



5

Molecular biology, bioinformatics and basic techniques

R. RAPLEY

- 5.1 Introduction
- 5.2 Structure of nucleic acids
- 5.3 Genes and genome complexity
- 5.4 Location and packaging of nucleic acids
- 5.5 Functions of nucleic acids
- 5.6 The manipulation of nucleic acids – basic tools and techniques
- 5.7 Isolation and separation of nucleic acids
- 5.8 Molecular biology and bioinformatics
- 5.9 Molecular analysis of nucleic acid sequences
- 5.10 The polymerase chain reaction (PCR)
- 5.11 Nucleotide sequencing of DNA
- 5.12 Suggestions for further reading

5.1 INTRODUCTION

The completion of the Human Genome Project (HGP) 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 the discipline 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 biology 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 which 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 carbons of nitrogenous bases of which there are two types, either a purine or a pyrimidine. 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 phosphate 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 end, each strand has a polarity or directionality. The polarity of the two strands of the molecule is 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 along a molecule that distinguishes one DNA (or RNA) 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, e.g. CGGATCT. 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.

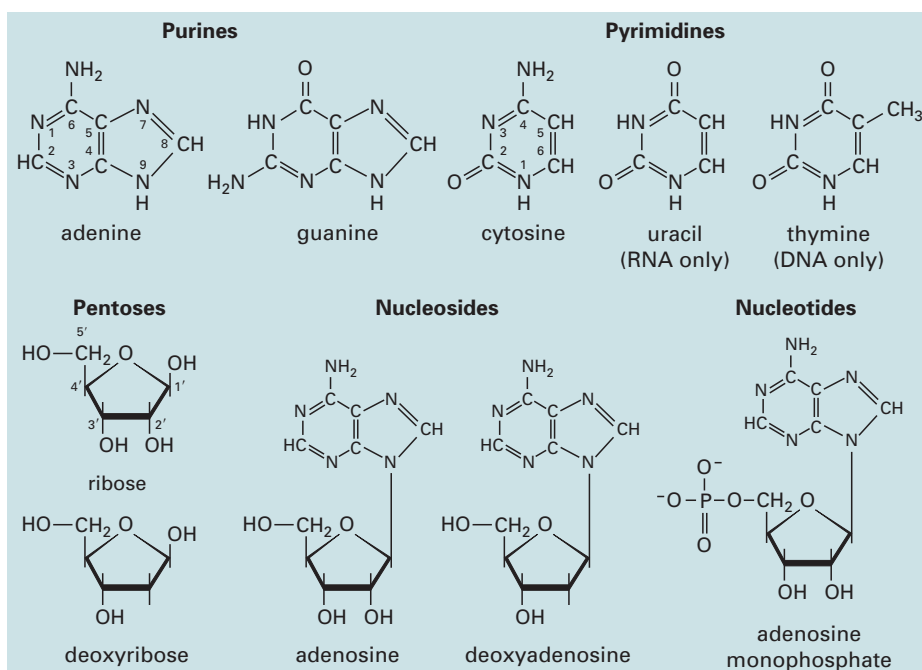


Fig. 5.1 Structure of bases, nucleosides and nucleotides.

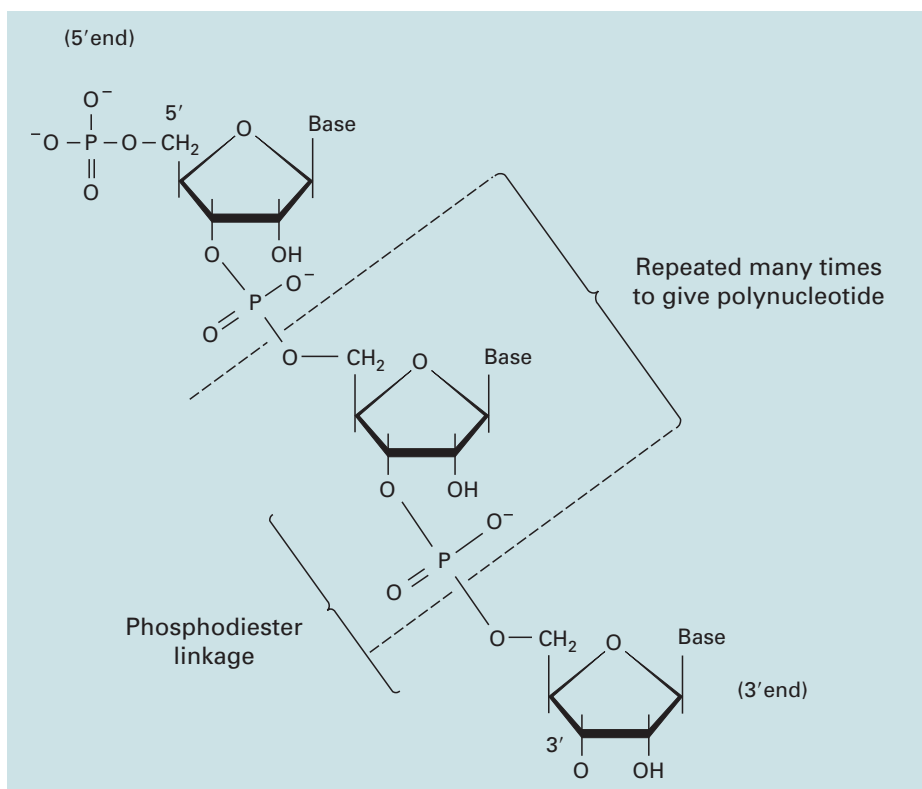


Fig. 5.2 Polynucleotide structure.

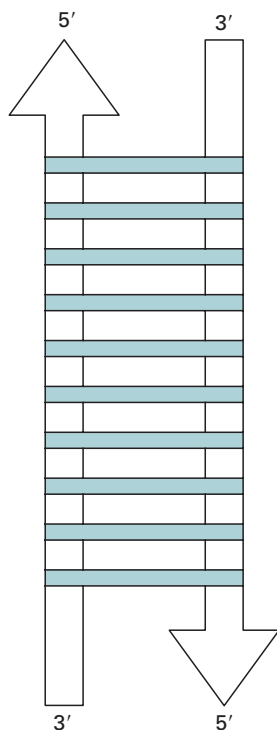


Fig. 5.3 The antiparallel nature of DNA. One strand in a double helix runs 5' to 3', whilst the other strand runs in the opposite direction 3' to 5'. The strands are held together by hydrogen bonds between the bases.

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 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, which are formed when DNA fibres are subjected to different relative humidities (Table 5.1).

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

Notes: 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.

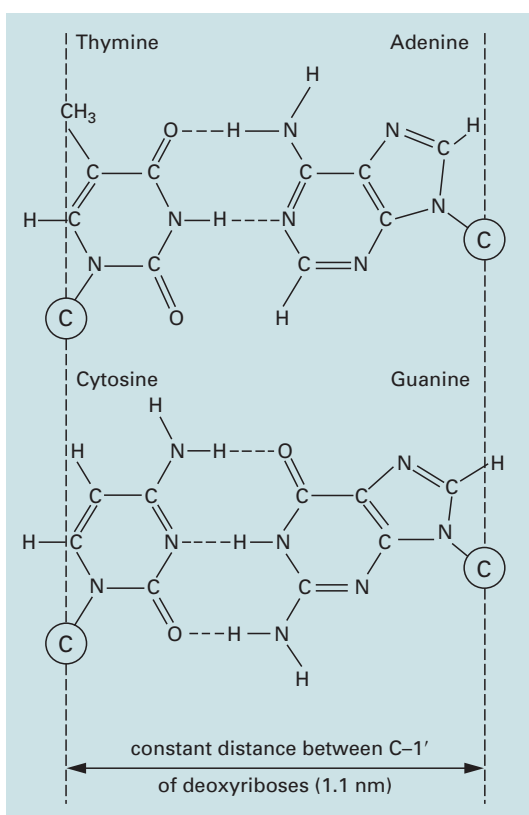


Fig. 5.4 Base-pairing in DNA. C in a circle represents carbon at the 1' position of deoxyribose.

The major distinguishing feature of B-DNA is that it has approximately 10 bases for one turn of the double helix; furthermore a distinctive major and minor groove 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.

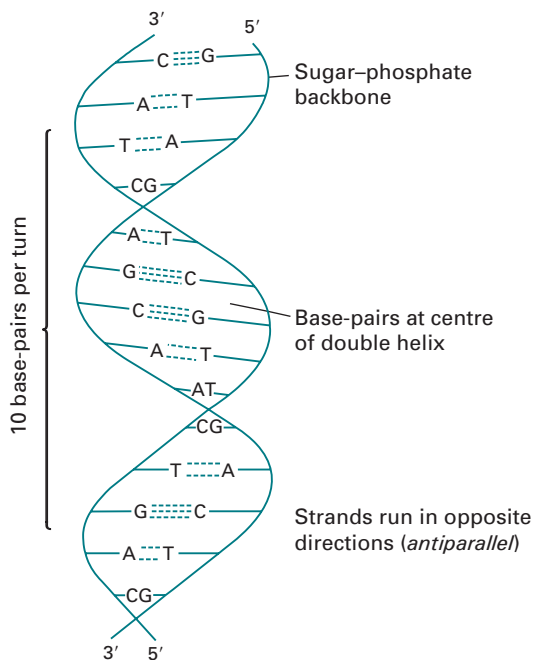


Fig. 5.5 The DNA double helix.

This form of DNA was first synthesised in the laboratory and is thought 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 that are self-complementary, and which can therefore base-pair if brought together by suitable folding of the 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 is termed **denaturation** and 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**.

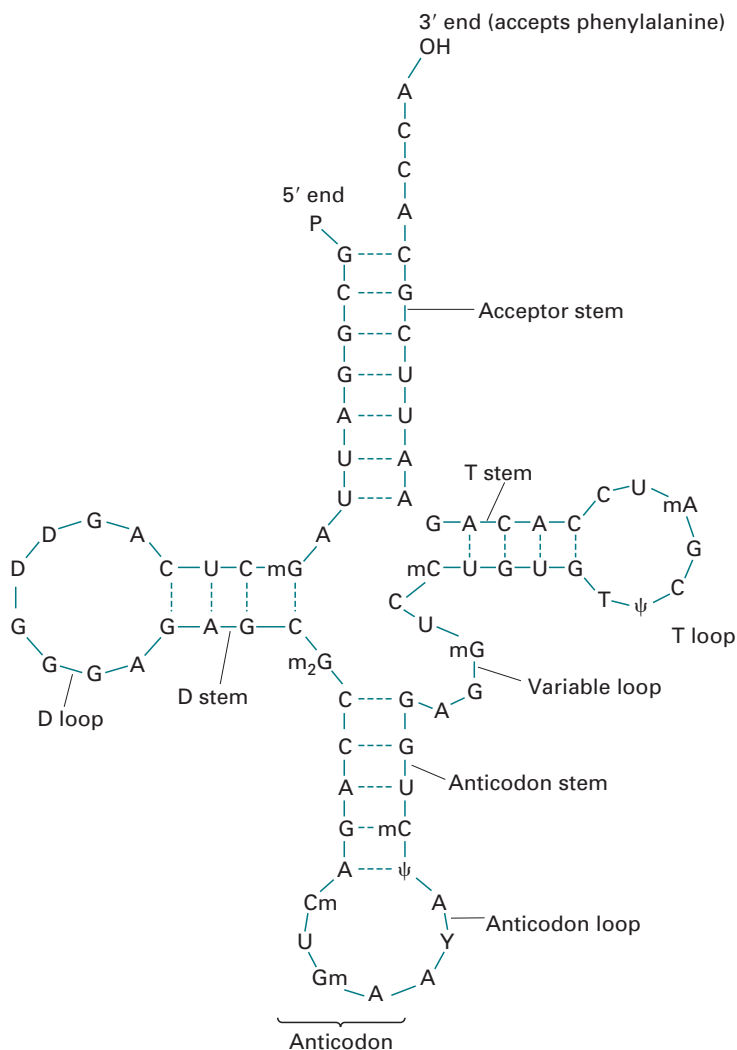


Fig. 5.6 Secondary structure of yeast tRNA^{Phe}. A single strand of 76 ribonucleotides forms four double-stranded 'stem' regions by base-pairing between complementary sequences. The anticodon will base-pair with UUU or UUC (both are codons for phenylalanine); phenylalanine is attached to the 3' end 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); X_m 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 DNA (Fig. 5.7). If several different samples of DNA are melted, it is found that the T_m is highest for those DNAs which 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

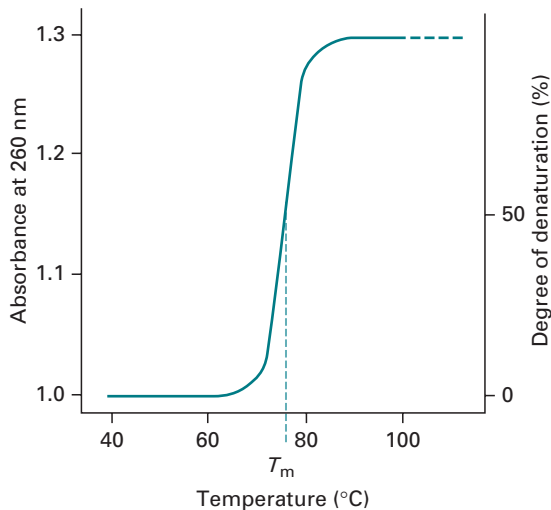


Fig. 5.7 Melting curve of DNA.

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 only be formed 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.

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. This hybridisation of complementary strands of nucleic acids is very useful for isolating a specific fragment of DNA from a complex mixture. 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 which codes for a single RNA or protein is called a gene, and the entire set of genes in a cell, organelle or virus forms its genome. Cells and organelles

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)
<i>Notes: bp, base-pairs; kb, kilobase-pairs.</i>		

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 which may vary (heterozygous). It is thought that there are approximately 20 000 genes present in the human genome, although not all will be expressed in a given cell at the same time. However various processing events such as alternative splicing or RNA editing can increase the number of actual proteins found in the cell in relation to the number of genes to nearly 1 million. 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 into so-called **classical satellite DNA**, **minisatellite** and **microsatellite DNA**, the latter being mainly composed 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.

5.3.2 Single nucleotide polymorphisms (SNPs)

A further important source of polymorphic diversity known to be present in genomes is termed **single nucleotide polymorphisms** or 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 every 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 polymorphisms may account for the differences in disease susceptibility, drug metabolism and response to environmental factors between individuals. Indeed there are now a number of initiatives to identify SNPs and produce genomic SNP maps. One initiative is the international **HapMap** project. This will enable a haplotype map of common sources of variations from groups of associated SNPs to be produced. This will potentially allow a set of so-called **tag SNPs** to be identified and potentially provide an association between the haplotype and a disease.

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. Perhaps the most well known example of this is the correlation of Down 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 5.9), 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 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

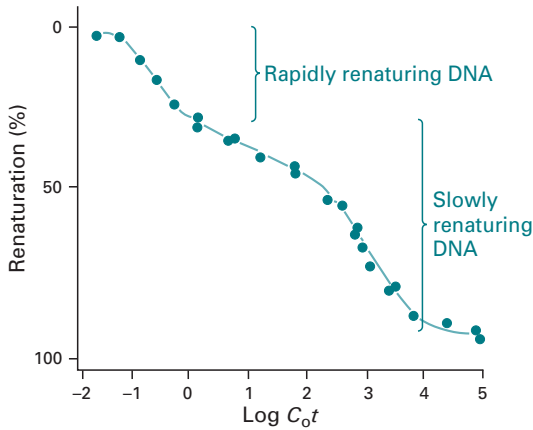


Fig. 5.8 C_0t curve of human DNA. DNA was allowed to renature at 60 °C after being completely dissociated by heat. Samples were taken at intervals and passed through a hydroxylapatite column to determine the percentage of double-stranded DNA present. This percentage was plotted against $\log C_0t$ (original concentration of DNA \times 3 time of sampling).

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 which 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.

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 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

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.

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).

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

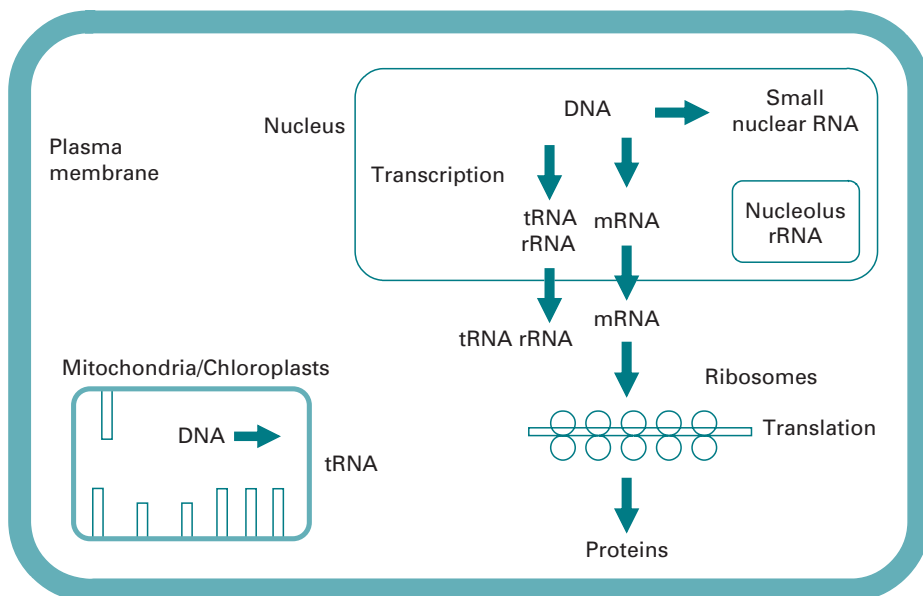


Fig. 5.10 Location of DNA and RNA molecules in eukaryotic cells and the flow of genetic information.

direct the synthesis of RNA molecules, which fall into three classes. Figure 5.10 indicates the locations of nucleic acids in prokaryotic and eukaryotic cells.

- *Messenger RNA (mRNA)* contains sequences of ribonucleotides which code for the amino acid sequences of proteins. A single mRNA codes for a single polypeptide chain in eukaryotes, but may code for several polypeptides in prokaryotes.
- *Ribosomal RNA (rRNA)* 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)* 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 which 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 it 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

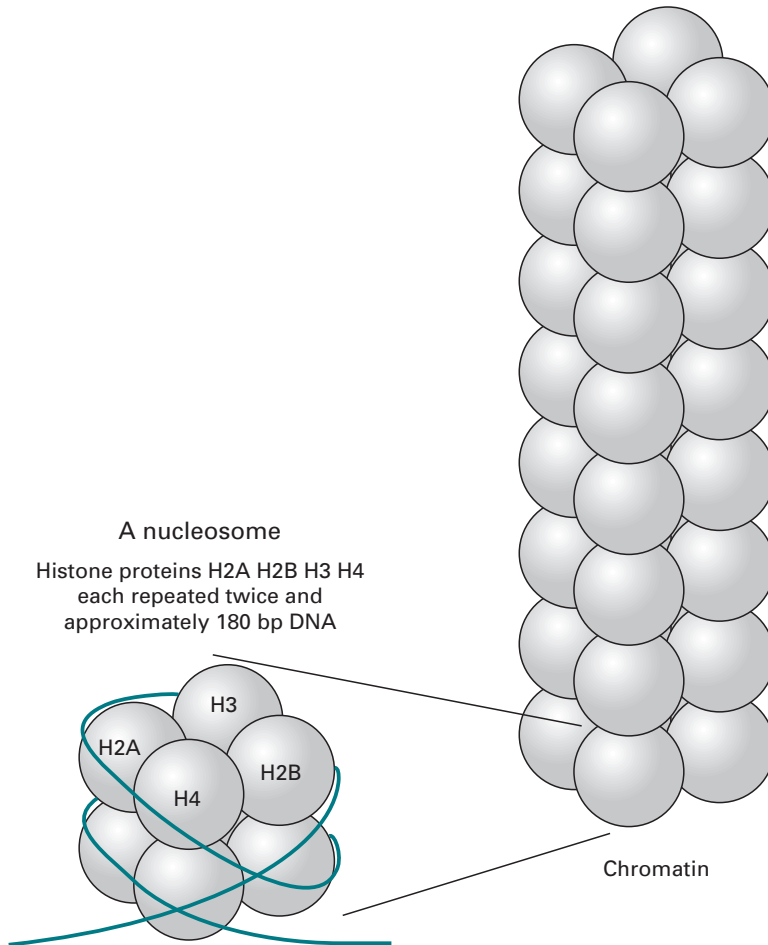


Fig. 5.11 Structure and composition of the nucleosome and chromatin.

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, 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 six.

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 seven (Fig. 5.11). These structures 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).

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** (DNApolIII) 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 only add new nucleotides 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 polI, 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

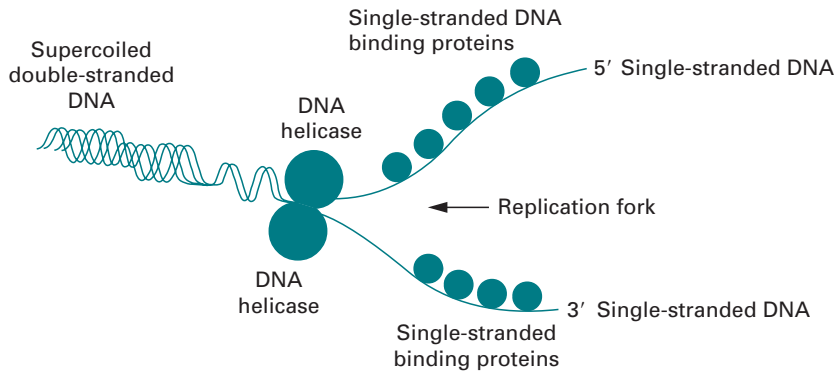


Fig. 5.12 Initial events at the replication fork involving DNA unwinding.

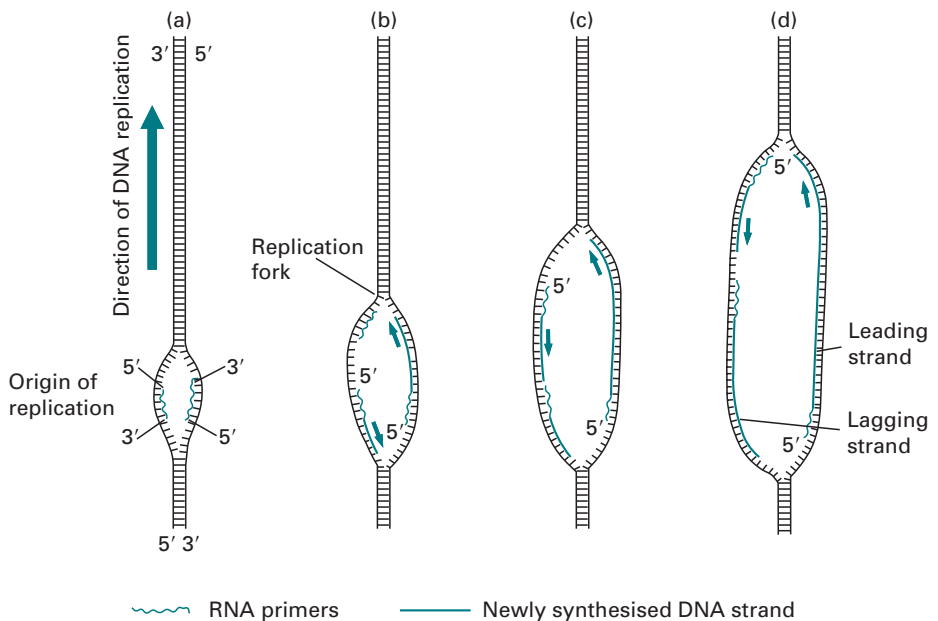


Fig. 5.13 DNA replication. (a) Double-stranded DNA separates at the origin of replication. RNA polymerase synthesises short DNA primer strands complementary to both DNA strands. (b) DNA polymerase III synthesises new DNA strands in a 5' to 3' direction, complementary to the exposed, old DNA strands, 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.

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 biology 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 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 5.9.

Repair systems allow the recognition of altered, mispaired or missing bases in double-stranded DNA and invoke an excision repair process. The systems characterised for bacterial systems are based on the length of repairable DNA during either 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 which 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 RNA polymerase from a DNA template encoding the gene. Most prokaryotic genes are made up of three regions. At the centre is the sequence which will be copied in the form of RNA, called the **structural gene**. To the 5' side (**upstream**) of the strand which will be copied (the plus (+) strand) lies a region called the **promoter**, and **downstream** of 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 + strand of 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 polymerase I and

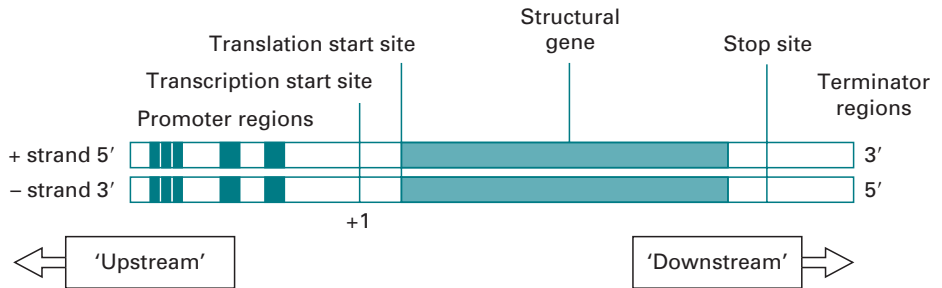


Fig. 5.14 Structure and nomenclature of a typical gene.

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 of 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 'GC-rich' sequence which is centred about –25 bp upstream from the TATA box. The GC element is thought to be important in the initial recognition and binding of RNA polymerase 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 is TFIID 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, compared to normally less than 100 bp for prokaryotic promoters. Particularly critical is the TATA box sequence, located approximately –35 bp upstream of the transcription initiation point in the majority of genes (Fig. 5.15b). This is analogous to the –10 sequence in prokaryotes. A number

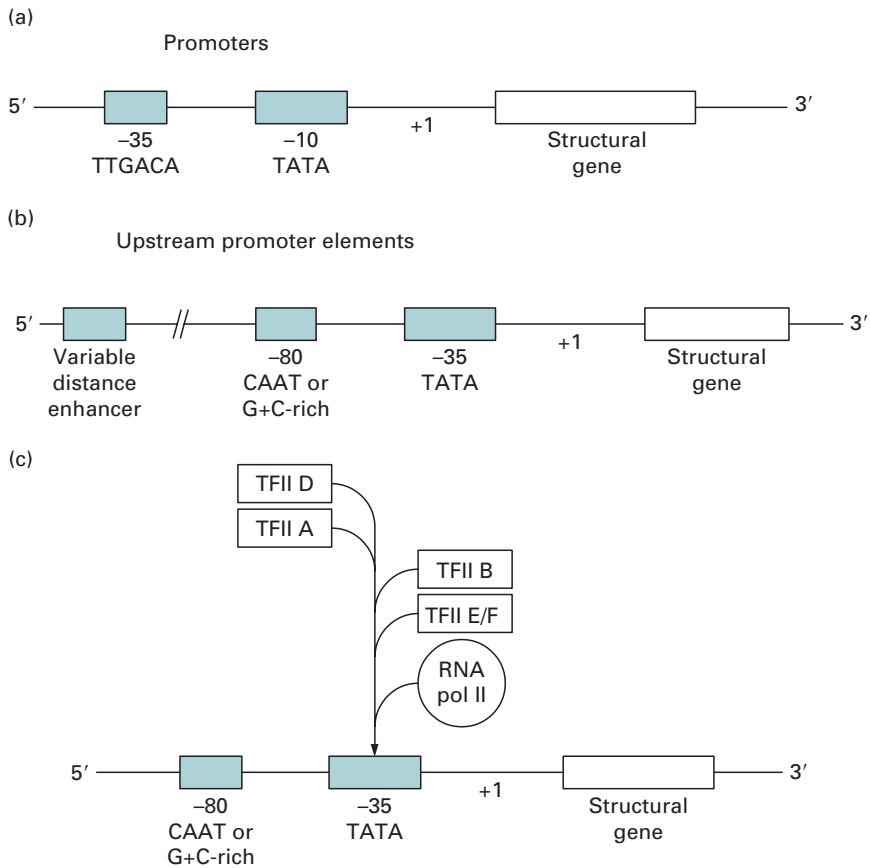


Fig. 5.15 (a) Typical promoter elements found in a prokaryotic cell (e.g. *E. coli*). (b) Typical promoter elements found in eukaryotic cells. (c) Generalised scheme of binding of transcription factors to the promoter regions of eukaryotic cells. Following the binding of the transcription factors IID, 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.

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, while others may be specific to particular genes or classes of genes.

Of particular interest is a class of promoter first investigated in the virus 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 of 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 they bind to 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 which are not interrupted. 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 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* gene and prevents RNA polymerase from transcribing the genes (Fig. 5.16a). However the presence of lactose 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**) which 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

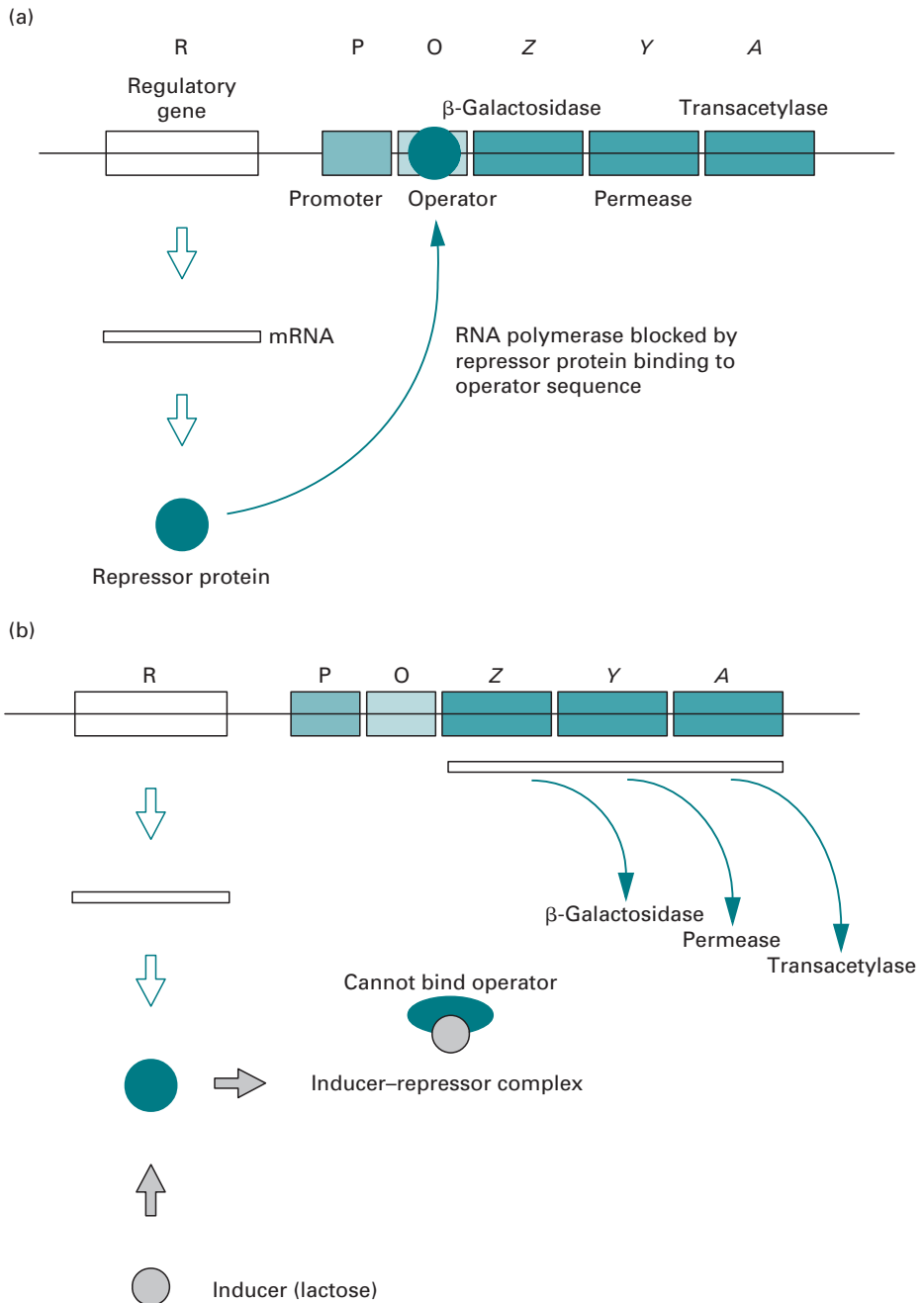


Fig. 5.16 Lactose operon (a) in a state of repression (no lactose present) and (b) following induction by lactose.

regions or **introns**). Intron–exon boundaries are generally determined by the sequence GU–AG 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

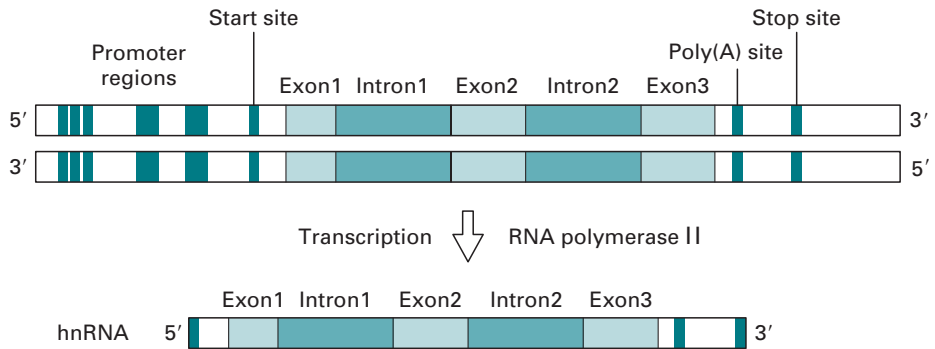


Fig. 5.17 Transcription of a typical eukaryotic gene to form heterogeneous nuclear RNA.

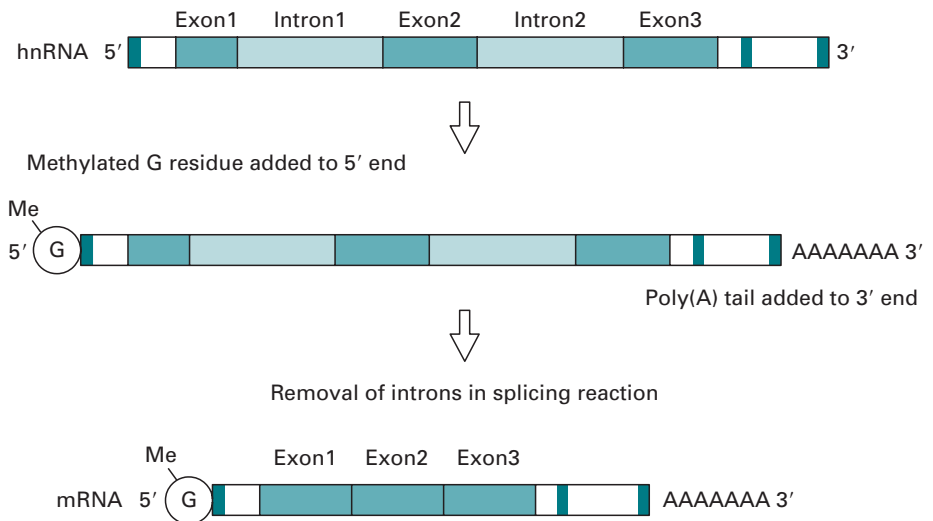


Fig. 5.18 Post-transcriptional modifications of heterogeneous nuclear RNA.

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 such as histone genes contain no introns 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 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 biology 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 70S or 80S depending on their sedimentation coefficient. Prokaryotic cells have 70S ribosomes whilst those of the eukaryotic cytoplasm are 80S. 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*: involving the assembly of the ribosome subunits and the binding of the mRNA.
- *Elongation*: where specific amino acids are used to form polypeptides, this being directed by the codon sequence in the mRNA.
- *Termination*: which involves the disassembly of the components of translation following the production of a polypeptide.

Transfer RNA molecules are also essential for translation. Each of these are covalently linked to a specific amino acid, forming an **aminoacyl tRNA**, and each has a triplet of bases exposed which 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 ‘adapter’ molecule, bringing together a codon and its corresponding amino acid. 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 a 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, and 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

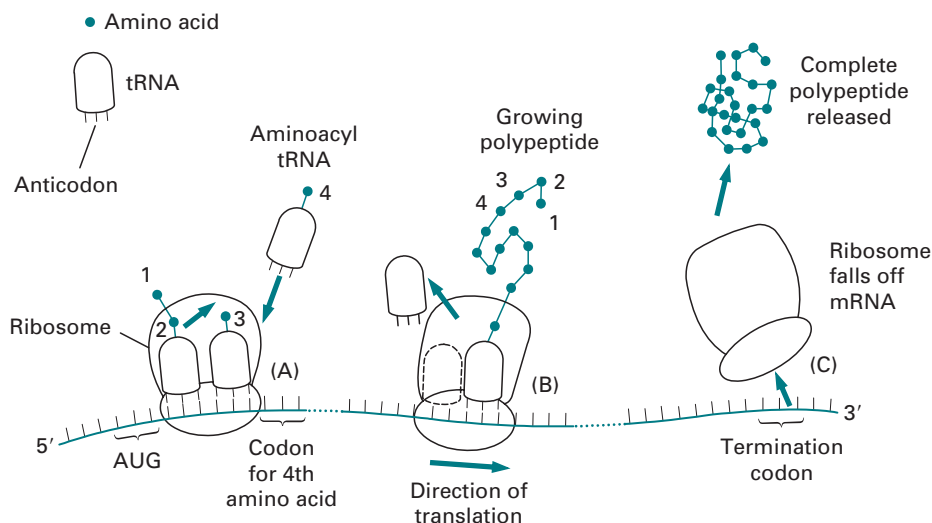


Fig. 5.19 Translation. Ribosome A has moved only a short way from the 5' end of the mRNA, and has built up a dipeptide (on one tRNA) that is about to be transferred onto the third amino acid (still attached to tRNA). Ribosome B has moved much further along the mRNA and has built up an oligopeptide that 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.

in prokaryotes by base-pairing between the Shine–Dalgarno sequence (Kozak sequence in eukaryotes) and a complementary sequence of one of the ribosome's rRNAs, 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 6.8.6.

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 either at the gene level or at 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 biology technique to aid in the modulation of mRNA is termed *RNA interference* (RNAi). This involves the synthesis of short double-stranded RNA molecules which 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 which, 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 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

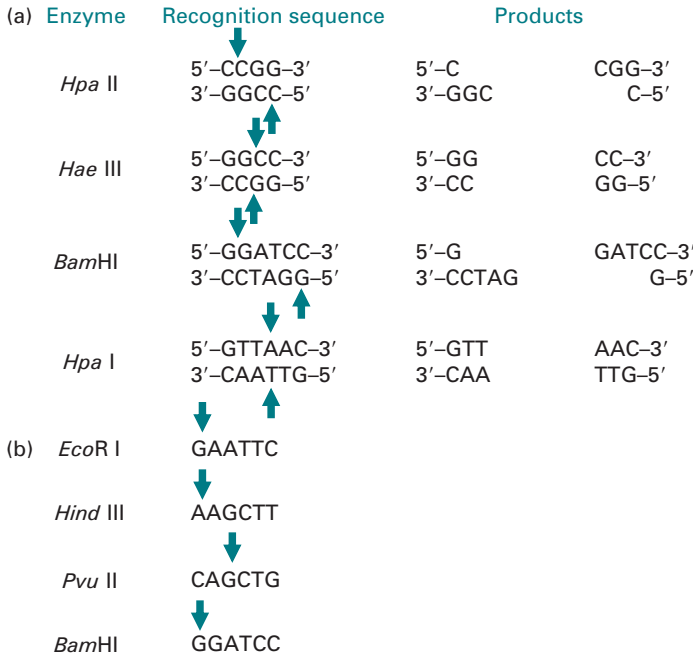


Fig. 5.20 Recognition sequences of some restriction enzymes showing (a) full descriptions and (b) conventional representations. Arrows indicate positions of cleavage. Note that all the information in (a) can be derived from knowledge of a single strand of the DNA, whereas in (b) only one strand is shown,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

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'→5' exonuclease activity
	Klenow	DNA pol I lacks 5'→3' exonuclease activity
	T4 DNA pol	Lacks 5'→3' exonuclease activity
DNA polymerases	<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
	T7 RNA pol	DNA-dependent RNA polymerase
RNA polymerases	T3 RNA pol	DNA-dependent RNA polymerase
	Q β replicase	RNA-dependent RNA polymerase, used in RNA amplification
	DNase I	Non-specific endonuclease that cleaves DNA
	Exonuclease III	DNA-dependent 3'→5' stepwise removal of nucleotides
Nucleases	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

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

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 is 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+} ions needed for enzymes that degrade DNA termed DNase. Ideally, cell walls, if present, should be digested enzymatically (e.g. lysozyme treatment of bacteria), and the cell 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 flow chart of DNA extraction is indicated in Fig. 5.21.

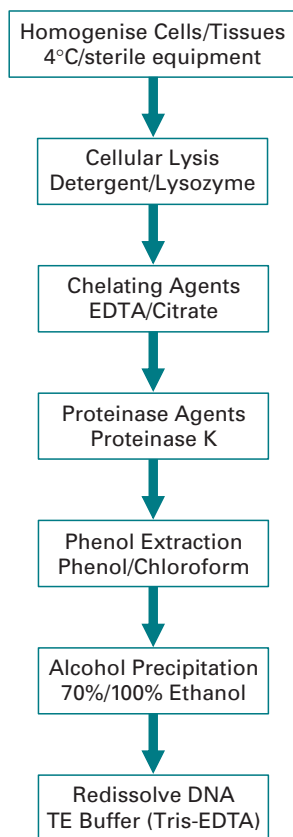


Fig. 5.21 General steps involved in extracting DNA from cells or tissues.

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 $50 \mu\text{g ml}^{-1}$ of DNA and so:

$$50 \times A_{260} = \text{concentration of DNA sample } (\mu\text{g ml}^{-1})$$

Contaminants may also be identified by scanning **UV spectrophotometry** from 200 nm to 300 nm. A ratio of 260 nm : 280 nm of approximately 1.8 indicates that the sample is free of 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

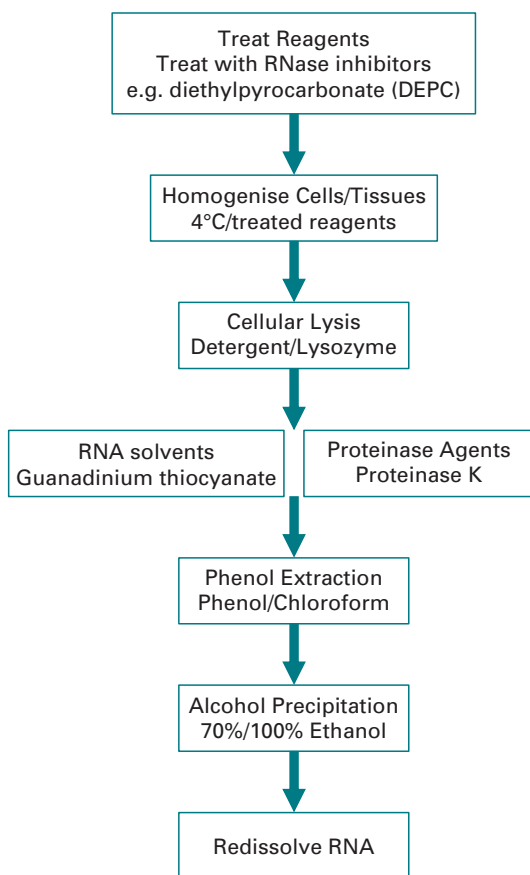


Fig. 5.22 General steps involved in extracting RNA from cells or tissues.

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 which is commonly used in RNA extraction is guanadinium thiocyanate which is both a strong inhibitor of RNase and a protein denaturant. A flow chart 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, the rRNA molecules 23S and 16S for prokaryotes and 18S and 28S for eukaryotes, appear as discrete bands on the agarose gel and thus indicate that the other RNA components are likely to be intact. This is usually carried out under denaturing conditions to prevent secondary structure formation in the RNA. The concentration of the RNA may be estimated by using UV spectrophotometry. At 260 nm 1 absorbance unit equates to $40 \mu\text{g ml}^{-1}$ of RNA and therefore:

$$40 \times A_{260} = \text{concentration of DNA sample } (\mu\text{g ml}^{-1})$$

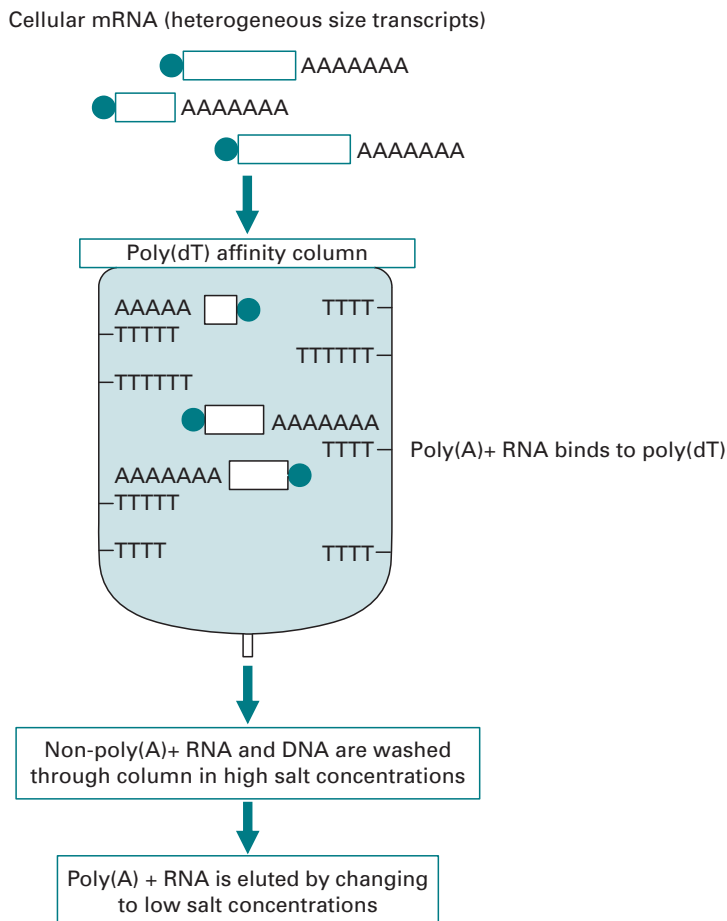


Fig. 5.23 Affinity chromatography of poly(A)+RNA.

Contaminants may also be identified in the same way as that for DNA by scanning UV spectrophotometry; however, in the case of RNA a 260 nm : 280 nm ratio of approximately 2 would be expected for a sample containing no protein (Section 5.8.1).

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 means such as electrophoretic or chromatographic separations based on differences in nucleic acid fragment sizes or physicochemical characteristics. Nanodrop spectrophotometer systems have also aided the analysis of nucleic acids in recent years in allowing the full spectrum of information whilst requiring only a very small (microlitre) sample volume.

5.7.3 Automated and kit-based extraction of nucleic acids

Most of the current reagents used in molecular biology and the most common techniques can now be found in kit form or can be automated, and the extraction of nucleic acids by these means is no exception. The advantage of their use lies in the fact that the reagents are standardised and quality control tested providing a high degree of reliability. For example glass bead preparations for DNA purification have been used increasingly and with reliable results. Small compact column-type preparations such as QIAGEN columns are also used extensively in research and in routine DNA analysis. 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. 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, e.g. 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 most 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 **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 give results quickly. 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 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

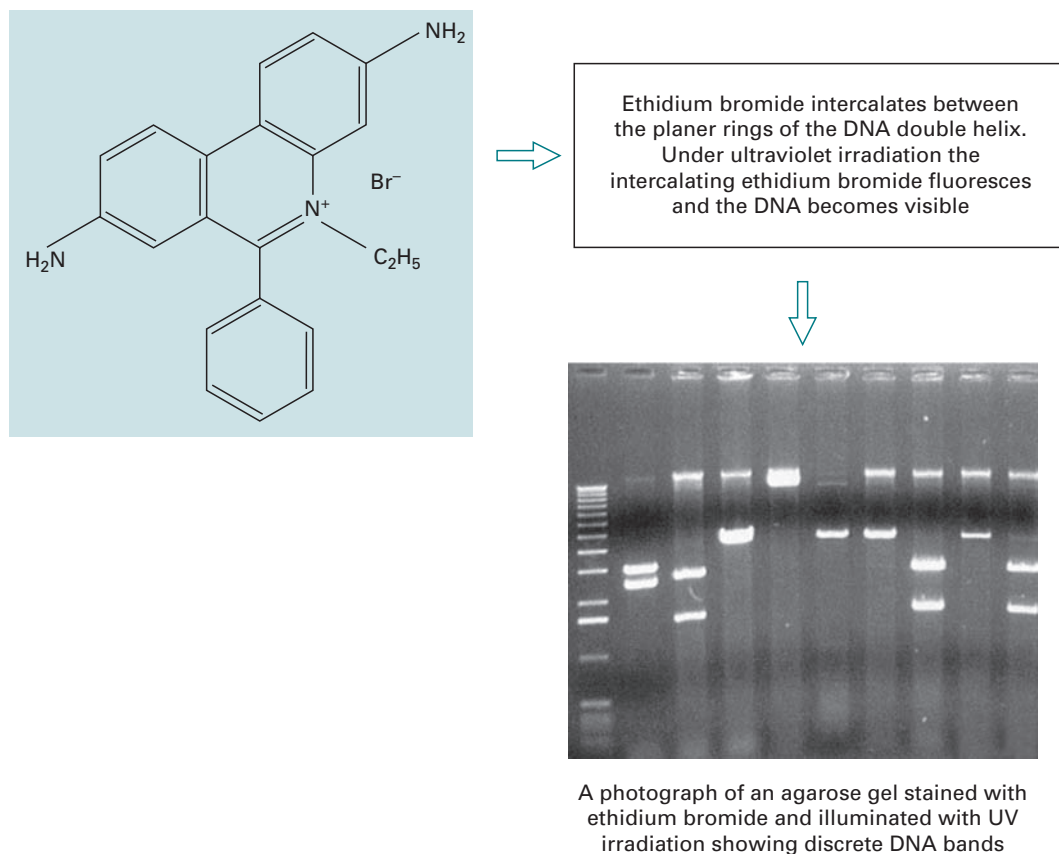


Fig. 5.24 The use of ethidium bromide to detect DNA.

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 and standardised reagents are available that are now very popular. This is especially useful in situations where a large number of samples or high-throughput analysis is required. In addition technologies such as the Agilent's Lab-on-a-chip have been developed that obviate the need to prepare electrophoretic gels. These employ **microfluidic circuits** constructed on small cassette units that contain interconnected micro-reservoirs. 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 the units, being computer controlled, allow data

to be captured within a very short timescale. More recently alternative methods of analysis including high performance liquid chromatography based approaches have gained in popularity, especially for DNA mutation analysis. Mass spectrometry is also becoming increasingly used for nucleic acid analysis.

5.8 MOLECULAR BIOLOGY AND BIOINFORMATICS

5.8.1 Basic bioinformatics

Bioinformatics is now an established and vital resource for molecular biology research and is also a mainstay of routine analysis of DNA. This increase in use of bioinformatics has been driven by the increase in genetic sequence information and the need to store, analyse and manipulate the data. 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 now an indispensable starting point for molecular biology research. The main primary databases include GenBank at the National Institutes of Health (NIH) in the USA, EMBL at the European Bioinformatics Institute (EBI) at Cambridge, UK and the DNA Database of Japan (DDBJ) at Mishima in Japan. These databases contain the nucleotide sequences which 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 found in cytochrome P450 domains, or DNA-binding domains. Importantly all of the databases may be freely accessed over the internet. A number of these important databases and internet resources are listed in Table 5.4. 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 internet 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 Map Viewer. Further resources such as ORF (open reading frame) finder allow a search to be undertaken for open reading frames, e.g. 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

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

Database or resource		URL (uniform resource locator)
<i>General DNA sequence databases</i>		
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 >
<i>Protein sequence databases</i>		
Swiss-Prot	European protein sequence database	< http://www.expasy.org >
UniProt TREMBL	European protein sequence database	< http://www.ebi.ac.uk/trembl >
<i>Protein structure databases</i>		
PDB	Protein structure database	< http://www.rcsb.org >
<i>Genome project databases</i>		
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 >

regions. A number of bioinformatics resources such as GRAIL can be used to identify such features in a DNA sequence.

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 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, so it has a single site 3 kb from one end; but it is not possible at this stage to say if 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, then A and B cut at opposite ends of the molecule; if they are 1, 2 and 6 kb, the sites are near each other. Not surprisingly, the mapping of real molecules is rarely

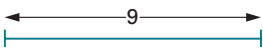
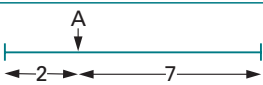
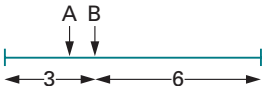


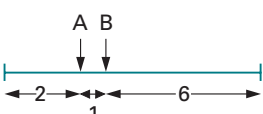
Treatment	Measured sizes of fragments (kb)	Interpretation
No digestion	9	
Enzyme A	2 + 7	
Enzyme B	3 + 6	<div> <div> EITHER  </div> <div> OR  </div> </div>
Enzymes A + B	2, 3 + 4	
	alternative result 1, 2 + 6	

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.

as simple as 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 to be carried out, 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 5.9.4). 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

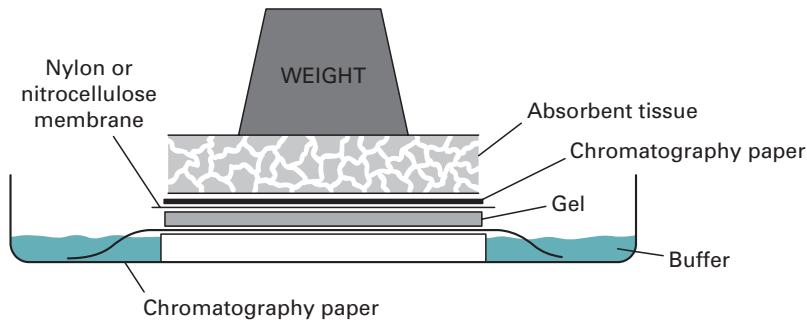


Fig. 5.26 Southern blot apparatus.

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 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 to not only detect specific mRNA molecules but it may also be used 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; it does not, however, 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 biology 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, that is 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

Polypeptide		Phe	Met	Pro	Trp	His	
Corresponding nucleotide sequences	5'	TTC	ATC	CCC A G	TGG	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).

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 DNA information 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 synthesise appropriate oligonucleotide sequences chemically. Due 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 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 which contains as many tryptophan and methionine residues as possible, since these have unique codons, and there will therefore be 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 biology 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 this has been carried out using **radioactive labels**, but gaining in popularity are **non-radioactive labels**.

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 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 their improved safety, have led to their greater acceptance.

The labelling systems are either termed 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 which has a label attached. At present three of the main labels in use are biotin, fluorescein and digoxigenin. These molecules are covalently linked 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 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 either be 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 products of either 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 ^{32}P , is added. This is usually carried out by using two enzymes; the first, alkaline phosphatase, is used to remove the existing phosphate group from the DNA. Following removal of the released 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).

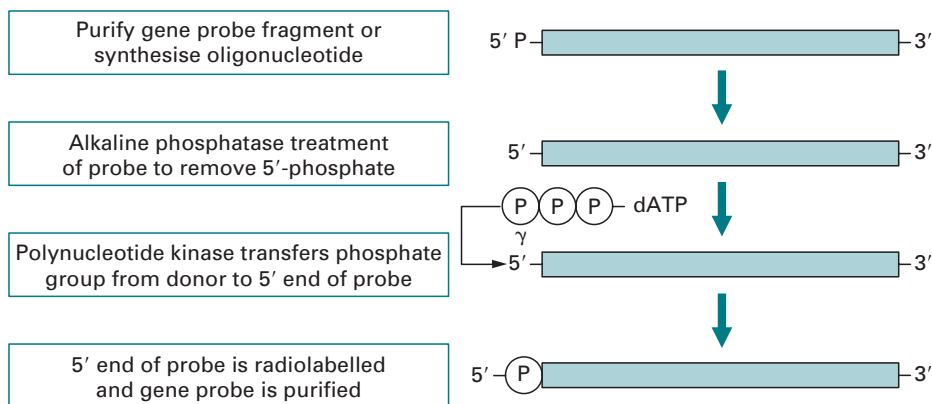


Fig. 5.28 End-labelling of a gene probe at the 5' end with alkaline phosphatase and polynucleotide kinase.

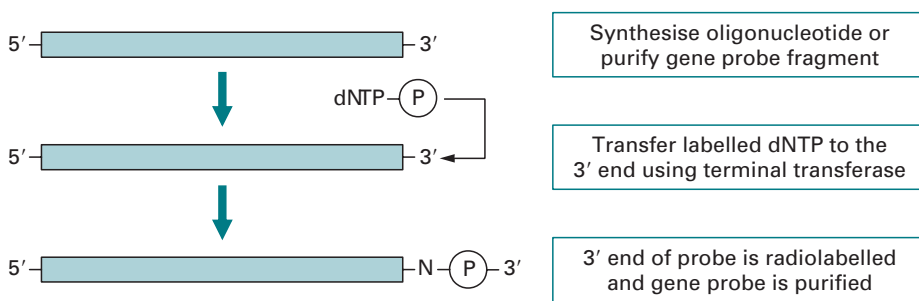


Fig. 5.29 End-labelling of a gene probe at the 3' end using terminal transferase. Note that the addition of a labelled dNTP at the 3' end alters the sequence of the gene probe.

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

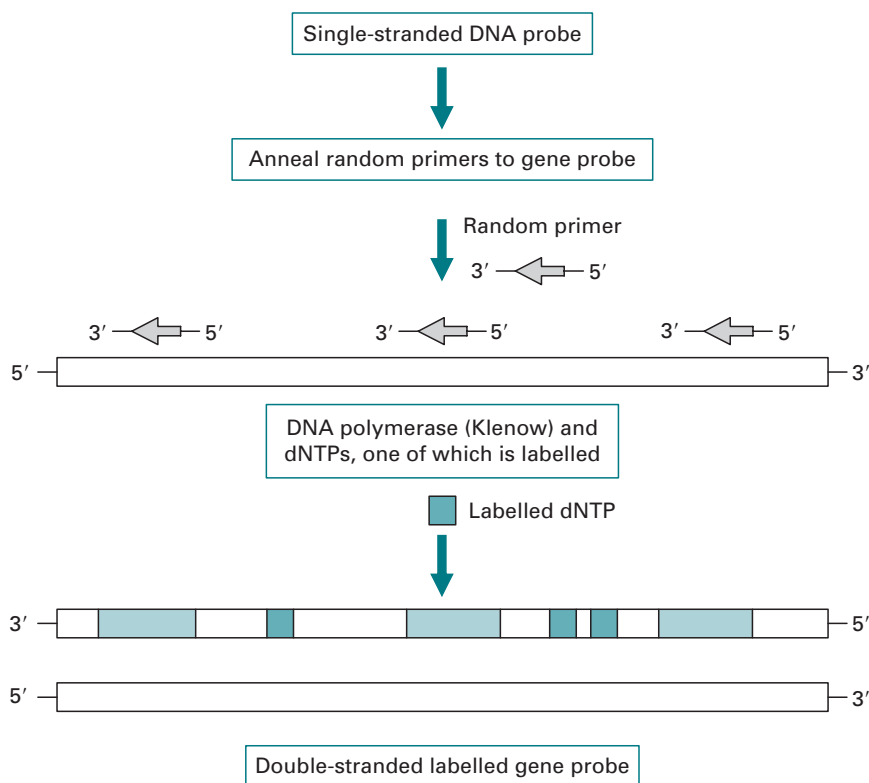


Fig. 5.30 Random primer gene probe labelling. Random primers are incorporated and used as a start point for Klenow DNA polymerase to synthesise a complementary strand of DNA whilst incorporating a labelled dNTP at complementary sites.

3' exonuclease activity. This is prepared by cleavage of DNA polymerase with subtilisin, giving a large enzyme fragment which has no 5' to 3' exonuclease activity, but which still acts as a 5' to 3' polymerase. Thus when the Klenow enzyme is mixed with the annealed DNA sample in the presence of dNTPs, including at least one which 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 5.11.4).

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 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

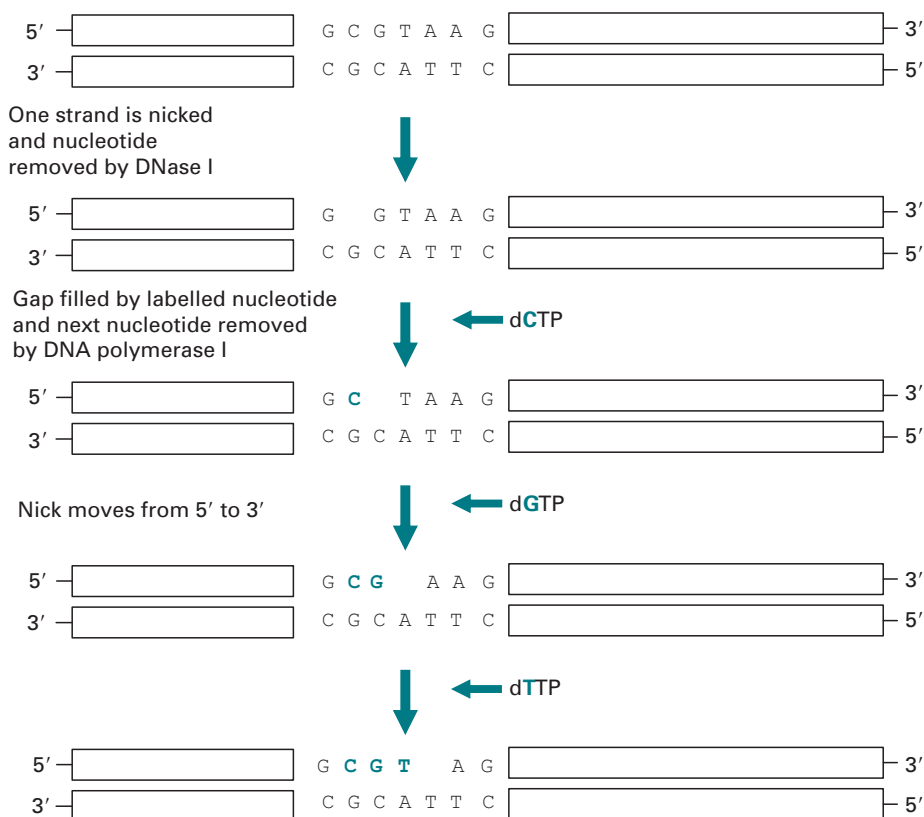


Fig. 5.31 Nick translation. The removal of nucleotides and their subsequent replacement with labelled nucleotides by DNA polymerase I increase the label in the gene probe as nick translation proceeds.

and a quencher molecule at the other. The oligonucleotide has a stem-loop structure where the stems place 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 Förster or 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 detect nucleic acid amplification system 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 (PCR)

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

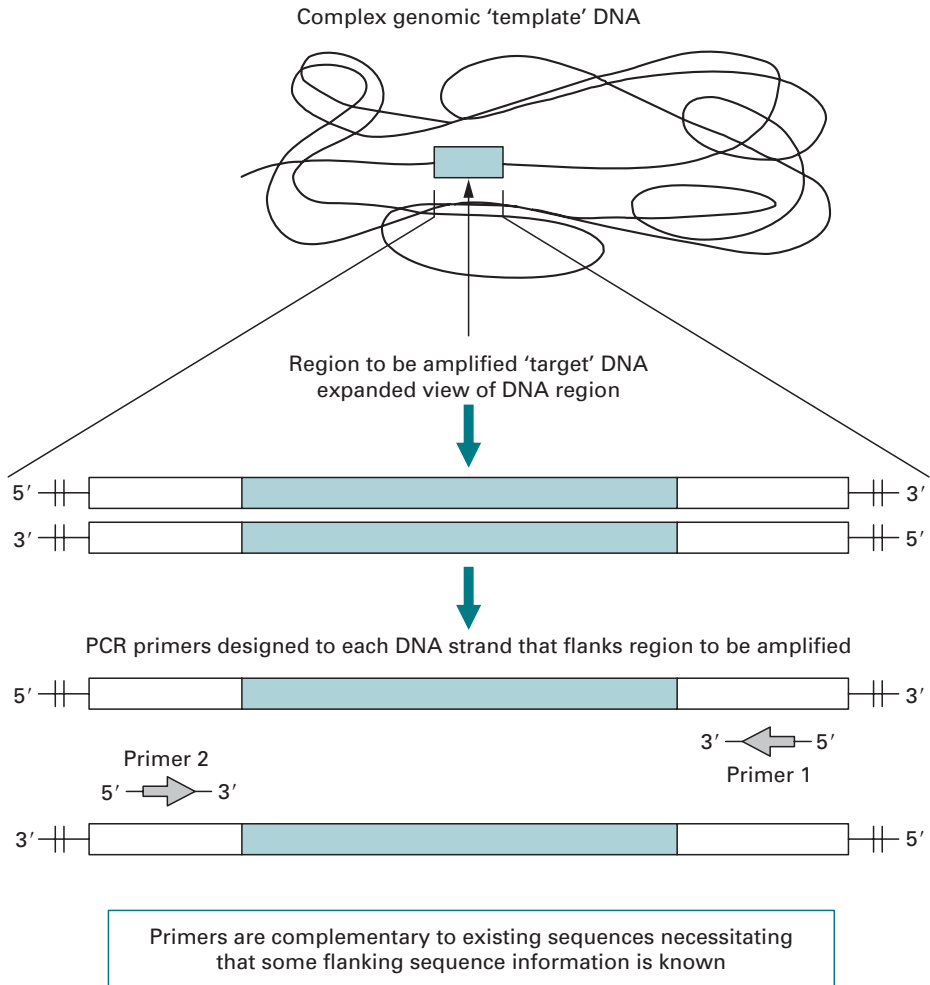


Fig. 5.32 The location of polymerase chain reaction (PCR) primers. PCR primers designed for sequences adjacent to the region to be amplified allow a region of DNA (e.g. a gene) to be amplified from a complex starting material of genomic template DNA.

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 DNA purification. It does require the knowledge of some DNA sequence information which flanks 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

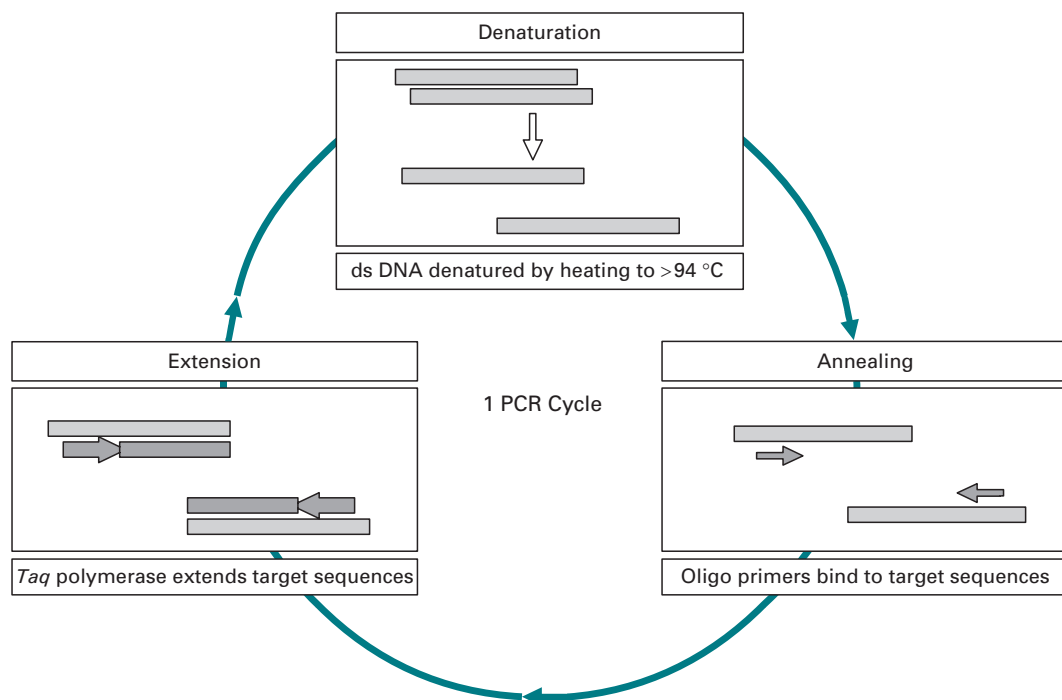


Fig. 5.33 A simplified scheme of one PCR cycle that involves denaturation, annealing and extension. ds, double-stranded.

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 which 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.

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 40–60 °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 the original strands will 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 have to not only be complementary to sequences flanking the target DNA but also must not be self-complementary or bind each other to form dimers since both prevent DNA amplification. They also have to be matched in their GC 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 amplified, primer length, product size, GC 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 of the primers or both are random. This gives rise to

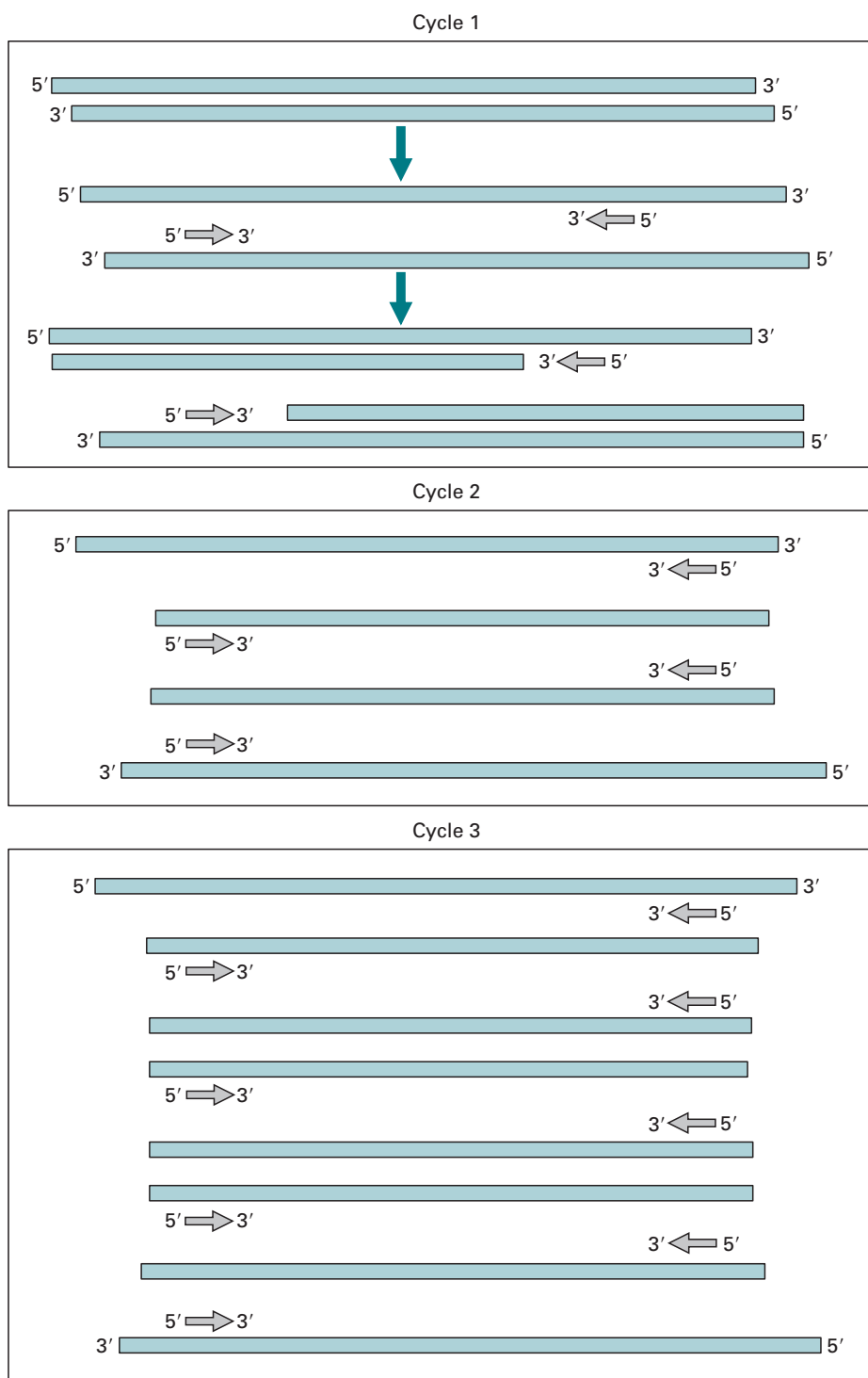


Fig. 5.34 Three cycles in the PCR. As the number of cycles in the PCR increases, the DNA strands that are synthesised and become available as templates are delimited by the ends of the primers. Thus specific amplification of the desired target sequence flanked by the primers is achieved. Primers are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

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 or species. This is sometimes referred to as **random amplified 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 using techniques such as qPCR (Section 5.10.7).

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 current 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 which converts the RNA to cDNA 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* 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 which replicate and thus 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 even 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 UV 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 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.

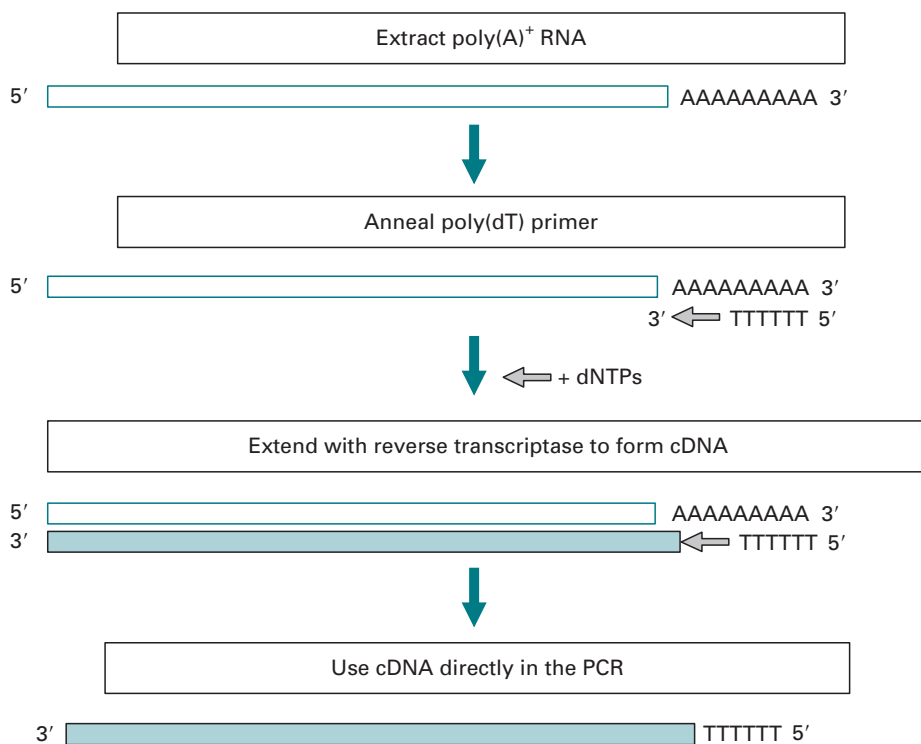


Fig. 5.35 Reverse transcriptase–PCR (RT–PCR): mRNA is converted to complementary DNA (cDNA) using the enzyme reverse transcriptase. The cDNA is then used directly in the PCR.

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 while 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 which 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 PCR (qPCR)

One of the most useful PCR applications is **quantitative PCR** or qPCR. This allows the PCR to be used as a means of identifying the initial concentrations of DNA or cDNA template used. Early qPCR methods involved the comparison of a standard or

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

control DNA template amplified with separate primers at the same time as the specific target DNA. However these types of quantitation rely on the fact that all the reactions are identical and so any factors affecting this may also affect the result. The introduction of thermal cyclers that incorporate the ability to detect the accumulation of DNA through fluorescent dyes binding to the DNA has rapidly transformed this area.

In its simplest form a PCR is set up that includes a DNA-binding cyanine dye such as **SYBR green**. This dye binds to the major groove of double-stranded DNA but not single-stranded DNA and so as amplicons accumulate during the PCR process SYBR green binds the double-stranded DNA proportionally and fluorescence emission of the dye can be detected following excitation. Thus the accumulation of DNA amplicons can be followed in real time during the reaction run. In order to quantitate unknown DNA templates a standard dilution is prepared using DNA of known concentration. As the DNA accumulates during the early exponential phase of the reaction an arbitrary point is taken where each of the diluted DNA samples cross. This is termed the **crossing threshold** or **Ct value**. From the various Ct values a log

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
<i>Target amplification methods</i>		
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
<i>Signal amplification methods</i>		
Branched DNA amplification (b-DNA)	Isothermal microwell format using hybridisation or target/capture probe and signal amplification	Mutation detection
<i>Note: HIV, human immunodeficiency virus.</i>		

graph is prepared from which an unknown concentration can be deduced. Since SYBR green and similar DNA-binding dyes are non-specific, in order to determine if a correctly sized PCR product is present most qPCR cyclers have a built-in melting curve function. This gradually increases the temperature of each tube until the double-stranded PCR product denatures or melts and allows a precise although not definitive determination of the product. Confirmation of the product is usually obtained by DNA sequencing.

5.10.8 The TaqMan system

In order to make qPCR specific a number of strategies may be employed that rely on specific hybridisation probes. One ingenious method is called the **TaqMan** assay or 5' nuclease assay. Here the probe consists of an oligonucleotide labelled with a fluorescent reporter at one end of the molecule and quencher at the other end.

The PCR proceeds as normal and the oligonucleotide probe binds to the target sequence in the annealing step. As the *Taq* polymerase extends from the primer its 5' exonuclease activity degrades the hybridisation probe and releases the reporter from the quencher. A signal is thus generated which increases in direct proportion to the number of starting molecules and fluorescence can be detected in real time as the PCR proceeds (Fig. 5.36). Although relatively expensive in comparison to other methods for determining expression levels it is simple, rapid and reliable and now in use in many research and clinical areas. Further developments in probe-based PCR systems have also been used and include scorpion probe systems, amplifluor and real-time LUX probes.

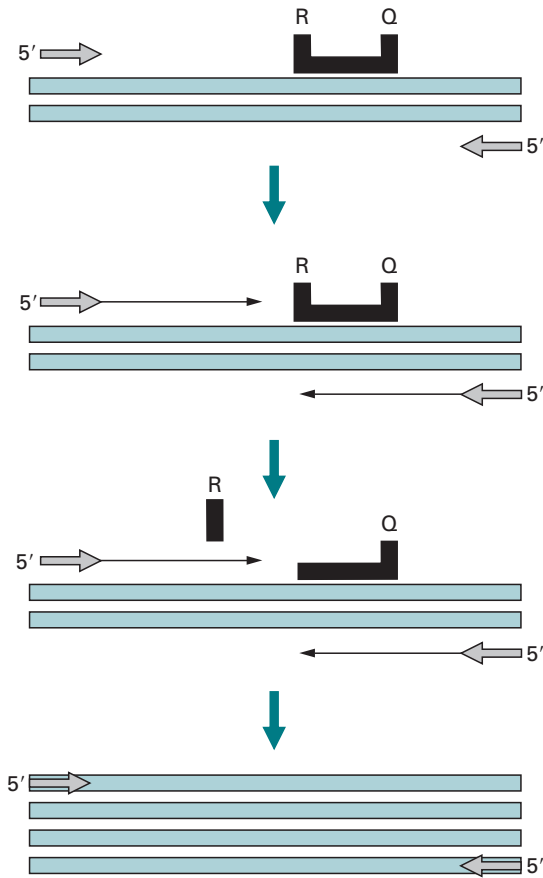


Fig. 5.36 5' Nuclease assay (TaqMan assay). PCR is undertaken with RQ probe (reporter/quencher dye). As R–Q are in close proximity, fluorescence is quenched. During extension by *Taq* polymerase the probe is cleaved as a result of *Taq* having 5' nuclease activity. This cleaves R–Q probe and the reporter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

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. The precise usage of codons, information regarding mutations and polymorphisms and the identification of gene regulatory control sequences are also only possible 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 and 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 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).

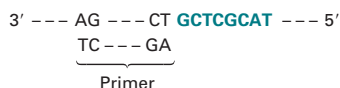
The Sanger method is simple and elegant and mimics in many ways the natural ability of DNA polymerase to extend a growing nucleotide chain based on an existing template. Initially the DNA to be sequenced is allowed to hybridise with an oligonucleotide primer, which is complementary to a sequence adjacent to the 3' side of 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. 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 radioactively labelled with ^{32}P or ^{35}S , and so the newly synthesised strand will be labelled.

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'-dideoxyadenosine triphosphate (ddATP)) which is similar to A but has no 3' hydroxyl group and so will terminate the growing chain 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 (Fig. 5.37).

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 at the same type of base. Thus four sets of DNA sequence are generated, each terminating at 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 which terminate at adenine, and those in the ddCTP terminate

Fragment to be sequenced, cloned in M13 phage



↓ DNA polymerase
4 dNTPs (radioactive)
ddGTP

Synthesis of complementary second strands:



Denature to give single strands

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

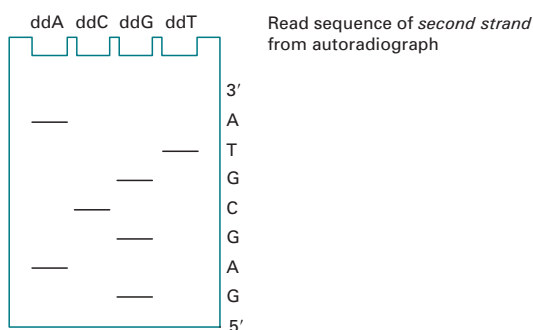


Fig. 5.37 Sanger sequencing of DNA.

at cytosine, etc., it is possible to read the sequence of the newly synthesised strand from the autoradiogram, 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.

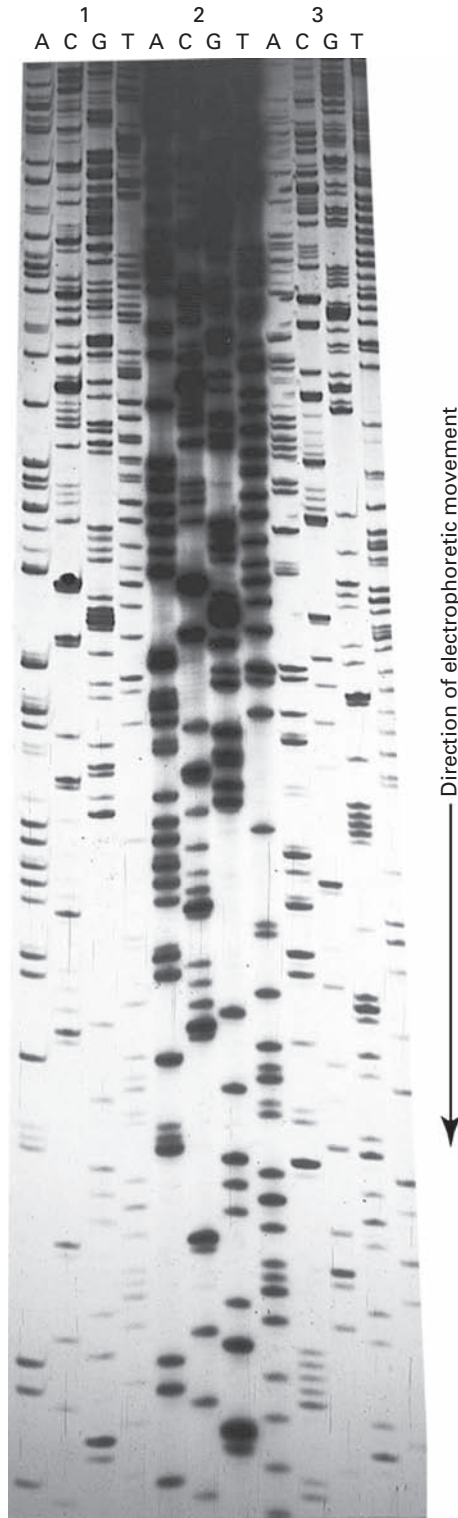


Fig. 5.38 Autoradiograph of a DNA sequencing gel. Samples were prepared using the Sanger dideoxy method of DNA sequencing. Each set of four samples was loaded into adjacent tracks, indicated by A, C, G and T, depending on the identity of the dideoxynucleotide used for that sample. Two sets of samples 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.

It is also possible to undertake nucleotide sequencing from double-stranded molecules such as plasmid cloning vectors and PCR amplicons directly. 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, 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 fluorescent-labelled dideoxynucleotides are then introduced in 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 have 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 its 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 (CCD) and the sequence is 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

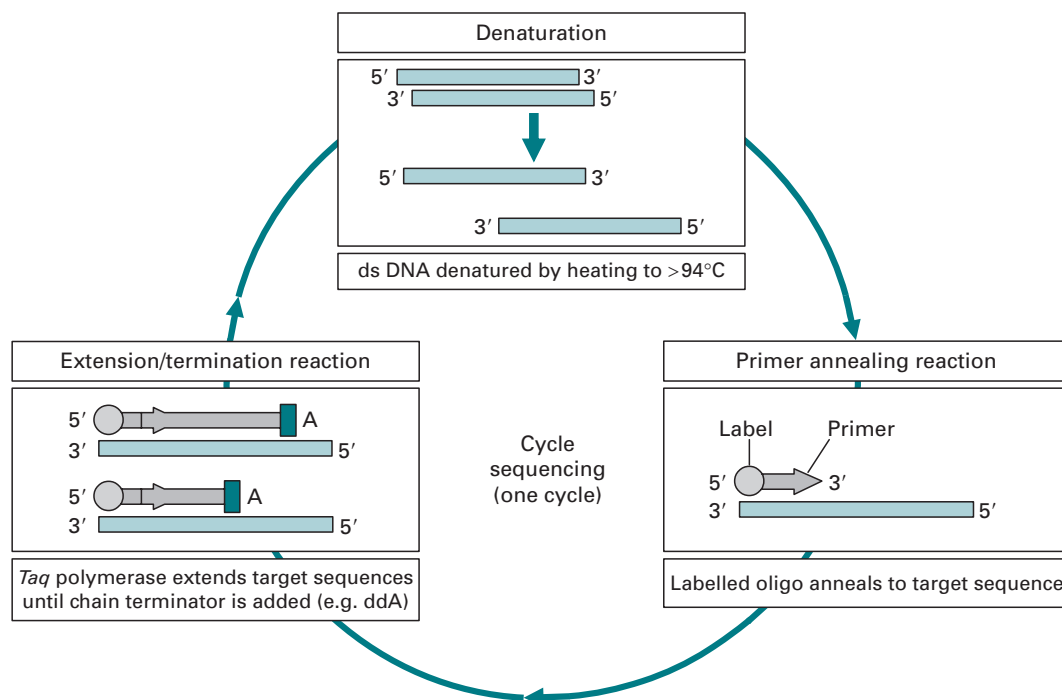


Fig. 5.39 Simplified scheme of cycle sequencing. Linear amplification takes place with the use of labelled primers. During the extension and termination reaction, the chain terminator dideoxynucleotides are incorporated into the growing chain. This takes place in four separate reactions (A, C, G and 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.

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.

5.11.6 Alternative DNA sequencing methods

Developments in the technology of DNA sequencing have made whole-genome sequencing projects a realistic proposition within achievable timescales; indeed the first diploid genome sequence to be completed was of Craig Venter who pioneered high-throughput sequencing. This makes studies on genome variation and evolution viable, as evidenced by the 1000 Genomes Project which is providing high-resolution sequence analysis of genomes. This has been made possible not only by refinements in traditional automated sequencing but also by new developments such as sequencing by synthesis and the development of sequencing by **hybridisation arrays**. These methods are changing the way genome analysis is undertaken and makes individual

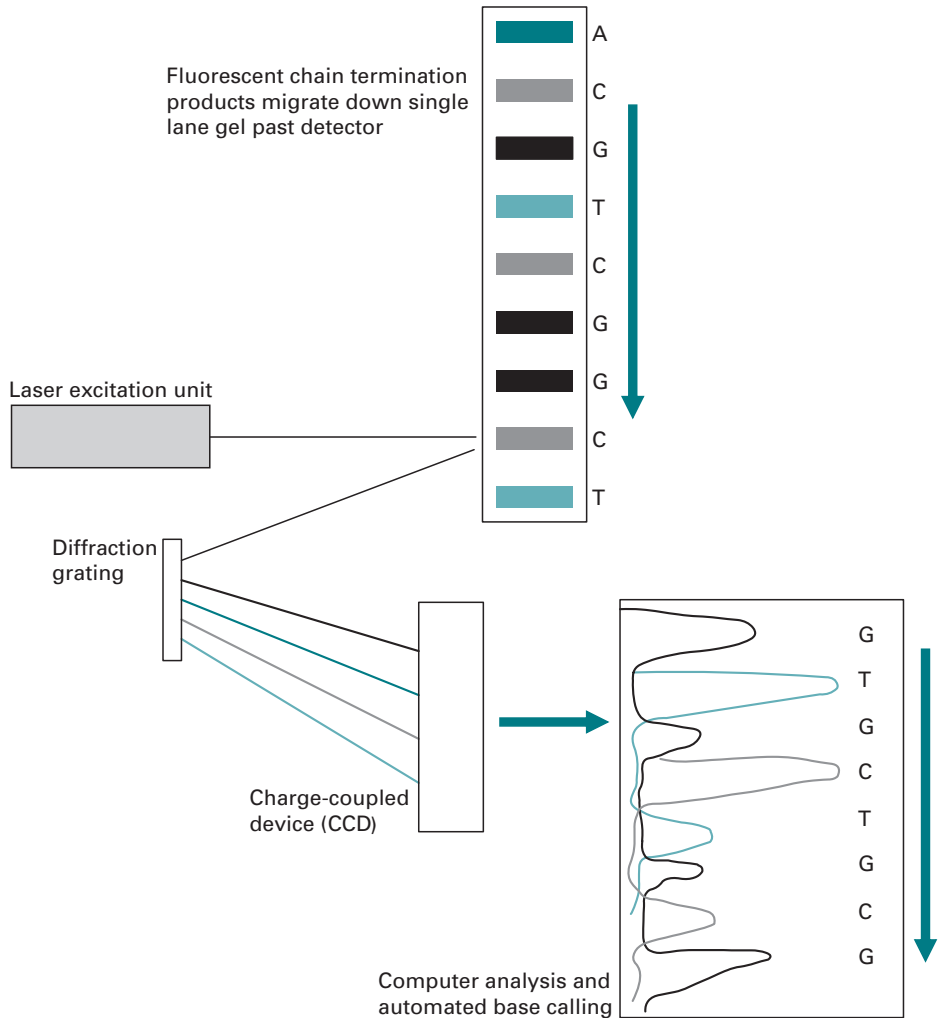


Fig. 5.40 Automated fluorescent sequencing detection using single-lane gel and charge-coupled device.

genome analysis a reality. Indeed more advanced methods using nanotechnology are in development and may provide an even more effective means of DNA sequencing.

5.11.7 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 Maxam and 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' ends of a double-stranded DNA sample (Fig. 5.41). The strands are then

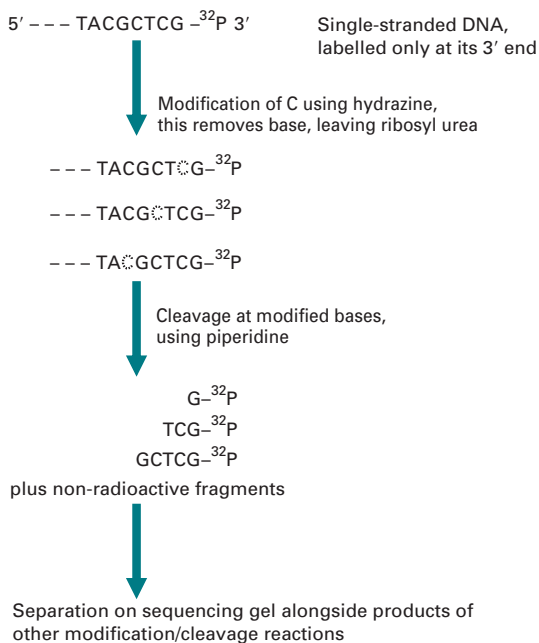


Fig. 5.41 Maxam and Gilbert sequencing of DNA. Only modification and cleavage of deoxycytidine is shown, but three more portions of the end-labelled DNA would be modified and cleaved at G, G+A, and T+C, and the products would be separated on the sequencing gel alongside those from the C reactions.

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 which 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. Following 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 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.

5.12 SUGGESTIONS FOR FURTHER READING

- Augen, J. (2005). *Bioinformatics in the Post-Genomic Era*. Reading, MA: Addison-Wesley.
- Brooker, R. J. (2005). *Genetics Analysis and Principles*, 2nd edn. New York: McGraw-Hill.
- Hartwell, L. et al. (2008). *Genetics: From Genes to Genomes*, 3rd edn. New York: McGraw-Hill.
- Lodish, H. et al. (2008). *Molecular Cell Biology*, 6th edn. San Francisco, CA: W. H. Freeman.
- Lewin, B. (2007). *Genes IX*. Sudbury, MA: Jones & Bartlett.
- Strachan, T. and Read, A. P. (2004). *Human Molecular Genetics*, 3rd edn. Oxford, UK: Bios.
- Walker, J. M. and Rapley, R. (2008). *Molecular Biomethods Handbook*, 2nd edn. Totowa, NJ: Humana Press.