CHAPTER 5

Nucleic Acids in Biotechnology

CONTENTS

5.1 DNA Sequence Determination 168
  5.1.1 Principles of DNA Sequencing 168
  5.1.2 Automated Fluorescent DNA Sequencing 169
  5.1.3 RNA Sequencing by Reverse Transcription 170

5.2 Gene Cloning 170
  5.2.1 Classical Cloning 170
  5.2.2 The Polymerase Chain Reaction 173

5.3 Enzymes Useful in Gene Manipulation 174
  5.3.1 Restriction Endonucleases 174
  5.3.2 Other Nucleases 175
  5.3.3 Polynucleotide Kinase 176
  5.3.4 Alkaline Phosphatase 176
  5.3.5 DNA Ligase 176

5.4 Gene Synthesis 177
  5.4.1 Classical Gene Synthesis 177
  5.4.2 Gene Synthesis by the Polymerase Chain Reaction 178

5.5 The Detection of Nucleic Acid Sequences by Hybridisation 178
  5.5.1 Parameters that Affect Nucleic Acid Hybridisation 179
  5.5.2 Southern and Northern Blot Analyses 180
  5.5.3 DNA Fingerprinting 181
  5.5.4 DNA Microarrays 184
  5.5.5 In Situ Analysis of RNA in Whole Organisms 188

5.6 Gene Mutagenesis 188
  5.6.1 Site-Specific In Vitro Mutagenesis 188
  5.6.2 Random Mutagenesis 191
  5.6.3 Gene Therapy 192

5.7 Oligonucleotides as Reagents and Therapeutics 193
  5.7.1 Antisense and Steric Block Oligonucleotides 193
  5.7.2 RNA Interference 197
  5.7.3 In Vitro Selection 198

5.8 DNA Footprinting 203

References 205
5.1 DNA SEQUENCE DETERMINATION

5.1.1 Principles of DNA Sequencing

There are two major ways of determining the sequence of a DNA molecule. These methods were developed in the laboratories of Sanger and of Gilbert for which each received a Nobel Prize in 1980. Both methods rely upon sequencing only one strand at a time.

5.1.1.1 Sanger DNA Sequencing. In the traditional method of Sanger DNA Sequencing, the DNA to be sequenced acts as a template and a new strand of DNA is synthesised enzymatically by use of either the Klenow fragment of DNA polymerase I, which lacks the 3'-5' exonuclease, or the DNA polymerase from bacteriophage T7 (Figure 5.1). The method depends on obtaining specific termination of the reaction at just one nucleotide base type to generate a mixture of shorter sequences.

To terminate the polymerisation at a specific point, a small amount of one of four 2',3'-dideoxynucleoside 5'-triphosphates is added. These can be incorporated into a growing DNA strand, but since they possess no 3'-hydroxyl group, they are unable to accept the addition of any extra nucleotides. They are thus chain terminators. The addition of a small amount of one of these, together with all four of the normal 2'-deoxyribonucleoside 5'-triphosphates to a polymerisation reaction gives rise to a series of oligonucleotides, each terminated by a dideoxynucleotide. Four reactions are carried out in parallel, each with a different dideoxynucleoside triphosphate (A, G, C and T). Separation of the oligonucleotide extension products from each individual reaction is achieved by polyacrylamide gel electrophoresis under denaturing conditions (Section 11.4.3) to generate a sequencing ladder. For visualisation by autoradiography, one of the unmodified deoxynucleoside triphosphates is radiolabelled with 32P (or with 35S via use of an α-thio triphosphate, Section 3.3.2). For example in Figure 5.1, there are two fragments generated in the ddATP reaction, two in the ddGTP reaction, one in the ddCTP reaction and three in the ddTTP reaction. The order of fragments up the gel represents the sequence of the extension product from 5' to 3'. The complement of this 'read' sequence is that of the template.

In practice, it is necessary to elongate a short primer that has already been annealed to the template, since DNA polymerases can only elongate existing hybrids. For this purpose, the DNA fragment to be sequenced is usually sub-cloned into a vector (Section 5.2) that has known sequences flanking the insertion site. Chemically synthesised oligonucleotides (typically 17–25 nucleotides in length) that correspond to one or the other side of the insert are annealed to the sub-clone of DNA and the dideoxy-sequencing reactions are carried out on these templates. The polymerisation reaction can proceed on double-stranded templates, with one strand being displaced by the elongated primer. More usually, single-stranded templates are used, such as the viral DNA from bacteriophage M13-derived recombinants. Two hundred to three hundred nucleotides can be sequenced routinely by this approach for each set of reactions.

Modern sequencing polymerases are derivatives of the thermostable polymerase from Thermus aquaticus (Taq). This allows sequence data to be obtained from a very few copies of DNA template by carrying out amplification cycle sequencing in a similar manner to PCR (Section 5.2.2).

5.1.1.2 Maxam and Gilbert Sequencing. This now rarely used method relies upon radioactive labelling of only one end of the DNA. The labelled DNA is then subjected to four separate, partial, base-selective, chemical modification (or for G + A, depurination) reactions (Table 5.1). These reactions allow

<table>
<thead>
<tr>
<th>3'-Cleavage adjacent to</th>
<th>Modification</th>
<th>Reagent</th>
<th>Strand breakage</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>Methylation</td>
<td>Dimethyl sulfate</td>
<td>1 M piperidine (at 90°C for 30 min)</td>
</tr>
<tr>
<td>G + A</td>
<td>Depurination</td>
<td>88% Formic acid</td>
<td>1 M piperidine (at 90°C for 30 min)</td>
</tr>
<tr>
<td>T + C</td>
<td>Base ring-opening</td>
<td>Hydrazine</td>
<td>1 M piperidine (at 90°C for 30 min)</td>
</tr>
<tr>
<td>C</td>
<td>Base ring-opening</td>
<td>Hydrazine, high salt</td>
<td>1 M piperidine (at 90°C for 30 min)</td>
</tr>
</tbody>
</table>
Figure 5.1  The principles of classical DNA sequencing. A primer oligodeoxynucleotide is annealed to the DNA template to be sequenced (top) and four separate extension reactions are carried out in the presence of DNA polymerase I, the four deoxynucleoside triphosphates (one usually $^{32}$P-labelled and a single 2'-3'-dideoxynucleoside triphosphate) to produce a series of truncated extension products (middle). The products of the reactions are separated by denaturing polyacrylamide gel electrophoresis and the gel autoradiographed to obtain a DNA sequence ladder (bottom).

the DNA to become sensitive at these sites to cleavage by alkaline hydrolysis. The fragments created are then separated by polyacrylamide gel electrophoresis in very much the same way as for Sanger sequencing.

5.1.2 Automated Fluorescent DNA Sequencing

Machines have now been developed to separate and identify the products of dideoxy-sequencing reactions. Here a fluorescent label is built into a set of four alternative dideoxynucleotide chain terminators
each with a different dye attached (Section 8.5.4, Figure 8.11). The four sequencing reactions are carried out together in one reaction and subjected to capillary electrophoresis in free solution. As each length of fluorescently labelled oligonucleotide emerges from the capillary column, it is detected by a fluorescence detector, the particular colour in each case corresponding to one of the four dideoxynucleotides. A computer analyses the data and produces a series of fluorescent signals that correspond to the read sequence. Such machines are capable of generating sequence reads of 500–1000 residues.

Sequencing machines have proved to be essential in large-scale DNA sequencing of genomes. For example, the DNA sequences of the yeast *Saccharomyces cerevisiae*, the fruitfly *Drosophila melanogaster* and *Homo sapiens* have been determined in this way. In genome sequencing, it is usual for the sequence to be determined three to ten times from different clones or fragments, including from both strands of the DNA, to obtain higher accuracy. Powerful computer programmes are then able to determine overlaps between fragments, and align the sequences against genome maps (Section 6.5.2). The positions of genes, introns and alternative splicing patterns can be predicted and genomes compared between different organisms to obtain knowledge of the RNA transcripts as well encoded proteins.

### 5.1.3 RNA Sequencing by Reverse Transcription

It is possible to carry out base-specific chemical treatments or digestions by nuclease enzymes to determine an RNA sequence directly. However, a more common procedure is to use a reverse transcriptase enzyme to make a cDNA copy of the single-stranded RNA. The reaction is initiated by use of an oligodeoxyribonucleotide primer from a known part of the sequence. This cDNA can then be amplified by the polymerase chain reaction (PCR) (Section 5.2.2) and sequence determined by standard DNA sequencing. The method only gives information regarding the base sequence and not regarding RNA modifications.

### 5.2 GENE CLONING

Cloning is the technique of growing large quantities of genetically identical cells or organisms that are derived from a single ancestor (clones). Gene cloning is an extension of this whereby a particular gene, group of genes or a fragment of DNA is selected from a mixed population (often a complete genome) and amplified to a huge extent. This can be carried out by insertion of the chosen DNA into a vector DNA and introduction of the hybrid (recombinant DNA) into cells by transformation (transfection). The cells containing the recombinant DNA are propagated and each cell in a colony contains an exact copy (or copies) of the gene ‘cloned’ in the vector. Cloning and recombinant DNA technology are well documented in established textbooks and manuals.6,7

Nowadays, a separate and complementary approach that uses PCR can achieve the same objective in a fraction of the time, of amplification of DNA segments without involving living cells.

### 5.2.1 Classical Cloning

#### 5.2.1.1 Vectors

Several classes of vector exist into which a foreign DNA can be inserted and amplified. The major classes are plasmids, bacteriophage and cosmids and bacterial or yeast artificial chromosomes (YACs). Prokaryotic plasmids are almost always circular double-stranded DNAs that contain antibiotic resistance genes as markers and a variety of restriction sites that can be used for insertion of the foreign DNA. A large number of plasmids are available for use in *E. coli*. The most useful bacteriophage is λ, which has been engineered in many ways to accept inserts of many different sizes and types, up to approximately 20,000 base pairs. Because the transformation frequency is high, screening is easy. Cosmids are large plasmids that contain the packaging site for bacteriophage λ DNA. Therefore they can either be packaged into phage particles or they can be replicated as plasmids. Since the amount of DNA that can be packaged in a λ phage particle is 50,000 base pairs, the potential size of cosmid inserts is very large. Cosmids have been used in chromosome walking (see later this section), but they are somewhat more difficult to manipulate than is λ. Artificial chromosomes are vectors containing the constituents of natural yeast chromosomes,
namely a prokaryotic origin of DNA replication in the case of bacterial artificial chromosomes (BACs) or a centromere (the region required for correct segregation of daughter chromosomes during mitosis and meiosis) and two telomeres (the ends of the chromosome) in the case of YACs. BACs accept inserts of up to about 150,000 bp and YACs can accept inserts of many hundreds or thousands of base pairs. Both types replicate inside their host cells in exactly the same way as a natural chromosome. This carries the disadvantage that only one YAC molecule can exist per yeast cell. BACs are preferred nowadays over YACs because YACs have problems with instability of the inserts and an unacceptably high occurrence of chimeric inserts derived from more than one genomic region.

The choice of vector is dependent on the ease of screening, the transformation efficiency, the insert size and the ease of isolating the DNA after cloning. If the average insert size is large, then less recombinants are needed to obtain a representative recombinant library. Vector DNA use allows not only an original cloning to be carried out, but also sub-cloning into more amenable fragments later in a project.

5.2.1.2 DNA Inserts. Often the source of the nucleic acid is a cDNA copy of messenger RNA (mRNA), which has been generated using the enzyme reverse transcriptase. Sometimes the DNA of interest can be synthesised chemically (Section 4.1). More often now it is the product of PCR (Section 5.2.2). Often, the DNA to be cloned is inserted as a duplex into a restriction site of the vector DNA, after treatment of the vector with that restriction enzyme, by use of the enzyme DNA ligase (Section 5.3.5, Figure 5.2). Usually the insert DNA is a restriction fragment with termini compatible with the vector ends. Sometimes oligonucleotide linkers need to be joined to the insert. These are self-complementary synthetic duplex oligonucleotides that specify a recognition site for a restriction enzyme. Linkers are ligated to the fragment to be cloned and then treated with the restriction enzyme, thus generating new termini, which are now identical or compatible for joining to the cleaved vector. In all cases, the joined DNA is then transfected into a host cell line.

5.2.1.3 Identification of Clones. The rate-limiting step in classical cloning is often the identification of the correct clone from a huge excess of other molecules. Typically, a vast number of visually identical bacterial colonies or bacteriophage plaques are generated on the surface of agar in Petri dishes. To identify the very few clones of interest, several approaches can be used. Often, a copy of the entire set of recombinants is made by touching a nitrocellulose or nylon filter membrane to the agar surface. The ‘master’ copy agar plate is stored away until the location of the required clone has been established on the filter copy.

Figure 5.2 Basic cloning procedure
In some cases, nucleic acid hybridisation (annealing of complementary strands, Section 5.5) is useful to find the desired clone. This is particularly true if a related sequence, such as that from another species, has already been cloned. Alternatively, one can deduce the DNA sequence from the corresponding protein sequence (if available) and chemically synthesize oligonucleotide probes complementary to part of the target DNA for screening of clones.

A complication here is that most amino acids are encoded by more than one nucleotide triplet. The result is that many different oligonucleotides need to be made to be sure of using the correct one. The number can be reduced by choosing a region of protein sequence containing the less ambiguous amino acids such as methionine and tryptophan which are specified by a single codon (TGG and ATG respectively). It is also possible to synthesize a mixture of oligonucleotides with two, three or four bases at the points where ambiguity is present, since the first two bases are often invariant for a particular amino acid (Figure 5.3; see also Figure 7.25). Lastly, several different regions of a protein can be used to derive a battery of probes all of which can be used to screen the library. In this way artefacts can be discounted. The probe is labelled (either by radioactivity or by use of a non-radioactive reporter molecule) and a solution of the probe is incubated with the DNA clones on the filter. After careful washing of the filter, only that probe which is exactly complementary to the desired sequence is left attached to the filter and positive clones can be identified by autoradiography or by visualisation of the reporter molecule. The hybridisation conditions used in such experiments are often crucial to a successful outcome.

In other cases, antibody screening of the filter copies can be used to detect the required clone. Of course this can only succeed if an antibody to the polypeptide product of the required gene is available and the vector into which the gene has been cloned contains appropriate transcriptional and translational regulatory sequences for the expression of the cloned gene as protein.

Another way of screening libraries is by the use of PCR. This is particularly powerful when screening artificial chromosome libraries of complete genomes. First, the individual clones in the library are arrayed into individual tubes (typically wells in a 24 × 16 multi-well plate). Next, a set of pooled subsets of the library is prepared. For example, a complete library of 10 multi-well plates might be split into ten pools, each comprising the complete contents of a single plate. PCR reactions on these ten samples would narrow down the target clone to a single multi-well plate. Further pools can be prepared, such as a particular row number for every plate (24 pools) or a particular column number (16 pools). In such a circumstance, these three sets of PCRs, totalling 10 + 24 + 16 = 50 reactions, would identify a single well in the library containing the clone producing the PCR product.

There are a number of shortcuts to molecular cloning which are sometimes useful.

5.2.1.4 Transposon Tagging. A previously cloned transposon (Section 6.6.5) is used to create mutations in the required gene (Figure 5.4). The transposon can then itself be used as a molecular ‘tag’ to isolate the gene by hybridisation (the transposon and its surrounding DNA must both be isolated by this method). Note that the detection method is based entirely upon the mutant phenotype and therefore no knowledge of the structure or biochemical function of the gene or gene product is needed.

5.2.1.5 Microdissection. It is possible physically to dissect and clone the required part of the chromosome (provided the chromosomal location of the gene of interest is known). Chromosomes may be separated from one another by pulsed-field gel electrophoresis or by fluorescence-activated sorting.

![Figure 5.3](Example of a mixed sequence oligonucleotide incorporating each alternative base which is used in gene cloning to probe for the gene encoding this peptide)
5.2.1.6 Chromosome Walking. If an overlapping series of clones can be isolated, it is possible to use one clone to isolate the next in line and thus ‘walk’ along the DNA to the required sequence. This is a very time-consuming process, but nevertheless has been used frequently.

5.2.1.7 Chromosome Jumping. This is an extension of chromosome walking that proceeds by larger steps and ignores the large DNA stretches in the middle of each step (hence the word ‘jumping’).

5.2.2 The Polymerase Chain Reaction

A DNA or gene of interest is best isolated now directly from the total DNA of the organism in question by use of PCR. Since the complete DNA sequences of many organisms (human, mouse, etc.) are now known, it is easy to design and synthesise chemically a pair of oligonucleotide primers of 20–30 residues flanking the region to be amplified, with each complementary to a different DNA strand (Figure 5.5). If it is required to clone the DNA into a vector, the oligonucleotide primers each can contain a restriction site at the 5'-ends, whereas the 3'-ends are complementary to the ends of the sequence to be cloned.

The target duplex DNA is denatured by heat and annealed to the primers, which are in vast excess to prevent the target DNA strands renaturing with each other. The two primers are next elongated on the separated target DNA template strands by use of a thermostable DNA polymerase, for example, Taq DNA polymerase from the thermophilic bacterium Thermus aquaticus, to give a twofold amplification. Since the primers are derived from different DNA strands, each newly synthesised strand now contains a binding site for the primer used for copying of the other strand. A second round of denaturation by heat, annealing and primer extension results in a fourfold amplification. After 20 rounds of amplification, \(2^{20}\) copies of the original target DNA are formed. Such a powerful technique can produce as much DNA as can be made by classical cloning methods. For cloning, the DNA is either treated with both restriction enzymes to create sticky ends, and the product joined to a similarly treated vector with T4 DNA ligase, just as in classical cloning (Section 5.2.1), or cloned without restriction digestion. In the latter case, the PCR products are often cloned into a special cleaved vector with single 3'-T overhanging bases at the cleavage sites, because PCR products usually contain a single A base overhang at both 3'-ends, which is added in a template-independent manner. But of course, it is not essential to clone the DNA, since the PCR reaction can be repeated again if further amounts of DNA are required.
The PCR method has also found widespread application in forensics for the detection and amplification of very tiny amounts of DNA, for example at crime scenes, and in DNA mutation detection.

5.3 Enzymes Useful in Gene Manipulation

Gene cloning would not have become possible without the discovery and isolation of a range of enzymes that act on DNA to enable manipulation of particular sections. One important class of enzymes is the DNA polymerases (Section 3.6.1). Other enzymes are involved in cutting the DNA, adding or removing a terminal phosphate or joining DNA fragments.

5.3.1 Restriction Endonucleases

Bacteria require a system to prevent foreign DNA from being replicated. This is provided by restriction endonucleases, which recognise and bind to DNA sequences at specific sites and make a double-stranded cleavage (Section 10.5.1). There are three types of restriction and modification systems, termed I, II and III in order of their discovery. The names of the enzymes are usually based on the names of the bacteria from which they are isolated. More than 3600 enzymes have been identified to date (http://rebase.neb.com/rebase/rebase.html).
Type I consists of a large enzyme complex containing subunits encoding endonuclease, methylase and several other activities. The recognition sequence comprises a trinucleotide and a tetranucleotide separated by about six non-specific base pairs (Table 5.2), but the endonucleolytic cleavage site can be up to 7000 base pairs distant.

Type II systems have independent endonucleases and methylases that act on the same DNA sequence. These sequences are generally palindromic (i.e. they have a twofold axis of symmetry) and the cleavage sites are usually within or very close to the recognition sites (Section 10.5.1). In some cases, the symmetrical recognition sequence is interrupted (e.g. BglI), while a few enzymes recognise an asymmetric sequence and cleave at a defined distance (e.g. MnlI). Restriction enzymes cleave both strands of the DNA either symmetrically to give blunt ends or asymmetrically to give sticky ends. A vast range of enzymes with different specificities has been isolated from a wide variety of organisms and type II restriction enzymes are highly useful tools in recombinant DNA research, and the products of cleavage (restriction fragments) can often be rejoined using DNA ligase (Section 5.3.5).

The type III system shares features in common with both type I and type II. There are two independent polypeptides, one of which acts independently as a methylase, but both are required for specific endonucleolytic activity. In the case of EcoPI, for example, the recognition sequence is an asymmetric pentanucleotide and the cleavage site is 25 bp downstream (Table 5.2).

### 5.3.2 Other Nucleases

Almost every organism contains a wide variety of nucleases, of which some are involved in the salvage of nucleotides and some feature as intrinsic activities of proteins used in replication and repair processes. Apart from non-specific nucleases, such as DNase I, and ribonucleases, there are several other nucleases that are used in the manipulation of DNA and RNA (Table 5.3).

---

Table 5.2 Some restriction endonucleases and their recognition sequences. N signifies any nucleotide. Cleavage sites indicated

<table>
<thead>
<tr>
<th>Type</th>
<th>Enzyme</th>
<th>Recognition site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>EcoK</td>
<td>A A C (N)_6 G T A C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T T G (N)_6 C A C G</td>
</tr>
<tr>
<td></td>
<td>EcoB</td>
<td>T G A (N)_8 T G C T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A C T (N)_8 A C G A</td>
</tr>
<tr>
<td>Type II</td>
<td>EcoRI</td>
<td>G T A A T T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T T A A G</td>
</tr>
<tr>
<td></td>
<td>SmaI</td>
<td>C C C G G G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G G A C C C</td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>C T G C A G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G A C G T C</td>
</tr>
<tr>
<td></td>
<td>Sau3AI</td>
<td>G A T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T A G</td>
</tr>
<tr>
<td></td>
<td>NotI</td>
<td>G C G G G C C G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G C C G G A C G</td>
</tr>
<tr>
<td></td>
<td>BglII</td>
<td>G C C N N N N N G G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G G N A N N N N C C G</td>
</tr>
<tr>
<td></td>
<td>MnlII</td>
<td>C C T C (N)_7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G A G (N)_7</td>
</tr>
<tr>
<td>Type III</td>
<td>EcoPI</td>
<td>A G A C C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T C T G G</td>
</tr>
</tbody>
</table>

---

Nucleic Acids in Biotechnology

[Table 5.2](#) Some restriction endonucleases and their recognition sequences. N signifies any nucleotide. Cleavage sites indicated

<table>
<thead>
<tr>
<th>Type</th>
<th>Enzyme</th>
<th>Recognition site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>EcoK</td>
<td>A A C (N)_6 G T A C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T T G (N)_6 C A C G</td>
</tr>
<tr>
<td></td>
<td>EcoB</td>
<td>T G A (N)_8 T G C T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A C T (N)_8 A C G A</td>
</tr>
<tr>
<td>Type II</td>
<td>EcoRI</td>
<td>G T A A T T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T T A A G</td>
</tr>
<tr>
<td></td>
<td>SmaI</td>
<td>C C C G G G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G G A C C C</td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>C T G C A G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G A C G T C</td>
</tr>
<tr>
<td></td>
<td>Sau3AI</td>
<td>G A T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T A G</td>
</tr>
<tr>
<td></td>
<td>NotI</td>
<td>G C G G G C C G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G C C G G A C G</td>
</tr>
<tr>
<td></td>
<td>BglII</td>
<td>G C C N N N N N G G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G G N A N N N N C C G</td>
</tr>
<tr>
<td></td>
<td>MnlII</td>
<td>C C T C (N)_7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G A G (N)_7</td>
</tr>
<tr>
<td>Type III</td>
<td>EcoPI</td>
<td>A G A C C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T C T G G</td>
</tr>
</tbody>
</table>

---

Nucleic Acids in Biotechnology

[Table 5.2](#) Some restriction endonucleases and their recognition sequences. N signifies any nucleotide. Cleavage sites indicated

<table>
<thead>
<tr>
<th>Type</th>
<th>Enzyme</th>
<th>Recognition site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>EcoK</td>
<td>A A C (N)_6 G T A C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T T G (N)_6 C A C G</td>
</tr>
<tr>
<td></td>
<td>EcoB</td>
<td>T G A (N)_8 T G C T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A C T (N)_8 A C G A</td>
</tr>
<tr>
<td>Type II</td>
<td>EcoRI</td>
<td>G T A A T T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T T A A G</td>
</tr>
<tr>
<td></td>
<td>SmaI</td>
<td>C C C G G G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G G A C C C</td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>C T G C A G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G A C G T C</td>
</tr>
<tr>
<td></td>
<td>Sau3AI</td>
<td>G A T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T A G</td>
</tr>
<tr>
<td></td>
<td>NotI</td>
<td>G C G G G C C G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G C C G G A C G</td>
</tr>
<tr>
<td></td>
<td>BglII</td>
<td>G C C N N N N N G G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G G N A N N N N C C G</td>
</tr>
<tr>
<td></td>
<td>MnlII</td>
<td>C C T C (N)_7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G A G (N)_7</td>
</tr>
<tr>
<td>Type III</td>
<td>EcoPI</td>
<td>A G A C C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T C T G G</td>
</tr>
</tbody>
</table>

---

Nucleic Acids in Biotechnology

[Table 5.2](#) Some restriction endonucleases and their recognition sequences. N signifies any nucleotide. Cleavage sites indicated

<table>
<thead>
<tr>
<th>Type</th>
<th>Enzyme</th>
<th>Recognition site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>EcoK</td>
<td>A A C (N)_6 G T A C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T T G (N)_6 C A C G</td>
</tr>
<tr>
<td></td>
<td>EcoB</td>
<td>T G A (N)_8 T G C T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A C T (N)_8 A C G A</td>
</tr>
<tr>
<td>Type II</td>
<td>EcoRI</td>
<td>G T A A T T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T T A A G</td>
</tr>
<tr>
<td></td>
<td>SmaI</td>
<td>C C C G G G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G G A C C C</td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>C T G C A G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G A C G T C</td>
</tr>
<tr>
<td></td>
<td>Sau3AI</td>
<td>G A T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T A G</td>
</tr>
<tr>
<td></td>
<td>NotI</td>
<td>G C G G G C C G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G C C G G A C G</td>
</tr>
<tr>
<td></td>
<td>BglII</td>
<td>G C C N N N N N G G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G G N A N N N N C C G</td>
</tr>
<tr>
<td></td>
<td>MnlII</td>
<td>C C T C (N)_7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G A G (N)_7</td>
</tr>
<tr>
<td>Type III</td>
<td>EcoPI</td>
<td>A G A C C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T C T G G</td>
</tr>
</tbody>
</table>

---

Nucleic Acids in Biotechnology

[Table 5.2](#) Some restriction endonucleases and their recognition sequences. N signifies any nucleotide. Cleavage sites indicated

<table>
<thead>
<tr>
<th>Type</th>
<th>Enzyme</th>
<th>Recognition site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>EcoK</td>
<td>A A C (N)_6 G T A C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T T G (N)_6 C A C G</td>
</tr>
<tr>
<td></td>
<td>EcoB</td>
<td>T G A (N)_8 T G C T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A C T (N)_8 A C G A</td>
</tr>
<tr>
<td>Type II</td>
<td>EcoRI</td>
<td>G T A A T T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T T A A G</td>
</tr>
<tr>
<td></td>
<td>SmaI</td>
<td>C C C G G G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G G A C C C</td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>C T G C A G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G A C G T C</td>
</tr>
<tr>
<td></td>
<td>Sau3AI</td>
<td>G A T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T A G</td>
</tr>
<tr>
<td></td>
<td>NotI</td>
<td>G C G G G C C G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G C C G G A C G</td>
</tr>
<tr>
<td></td>
<td>BglII</td>
<td>G C C N N N N N G G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G G N A N N N N C C G</td>
</tr>
<tr>
<td></td>
<td>MnlII</td>
<td>C C T C (N)_7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G A G (N)_7</td>
</tr>
<tr>
<td>Type III</td>
<td>EcoPI</td>
<td>A G A C C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T C T G G</td>
</tr>
</tbody>
</table>

---

Nucleic Acids in Biotechnology

[Table 5.2](#) Some restriction endonucleases and their recognition sequences. N signifies any nucleotide. Cleavage sites indicated

<table>
<thead>
<tr>
<th>Type</th>
<th>Enzyme</th>
<th>Recognition site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>EcoK</td>
<td>A A C (N)_6 G T A C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T T G (N)_6 C A C G</td>
</tr>
<tr>
<td></td>
<td>EcoB</td>
<td>T G A (N)_8 T G C T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A C T (N)_8 A C G A</td>
</tr>
<tr>
<td>Type II</td>
<td>EcoRI</td>
<td>G T A A T T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T T A A G</td>
</tr>
<tr>
<td></td>
<td>SmaI</td>
<td>C C C G G G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G G A C C C</td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>C T G C A G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G A C G T C</td>
</tr>
<tr>
<td></td>
<td>Sau3AI</td>
<td>G A T C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C T A G</td>
</tr>
<tr>
<td></td>
<td>NotI</td>
<td>G C G G G C C G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G C C G G A C G</td>
</tr>
<tr>
<td></td>
<td>BglII</td>
<td>G C C N N N N N G G C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C G G N A N N N N C C G</td>
</tr>
<tr>
<td></td>
<td>MnlII</td>
<td>C C T C (N)_7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G G A G (N)_7</td>
</tr>
<tr>
<td>Type III</td>
<td>EcoPI</td>
<td>A G A C C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T C T G G</td>
</tr>
</tbody>
</table>
5.3.3 Polynucleotide Kinase

A polynucleotide kinase isolated from bacteriophage T4 catalyses the transfer of the $\gamma$-phosphate of ATP to the 5′-hydroxyl terminus of DNA, RNA or an oligonucleotide in a reaction that requires magnesium ions. The enzyme is particularly useful for introducing a radioactive label on to the end of a polynucleotide, where the phosphate donor is $\gamma^{32}$P-ATP. Both single- and double-stranded polynucleotides can be phosphorylated, although recessed 5′-hydroxyl groups in double-stranded DNA, such as those obtained by cleavage with certain restriction enzymes, are poorly phosphorylated. This sort of polynucleotide kinase activity, though not found in bacteria, has been found in some mammalian cells. The T4 enzyme is the only well-characterised kinase that has polynucleotides as substrates. This T4 protein also has a 5′-phosphatase activity which is unusually specific for a 3′-phosphate of a nucleoside or polynucleotide.

5.3.4 Alkaline Phosphatase

Phosphatases catalyse the hydrolysis of phosphate monoesters to produce inorganic phosphate and the corresponding alcohol. Most phosphatases are non-specific. Alkaline phosphatases are found in bacteria, fungi and higher animals (but not plants) and will remove terminal phosphates from polynucleotides, carbohydrates and phospholipids. The *E. coli* enzyme is a dimer of molecular weight about 89 kDa, requires a zinc (II) ion, and is allosterically activated by magnesium ions. During dephosphorylation of the substrate, its phosphate is transferred to a serine residue on the enzyme located in the sequence Asp-Ser-Ala. This same sequence is found in mammalian alkaline phosphatases (the calf intestinal enzyme is particularly well characterised) and it is similar to the active centre of serine proteases. Acidic phosphatases are also common, but these do not usually operate on polynucleotides as substrates.

5.3.5 DNA Ligase

A ligase is an enzyme that catalyses the formation of a phosphodiester linkage between two polynucleotide chains. In the case of DNA ligases, a 5′-phosphate group is esterified by an adjacent 3′-hydroxyl group and there is concomitant hydrolysis of pyrophosphate in NAD$^+$ (bacterial enzymes) or ATP (phage and eukaryotic enzymes). Particularly efficient joining takes place when the phosphate and hydroxyl groups are held close together within a double helix, typically where the joining process seals a ‘nick’ and creates a perfect duplex (Figure 5.6). This situation occurs both in gene synthesis (Section 5.4) and in recombinant DNA technology (Section 5.2) in ligation of identical ‘sticky ends’ formed by cleavage with a restriction endonuclease. *E. coli* and phage T4 DNA ligases are well-characterised enzymes which have an important role in DNA replication (Section 6.6.4). T4 DNA ligase will join blunt DNA duplex ends when used at high concentrations and...
it will also catalyse the joining of two oligoribonucleotides in the presence of a complementary splint oligodeoxyribonucleotide.

5.4 GENE SYNTHESIS

5.4.1 Classical Gene Synthesis

The principles of gene assembly were developed 35 years ago by Khorana and his colleagues.10

5.4.1.1 5'-Phosphorylation. To join the 3'-end of one oligonucleotide to the 5'-end of another, a phosphate group must be attached to one of the ends. This is most easily accomplished at a 5'-end either chemically or enzymatically. The chemical procedure involves reaction of the 5'-hydroxyl group of a protected oligonucleotide, while still attached to a solid support with a special phosphoramidite derivative (e.g. DMTO(CH2)2SO2(CH2)2OP(NiPr)2OCH2CH2CN) (Section 4.4.2, Figure 4.19). The DMT group is removed by acidic treatment and during subsequent ammonia deprotection both the 2-cyanoethyl and hydroxyethylsulfonylethyl groups are removed to liberate the 5'-phosphate. Alternatively, and so as to introduce a 32P-radiolabel, phosphorylation is carried out enzymatically using T4 polynucleotide kinase (Section 5.3.3) to transfer the γ-phosphate of ATP to the 5'-end of an oligonucleotide.

5.4.1.2 Gene Assembly. Figure 5.7 shows schematically the construction of a gene coding for a small bovine protein caltrin (a protein believed to inhibit calcium transport into spermatozoa).11 Each synthetic oligonucleotide is denoted by the position of the arrows. These are arranged such that annealing (heating to 90°C and slow cooling to ambient temperature) of all ten oligonucleotides simultaneously gives rise to a contiguous section of double-stranded DNA, the sequence of which corresponds to the desired protein sequence. In this example, the oligonucleotides are 24–38 residues long, but chains of 80 residues or more have been used in gene synthesis.

Oligonucleotides C2–C9 are previously phosphorylated such that, for example, the 5'-phosphate group of C3 lies adjacent to the 3'-hydroxyl group of C1. The duplex is only held together by virtue of the complementary base pairing between strands. The enzyme T4 DNA ligase (Section 5.3.5) is now used to join the juxtaposed 5'-phosphate and 3'-hydroxyl groups (the caret marks denote the joins). The overlaps are such that each oligonucleotide acts as a splint for joining of two others.

Note that oligonucleotides C1 and C10 are not phosphorylated. Each end corresponds to a sequence that would be generated by cleavage by a restriction enzyme (Section 5.3.1). Lack of a phosphate group prevents these self-complementary ends from joining to themselves during ligation. The ends are later joined to a vector DNA, previously cleaved by the same two restriction enzymes, to give a closed circular duplex ready for transformation and cloning in E. coli (Section 5.2).
Other features of the synthetic gene are internal restriction enzyme sites (shaded), which can be introduced artificially merely by judicious choice of codons specifying the required amino acid sequence. This particular gene is designed without a methionine initiation codon, since the protein is intended for expression as a fusion with another vector-encoded protein. This fusion can be cleaved to generate caltrin by treatment with the proteolytic enzyme factor $X_a$ (an enzyme important in the blood-clotting cascade and whose natural substrate is prothrombin), since the synthetic gene has been designed to include a section encoding the tetrapeptide recognition sequence for this enzyme.

**5.4.2 Gene Synthesis by the Polymerase Chain Reaction**

There are numerous procedures for gene synthesis that involve use of PCR (Section 5.2). A particularly simple version known as recursive PCR has been used for the preparation of large genes such as that for human lysozyme. Oligonucleotides are synthesised 50–90 residues long but, unlike the classical approach, only their ends have complementarity (Figure 5.8). Overlaps of 17–20 bp are designed to have annealing temperatures calculated to be in the range 52–56°C. A computer search ensures that no two ends are similar in sequence. Recursive PCR is carried out in the presence of all oligonucleotides simultaneously with cycles of heating to 95°C, cooling to 56°C and primed DNA synthesis at 72°C using the four deoxynucleoside triphosphates and the thermostable Vent DNA polymerase derived from *Thermococcus litoralis*. In initial cycles (step 1), each 3'–5' end is extended using the opposite strand as a template to yield sections of duplex DNA. In further cycles (steps 2–5), one strand of a duplex is displaced by a primer oligonucleotide derived from one strand of a neighbouring duplex. Finally (step 6), a high concentration of the two terminal oligonucleotides drives efficient amplification of the complete duplex. Success is due to the useful characteristics of Vent DNA polymerase, which has both a strand displacement activity and an active 3'–5' proofreading activity that reduces the chances of incorrect nucleotide incorporation.

**5.5 THE DETECTION OF NUCLEIC ACID SEQUENCES BY HYBRIDISATION**

Molecular cloning is only the beginning of the study of a gene. Often it is important to study the same gene from a variety of different individuals. For example, much can be learned from structural analysis of a series of mutants in the gene. While it is possible to molecularly clone the gene from each mutant individual, it is often much easier simply to analyse the uncloned nucleic acid, for example by PCR amplification (Section 5.2.2) and DNA sequencing (Section 5.1). It is also important to be able to detect the RNA
encoded by a gene and to determine its levels of transcription and tissue specificity. Methods exist for both of these purposes and both depend upon the ability of a single-stranded nucleic acid to pair specifically with its complementary strand.

5.5.1 Parameters that Affect Nucleic Acid Hybridisation

Hybridisation is the annealing of two single strands of a nucleic acid to form a duplex. Duplex strength is measured by observation of the **melting temperature** \(T_m\) and is affected by several parameters.

5.5.1.1 Base Composition (%GC). Since G–C base pairs have three hydrogen bonds, they are stronger than A–T base pairs, which have only two. Thus duplexes with higher G–C content have higher melting temperatures.

5.5.1.2 Temperature (\(T\)). The rate of association of single-stranded DNA into a duplex varies markedly with temperature (Figure 5.9: see also Section 2.5.1). The shape of this curve is governed by two

---

**Figure 5.8** Gene synthesis by recursive PCR. Bars represent oligonucleotides and their extension products after PCR.
factors. At low temperatures, the re-association rate is determined by the difference in free energy between the unassociated and the transition state.

\[ k = Z e^{-E_a/RT} \]  

(5.1)

where \( k \) is the re-association rate constant, \( E_a \) the activation free energy, \( R \) the gas constant and \( T \) the absolute temperature. At higher temperatures, the stability of the duplex is markedly reduced until eventually it is unstable and the hybrid melts. Thus there is a fall off in re-association rate as this point is approached.

5.5.1.3 Monovalent Cation Concentration (\( M \)). The melting temperature of a hybrid (Section 2.5.1) is reduced at lower salt concentration because cations help to stabilise a duplex. Divalent cations such as magnesium are much more effective in stabilisation of hybrids, but are less frequently used in hybridisation studies (Section 9.3).

5.5.1.4 Duplex Length (\( L \)). The melting temperature of a duplex shorter than a few hundred base pairs is length dependent. In practice, these four factors can be combined into an empirical equation giving the melting temperature \( T_m \) of a hybrid DNA.

\[ T_m = 69.3 + 0.41(\%GC) + 18.5 \log_{10} M - 500 L^{-1}^\circ C \]  

(5.2)

Web-based algorithms are available now for calculating \( T_m \) from knowledge of the various parameters (e.g. see http://www.basic.northwestern.edu/biotools/OligoCalc.html). Use of hybridisation temperatures from 10 to 20°C below the calculated \( T_m \) of the hybrid is optimal in practice to ensure annealing of strands.

For synthetic oligonucleotide probes of 15–20 residues, the calculation of \( T_m \) is simplified to 2°C per dA·dT and 4°C per dG·dC base pair in 1 M sodium chloride solution. This is known as the Wallace rule. In the case of the quaternary ammonium salt tetramethylammonium chloride (3 M TMAC), the \( T_m \) of a duplex is independent of base composition and is thus directly proportional to its length. This is of practical value for example in cloning applications that involve hybridisation of mixed sequence oligonucleotides (Section 5.2.2).

5.5.2 Southern and Northern Blot Analyses

It is possible to use nucleic acid hybridisation to detect uncloned genomic DNA. Genomic DNA is immobilised on a nitrocellulose or nylon filter, in basically the same way as described for gene cloning (Section 5.2.1). The gene of interest is detected on the filter by hybridising a complementary nucleic acid strand labelled either with radioactivity or an affinity label such as biotin, which can be detected with great sensitivity. Of course, if the DNA is just spotted onto the filter, all that is seen is a spot whose intensity reflects the
concentration of the corresponding gene in the sample. This latter technique is called dot blotting and is very useful in this limited respect. However, if the DNA is fractionated before transfer, much more information is acquired. Southern blot analysis (named after its inventor Ed Southern) involves fractionation of DNA by gel electrophoresis, followed by transfer of the DNA out of the gel onto a filter (Figure 5.10). The filter is then probed for the gene of interest as before. In the commonest case, the DNA has been digested with restriction enzymes and the result is the detection of those restriction fragments that are homologous to the gene probe. In this way, restriction maps of genes can be derived from genomic DNA without resort to cloning.

This technology can be extended to the study of RNA (Northern blot analysis). RNA can be electrophoresed in gels and immobilised on filters, provided it is denatured by treatment with formaldehyde. It can then be detected in the same way as DNA. Unfortunately, RNA cannot be cut into large defined fragments with the same ease as DNA, so such an approach is more limited. It is particularly useful for the determination of the sizes of RNAs and their tissue specificities, the latter approach relying upon the isolation of RNAs from different tissues. Northern blot analysis is often used to determine the transcribed regions in a stretch of DNA. By this approach, a battery of different restriction fragments, which together span the DNA of interest, are separately used to probe a Northern blot. Those DNA fragments that are transcribed detect bands in the Northern blot. The complementary approach (by use of radioactive RNA to probe restriction-digested DNA) is only possible if the transcripts arising from the DNA are particularly abundant.

5.5.3 DNA Fingerprinting

The notion that human characteristics can be inherited is long established. However, the ABO blood-group system can still only be used to classify people into just four types (groups A, B, AB and O). Moreover, such serological and protein markers are all too readily degraded in aged forensic samples. Clearly, the solution to such limitations lies in the direct examination of the genetic material itself. Even before the DNA revolution, it was evident that the 3 billion base pairs that make up the human genome must contain a huge number of sites of heritable variation and ought to support truly positive biological identification. Moreover, DNA is surprisingly tough and bits can survive in typeable form for remarkably long periods.

Genetic fingerprinting was developed in 1984 by accident. It was at first an academic curiosity, but then moved speedily into real-life casework where it established that molecular genetics could really provide an entirely new dimension to biological identification. This technology has changed the lives of thousands of people involved in criminal investigations, paternity disputes, immigration challenges, identification of victims of mass disasters and the like. The analysis of human DNA has been of prime importance though there are tremendous applications in non-human DNA analysis, in particular the use of animal and plant DNA-typing and the field of ‘microbial forensics’, which has expanded as a response to the threat of bio-terrorism.

5.5.3.1 Super Markers. Alec Jeffreys started a search for hypervariable regions in human DNA in the 1980s. He found the answer in minisatellites. These are regions of DNA consisting approximately 30 base pairs repeated over and over again tens or hundreds of times, and with different alleles varying in the

Figure 5.10 Southern blot analysis
number of stutters. The problem was how to access them. Jeffreys observed that a chance-studied minisatellite tucked away inside a human gene looked rather familiar, not unlike the stutters in the few other minisatellites described in the literature. The implications were clear – a hybridisation probe consisting of this DNA sequence motif shared by different minisatellites should latch onto many different minisatellites simultaneously, giving unlimited access to these potentially extremely informative genetic markers. Minisatellites are simply detected by hybridisation of probes to Southern blots of restriction-enzyme-digested genomic DNA.

5.5.3.2 Stumbling upon DNA Fingerprinting. In September 1984, Jeffreys tested a range of samples that included DNA from a human father/mother/child trio. The results provided multiple, highly variable DNA fragments. While mother and father were obviously different, the child seemed to be a union of the DNA patterns of the parents. Improved technology was able to resolve large numbers of extremely variable DNA fragments containing these minisatellites (Figure 5.11), not just in humans but in other organisms as well. In humans, the banding patterns are individual-specific, with essentially zero chance of matching even between close relatives or members of an isolated inbred community. For any individual, the patterns are constant, irrespective of the source of DNA. The multiple markers that make up a DNA fingerprint are inherited in a simple Mendelian fashion, with each child receiving a random selection of about half of the father’s bands and half of the mother’s. Happily, the term ‘DNA fingerprint’ was chosen rather than the more accurate description ‘idosyncratic Southern blot minisatellite hybridisation profile’ (Section 5.5.2).

5.5.3.3 The Evolution of Forensic Genetics. The amount of variation currently accessible in DNA is extremely informative. Sequence variations between different minisatellite loci allows probes to detect many independent minisatellites simultaneously, yielding the hypervariable multi-band patterns known as DNA fingerprints. By use of only a single probe, the match probability is estimated to be $<3 \times 10^{-11}$, while two probes together give a value of $<5 \times 10^{-19}$. This is so low that the only individuals sharing DNA fingerprints are monozygotic twins. At the same time, a method known as differential lysis was developed that selectively enriches sperm concentration in vaginal fluid/semen mixtures, thereby avoiding the problem of the victim’s DNA (which is in great excess) masking that of a rapist.

5.5.3.4 Single-Locus Probes. Although use of DNA fingerprinting persisted for some years in paternity testing, criminal casework soon concentrated on the use of specific cloned minisatellites. Each of these ‘single-locus probes’ (SLPs) revealed only a single, highly polymorphic, restriction fragment length polymorphism, thereby simplifying interpretation. Typically, four SLPs were used successively to probe a Southern blot, yielding eight hyper-variable fragments per individual. SLPs were used in the first DNA-based criminal investigation in the UK in 1986.

5.5.3.5 Profiling DNA. DNA fingerprints are excellent for some applications, but not for forensic investigations that have to identify the origin of a biological sample with as much certainty as possible. This is because fingerprint patterns are complex and their interpretation is readily open to challenge in court, they are not easy to computerise, and they require significant amounts of good quality DNA, equivalent to that obtained from a drop of fresh blood. The solution to these problems was simple – the isolation and cloning of minisatellites. Each cloned minisatellite, used as a hybridisation probe, produces a much simpler pattern of just two bands per person, corresponding to the two alleles in an individual (Figure 5.11c). Such simple profiles can be obtained using considerably less DNA (one hair root is enough), and the estimated lengths of the DNA fragments easily support database construction. These DNA profiles have exposed the true variability of human minisatellites, some showing 100 or more different length alleles in human populations.

DNA profiles are not individual-specific no matter how variable the minisatellite is between unrelated people. This is particularly true for siblings, who have a one in four chance of sharing exactly the same profile. Nevertheless, by typing DNA sequentially, typically with a battery of five different minisatellites, excellent levels of individual specificity are obtained, leading to routine match frequencies of one in a billion with DNA profiling.
5.5.3.6 PCR-Based Methods. The discovery of short tandem repeats (STRs) together with the introduction of automated sequencing technology has led to the current powerful systems for the identification of individuals. Human forensic casework is now carried out using commercially developed autosomal STR multiplexes (single-tube PCR (Section 5.2.2) that amplify multiple loci) and is established worldwide because of its advantages of high discriminating power, sensitivity, ability to resolve simple mixtures, speed and automation. The resulting reduced cost has paved the way for the creation of national STR DNA databases (http://www.cstl.nist.gov/div831/strbase/). For example, the UK National DNA Database contained some 2.5 million reference profiles and about 200,000 crime-scene profiles as at July 2004 (http://www.forensic.gov.uk/forensic/news/press_releases/2003/NDNAD_Annual_Report_02-03.pdf). Automated sequencing equipment for multiplex analysis typically uses multi-channel capillary electrophoresis systems that detect fluorescently labelled PCR products. These are combined with robotics.
and laboratory information management systems, including bar coding of samples, to reduce operator errors. A typical electropherogram output is illustrated (Figure 5.12).

A ‘second-generation multiplex’ (SGM) has further included a PCR assay targeted at the XY-homologous amelogenin genes\(^{17}\) to reveal the sex of a sample donor. An additional four loci were added to the multiplex, now renamed ‘SGM Plus10’ (Figure 5.12), giving it a match probability of less than \(10^{-13}\).

Although some differences in practice between individual national jurisdictions remain, there has been rapid development and near-universal acceptance of this new DNA-based technology in forensic genetics.\(^{18}\)

### 5.5.4 DNA Microarrays

DNA microarrays are now one of the most widely used tools in functional genomics.\(^{19}\) They are providing biology with the equivalent to the chemist’s periodic table – a classified inventory of all the genes for a living organism. Oligonucleotide microarrays, also known as DNA chips, are miniature parallel analytical devices containing libraries of oligonucleotides robotically spotted (printed) or synthesised in situ on solid supports (glass, coated glass, silicon or plastic). The major DNA-chip technologies are distinguished by the sizes of the DNA fragments arrayed, by methods of arraying, by their chemistry and linkers for attaching DNA to the chip, and by hybridisation and detection methods.

Microarrays work by exploiting the ability of a given cDNA or mRNA test sample to hybridise to the DNA template from which it originated. By use of a two-dimensional (2-D) array containing very many DNA samples, the expression levels of hundreds or thousands of genes within a cell can be determined quickly by measuring the amount of cDNA or mRNA bound to individual sites on the array. The precise amount of mRNA bound to each locus gives a profile of gene expression in the cell. Alternatively, comparative binding
of a test and a standard probe provides an immediate signal of the presence or absence of a particular sequence. Ultimately, such studies promise to expand the size of existing gene families, reveal new patterns of coordinated gene expression across gene families, and uncover entirely new categories of genes.

5.5.4.1 Technical Foundations. Two technologies are central to the production and use of DNA microarrays. The first is the fabrication of tens to hundreds of thousands of polynucleotides at high spatial resolution in precise locations on a 2-D surface. The second involves the measurement of molecular hybridisation events on the array using laser fluorescence scanning. By use of one of three different methodologies, DNA is synthesised, spotted or printed onto the support, which is usually a glass microscope slide, but can also be a silicon chip or a nylon membrane. The DNA sequences in a microarray are attached to the support in a fixed way, so that the location of each spot in the 2-D grid identifies a particular sequence. The spots themselves are either oligodeoxynucleotides, DNA or cDNA.

5.5.4.2 Use of DNA Microarrays. The five steps for carrying out a microarray experiment are

- DNA chip preparation using the chosen target DNAs,
- making a hybridisation solution containing a mixture of fluorescently labelled cDNAs,
- incubating the hybridisation mixture of fluorescently labelled cDNAs with the DNA chip,
- detecting bound cDNA using laser technology and data storage in a computer, and
- data analysis using computational methods.

5.5.4.3 Microarray Preparation. The first chip technology came in 1984 from the work of Stephen Fodor in the California-based company, Affymetrix, and is based on photolithography. A synthetic linker with a photochemically removable protecting group is bonded to a flat glass substrate. Light is then directed through a photolithographic mask to specific areas on the surface to produce localised photodeprotection (Figure 5.13). The first of a series of DNA phosphoramidite monomers (Section 4.1.2), also having a 5′-(α-methyl-2-nitropiperonyloxy carbonyl), photochemically labile protecting group\(^{20}\) (Figure 5.14a) is

![Figure 5.13](https://pubs.rsc.org/en/content/articlelanding/2007/en/DP/9781847553803-00167/page/185)

**Figure 5.13** Light directed oligonucleotide synthesis. Derivatised solid support has hydroxyl groups protected with a photolabile group. Light is directed through a mask to effect selective deprotection. The first dT-phosphoramidite with 5′-photolabile protection is introduced. A new mask enables deprotection of a second set of spots on the array which are then linked to the second nucleotide, dA. Repetition of this procedure for next dC and finally dG completes the cycle for the first nucleotides in the oligomer array. The cycle is then repeated with new masks to install the second nucleotides in the array, and so on.
incubated with the surface and chemical coupling occurs only at those sites that have been irradiated in the preceding step. Light is next directed at further regions of the substrate by a new mask, and the reaction sequence is successively repeated for the second, third and fourth of the four monomers, A, C, G and T to complete the first cycle. A second complete cycle lays down the second nucleotide in the oligomer. Further repetitions of this cycle provide the full set of $4^N$ polydeoxyribonucleotides of length $N$, or any subset, in just $N$ complete cycles. Thus, for a given reference sequence, a DNA array can be designed that consists of a highly dense collection of DNA single-stranded oligomers, usually around 25 residues long. This photolithographic process enables construction of arrays with extremely high information content. Large-scale commercial methods permit approximately 300,000 oligodeoxynucleotides to be synthesised on small $1.28 \times 1.28$ cm arrays, while versions with $10^6$ probes per array are being developed.

In a separate development, Patrick Brown at Stanford University developed a cDNA spotting method that is suited to the display of single-DNA fragments, often greater than several hundred base pairs in length. C DNA samples (about 15 ng) are micro-spotted robotically onto a glass (or nylon membrane) surface that has been treated chemically to provide primary amino groups. Droplets ($\approx 1$ nL) are located $\approx 200 \mu$m apart and the DNA in the spots is covalently bonded to the surface by UV irradiation to link the surface amino groups to thymidine residues.

A third robotic methodology has been developed by Rosetta Inpharmatics that uses ink-jet printer technology to perform classical oligodeoxyribonucleotide synthesis based on the four-step dimethoxytrityl protecting group chemistry (Section 4.1.2). In a fourth approach, Agilent Labs in collaboration with Marvin Caruthers have developed a two-step microarray synthesis cycle to halve the number of steps required to
build oligodeoxyribonucleotides on glass surfaces. An entirely new carbonate-based protecting group chemistry enables deprotection and oxidation in a single step, reducing time and cost for microarray synthesis.23

5.5.4.4 Microarray Analysis. How does one analyse the information encoded in thousands of individual gene sequences on a small glass or silicon chip? The process is based on hybridisation probing, a technique that uses fluorescently labelled nucleic acid molecules as ‘mobile probes’ to identify complementary DNA sequences using base pair recognition. The DNA probes to be hybridised to the array are labelled by incorporating fluorescently tagged nucleotides (such as Cy3-dUTP, Figure 5.14b) during oligo-primed reverse transcription of mRNA. Alternatively, they can be chemically tagged by 5′-end labelling (Figure 5.14c). Different green and red fluorophores are used to label cDNAs from control (reference) and experimental (test) RNAs. The labelled cDNAs are then mixed together prior to hybridisation to the array so that relative amounts of a particular gene transcript in the two samples are determined by measuring the signal intensities detected for both green and red fluorophores. Because the arrays are constructed on a rigid surface (glass), they can be inverted and mounted in a temperature-controlled hybridisation chamber. When the fluorescent mobile probe, DNA, cDNA or mRNA, locates a complementary sequence on the chip, it will lock onto that immobilised target, and the probe is identified by fluorescence microscopy. The fluorescent tag on the probe is excited by a laser and the digital image of the array is captured. These data are then stored in a computer for analysis. Thus, for example, cDNA from a normal cell and a diseased cell can be separately labelled with green and red fluorescent markers to enable comparative analysis. The location and intensity of both colours shows whether the gene, or a mutant, is present in either the control and/or sample DNA (Figure 5.15). It can also provide an estimate of the expression level of the gene(s) in the sample and control DNA.

5.5.4.5 Types of Microarray. There are three basic types of samples used to construct DNA microarrays, two are genomic and the other is ‘transcriptomic’, for measuring mRNA levels. They differ in the kind of immobilised DNA used to generate the array and, ultimately, the kind of information that is derived from the chip. The target DNA used will also determine the type of control and sample DNA that is used in the hybridisation solution.

5.5.4.5.1 Changes in Gene Expression Levels. Determining the level, or volume, at which a particular gene is expressed is called microarray expression analysis, and the arrays used in this kind of analysis are called ‘expression chips’. The immobilised DNA is cDNA derived from the mRNA of known
genes, and the control and sample DNA hybridised to the chip is cDNA derived from the mRNA of for example, normal and diseased tissue. If a gene is overexpressed in a certain disease state, then more sample cDNA, as compared to control cDNA, will hybridise to the spot representing that expressed gene. Expression analysis is valuable in drug development, drug response and therapy development.

5.5.4.5.2 Genomic Gains and Losses. A technique called microarray comparative genomic hybridisation (CGH) is used to look for genomic gains and losses or for a change in the number of copies of a particular gene involved in a disease state. In microarray CGH, large pieces of genomic DNA provide the target DNA, and each spot of target DNA in the array has a known chromosomal location. The hybridisation mixture contains fluorescently labelled genomic DNA harvested from both normal (control: green) and diseased (sample: red) tissue. If the number of copies of a particular target gene has increased, a large amount of sample DNA will hybridise to the corresponding loci on the microarray, whereas comparatively small amounts of control DNA will hybridise to the same spots. As a result, those spots containing the disease gene will fluoresce red with greater intensity than they will fluoresce green. CGH is used clinically for tumour classification, risk assessment and prognosis prediction.

5.5.4.5.3 Mutations in DNA. Detection of mutations or polymorphisms in a gene sequence employs the DNA of a single gene as the immobilised target. In such arrays, the target sequence at a given locus will differ from that of other spots in the same microarray sometimes by only one or a few specific bases. A type of sequence commonly used in such analyses are single nucleotide polymorphisms (SNPs). SNPs have a single genetic change within a person’s DNA sequence. The analysis of such a target microarray requires genomic DNA derived from a normal sample for use in the hybridisation mixture. An SNP pattern associated with a particular disease can be used to test an individual to determine whether he or she is susceptible to that disease. Such ‘mutation/polymorphism analysis’ is commonly used in drug development, therapy development and tracking disease progression.

5.5.4.6 Microarray Data Management. Data management technology is critical for the efficient use of microarray results, but is beyond the scope of this book. The Gene Expression Omnibus (GEO: www.ncbi.nlm.nih.gov/geo/) is an online resource for the storage and retrieval of gene expression data from any organism or artificial source.

Personalised drugs, molecular diagnostics, integration of diagnosis and therapeutics are the long-term medical promises of microarray technology. For the future, DNA microarrays offer hope for obtaining global views of biological processes – simultaneous readouts of all the body’s components – by providing a systematic way to survey DNA and RNA variation.

5.5.5 In Situ Analysis of RNA in Whole Organisms

Hybridisation can be used to detect transcripts in a cell or organism. Cells and organisms smaller than about 1 mm are fixed (the macromolecules are immobilised) by treatment with a fixative, such as formaldehyde, glutaraldehyde or methanol/acetic acid. Larger organisms are normally sliced into thin sections before fixation. The fixed specimens are then probed with radioactively or fluorescently labelled nucleic acid in the same way as for a Southern blot. Synthetic 2’-O-methyloligoribonucleotides are particularly good probes of mRNA in cells, because they are resistant to cellular nucleases (Sections 3.1.4.2 and 4.4.3.6). By use of microscopy (Section 11.5), the locations of RNAs can be determined at the cellular or even the sub-cellular level.

5.6 GENE MUTAGENESIS

5.6.1 Site-Specific In Vitro Mutagenesis

The process of engineering specific changes in a DNA sequence is termed as in vitro mutagenesis. It is an invaluable tool for modification of a DNA sequence in a pre-determined manner to study its biological function. In classical mutagenesis, alterations are created randomly and the effects of each mutation need
to be screened separately, which is time consuming. Now more directed methods are standard, where the DNA is first cloned for ease of manipulation and then deletions, insertions or replacements made.

5.6.1.1 Deletions. Deletions can be created at restriction sites (Section 5.3.1) by cleavage with the corresponding enzyme and then by treatment for a short period with an exonuclease enzyme. For example, the exonuclease Bal 31 is used to remove both double- and single-stranded DNA from both ends. Alternatively, the enzyme Exo III is used to generate single-stranded ends followed by treatment with SI nuclease to trim the created single strands (Table 5.3). Re-ligation of the two new double-stranded ends generates deletion mutations of the parent DNA. This method has the serious limitation that deletions can only be made around restriction sites.

A more general deletion method involves use of synthetic oligonucleotides (Figure 5.16). In this procedure, an oligonucleotide complementary to the desired site of deletion on the DNA, but not containing the nucleotides required for deletion, is used as a primer for synthesis of a second DNA strand. In the process of cloning, mutant DNA segregates from wild type DNA and clones containing mutant the deletion can be selected.

One problem associated with this technique is that bacteria will often attempt to repair the mutagenised strand because the \textit{in vivo}-generated DNA strand is methylated. This can result in low yields of the mutated sequence. Eckstein has developed a reliable method that involves incorporation of \textit{phosphorothioate}-modified nucleotides (Section 4.4.3) into the \textit{in vitro}-generated strand. Such nucleotides are more resistant to nuclease degradation, with the result that the unmutagenised DNA strand can be removed by exonuclease digestion and the gap filled to generate the mutation in both strands (Figure 5.17). Deletion mutants can also be generated by this method by use of PCR (Section 5.2.2).

Deletion mutants can also be generated by PCR (Figure 5.18). This method relies upon the fact that PCR primers are tolerant of primer-template mismatch to create a mutation at the priming site in an analogous way to that shown in Figure 5.17. Unfortunately, this raises the problem that PCR-based mutagenesis can only make a mutated site at an end of the PCR fragment. However, this problem can be solved by generating two PCR products sharing a common central mutated region (Figure 5.18). Denaturation and annealing of these two products, followed by extension of the duplex with \textit{Taq} DNA polymerase yields a larger product with the mutated site in the centre.

5.6.1.2 Insertions. Insertions may be generated by ligation of a synthetic oligonucleotide duplex into a restriction site after cutting with the appropriate restriction enzyme. Sequence additions at other sites can

Figure 5.16 Oligonucleotide site-directed deletion mutagenesis
be achieved by means of site-directed mutagenesis using oligonucleotides in an analogous way to that described for deletions (Figures 5.17 and 5.18), but in this case the synthetic oligonucleotide primer contains the desired additional nucleotide(s).

5.6.1.3 Replacements. A common type of mutation is that which maintains the same number of nucleotides but where part of a sequence is replaced. This is particularly useful for single-base alterations that lead to a change of amino acid codon. Expression of the mutated gene leads to the production of a protein with a single amino acid alteration (protein engineering). One replacement method is to introduce a small deletion at a restriction site followed by ligation into the gap of an oligonucleotide duplex of the same size but of different sequence. A more general approach involves use of a synthetic oligonucleotide primer in an analogous way to the introduction of deletions and insertions, but with the same number of nucleotides in the mutant strand as wild type (Figures 5.17 and 5.18).
5.6.2 Random Mutagenesis

Random mutagenesis is a method of introduction of multiple mutations into a DNA sequence but in arbitrary sequence positions. Random mutagenesis was achieved classically by reaction of DNA with chemicals (Chapter 8) to establish its biological role. More recently, random mutagenesis has been used to evolve DNA sequences that code for proteins with new or enhanced properties and in the evolution of new DNA catalysts (Section 5.7.3). The introduction of a number of random mutations into a DNA sequence generates a library of new DNA entities, which can then be used to identify those sequences with the required functionality. There are two main methods of random mutagenesis.

In error-prone PCR, the aim is to modify the usual PCR protocol (Section 5.2.2) in order to deliberately introduce mutations. There are several ways in which this may be done. For example, the use of a polymerase which lacks proof-reading ability (such as Taq DNA polymerase) allows errors inherent in DNA synthesis to go uncorrected. Other changes to help reduce the fidelity of the polymerase include a lower annealing temperature in the PCR cycle, low or unequal dNTP concentrations and/or use of a large number of PCR cycles (60–80), which allows amplification of erroneous copies. Another common method of increasing polymerase infidelity is to increase the Mg$^{2+}$ concentration (up to 10 mM) in the PCR reaction or to replace Mg$^{2+}$ by Mn$^{2+}$ (typically 0.05–0.5 mM). Mutated products can be amplified by further rounds of PCR.

A second method for introduction of mutations is the use of nucleotide analogues into the nascent DNA during PCR. When the nucleotide analogue is copied in subsequent rounds of PCR, it is not recognised by the polymerase as a normal nucleotide and an incorrect deoxyribonucleotide is inserted opposite the analogue. Examples of such analogues includes 2'-deoxyinosine, 5-fluoro-2'-deoxyuridine, 8-oxo-2'-deoxyguanosine and the degenerate pyrimidine analogue dP (Figure 5.19).
5.6.3 Gene Therapy

The ability to introduce new or altered genes into a mammalian genome has tremendous implications. For example, it may prove possible to cure some genetic diseases by introducing a healthy copy of a gene into an afflicted individual. It is already possible to introduce into mammals genes that encode economically or medically important polypeptides such as insulin growth hormones and interferon. The intention is that the animal either grows faster or produces large amounts of protein, which can be harvested. We will address the ways in which this can be carried out, leaving the ethical questions raised by this issue to others.

There are three major ways of introducing DNA into mammalian germ tissue such that the progeny of the recipient will carry the gene. The first involves microinjection of DNA solutions into the nucleus of an egg by means of an extremely fine capillary. Such a technique works very well with a mouse egg but is more difficult with other mammals, such as sheep, where it is extremely hard to see the nucleus. In this way, transgenic animals have been created which carry functioning genes from another organism.

The second method involves the use of retrovirus-based vectors (Figure 5.20). As described in Section 6.4.6, retroviruses can infect a cell and then insert their DNA into its chromosomes. The gene to be introduced into the host is ligated into the genome of the retrovirus. The retroviral DNA is then introduced into a cultured cell line, which is capable of producing all components of a retrovirus except for the viral RNA (such a cell culture is called a helper cell line). This cell line will then package the recombinant virus stock into virus particles that can be harvested from the culture medium. Helper cells are necessary because the presence of the insert in the retroviral genome disrupts some of the normal retroviral genes needed for viral production. The harvested recombinant virus stock is then used to infect an early embryo, which is then replaced into a donor mother. During growth, some cells of the embryo become infected by the virus and the retroviral gene, including the gene insert, becomes stably inserted into the DNA of these cells. Because not all cells become infected, the animal is a chimera. However, if the germ cells of this animal contain proviral DNA then its offspring will retain the recombinant in every cell of its body.

In addition to introduction of a new gene coding for a protein, it is also possible to introduce via a retroviral vector a gene that codes for a specialised RNA (e.g., antisense RNA (Section 5.7.1), short interference RNA (Section 5.7.2) or an RNA that folds into ribozyme or an aptameric structure (Section 5.7.3) that can act in trans to interact with and block the function of another RNA.

The third method for introducing DNA into the mammalian germ line relies upon the existence of cultured cell lines, which can become germ cells if injected into early embryos. This approach is particularly useful in the mouse, where such cells, embryonal carcinoma cells, can be grown in dishes. The gene of interest can be introduced into these cells, which are then injected into embryos.
5.7 OLIGONUCLEOTIDES AS REAGENTS AND THERAPEUTICS

The ability to synthesise DNA and RNA oligonucleotides of defined sequence rapidly (Sections 4.1 and 4.2), including a range of nucleotide analogues (Section 4.4), has led to a large number of applications as therapeutic and diagnostic agents. Many applications involve the principle of recognition of a linear sequence of RNA or DNA by the oligonucleotide. For example, antisense, steric block and short interfering RNA (siRNA) all involve targeting of RNA within cells to form duplexes as a means of control of gene expression (Sections 5.7.1 and 5.7.2). Synthetic oligonucleotides have been used also to form triplexes (Sections 2.3.6 and 9.10.1) with double-stranded DNA to block gene expression, but this principle has so far not led to therapeutic products. Other types of application include in vitro selection and design of oligonucleotides that recognise and bind to nucleic acids structures, to proteins or to small molecule ligands (Section 5.7.3).

5.7.1 Antisense and Steric Block Oligonucleotides

In 1979, Zamecnik and Stephenson were the first to show that a synthetic oligonucleotide could be used to block specific gene expression in Rous Sarcoma Virus. This pioneering work led to the study of oligonucleotides and their analogues as therapeutic agents. This field is commonly known as ‘antisense’, since the principle of biological activity usually involves either degradation or steric blocking of the sense
strand (the coding strand) of RNA (commonly mRNA or viral RNA) through formation of an exactly base paired duplex between the target RNA and an added complementary strand (the antisense strand) (Figure 5.21). Formation of the duplex causes inhibition of expression of a particular gene within cells or in vivo and the aim is to do this without affecting any other gene.30,31

5.7.1.1 Basic Mechanisms

5.7.1.1.1 Steric Block. This mechanism involves formation of an RNA–DNA duplex to physically block the RNA and to prevent recognition by a protein or other cellular machinery. For example, binding of an oligonucleotide close to the 5′-cap site32 (Section 7.2.1) or at the site of initiation of translation in mRNA33 (Section 7.3.3) may prevent the ribosome or associated machinery from binding to the RNA and initiating translation (Figure 5.22). Other RNA processing events that can be interfered with stericly by duplex formation include nuclear splicing34 and polyadenylation35 (Section 7.2.1), both of which are required for the processing of most mammalian gene transcripts and which involve numerous steps of RNA–protein recognition. In the case of viral RNA, it is possible to block recognition of essential RNA binding proteins that are required for virus-specific gene regulation.

5.7.1.1.2 Induction of RNase H. Although steric block activity requires stoichiometric amounts of complementary added oligonucleotide, a more potent inhibitory effect can often be obtained through recognition of an RNA–oligonucleotide duplex by the ubiquitous cellular enzyme Ribonuclease H (RNase H). The normal function of this enzyme is to help the removal of RNA primers in DNA replication (Section 6.6.3). However, when an RNA sequence is targeted in cells by a complementary oligodeoxynucleotide, RNase H-induced cleavage can occur rapidly, usually close to the centre of the targeted RNA section.36 The loss of intact RNA leads to rapid degradation of the RNA. Thus in the case of mRNA, there is a concomitant reduction in the level of the encoded protein expressed. Most regions of an mRNA can usually be targeted by such oligonucleotides, including 3′- and 5′-untranslated regions.

![Figure 5.21](image1)

**Figure 5.21** Duplex formed by an antisense oligodeoxyribonucleotide and a target mRNA

![Figure 5.22](image2)

**Figure 5.22** Three alternative mechanisms of steric block action of antisense oligonucleotides acting upon RNA
5.7.1.2 Optimal Oligonucleotide Characteristics. There are many factors that influence cellular or in vivo antisense activity. In practice, oligonucleotide optimisation is carried out by experimentation through use of in vitro, cell-based and ultimately in vivo assays, although some general principles can be used in oligonucleotide design.

5.7.1.2.1 Duplex Stability. For intracellular antisense activity, an oligonucleotide must be of sufficient length to form a strong duplex with its RNA target at 37°C under cellular conditions. In general, binding strength increases as a function of length as well as the number of G:C pairs (Section 5.5.1). In addition, the type of nucleotide analogues used and their placement within the oligonucleotide are also crucial and those nucleoside analogues that adopt an RNA-like, 3'-endo sugar conformation (such as 2'-O-methylribonucleosides) tend to result in increased binding strength, since there is a tendency to form a more compact A-helix (Section 2.2.3). It is important also that the oligonucleotide does not form unusual secondary structures (such as G quadruplexes, Section 2.3.7) that may hinder duplex formation. Another important consideration is whether the target RNA site is easily accessible, i.e. does not exhibit tight RNA secondary or tertiary structure or is not strongly bound by cellular proteins. In this regard, experimental approaches to target choice are often more reliable than RNA structure prediction.

5.7.1.2.2 Specificity. For unique sequence recognition within the human genome (i.e. no other likely exact match for an oligonucleotide of typically mixed composition), a minimum length of around 12 nucleotides is usually required. However, the longer the chosen sequence, the greater the chance for the oligonucleotide to form a mismatched duplex with an incorrect RNA sequence. This is particularly of concern in the case of RNase H induction where an incorrect RNA may be cleaved in addition to that targeted, leading to side effects. In practice, a compromise between duplex stability and target specificity limits oligonucleotide length usually to 12–25 residues.

5.7.1.2.3 Nuclease Stability. Unmodified single-stranded DNA and RNA oligonucleotides are degraded very fast by cellular nucleases in cells and serum. 3'-Exonucleases are the most prevalent, such that minimally the 3'-end of an antisense oligonucleotide must be protected, usually by chemical modification. But 5'-exonucleases as well as endonucleases are also present in cells, and thus for therapeutic applications, nuclease protection of each internucleotide linkage by inclusion of analogues is often thought necessary.

5.7.1.2.4 Cellular Uptake. A significant difficulty is that oligonucleotides and their analogues rarely penetrate cells in culture without co-addition of a carrier or cell delivery agent. For example, popular delivery agents for many cultured mammalian cell lines are cationic lipids, which can form complexes with negatively charged oligonucleotides to help cell association, uptake through the endosomal pathway and subsequent release into the cytosol by endosome destabilisation. Oligonucleotides are able to enter cell nuclei readily once they have been released into the cytosol. Oligonucleotides are usually administered in vivo without carrier, and here there may be special mechanisms available for cell uptake, but this remains a difficult and controversial subject of study.

5.7.1.2.5 Pharmaceutical Considerations. One positive feature of many clinically investigated oligonucleotides to date is their relative lack of toxicity during systemic delivery into animals and man. By contrast, a major concern has been the frequently observed, rapid clearance through the kidney, which is typical of many macromolecules. It is thus not surprising that the greatest success to date for therapeutic oligonucleotides has been in local or topical administration. Pharmaceutical development remains a significant challenge in terms of reaching the required tissue or organ, efficacy of action at the site and the maintenance of a therapeutic dose at manageable and affordable concentration levels. Many studies continue that focus on investigations of new nucleotide analogue types and combinations, conjugates and formulations.

5.7.1.3 Nucleotide Analogues Used in Antisense Applications. The most potent antisense oligonucleotides to date have been those shown to induce RNase H cleavage within cellular models and several have been taken to clinical trials. Strong recognition by RNase H requires there to be a contiguous stretch of minimally 6–10 residues of 2'-deoxyribonucleosides where internucleotide linkages are phosphodiester
or the close analogue phosphorothioate (Section 4.4.3). Phosphorothioates are considerably more resistant than phosphodiesters to nuclease degradation and are well tolerated in humans. Thus, first generation therapeutic oligonucleotides contained only 2′-deoxyribonucleotide phosphorothioates, such as the clinically approved drug Vitravene, which is a 26-mer used for treatment of CMV-induced retinitis in AIDS patients.

Second-generation oligonucleotides employ the principle of a gapmer. Such oligonucleotides contain a section of 6–10 residues (usually centrally placed) of 2′-deoxyribonucleoside phosphorothioates, but the flanking regions on each side comprised of other analogues that enhance binding to the RNA target and further increase the oligonucleotide stability to nuclease, but which do not direct RNase H cleavage. Such analogues are generally ribonucleoside analogues, such as 2′-O-methyl, 2′-O-methoxyethyl or locked nucleic acids (LNA), where the sugar conformation is 3′-endo. Overall, gapmers are recognised and direct RNA cleavage by RNase H. Whereas several first generation antisense oligonucleotides, such as ISIS 3521 (Affinata) targeted to the mRNA for protein kinase Cα,37 failed clinical trials, there is more hope of clinical benefit for some higher potency gapmers against viral and cancer targets (Figure 5.23).26,31

For steric block applications, there is no restriction in principle to the type or placement of an analogue within a sequence as long as other antisense considerations are met. The variety of analogues that have been investigated is very large. In some cases analogues can be combined in one oligonucleotide to give mixmers. In addition to 2′-O-methyl and 2′-O-methoxyethyl ribonucleotides described above, other important analogues used in steric block applications fall into two classes: (a) those that contain a phosphate group, such as LNA, tricyclo DNA, 3′-amino phosphoroamidate and phosphorothioamidate, and (b) non-phosphate containing analogues such as peptide nucleic acids (PNA) and morpholinodiamidates. In class (b) it was hoped that the absence of the negative charges on the oligonucleotide would enhance cell uptake, but this is not the case. Attachment of a cationic or other cell penetrating peptide appears to improve cell uptake, but the universality of this approach is still under study. A steric blocking, phosphorothioamidate oligonucleotide targeted to the essential RNA involved in the enzyme telomerase (Section 6.6.5) is moving close to clinical trials as an anti-cancer agent.38

5.7.1.4 Non-Duplex Therapeutic Activities of Single-Stranded Oligonucleotides. Recently, other biological activities of oligonucleotides have been found that are sequence-dependent but which are independent of duplex formation with an RNA target.

5.7.1.4.1 Immune Modulation. Single-stranded oligodeoxynucleotide phosphodiesters and phosphorothioates that contain the dinucleotide sequence CpG can trigger an immune response when administered to humans and animals.39,40 The response is mediated through binding to a ‘toll-like receptor’ TLR9 that is present in cytosolic vesicles and the binding stimulates signalling pathways that activate transcription factors. By contrast, double-stranded RNA and siRNA (Section 5.7.2) binds to another receptor TLR7 and may stimulate a different immune response. The context of the CpG determines the immune modulation specificity, such that mouse TLR9 prefers CpG when flanked at 5′ by two purines and at 3′ by two pyrimidines, while human TLR9 is recognised optimally by GTCGTT and TTCGTT sequences. Such activities are now recognised to have contributed being harnessed for therapeutic applications and as vaccine

Table: First and second generation phosphorothioate-containing oligodeoxyribonucleotides

<table>
<thead>
<tr>
<th>First Generation</th>
<th>Second Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISIS 3521 (Affinata)</td>
<td>ISIS 9606</td>
</tr>
<tr>
<td>GsTsTsCsTsCsGsCsTsGsGsTsGsAsGsTsTsTsCsA</td>
<td>GsTsTsCsTsCsGsCsTsGsGsTsGsAsGsTsTsTsCsA</td>
</tr>
</tbody>
</table>

s = phosphorothioate linkage
underline = 2′-O-methoxyethyl nucleoside

Figure 5.23 First and second generation clinically used phosphorothioate-containing oligodeoxyribonucleotides
adjuvants. Immune stimulation of this or similar type has also been used to explain \textit{in vivo} activities of some supposed antisense oligonucleotides that contain CpG sequences.

5.7.1.4.2 DNA Aptamers. Oligodeoxynucleotides have been selected (Section 5.7.3) that bind specific cellular proteins. Such \textit{aptamers} can fold into unusual structures (such as a G quadruplex, Section 2.3.7). A chemically synthesised DNA aptamer that binds strongly to vascular endothelial growth factor has been formulated as a conjugate with polyethylene glycol and has recently been given regulatory approval (Macugen) for treatment of patients with the wet form of age-related macular degeneration, a common cause of blindness due to abnormal blood vessel growth.\(^{41}\)

5.7.2 RNA Interference

In the late 1990s, gene silencing by double-stranded RNA was observed in plants in the laboratory of David Baulcombe\(^{42}\) and in the worm \textit{C. elegans} in that of Craig Mello.\(^{43}\) Within a very few years gene silencing activities were found in many diverse organisms and have gone on to be harnessed as powerful diagnostic reagents in genome research and as potential therapeutics.\(^{44-46}\) \textbf{RNA interference (RNAi)} probably evolved from the need for eukaryotic cellular defence against foreign (e.g. viral) duplex RNA or DNA that is transcribed into RNA. Distinct but overlapping pathways have been elucidated that represent different forms of genetic regulation (Figure 5.24).

Primary RNA transcripts in the nucleus has been found to contain hundreds of endogenous sequences known as \textbf{microRNAs (miRNAs)} that can fold into hairpins that contain imperfect matches (Section 7.5.3). The transcripts are processed by the nuclear complex \textit{Drosha}, which contains an RNase III activity that cleaves such RNAs to produce hairpins of about 70 residues (pre-microRNAs). After export to the cytosol, the

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.24.png}
\caption{Mechanisms of action involved in RNA interference (RNAi). Pathway (a) shows steps in the processing of microRNA (miRNA) eventually leading to inhibition of translation. Pathway (b) shows the processing of double-stranded viral RNA to form short interfering RNA (siRNA) and eventual cleavage of mRNA by the RNA-induced silencing complex (RISC) complex. ShortRNA (shRNA) (centre) can in principle enter either pathway.}
\end{figure}
Hairpins are further processed by the enzyme complex **Dicer** (Dcr-1), a multi-subunit protein complex that also contains an RNase III activity, to give imperfectly paired RNA duplexes of about 21–23 residues (miRNAs). These are recognised by the RNA-induced silencing complex (**RISC**), which directs one of the two RNA strands to bind to a selected sequence in the 3′-untranslated region of a gene (a microRNA recognition element) to form an imperfect complement, which results in a block to translation (Figure 5.24, pathway a).

A second pathway is triggered by the introduction into a cell of double-stranded RNA, such as viral RNA. A second DICER variant, Dcr-2, is responsible for processing this duplex RNA into 21–23 residue perfect duplexes (**siRNA**). SiRNAs are then utilised by the RISC complex to direct one strand (antisense or **guide**) to form a duplex with an exact complement on a mRNA and cleave the phosphodiester bond precisely between residue 10 and 11 counting from the 5′-end of the complement, as directed by a member of the Argonaute family (Ago 2), an RNA endonuclease within RISC (Figure 5.24, pathway b). The other strand (sense or **passenger**) is discarded and then degraded.

Thomas Tuschl and colleagues found that when synthetic siRNAs of 21 residues were transfected into mammalian cells, the RISC-dependent cleavage pathway could be triggered, thus allowing site-specific cleavage of mRNA and subsequent inhibition of gene expression. Synthetic siRNAs have now become used widely as reagents for specific gene inhibition in many cell types and seem to be applicable to almost all genes. Generally two-nucleotide 3′-overhangs are added on each strand for optimal activity, similar to those found following natural DICER cleavage of duplex RNA. Although siRNA duplexes appears to be stable to nuclease degradation for hours to days within cells, unlike single-stranded RNA, much effort has been expended on investigation of the tolerance to incorporation of analogues or conjugates that might have advantages in vivo to aid stability or pharmacology. The sense strand appears to be highly tolerant of chemical modification, but the antisense strand, which is the one introduced by the RISC complex to pair with mRNA, is less so. A recent demonstration of efficacy in a transgenic mouse model promises that modified siRNAs may have therapeutic value.

A third RNAi pathway is triggered by introduction into cells of short hairpin RNAs (**shRNAs**) of around 29-base pairs (Figure 5.24). Such shRNAs are recognised by DICER and processed to give siRNAs of high potency, presumably because they are generated endogenously and may be more readily utilised by RISC. ShRNAs are also potential precursors of imperfectly matched miRNA.

Although siRNAs have been shown to generate some immune response effects, it is not clear at present whether these will present significant problems or not for their in vivo use.

### 5.7.3 In Vitro Selection

#### 5.7.3.1 Principles of In Vitro Selection (SELEX).

The advent of in vitro selection or **SELEX** (systematic evolution of ligands by exponential enrichment) in the early 1990s by the groups of Gold, Szostak and Joyce marked the beginning of a new age in the design of functional nucleic acid molecules as both ligands for given targets and as catalysts. SELEX is a combinatorial technique in which nucleic acids with specific properties, such as binding with high affinity to a given target molecule (an aptamer) or catalysis of a chemical reaction, are selected from a pool of typically $10^{12}$–$10^{15}$ RNA or DNA molecules of randomised sequence. The technique exploits the wide range of structures that single-stranded nucleic acids can adopt and mimics the natural processes of evolution.

#### 5.7.3.2 Selection of Aptamers.

The basic principle of SELEX (Figure 5.25) involves the creation by automated chemical synthesis of an initial oligonucleotide library consisting of an internal random nucleotide sequence flanked by 5′- and 3′-tails of a constant sequence which act as primer binding sites for subsequent amplification of the library by PCR (Section 5.2.2). The random sequence, typically 10–100 nucleotides long, is generated by delivering a mixture of all four nucleoside phosphoramidites simultaneously during automated synthesis (Section 4.1). Since the library contains just a few copies of each sequence, it is first amplified by PCR in which one of the oligonucleotide primers carries a biotin modification. To act as ligands to a specific target molecule, the nucleic acids within the library must be free to fold into a wide
range of 3-D structures and hence must be single stranded. Thus the selection of RNA aptamers, the duplex DNA library contains a T7 RNA promoter sequence upstream of the 5′-constant sequence. This enables a single-stranded RNA pool to be generated through transcription by use of T7 RNA polymerase (Figure 5.25). A single-stranded DNA pool required for the selection of DNA aptamers is produced by capture of only the 5′-biotinylated strand from the DNA duplex on streptavidin-derivatised beads (Figure 5.26).

The selection of DNA or RNA aptamers, which can bind to a given target molecule is achieved by use of affinity chromatography. Thus the target molecule is immobilised on a solid support usually within a small column and a solution of the nucleic acid pool is passed through. Unbound nucleic acids are eluted by simple washing of the column with a suitable buffer. Sequences that have some affinity for the target are bound by the column and subsequently eluted by washing with a buffer that contains the free target molecule. For RNA aptamers, a cDNA library is then generated from the bound fraction using enzyme reverse transcriptase, which is then amplified by PCR (Figure 5.25). For DNA aptamers, the aptamer-containing fraction is subjected directly to PCR (Figure 5.26). In subsequent rounds of SELEX, the stringency of the washing protocols is increased or the concentration of immobilised target is reduced such that the affinity chromatography step leads to an enrichment of high-affinity binding sequences within the library. Typically about ten cycles of SELEX are carried out after which perhaps about 100 or so different sequences remain. These aptamers are then cloned into a vector and characterised by DNA sequencing.

SELEX has been used to identify a wide range of aptamers to a variety of diverse molecular targets, for example ions, small molecules such as organic dyes, nucleotides and their bases, amino acids, co-factors, antibiotics, transition state analogues, as well as peptides and proteins. The remarkable selectivity of
aptamers for their target molecule is nicely illustrated by one such example that binds with high affinity ($K_D$ 0.6 μM) to theophylline (1,3-dimethylxanthine) but binds the highly similar structure caffeine (1,3,7-trimethylxanthine) about 10^4 fold less efficiently.

The selection of an aptamer with high affinity for the blood-clotting protein thrombin was the first example of a nucleic acid ligand designed to bind to a protein target that does not normally interact with DNA. Such aptamers show affinities between 25 and 200 nM and contain a highly conserved 14–17 base consensus sequence. NMR of the 15-mer aptamer d(GGTTGGTGTGGTTGG) and X-ray crystallography as a complex with thrombin have revealed that the oligonucleotide forms a DNA quadruplex structure (Sections 2.3.7 and 9.10.2). Such aptamers have potential clinical application as anti-coagulants.

5.7.3.3 Selection of Nucleic Acid Catalysts. SELEX has also been exploited to generate nucleic acid-based catalysts for a wide range of chemical reactions. Examples including RNA cleavage, DNA cleavage, DNA ligation, DNA phosphorylation, porphyrin metalation, DNA capping, DNA depurination, amide bond formation and the Diels–Alder reaction and the ability for stereochemical control during catalysis highlight the potential of SELEX in the area of synthetic organic chemistry. Initial selections require protocols in which the chemical reaction is intramolecular (in cis), for example, self-cleavage, self-alkylation or self-phosphorylation. However, the analogous intermolecular reaction (in trans) with a separate substrate molecule is generally of more practical value.

In an example derived from the work of Santoro and Joyce, a DNA catalyst (DNAzyme) capable of the specific cleavage of an HIV RNA target sequence was identified (Figure 5.27). Here a synthetic DNA library containing a central randomised region of 50 nucleotides flanked by 5’ and 3’ constant sequences...
was first copied using a synthetic 5′-biotinylated mixed DNA/RNA primer. The primer contained an embedded 12 nucleotide RNA sequence and a 3′-DNA tail complementary in sequence to the 3′-terminus of the DNA library. Single-stranded DNA containing both enzyme (within the 50 nucleotide random sequence) and target RNA sequence was then obtained following capture by streptavidin-coated beads, denaturation and removal of the non-biotinylated template. This allows folding and interaction of the enzyme and substrate portions of the immobilised sequence. In the presence of magnesium ion co-factor, the active sequences undergo self-cleavage within the RNA target section. The released nucleic acid sequences are then amplified by PCR where one primer is biotinylated. Single-stranded templates are produced by streptavidin capture of the biotinylated strand ready for the next round of SELEX.

The DNAzyme sequences identified in this work can be simplified and shortened such that cleavage of a separate substrate strand (trans cleavage) can occur. The ‘10–23’ enzyme derived from this work comprises of a catalytic core of 15 nucleotides flanked by two substrate recognition domains in which Watson–Crick base pairing occurs (Figure 5.28). The structure of the ‘10 –23’ enzyme can be modified to recognise and
cleave different target sequences providing that Watson–Crick recognition between the enzyme and substrate is maintained. Recently a ‘10–23’ DNAzyme capable of achieving cleavage rates of up to 10 min\(^{-1}\) under certain metal ion, concentration and pH conditions has been identified, while a related DNAzyme also reported by Joyce\(^53\), the ‘8–17’ motif, has been exploited as a biosensor for Pb\(^{2+}\) ions.

**5.7.3.4 Modified Nucleotides Used in SELEX.** The diversity of functional groups available for binding and catalysis offered by the four natural nucleotides is limited. Furthermore, the pK\(_a\) values of the nucleobases are not ideal to permit electrostatic transition-state stabilisation or general acid–base mechanisms at appropriate neutral pH. Consequently, both sugar and base modified nucleotides have been exploited in SELEX, but such modified nucleotides must be substrates for T7 RNA polymerase or a suitable DNA polymerase to replace their natural analogues during replication or transcription of the template DNA or RNA pool respectively.

Analogues such as 5-(1-pentynyl)-2′-deoxyuridine (Figure 5.29, structure c) have been used in the selection of aptamers designed to bind thrombin. These aptamers display similar binding to thrombin as the unmodified 15-mer aptamers, but have different structures and do not function if the analogue is replaced by 2′-deoxyuridine or thymidine. The potential use of aptamers as therapeutics has led to attempts to increase their stability towards degradation by nuclease enzymes. Thus 2′-fluoro- and 2′-amino modified ribonucleotides have been employed in SELEX.\(^54\) These analogues impart a considerably increased stability to RNA in respect of both chemical and enzymatic hydrolysis. Complete resistance to nucleases can be achieved if the aptamers are comprised of mirror image or L-RNA (spiegelmers).\(^61\)

A recent variation suitable for the isolation of very high affinity nucleic acid ligands is photo-SELEX.\(^62\) Here a 5-bromo-UTP, -2′-fluoro-UTP or -dUTP is employed as one of the nucleotide units and the affinity chromatography step is followed by a brief UV laser irradiation step at 308 nm, during which the single-stranded nucleic acid ligand is cross-linked to a proximal electron-rich amino acid in a protein target. The complexes are purified by SDS–PAGE and following proteolysis with proteinase K, the aptamer functions as

![Figure 5.29 A selection of modified nucleoside triphosphates that have been used in the selection of nucleic acid aptamers and catalysts](image-url)
a template for Taq DNA polymerase (DNA aptamer) or reverse transcriptase (RNA aptamer) in further rounds of SELEX. The technique has been used to isolate aptamers with nM and pM affinities to the HIV-1 Rev protein and basic fibroblast growth factor (bFGF) respectively. Only high affinity aptamers have the precise orientation of functional groups to permit cross-linking, and a harsh washing step removes all non-cross-linked proteins interacting with the immobilised aptamer. This reduces background signals due to binding of non-cognate proteins and allows a highly reliable diagnostic assay for proteins attached to a micro-chip.

The incorporation of base modified nucleosides with potential catalytic groups was first demonstrated by Eaton in which a C-5-modified UTP bearing an appended pyridyl or imidazole function (Figure 5.29, structures a and b) was employed in the selection of an RNA capable of acting as a Diels Alderase or an amide synthase. More recently imidazole- and amine-modified nucleotides have been employed to select DNAzymes capable of the sequence-specific cleavage of RNA (Figure 5.29, structures d, f and g). For example, Perrin and Williams selected DNAzymes functionalised with imidazolyl and amino functional groups that catalyse the sequence specific cleavage of RNA in the absence of metal ions with rate enhancements of about 10^5 compared to the uncatalysed reaction. Such DNAzymes display the functional side chains that are utilised by protein ribonuclease RNaseA in metal-independent RNA cleavage.

### 5.7.3.5 Riboswitches

While the catalytic potential of natural RNA has been known since the mid-1980s, it is only recently that a natural biological role for RNA aptamers has been revealed. Such naturally occurring aptamers or ‘riboswitches’ have been found within the leader sequences of several metabolic genes where they have important roles in regulating both transcription and translation of the respective gene. The function of these riboswitches is to assess cellular levels of certain metabolites, which in turn control expression of that gene. Thus the riboswitch functions in much the same way as synthetic aptamer that can recognise and bind to a small molecule target. The flavin mononucleotide (FMN)-sensing riboswitch found in bacteria was one of the first such examples. FMN and flavin adenine dinucleotide (FAD) are synthesised from riboflavin (vitamin B2). The enzymes that are responsible for riboflavin biosynthesis from GTP are derived from five genes that comprise the riboflavin operon. The first of these genes contains a 300 nucleotide untranslated region which, upon binding of FMN or FAD, changes conformation so as to cause the termination of transcription. In this work, a number of other riboswitches have been described, including examples, which recognise thiamine pyrophosphate, adenosylcobalamin, S-adenosyl methionine, lysine and guanine.

### 5.8 DNA FOOTPRINTING

**Footprinting** is a method for determining the precise DNA sequence of bases that is the site for attachment of a particular DNA enzyme or binding-protein or of a DNA-binding drug. DNA footprinting utilises a DNA cleaving agent, which can be either a nuclease or a chemical reagent. The agent must be able to cut DNA non-selectively at every exposed base pair while such DNA cleavage is inhibited at the site where the protein or drug binds to DNA. Thus a ‘footprint’ of the target sequence is identified as the region where no cutting is observed. The steps in a DNA ‘footprinting’ experiment are

- a fragment of dsDNA containing the target sequence (usually 200–300 base pairs) is labelled at the 5’-ends with 32P and then the label is removed preferentially from one end (e.g. the 3’-end of a gene) by a suitable restriction endonuclease (Section 5.3.1);
- this dsDNA fragment is incubated with the DNA binding-protein so that the protein protects the target region of DNA from DNase I digestion (Section 5.3.2);
- limited DNase I digestion is carried out, so that there is about one cut per strand and the sites of cleavage are randomly distributed among the accessible sites;
- the resulting DNA fragments are analysed by gel electrophoresis (Section 11.4.3), and give a ladder that has a ‘footprint’ region where there are no cuts, corresponding to the binding site (Figure 5.30). If a control track is generated from a Maxam–Gilbert G + A chemical sequencing reaction (Section 5.1.1) using the same probe as template, then the exact footprint sequence can be read out by comparing the location of the blank with the sequencing reaction.
Footprinting was first used by Galas\textsuperscript{67} to determine the binding sequence for the \textit{lac} repressor protein that established the operator sequence: d(CACCTTAACACTAACCTCTTGTTAAAG)-5'. It is now possible to identify stronger and weaker protein binding and to differentiate between affinities for each of the two DNA strands.

Since protein binding \textit{in vitro} may not accurately reflect binding-site occupancy in the cell nucleus, methods have been developed for DNA footprinting \textit{in vivo}. The GA-LMPCR \textit{in vivo} footprinting system employs \textit{dimethyl sulfate} (0.3–1.5%) to methylate nuclear DNA in whole cells suspended in phosphate buffer. Methylation occurs mainly at guanine N-7 in the major groove (Section 8.5.3) with further methylation at adenine N-3 in the minor groove. Incubation of the protein-free genomic DNA at 90°C and pH 7.0 for 15 min followed by treatment with 1 M NaOH for 30 min at 90°C leads to specific cleavage at methylated G and A sites (Maxam–Gilbert G > A procedure) (Section 5.1.1) Guanine-specific cleavage can also be accomplished by piperidine treatment of methylated DNA. The cleaved strands are then amplified using ligation-mediated PCR and analysed by PAGE, as above. Such methods have been used for the detection of upstream regulatory sequences, known as locus control regions.\textsuperscript{68,69}

For some purposes, cleavage of the DNA is better achieved by chemical means and one of the most successful reagents has been the \textit{hydroxyl radical}: Fenton’s reagent. The cleavage system used is ferrous ammonium sulfate (1 mM) in conjunction with ascorbic acid (10 mM) and hydrogen peroxide (0.3%) at room temperature for 2 min. This works by generating hydroxyl radicals that abstract a hydrogen atom from the deoxyribose leading to phosphate diester cleavage at that residue (Section 8.9.1).\textsuperscript{70}

In addition to its use in studies of protein binding to DNA, footprinting has been widely employed, for example, for investigating the selectivity of drug binding to DNA (Chapter 9) and for conformational analysis of triple helix formation.\textsuperscript{71}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{DNA_footprinting_scheme.png}
\caption{Scheme illustrating DNA footprinting for \textit{lac} repressor protein binding to dsDNA containing the \textit{lac} operator sequence. DNase I cuts DNA molecules randomly. Only one strand is 5'-end labelled with \textit{\textsuperscript{32}P}. For polyacrylamide gel electrophoresis see Section 11.4.3}
\end{figure}
REFERENCES


