Contents

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.

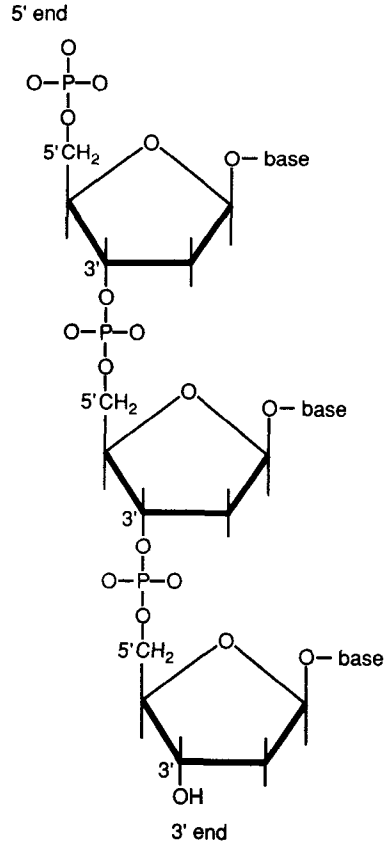

Figure 2.1 DNA backbone

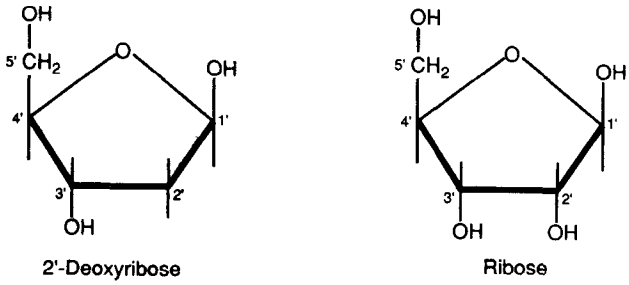

Figure 2.2 Nucleic acid sugars

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

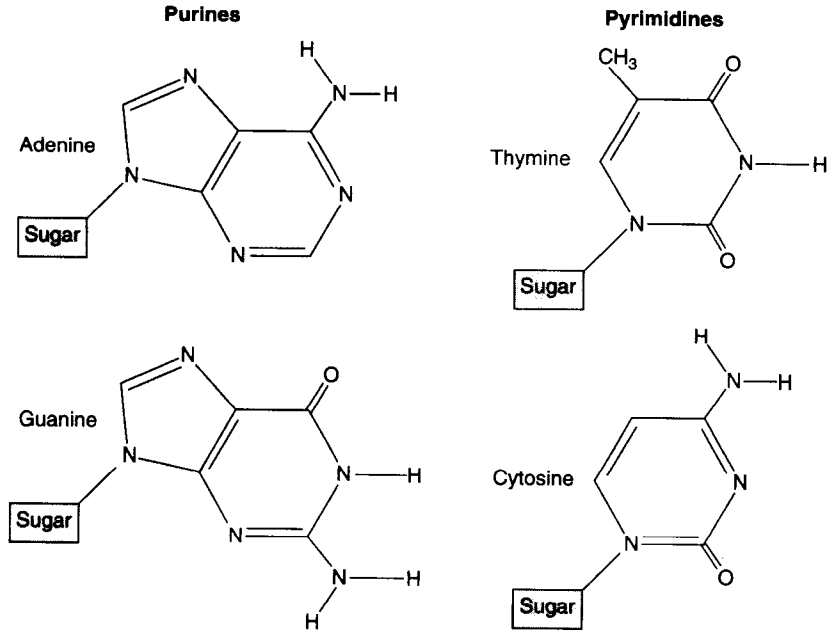

Figure 2.3 Nucleic acid bases

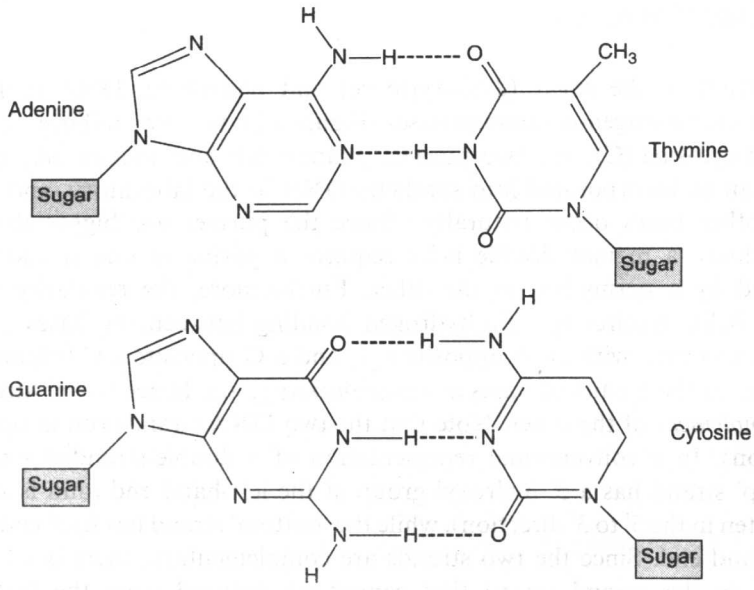

Figure 2.4 Base-pairing in DNA

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

5′-AGGCTG-3′
3′-TCCGAC-5′

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 byexposure 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,

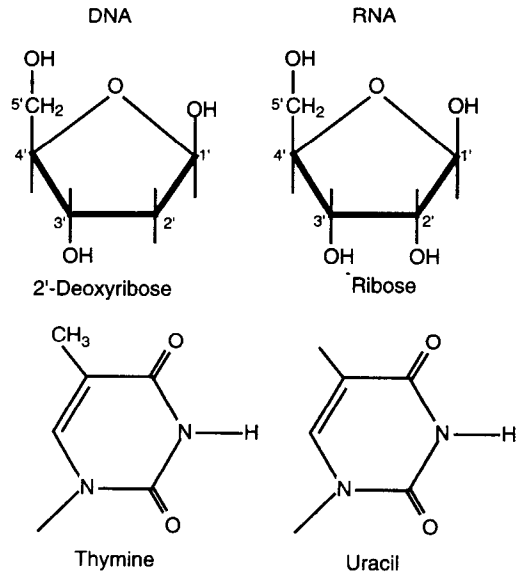

Figure 2.5 Differences between DNA and RNA

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.