

# Contents

## Preface

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Basic Molecular Biology</b>	<b>5</b>
2.1	Nucleic Acid Structure	5
2.1.1	The DNA backbone	5
2.1.2	The base pairs	7
2.1.3	RNA structure	10
2.1.4	Nucleic acid synthesis	11
2.1.5	Coiling and supercoiling	12
2.2	Gene Structure and Organization	14
2.2.1	Operons	14
2.2.2	Exons and introns	15
2.3	Information Flow: Gene Expression	16
2.3.1	Transcription	16
2.3.2	Translation	19
<b>3</b>	<b>How to Clone a Gene</b>	<b>21</b>
3.1	What is Cloning?	21
3.2	Overview of the Procedures	22
3.3	Gene Libraries	25
3.4	Hybridization	26
3.5	Polymerase Chain Reaction	28
<b>4</b>	<b>Purification and Separation of Nucleic Acids</b>	<b>31</b>
4.1	Extraction and Purification of Nucleic Acids	31
4.1.1	Breaking up cells and tissues	31
4.1.2	Enzyme treatment	32
4.1.3	Phenol-chloroform extraction	32
4.1.4	Alcohol precipitation	33
4.1.5	Gradient centrifugation	34
4.1.6	Alkaline denaturation	34
4.1.7	Column purification	35
4.2	Detection and Quantitation of Nucleic Acids	36

4.3	Gel Electrophoresis	36
4.3.1	Analytical gel electrophoresis	37
4.3.2	Preparative gel electrophoresis	39
<b>5</b>	<b>Cutting and Joining DNA</b>	<b>41</b>
5.1	Restriction Endonucleases	41
5.1.1	Specificity	42
5.1.2	Sticky and blunt ends	45
5.1.3	Isoschizomers	47
5.1.4	Processing restriction fragments	48
5.2	Ligation	49
5.2.1	Optimizing ligation conditions	51
5.3	Alkaline Phosphate	53
5.4	Double Digests	54
5.5	Modification of Restriction Fragment Ends	55
5.5.1	Trimming and filling	56
5.5.2	Linkers and adapters	57
5.5.3	Homopolymer tailing	58
5.6	Other Ways of Joining DNA Molecules	60
5.6.1	TA cloning of PCR products	60
5.6.2	DNA topoisomerase	61
5.7	Summary	63
<b>6</b>	<b>Vectors</b>	<b>65</b>
6.1	Plasmid Vectors	65
6.1.1	Properties of plasmid vectors	65
6.1.2	Transformation	71
6.2	Vectors Based on the Lambda Bacteriophage	73
6.2.1	Lambda biology	73
6.2.2	<i>In vitro</i> packaging	78
6.2.3	Insertion vectors	79
6.2.4	Replacement vectors	80
6.3	Cosmids	83
6.4	M13 Vectors	84
6.5	Expression Vectors	86
6.6	Vectors for Cloning and Expression in Eukaryotic Cells	90
6.6.1	Yeasts	90
6.6.2	Mammalian cells	92
6.7	Supervectors: YACs and BACs	96
6.8	Summary	97
<b>7</b>	<b>Genomic and cDNA Libraries</b>	<b>99</b>
7.1	Genomic Libraries	99
7.1.1	Partial digests	101
7.1.2	Choice of vectors	103
7.1.3	Construction and evaluation of a genomic library	106

---

7.2	Growing and Storing Libraries	109
7.3	cDNA Libraries	110
7.3.1	Isolation of mRNA	111
7.3.2	cDNA synthesis	112
7.3.3	Bacterial cDNA	116
7.4	Random, Arrayed and Ordered Libraries	116
<b>8</b>	<b>Finding the Right Clone</b>	<b>121</b>
8.1	Screening Libraries with Gene Probes	121
8.1.1	Hybridization	121
8.1.2	Labelling probes	125
8.1.3	Steps in a hybridization experiment	126
8.1.4	Screening procedure	127
8.1.5	Probe selection	129
8.2	Screening Expression Libraries with Antibodies	132
8.3	Rescreening	135
8.4	Subcloning	136
8.5	Characterization of Plasmid Clones	137
8.5.1	Restriction digests and agarose gel electrophoresis	138
8.5.2	Southern blots	139
8.5.3	PCR and sequence analysis	140
<b>9</b>	<b>Polymerase Chain Reaction (PCR)</b>	<b>143</b>
9.1	The PCR Reaction	144
9.2	PCR in Practice	148
9.2.1	Optimization of the PCR reaction	149
9.2.2	Analysis of PCR products	149
9.3	Cloning PCR Products	151
9.4	Long-range PCR	152
9.5	Reverse-transcription PCR	153
9.6	Rapid Amplification of cDNA Ends (RACE)	154
9.7	Applications of PCR	157
9.7.1	PCR cloning strategies	157
9.7.2	Analysis of recombinant clones and rare events	159
9.7.3	Diagnostic applications	159
<b>10</b>	<b>DNA Sequencing</b>	<b>161</b>
10.1	Principles of DNA Sequencing	161
10.2	Automated Sequencing	165
10.3	Extending the Sequence	166
10.4	Shotgun Sequencing: Contig Assembly	167
10.5	Genome Sequencing	169
10.5.1	Overview	169
10.5.2	Strategies	172
10.5.3	Repetitive elements and gaps	173

---

<b>11</b>	<b>Analysis of Sequence Data</b>	<b>177</b>
11.1	Analysis and Annotation	177
11.1.1	Open reading frames	177
11.1.2	Exon/intron boundaries	181
11.1.3	Identification of the function of genes and their products	182
11.1.4	Expression signals	184
11.1.5	Other features of nucleic acid sequences	185
11.1.6	Protein structure	188
11.1.7	Protein motifs and domains	190
11.2	Databanks	192
11.3	Sequence Comparisons	195
11.3.1	DNA sequences	195
11.3.2	Protein sequence comparisons	199
11.3.3	Sequence alignments: CLUSTAL	206
<b>12</b>	<b>Analysis of Genetic Variation</b>	<b>209</b>
12.1	Nature of Genetic Variation	209
12.1.1	Single nucleotide polymorphisms	210
12.1.2	Large-scale variations	212
12.1.3	Conserved and variable domains	212
12.2	Methods for Studying Variation	214
12.2.1	Genomic Southern blot analysis – restriction fragment length polymorphisms (RFLPs)	214
12.2.2	PCR-based methods	217
12.2.3	Genome-wide comparisons	222
<b>13</b>	<b>Analysis of Gene Expression</b>	<b>227</b>
13.1	Analysing Transcription	227
13.1.1	Northern blots	228
13.1.2	RNase protection assay	229
13.1.3	Reverse transcription PCR	231
13.1.4	<i>In situ</i> hybridization	234
13.1.5	Primer extension assay	235
13.2	Comparing Transcriptomes	236
13.2.1	Differential screening	237
13.2.2	Subtractive hybridization	238
13.2.3	Differential display	240
13.2.4	Array-based methods	241
13.3	Methods for Studying the Promoter	244
13.3.1	Reporter genes	244
13.3.2	Locating the promoter	245
13.3.3	Using reporter genes to study regulatory RNA elements	248
13.3.4	Regulatory elements and DNA-binding proteins	248
13.3.5	Run-on assays	252
13.4	Translational Analysis	253
13.4.1	Western blots	253

13.4.2 Immunocytochemistry and immunohistochemistry	254
13.4.3 Two-dimensional electrophoresis	255
13.4.4 Proteomics	256
<b>14 Analysis of Gene Function</b>	<b>259</b>
14.1 Relating Genes and Functions	259
14.2 Genetic Maps	259
14.2.1 Linked and unlinked genes	259
14.3 Relating Genetic and Physical Maps	262
14.4 Linkage Analysis	263
14.4.1 Ordered libraries and chromosome walking	264
14.5 Transposon Mutagenesis	265
14.5.1 Transposition in <i>Drosophila</i>	268
14.5.2 Other applications of transposons	270
14.6 Allelic Replacement and Gene Knock-outs	272
14.7 Complementation	274
14.8 Studying Gene Function through Protein Interactions	274
14.8.1 Two-hybrid screening	275
14.8.2 Phage display libraries	276
<b>15 Manipulating Gene Expression</b>	<b>279</b>
15.1 Factors Affecting Expression of Cloned Genes	280
15.2 Expression of Cloned Genes in Bacteria	284
15.2.1 Transcriptional fusions	284
15.2.2 Stability: conditional expression	286
15.2.3 Expression of lethal genes	289
15.2.4 Translational fusions	290
15.3 Expression in Eukaryotic Host Cells	292
15.3.1 Yeast expression systems	293
15.3.2 Expression in insect cells: baculovirus systems	294
15.3.3 Expression in mammalian cells	296
15.4 Adding Tags and Signals	297
15.4.1 Tagged proteins	297
15.4.2 Secretion signals	298
15.5 <i>In vitro</i> Mutagenesis	299
15.5.1 Site-directed mutagenesis	300
15.5.2 Synthetic genes	303
15.5.3 Assembly PCR	304
15.5.4 Protein engineering	304
<b>16 Medical Applications, Present and Future</b>	<b>307</b>
16.1 Vaccines	307
16.1.1 Subunit vaccines	309
16.1.2 Live attenuated vaccines	310
16.1.3 Live recombinant vaccines	312
16.1.4 DNA vaccines	314

16.2	Detection and Identification of Pathogens	315
16.3	Human Genetic Diseases	316
16.3.1	Identifying disease genes	316
16.3.2	Genetic diagnosis	319
16.3.3	Gene therapy	320
<b>17</b>	<b>Transgenics</b>	<b>325</b>
17.1	Transgenesis and Cloning	325
17.2	Animal Transgenesis and its Applications	326
17.2.1	Expression of transgenes	328
17.2.2	Embryonic stem-cell technology	330
17.2.3	Gene knock-outs	333
17.2.4	Gene knock-in technology	334
17.2.5	Applications of transgenic animals	334
17.3	Transgenic Plants and their Applications	335
17.3.1	Gene subtraction	337
17.4	Summary	338
	<b>Bibliography</b>	<b>339</b>
	<b>Glossary</b>	<b>341</b>
	<b>Index</b>	<b>353</b>

# 1 Introduction

This book is about the study and manipulation of nucleic acids, and how this can be used to answer biological questions. Although we hear a lot about the commercial applications, in particular (at the moment) the genetic modification of plants, the real revolution lies in the incredible advances in our understanding of how cells work. Until about 30 years ago, genetics was a patient and laborious process of selecting variants (whether of viruses, bacteria, plants or animals), and designing breeding experiments that would provide data on how the genes concerned were inherited. The study of human genetics proceeded even more slowly, because of course you could only study the consequences of what happened naturally. Then, in the 1970s, techniques were discovered that enabled us to cut DNA precisely into specific fragments, and join them together again in different combinations. For the first time it was possible to isolate and study specific genes. Since this applied equally to human genes, the impact on human genetics was particularly marked. In parallel with this, *hybridization* techniques were developed that enabled the identification of specific DNA sequences, and (somewhat later) methods were introduced for determining the sequence of these bits of DNA. Combining those advances with automated techniques and the concurrent advance in computer power has led to the determination of the full sequence of the human genome.

This revolution does not end with understanding how genes work and how the information is inherited. Genetics, and especially modern molecular genetics, underpins all the biological sciences. By studying, and manipulating, specific genes, we develop our understanding of the way in which the products of those genes interact to give rise to the properties of the organism itself. This could range from, for example, the mechanism of motility in bacteria to the causes of human genetic diseases and the processes that cause a cell to grow uncontrollably giving rise to a tumour. In many cases, we can identify precisely the cause of a specific property. We can say that a change in one single base in the genome of a bacterium will make it resistant to a certain antibiotic, or that a change in one base in human DNA could cause debilitating disease. This only scratches the surface of the power of these techniques, and indeed this book can only provide an introduction to them. Nevertheless, we hope that by the time

you have studied it, you will have some appreciation of what can be (and indeed has been) achieved.

Genetic manipulation is traditionally divided into *in vitro* and *in vivo* work. Traditionally, investigators will first work *in vitro*, using enzymes derived from various organisms to create a *recombinant DNA molecule* in which the DNA they want to study is joined to a *vector*. This recombinant vector molecule is then processed *in vivo* inside a *host* organism, more often than not a strain of the *Escherichia coli* (*E. coli*) bacterium. A *clone* of the host carrying the foreign DNA is grown, producing a great many identical copies of the DNA, and sometimes its products as well. Today, in many cases the *in vivo* stage is bypassed altogether by the use of PCR (polymerase chain reaction), a method which allows us to produce many copies of our DNA *in vitro* without the help of a host organism.

In the early days, *E. coli* strains carrying recombinant DNA molecules were treated with extreme caution. *E. coli* is a bacterium which lives in its billions within our digestive system, and those of other mammals, and which will survive quite easily in our environment, for instance in our food and on our beaches. So there was a lot of concern that the introduction of foreign DNA into *E. coli* would generate bacteria with dangerous properties. Fortunately, this is one fear that has been shown to be unfounded. Some natural *E. coli* strains *are* pathogenic – in particular the O157:H7 strain which can cause severe disease or death. By contrast, the strains used for genetic manipulation are harmless disabled laboratory strains that will not even survive in the gut. Working with genetically modified *E. coli* can therefore be done very safely (although work with *any* bacterium has to follow some basic safety rules). However, the most commonly used type of vector, plasmids, are shared readily between bacteria; the transmission of plasmids between bacteria is behind much of the natural spread of antibiotic resistance. What if our recombinant plasmids were transmitted to other bacterial strains that *do* survive on their own? This, too, has turned out not to be a worry in the majority of cases. The plasmids themselves have been manipulated so that they cannot be readily transferred to other bacteria. Furthermore, carrying a gene such as that coding for, say, dogfish insulin, or an artificial chromosome carrying 100 000 bases of human genomic DNA is a great burden to an *E. coli* cell, and carries no reward whatsoever. In fact, in order to make them accept it, we have to create conditions that will kill all bacterial cells *not* carrying the foreign gene. If you fail to do so when you start your culture in the evening, you can be sure that your bacteria will have dropped the foreign gene the next morning. Evolution in progress!

Whilst nobody today worries about genetically modified *E. coli*, and indeed diabetics have been injecting genetically modified insulin produced by *E. coli* for decades, the issue of genetic engineering is back on the public agenda, this time pertaining to higher organisms. It is important to distinguish the *genetic*



*modification* of plants and animals from *cloning* plants and animals. The latter simply involves the production of genetically identical individuals; it does not involve any genetic modification whatsoever. (The two technologies can be used in tandem, but that is another matter.) So, we will ignore the cloning of higher organisms here. Although it is conceptually very similar to producing a clone of a genetically modified *E. coli*, it is really a matter of reproductive cell biology, and frankly relatively uninteresting from the molecular point of view. By contrast, the genetic modification of higher organisms is both conceptually similar to the genetic modification of bacteria, and also very pertinent as it is a potential and, in principle, fairly easy application following the isolation and analysis of a gene.

At the time of writing, the ethical and environmental consequences of this application are still a matter of vivid debate and media attention, and it would be very surprising if this is not still continuing by the time you read this. Just as in the laboratory, the genetic modification as such is not necessarily the biggest risk here. Thus, if a food crop carries a gene that makes it tolerant of herbicides (weedkillers), it would seem reasonable to worry more about increased levels of herbicides in our food than about the genetic modification itself. Equally, the worry about such an organism escaping into the wild may turn out to be exaggerated. Just as, without an evolutionary pressure to keep the genetic modification, our *E. coli* in the example above died out overnight, it appears quite unlikely that a plant that wastes valuable resources on producing a protein that protects it against herbicides will survive long in the wild in the absence of herbicide use.

Nonetheless, this issue is by no means as clear-cut as that of genetically modified bacteria. We cannot test these organisms in a contained laboratory. They take months or a year to produce each generation, not 20 minutes as *E. coli* does. And even if they should be harmless in themselves, there are other issues as well, such as the one exemplified above. Thus, this is an important and complicated issue, and to understand it fully you need to know about evolution, ecology, food chemistry, nutrition, and molecular biology. We hope that reading this book will be of some help for the last of these. We also hope that it will convey some of the wonder, excitement, and intellectual stimulation that this science brings to its practitioners. What better way to reverse the boredom of a long journey than to indulge in the immense satisfaction of constructing a clever new screening algorithm? Who needs jigsaw and crossword puzzles when you can figure out a clever way of joining two DNA fragments together? And how can you ever lose the fascination you feel about the fact that the drop of enzyme that you're adding to your test tube is about to manipulate the DNA molecules in it with surgical precision?



# 2 Basic Molecular Biology

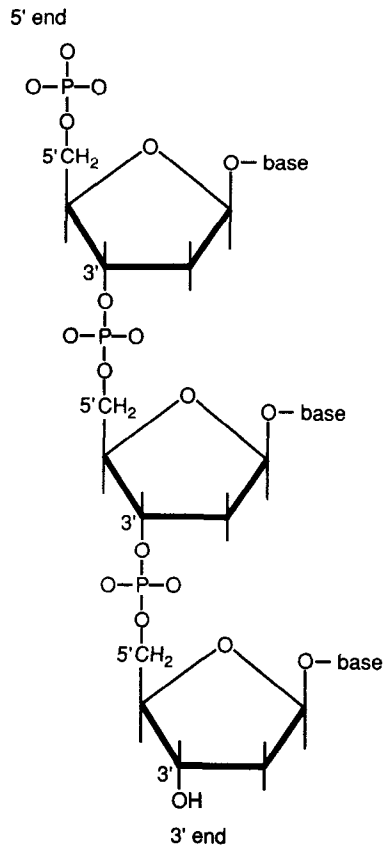
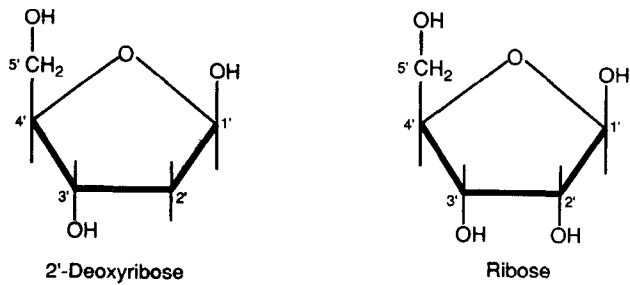
In this book, we assume you already have a working knowledge of the basic concepts of molecular biology. This chapter serves as a reminder of the key aspects of molecular biology that are especially relevant to this book.

## 2.1 Nucleic Acid Structure

### 2.1.1 The DNA backbone

Manipulation of nucleic acids in the laboratory is based on their physical and chemical properties, which in turn are reflected in their biological function. Intrinsically, DNA is a very stable molecule. Scientists routinely send DNA samples in the post without worrying about refrigeration. Indeed, DNA of high enough quality to be cloned has been recovered from frozen mammoths and mummified Pharaohs thousands of years old. This stability is provided by the robust repetitive phosphate-sugar backbone in each DNA strand, in which the phosphate links the 5' position of one sugar to the 3' position of the next (Figure 2.1). The bonds between these phosphorus, oxygen, and carbon atoms are all *covalent bonds*. Controlled degradation of DNA requires enzymes (nucleases) that break these covalent bonds. These are divided into *endonucleases*, which attack internal sites in a DNA strand, and *exonucleases*, which nibble away at the ends. We can for the moment ignore other enzymes that attack for example the bonds linking the bases to the sugar residues. Some of these enzymes are non-specific, and lead to a generalized destruction of DNA. It was the discovery of *restriction endonucleases* (or *restriction enzymes*), which cut DNA strands at specific positions, that opened up the possibility of *recombinant DNA technology* ('*genetic engineering*'), coupled with *DNA ligases*, which can join two double-stranded DNA molecules together.

RNA molecules, which contain the sugar ribose (Figure 2.2), rather than the deoxyribose found in DNA, are less stable than DNA. This is partly due to their greater susceptibility to attack by nucleases (*ribonucleases*), but they are also more susceptible to chemical degradation, especially by alkaline conditions.

**Figure 2.1** DNA backbone**Figure 2.2** Nucleic acid sugars

### 2.1.2 The base pairs

In addition to the sugar (2'deoxyribose) and phosphate, DNA molecules contain four nitrogen-containing bases (Figure 2.3): two pyrimidines, thymine (T) and cytosine (C), and two purines, guanine (G) and adenine (A). (Other bases can be incorporated into synthetic DNA in the laboratory, and sometimes other bases occur naturally.) Since the purines are bigger than the pyrimidines, a regular double helix requires a purine in one strand to be matched by a pyrimidine in the other. Furthermore, the regularity of the double helix requires specific hydrogen bonding between the bases so that they fit together, with an A opposite a T, and a G opposite a C (Figure 2.4). We refer to these pairs of bases as *complementary*, and hence to one strand as the *complement* of the other. Note that the two DNA strands run in opposite directions. In a conventional representation of a double-stranded sequence the 'top' strand has a 5' hydroxyl group at the left-hand end (and is said to be written in the 5' to 3' direction), while the 'bottom' strand has its 5' end at the right-hand end. Since the two strands are complementary, there is no information in the second strand that cannot be deduced from the first one. Therefore, to save space, it is common to represent a double-stranded DNA sequence by showing the sequence of only one strand. When only one strand is

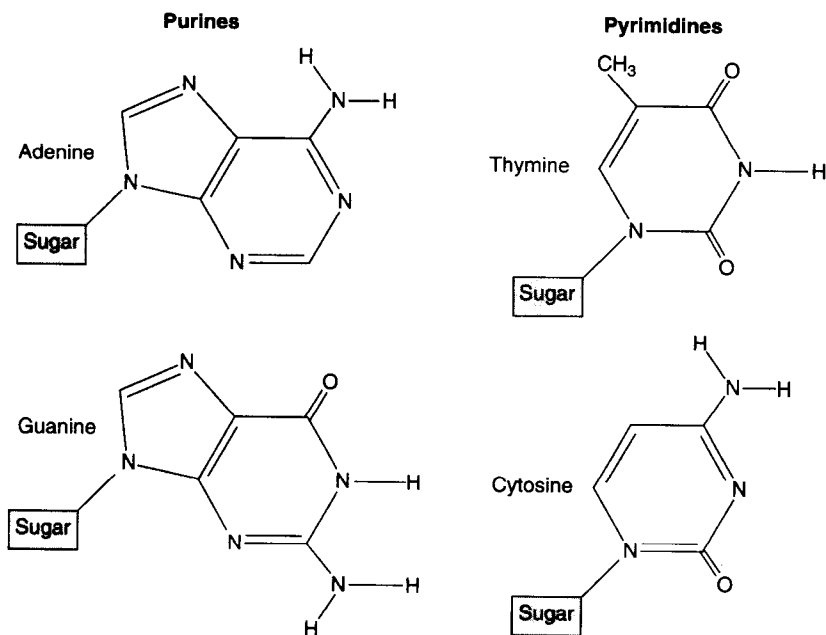
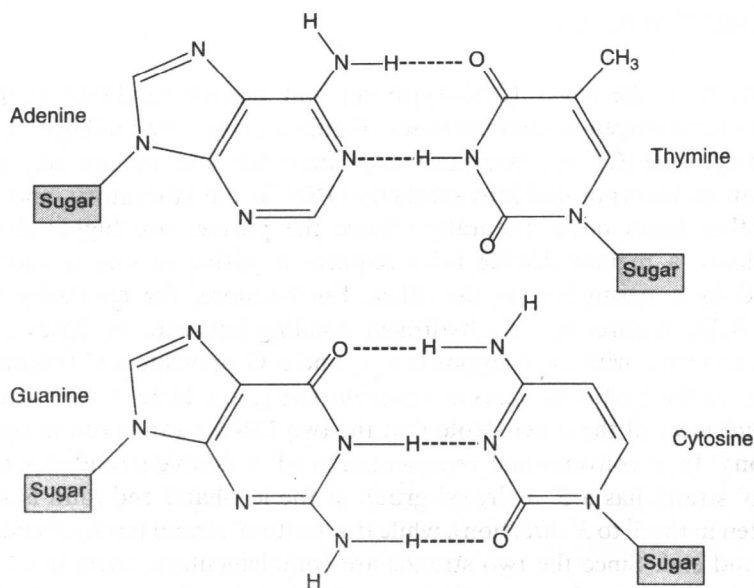


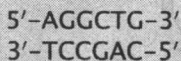
Figure 2.3 Nucleic acid bases



**Figure 2.4** Base-pairing in DNA

### Box 2.1 Complementary sequences

DNA sequences are often represented as the sequence of just one of the two strands, in the 5' to 3' direction, reading from left to right. Thus the double-stranded DNA sequence



would be shown as AGGCTG, with the orientation (i.e., the position of the 5' and 3' ends) being inferred.

To get the sequence of the other (complementary) strand, you must not only change the A and G residues to T and C (and vice versa), but you must also reverse the order.

So in this example, the complement of AGGCTG is CAGCCT, reading the lower strand from right to left (again in the 5' to 3' direction).

shown, we use the 5' to 3' direction; the sequence of the second strand is inferred from that, and you have to remember that the second strand runs in the opposite direction. Thus a single strand sequence written as AGGCTG (or more fully 5'AGGCTG3') would have as its complement CAGCCT (5'CAGCCT3') (see Box 2.1).

Thanks to this base-pairing arrangement, the two strands can be safely separated – both in the cell and in the test tube – under conditions which disrupt the hydrogen bonds between the bases but are much too mild to pose any threat to the covalent bonds in the backbone. This is referred to as *denaturation* of DNA and, unlike the denaturation of many proteins, it is reversible. Because of the complementarity of the base pairs, the strands will easily join together again and *renature*. In the test tube, DNA is readily denatured by heating, and the denaturation process is therefore often referred to as *melting* even when it is accomplished by means other than heat (e.g. by NaOH). Denaturation of a double-stranded DNA molecule occurs over a short temperature range, and the midpoint of that range is defined as the *melting temperature* ( $T_m$ ). This is influenced by the base composition of the DNA. Since guanine:cytosine (GC) base pairs have three hydrogen bonds, they are stronger (i.e. melt less easily) than adenine:thymine (AT) pairs, which have only two hydrogen bonds. It is therefore possible to estimate the melting temperature of a DNA fragment if you know the sequence (or the base composition and length). These considerations are important in understanding the technique known as *hybridization*, in which *gene probes* are used to detect specific nucleic acid sequences. We will look at hybridization in more detail in Chapter 8.

Although the normal base pairs (A–T and G–C) are the only forms that are fully compatible with the Watson–Crick double helix, pairing of other bases can occur, especially in situations where a regular double helix is less important (such as the folding of single-stranded nucleic acids into secondary structures – see below).

In addition to the hydrogen bonds, the double stranded DNA structure is maintained by *hydrophobic interactions* between the bases. The hydrophobic nature of the bases means that a single-stranded structure, in which the bases are exposed to the aqueous environment, is unstable. Pairing of the bases enables them to be removed from interaction with the surrounding water. In contrast to the hydrogen bonding, hydrophobic interactions are relatively non-specific. Thus, nucleic acid strands will tend to stick together even in the absence of specific base-pairing, although the specific interactions make the association stronger. The specificity of the interaction can therefore be increased by the use of chemicals (such as formamide) that reduce the hydrophobic interactions.

What happens if there is only a single nucleic acid strand? This is normally the case with RNA, but single-stranded forms of DNA also exist. For example, in some viruses the genetic material is single-stranded DNA. A single-stranded nucleic acid molecule will tend to fold up on itself to form localized double-stranded regions, including structures referred to as hairpins or stem-loop structures. This has the effect of removing the bases from the surrounding water. At room temperature, in the absence of denaturing agents,

a single-stranded nucleic acid will normally consist of a complex set of such localized secondary structure elements, which is especially evident with RNA molecules such as transfer RNA (tRNA) and ribosomal RNA (rRNA). This can also happen to a limited extent with double stranded DNA, where short sequences can tend to loop out of the regular double helix. Since this makes it easier for enzymes to unwind the DNA, and to separate the strands, these sequences can play a role in the regulation of gene expression, and in the initiation of DNA replication.

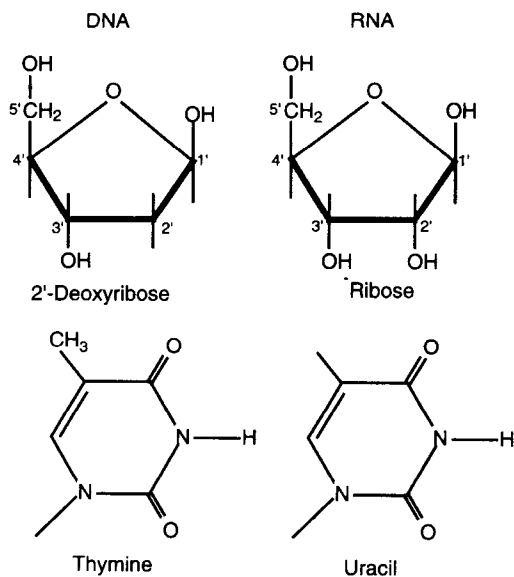
A further factor to be taken into account is the negative charge on the phosphate groups in the nucleic acid backbone. This works in the opposite direction to the hydrogen bonds and hydrophobic interactions; the strong negative charge on the DNA strands causes electrostatic repulsion that tends to repel the two strands. In the presence of salt, this effect is counteracted by the presence of a cloud of counterions surrounding the molecule, neutralizing the negative charge on the phosphate groups. However, if you reduce the salt concentration, any weak interactions between the strands will be disrupted by electrostatic repulsion – and therefore we can use low salt conditions to increase the specificity of hybridization (see Chapter 8).

### 2.1.3 RNA structure

Chemically, RNA is very similar to DNA. The fundamental chemical difference is that the RNA backbone contains ribose rather than the 2'-deoxyribose (i.e. ribose without the hydroxyl group at the 2' position) present in DNA (Figure 2.5). However, this slight difference has a powerful effect on some properties of the nucleic acid, especially on its stability. Thus, RNA is readily destroyed by exposure to high pH. Under these conditions, DNA is stable: although the strands will separate, they will remain intact and capable of renaturation when the pH is lowered again. A further difference between RNA and DNA is that the former contains uracil rather than thymine (Figure 2.5).

Generally, while most of the DNA we use is double stranded, most of the RNA we encounter consists of a single polynucleotide strand – although we must remember the comments above regarding the folding of single-stranded nucleic acids. However, this distinction between RNA and DNA is not an inherent property of the nucleic acids themselves, but is a reflection of the natural roles of RNA and DNA in the cell, and of the method of production. In all *cellular* organisms (i.e. excluding viruses), DNA is the inherited material responsible for the genetic composition of the cell, and the replication process that has evolved is based on a double-stranded molecule; the roles of RNA in the cell do not require a second strand, and indeed the presence of a second, complementary, strand would preclude its role in protein synthesis. However, there are some viruses that have double-stranded RNA as their genetic material,





**Figure 2.5** Differences between DNA and RNA

as well as some with single-stranded RNA, and some viruses (as well as some plasmids) replicate via single-stranded DNA forms.

### 2.1.4 Nucleic acid synthesis

We do not need to consider all the details of how nucleic acids are synthesized. The basic features that we need to remember are summarized in Figure 2.6, which shows the addition of a nucleotide to the growing end (3'-OH) of a DNA strand. The substrate for this reaction is the relevant deoxynucleotide triphosphate (dNTP), i.e. the one that makes the correct base-pair with the corresponding residue on the template strand. The DNA strand is always extended at the 3'-OH end. For this reaction to occur it is essential that the residue at the 3'-OH end, to which the new nucleotide is to be added, is accurately base-paired with its partner on the other strand.

RNA synthesis occurs in much the same way, as far as this description goes, except that of course the substrates are nucleotide triphosphates (NTPs) rather than the deoxynucleotide triphosphates (dNTPs). There is one very important difference though. DNA synthesis only occurs by extension of an existing strand – it always needs a *primer* to get it started. RNA polymerases on the other hand are capable of starting a new RNA strand from scratch, given the appropriate signals.