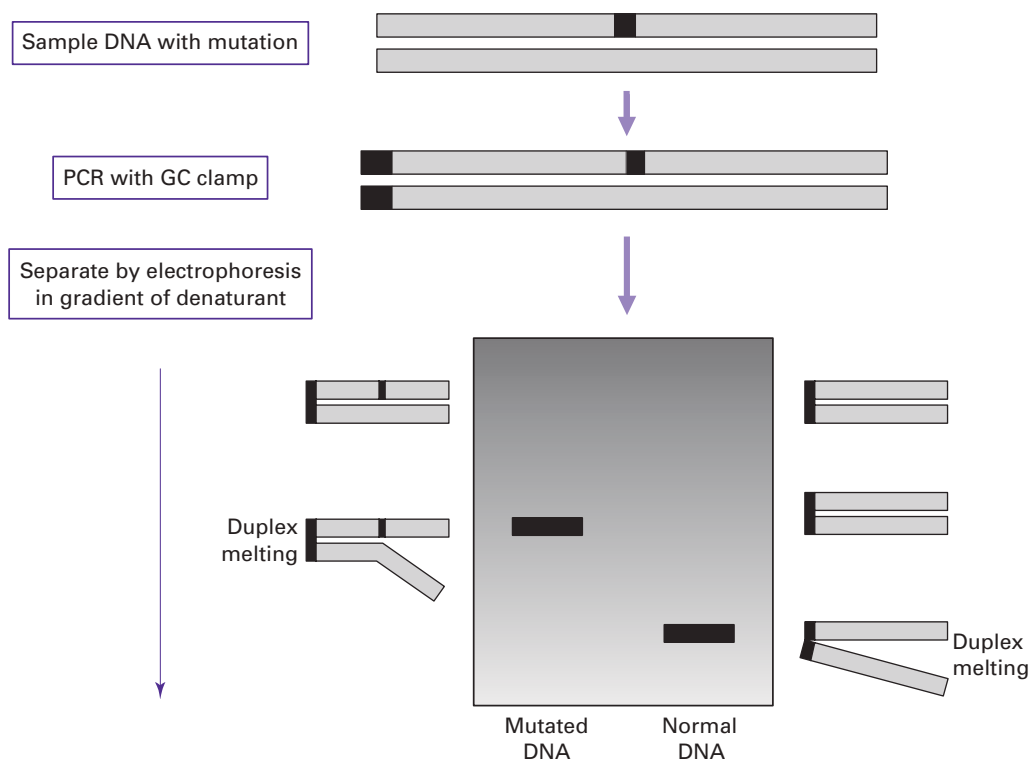


Fig. 6.46. Detection of mutations using denaturing gradient gel electrophoresis (DGGE).

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

DNA fingerprinting is a 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, 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 that under specific conditions detect only 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 that 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. The genetic detection of polymorphisms has been used in many cases of paternity testing and immigration control, and is becoming the central factor in many criminal investigations. It is also a valuable tool in plant biotechnology for cereal typing and in the fields of pedigree analysis and animal breeding.

6.8.8 Microarrays and DNA microchips

One exciting area of current development in molecular biology is in the development and refinement of microarrays or **DNA microchips** (Section 8.5). These provide a radically different approach to current laboratory molecular biological research strategies in that large-scale analysis and quantification of genes and gene expression are possible simultaneously. A microarray consists of an ordered arrangement of thousands of DNA sequences such as oligonucleotides or cDNAs deposited onto a solid surface approximately 1.2 cm × 1.2 cm. The solid support is usually glass, although silicon wafers have also been used successfully. Currently the arrays are synthesised on or off the glass and require complex fabrication methods similar to that used in producing microchips. They may also be spotted by robotic ultrafine microarray deposition instruments that dispense volumes as low as 30 pl. Alternatively on-chip fabrication as used by Affymetrix builds up layers of nucleotides using a process, borrowed from the microchip 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. One commercial example uses a 50 000 oligonucleotide array that represents known mutations in a tumour suppressor gene called *p53*, a

protein known to be mutated in many human cancers. Thus patient sample DNA is incubated on the array and any unhybridised DNA washed off. The array is then analysed and scanned for patterns of hybridisation by detection of a fluorescence signal. Any mutations in the *p53* gene may be rapidly analysed by computer interpretation of the resulting hybridisation pattern and mutation 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, 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 be used for analysing not only single genes but large numbers of genes or DNA representing microorganisms, viruses, etc. Since the potential for quantification 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 re-sequencing 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. SNPs are therefore clearly a candidate for microarray analysis and developments such as the Affymetrix HuSNP chip enables the simultaneous analysis of more than 10 000 SNPs on one gene chip. In order to simplify the problem of the vast numbers of SNPs that need to be analysed a haplotype mapping or HapMap project is underway to analyse SNPs that are inherited as a block; 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 the biosciences is the initiative to map and completely sequence a number of genomes from various organisms. The mapping

Table 6.6 Current selected genome-sequencing projects

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

and sequencing of a number of organisms indicated in Table 6.6 has been completed and many more are due for completion. Some 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 biological 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, such as the Genome Web (NCBI), that may be analysed and studied in detail chromosome by chromosome 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 this that it is now inconceivable to work without them.

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 10 than the longest inserts in the clones used in early mapping studies. The development of vectors with large insert capacity has enabled 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 that 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, owing to the impracticality of this approach a fingerprint of

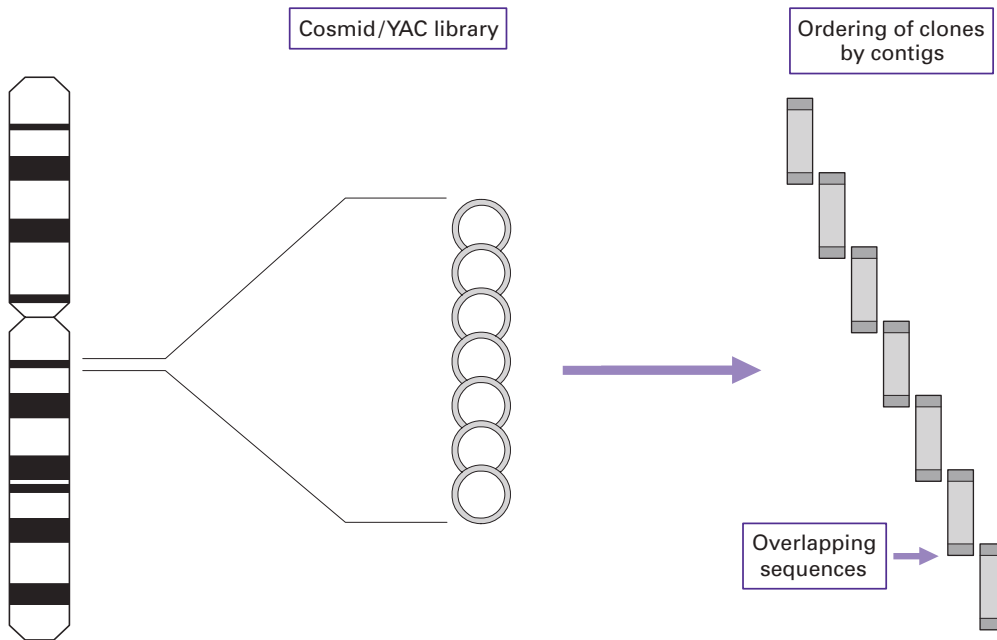


Fig. 6.47. Physical mapping using continuous overlapping cloned fragments (contigs). In order to assign the position of cloned DNA fragments resulting from the construction of a library in a YAC or cosmid vector, overlaps are detected between the clone fragments. These are created because of the use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

each clone is made 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 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 PCR has been developed based on *sequence-tagged sites* (STS). This is a small unique sequence of between 200 and 300 bp that is amplified by PCR (Fig. 6.48). The uniqueness of the STS is defined by the PCR primers that flank it. A PCR with those primers is performed and if the PCR results in selected amplification of the 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.

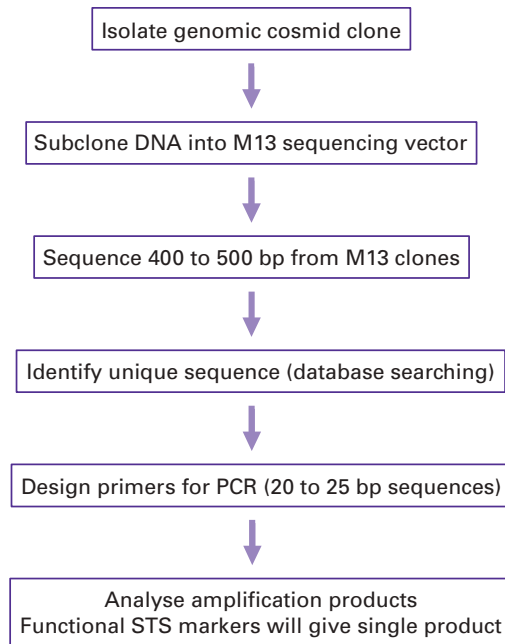


Fig. 6.48. General scheme of the production of a functional STS marker.

The same STS that occurs in two clones 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 onto a nylon membrane.

Alternatively a library of clones may be divided into pools and each pool PCR screened. This is usually a more rapid method of identifying an STS within a clone and further refinements 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 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 (Section 6.8.2). In fact a number of genes have been identified and their proteins determined where little was initially known about the genes except for their location. This method of gene discovery is known as **positional cloning** and was instrumental in the isolation of the *CFTR* gene (Fig. 6.50).

The number of genes that are actively expressed in a cell at any one time is estimated to be as little as 10% of the total. The remaining DNA is packaged and serves

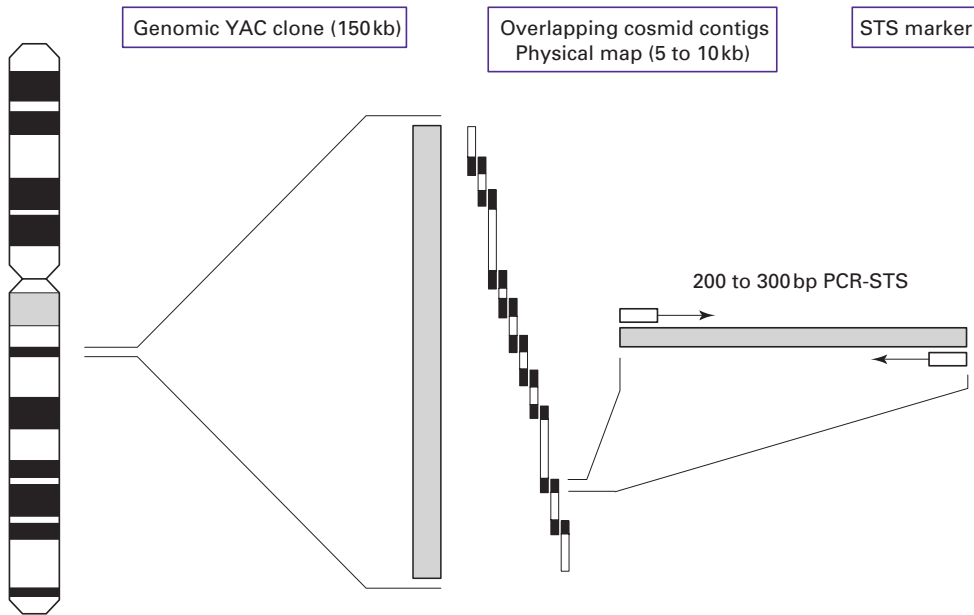


Fig. 6.49. The derivation of an STS marker. An STS is small unique sequence of between 200 and 300 bp that is amplified by PCR and allows ordering along a contig map. Such sequences are definable landmarks with which to order clones produced in genome libraries and usually lie approximately 100 000 bp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

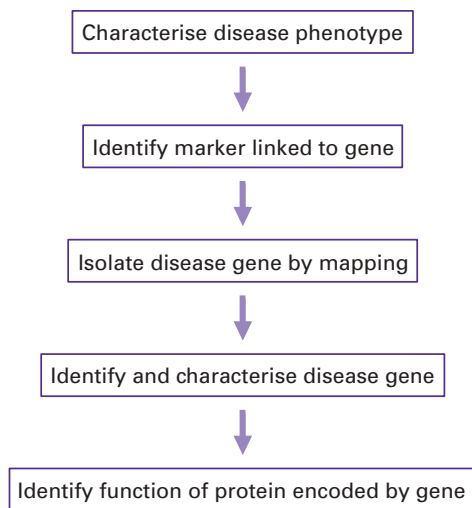


Fig. 6.50. The scheme of identification of a disease gene by positional cloning.

an as-yet unknown function. Recent investigations have found that certain active genes may be identified by the presence of so-called **HTF** (*HpaII* 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 *HpaII*. 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 that bind to sequences at the 5' end of the gene. Such PCRs may subsequently be used to map the putative gene to a chromosomal region or be used as probes to search a genomic DNA library for the remaining parts of the gene. Much interest currently lies in ESTs, since they may represent a shortcut 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 Human Genome Project

There is no doubt that the mapping and sequencing of the human genome was one of the most ambitious projects in contemporary science. Completed ahead of schedule, it has provided many new insights into gene function and gene regulation. It was also a multicollaborative effort that has engaged scientific research groups around the world and 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. Estimates indicate that as little as 10% of the genome appears to encode enzymes and proteins. Nevertheless this still corresponds to approximately 22 000 genes, although the understanding of the complete function of many remains a challenge. There is an extensive use of alternative splicing, where exons are essentially mixed and matched to form different mRNAs and thus different proteins. The study further aims to understand and possibly provide the eventual means of treating some of the 4000 known 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 that aim to understand the mutations that arise 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 that react in a specific way to drug treatment by identifying their genetic make-up. In particular SNPs (**single nucleotide polymorphisms**) may provide a key marker of potential disease development and

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

RBS, ribosome binding site; cDNA, complementary DNA.

reaction to a particular treatment. A simple example that has been known for some time is the reaction to a drug used to treat childhood acute lymphoblastic 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 as indicated in Section 6.8.7. It is also now possible to analyse SNPs that 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. This is available at present, although in a limited form, and commercial operations such as the Icelandic genetics company deCode may provide information based on analysis of disease genes in large population studies.

6.11 MOLECULAR BIOTECHNOLOGY AND ITS APPLICATIONS

It is a relatively short 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

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

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

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

Disorder	Defect	Gene target	Target cell
Emphysema	Deficiency (α 1-AT)	α 1-Antitrypsin (α 1-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
Soybean	Tolerance to herbicide, modified seed storage protein
Tomato	Modified ripening, resistance to insects and viruses
Sunflower	Modified seed storage protein

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 that reflect 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 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 that play key roles in such events. These genes and their proteins are also active at major points in the cell cycle and are intimately involved in cell regulation. A number of these are indicated in Table 6.8. In some cancers, well-defined molecular events have been correlated with mutations in oncogenes and therefore in the corresponding proteins. 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 in both normal and disease states, great promise is also evident 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 are also three of the most exciting developments of molecular biology. Genetic manipulation 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 of 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

- ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERT, K. and WATER, P. (2002). *Molecular Biology of the Cell*, 4th edn. Garland Publishing, London. (A comprehensive all round text.)
- BROWN, T. A. (2001). *Gene Cloning and DNA Analysis: An Introduction*. Blackwell Science, Oxford. (A very good introduction to genetics and molecular biology.)
- BROWN, T. A. (2002). *Genomes 2*. Bios Scientific Publishers, Oxford. (A very good introduction to the concepts of the genome.)
- RAPLEY, R. and HARBRON, S. (2004). *Molecular Analysis and Genome Discovery*. John Wiley & Sons, Chichester. (An up-to-date collection of key nucleic acid and proteins techniques in analysis and drug discovery.)
- STRACHEN, T. and READ, A. P. (2004). *Human Molecular Genetics*. Garland Science, London. (An excellent and comprehensive textbook with very good illustrations.)
- TURNER, P. C. C., BATES, A. D. and MCLENNAN, A. G. (2000). *Instant Notes in Molecular Biology*. Barnes and Noble/Bios Scientific Publishers Ltd, Oxford. (A clear treatment of key molecular biology concepts.)