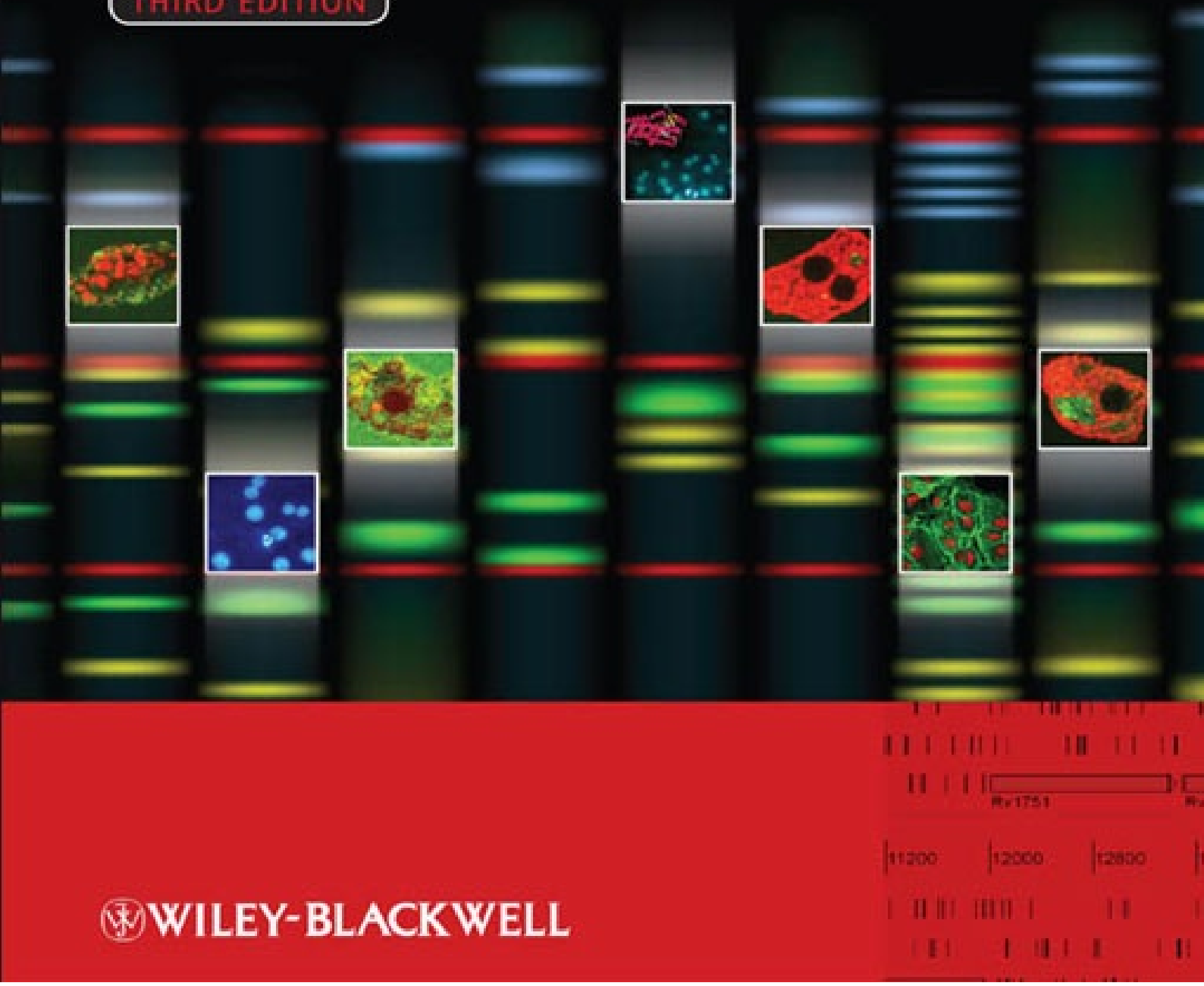
JEREMY W. DALE | MALCOLM VON SCHANTZ | NICK PLANT

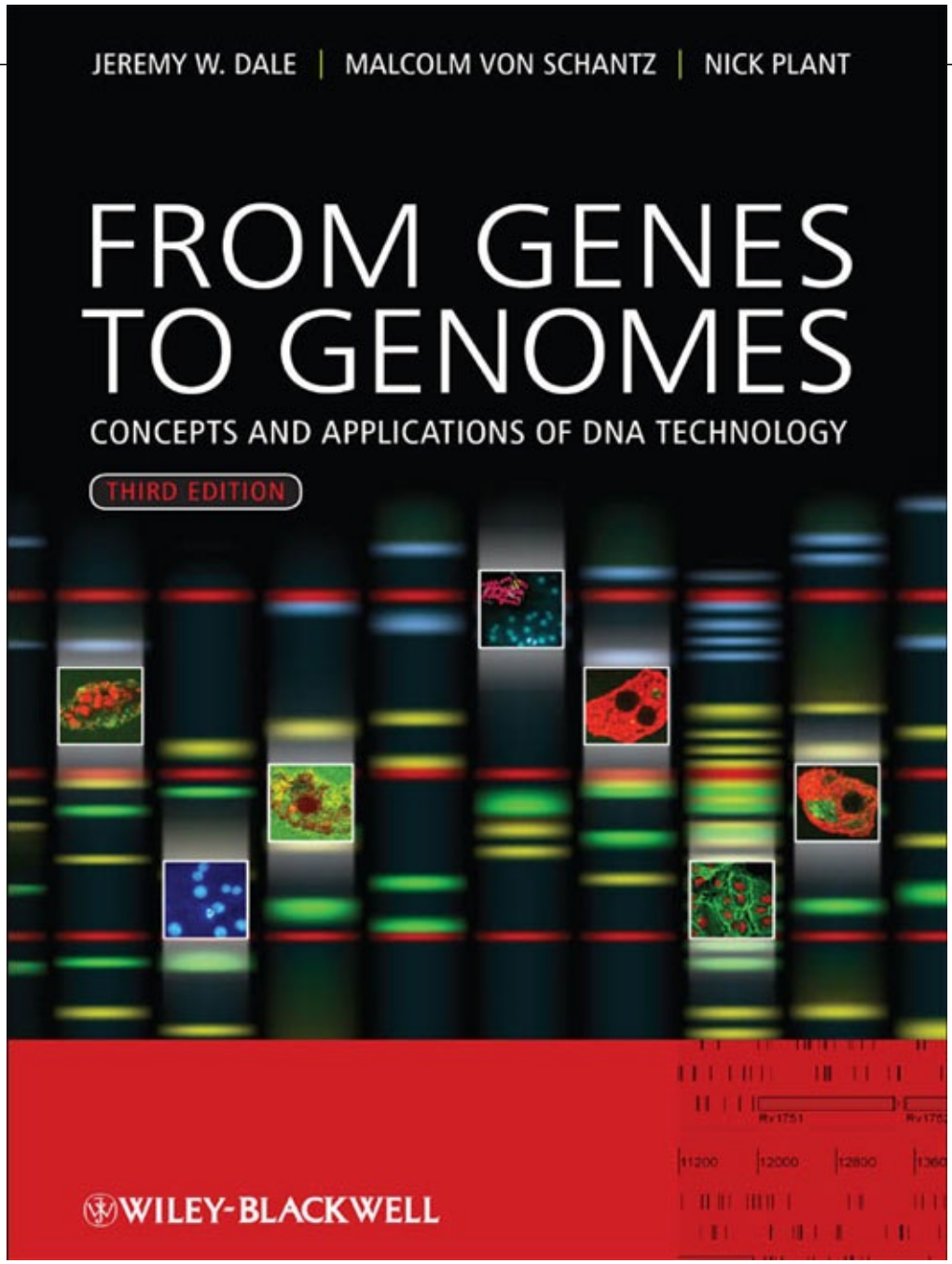
FROM GENES TO GENOMES

CONCEPTS AND APPLICATIONS OF DNA TECHNOLOGY

THIRD EDITION



WILEY-BLACKWELL


JEREMY W. DALE | MALCOLM VON SCHANTZ | NICK PLANT
FROM GENES TO GENOMES
CONCEPTS AND APPLICATIONS OF DNA TECHNOLOGY
THIRD EDITION
WILEY-BLACKWELL

Contents

From Genes to Genomes

Third Edition

Concepts and Applications of DNA Technology

Jeremy W. Dale, Malcolm von Schantz and **Nick Plant**

University of Surrey, UK

WILEY-BLACKWELL

A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial offices
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data

Dale, Jeremy, Professor.
From genes to genomes : concepts and applications of DNA technology / Jeremy W. Dale, Malcolm von Schantz, and Nick Plant. – 3rd ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-68386-6 (cloth) – ISBN 978-0-470-68385-9 (pbk.)
I. Schantz, Malcolm von. II. Plant, Nick. III. Title.
[DNLM: 1. Genetic Engineering. 2. Cloning, Molecular. 3. DNA, Recombinant. QU 450]
LC classification not assigned
660.6′5–dc23

2011030219

A catalogue record for this book is available from the British Library.

This book is published in the following electronic formats: ePDF 9781119953159; ePub 9781119954279; Mobi 9781119954286

Preface

The first edition of this book was published in 2002. By the time of the second edition (2007) th emphasis had moved away from just cloning genes, to embrace a wider range of technologie especially genome sequencing, the polymerase chain reaction and microarray technology. Th revolution has continued unabated, indeed even accelerating, not least with the advent of hig throughput genome sequencing. In this edition we have tried to introduce readers to the exciteme engendered by the latest developments – but this poses a considerable challenge. Our aim has been keep the book to an accessible size, so including newer technologies inevitably means discardi some of the older ones. Some might maintain that we could have gone further in that direction. Son methods that have been kept are no longer as important as they once were, and maybe there is a element of sentimentality in keeping them – but there is some virtue in retaining a balance so that v can maintain a degree of historical perspective. There is a need to understand, to some extent, how v got to the position we are now in, as well as trying to see where we are going.

The main title of the book, *From Genes to Genomes*, is derived from the progress of this revolutio It also indicates a recurrent theme within the book, in that the earlier chapters deal with analysis ar investigation at the level of individual genes, and then later on we move towards genome-wide studi – ending up with a chapter directed at the whole organism.

Dealing only with the techniques, without the applications, would be rather dry. Some of th applications are obvious – recombinant product formation, genetic diagnosis, transgenic plants ar animals, and so on – and we have attempted to introduce these to give you a flavour of the advanc that continue to be made, but at the same time without burdening you with excessive detail. Equal important, possibly more so, are the contributions made to the advance of fundamental knowledge areas such as developmental studies and molecular phylogeny.

The purpose of this book is to provide an introduction to the concepts and applications of th rapidly moving and fascinating field. In writing it, we had in mind its usefulness for undergradua students in the biological and biomedical sciences (who we assume will have a basic grounding molecular biology). However, it will also be relevant for many others, ranging from research worke and teachers who want to update their knowledge of related areas to anyone who would like understand rather more of the background to current controversies about the applications of some these techniques.

Jeremy W. Da
Malcolm von Schan
Nick Pla

From Genes to Genomes

1.1 Introduction

The classical approach to genetics starts with the identification of variants that have a specif *phenotype*, i.e., they differ from the *wildtype* in some way that can be seen (or detected in other way and defined. For Gregor Mendel, the father of modern genetics, this was the appearance of his pe (e.g., green versus yellow, or round versus wrinkled). One of the postulates he arrived at was that the characteristics assorted independently of one another. For example, when crossing one type of pea th produces yellow, wrinkled peas with another that produces green, round peas, the first generation (F are all round and yellow (because round is dominant over wrinkled, and yellow is dominant ov green). In the second (F_2) generation, there is a 3 : 1 mixture of round versus wrinkled peas, a independently a 3 : 1 mixture of yellow to green peas.

Of course Mendel did not know why this happened. We now know that if two genes are located o different chromosomes, which will segregate independently during meiosis, the genes will b distributed independently amongst the progeny. Independent assortment can also happen if the tw genes are on the same chromosome, but only if they are so far apart that any recombination betwe the homologous chromosomes will be sufficient to reassort them independently. However, if they a quite close together, recombination is less likely, and they will therefore tend to remain associate during meiosis. They will therefore be inherited together. We refer to genes that do *not* segrega independently as *linked*; the closer they are, the greater the degree of linkage, i.e., the more likely th are to stay together during meiosis. Measuring the degree of linkage (*linkage analysis*) is a centr tool in classical genetics, in that it provides a way of mapping genes, i.e., determining their relati position on the chromosome.

Bacteria and yeasts provide much more convenient systems for genetic analysis, because they gro quickly, as unicellular organisms, on defined media. You can therefore use chemical or physica mutagens (such as ultraviolet irradiation) to produce a wide range of mutations, and can sele specific mutations from very large pools of organisms – remembering that an overnight culture *Escherichia coli* will contain some 10^9 bacteria per millilitre. So we can use genetic techniques investigate detailed aspects of the physiology of such cells, including identifying the relevant genes b mapping the position of the mutations.

For multicellular organisms, the range of phenotypes is even greater, as there are then questio concerning the development of different parts of the organism, and how each individual pa influences the development of others. However, animals have much longer generation times tha bacteria, and using millions of animals (especially mammals) to identify the mutations you a interested in is logistically impossible, and ethically indefensible. Human genetics is even mo difficult as you cannot use selected breeding to map genes; you have to rely on the analysis of re families, who have chosen to breed with no consideration for the needs of science. Neverthele classical genetics has contributed extensively to the study of developmental processes, notably in th

fruit fly *Drosophila melanogaster*, where it is possible to study quite large numbers of animals, due their relative ease of housing and short generation times, and to use mutagenic agents to enhance th rate of variation.

However, these methods suffered from a number of limitations. In particular, they could only b applied, in general, to mutations that gave rise to a phenotype that could be defined in some wa including shape, physiology, biochemical properties or behaviour. Furthermore, there was no easy wa of characterizing the nature of the mutation. The situation changed radically in the 1970s with th development of techniques that enabled DNA to be cut precisely into specific fragments, and to b joined together, enzymatically –techniques that became known variously as genetic manipulatio genetic modification, genetic engineering or recombinant DNA technology. The term 'gene cloning' also used, since joining a fragment of DNA with a vector such as a plasmid that can replicate bacterial cells enabled the production of a bacterial strain (a clone) in which all the cells contained copy of this specific piece of DNA. For the first time, it was possible to isolate and study specif genes. Since such techniques could be applied equally to human genes, the impact on human geneti was particularly marked.

The revolution also depended on the development of a variety of other molecular techniques. Th earliest of these (actually predating gene cloning) was *hybridization*, which enabled the identificatio of specific DNA sequences on the basis of their sequence similarity. Later on came methods fo determining the sequence of these DNA fragments, and the polymerase chain reaction (PCR), whic provided a powerful way of amplifying specific DNA sequences. Combining those advances wi automation, plus the concurrent advance in computer power, led to the determination of the fu genome sequence of many organisms, including the human genome, and thence to enormous advanc in understanding the roles of genes and their products. In recent years, sequencing technology ha advanced to a stage where it is now a routine matter to sequence the full genome of many individual and thus attempt to pinpoint the causes of the differences between them, including some genet diseases.

Furthermore, since these techniques enabled the cloning and expression of genes from any or organism (including humans) into a more amenable host, such as a bacterium, they allowed the use genetically modified bacteria (or other hosts) for the production of human gene products, such hormones, for therapeutic use. This principle was subsequently extended to the genetic modificatio of plants and animals – both by inserting foreign genes and by knocking out existing ones – to produ plants and animals with novel properties.

As is well known, the construction and use of genetically modified organisms (GMOs) is n without controversy. In the early days, there was a lot of concern that the introduction of foreign DN into *E. coli* would generate bacteria with dangerous properties. Fortunately, this is one fear that ha been shown to be unfounded. Due to careful design, genetically modified bacteria are, generally, n well able to cope with life outside the laboratory, and hence any GM bacterium released into th environment (deliberately or accidentally) is unlikely to survive for long. In addition, one mu recognize that nature is quite capable of producing pathogenic organisms without our assistance which history, unfortunately, has repeatedly demonstrated through disease outbreaks.

The debate on GMOs has now largely moved on to issues relating to genetically modified plants an animals. 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; does not involve any genetic modification whatsoever. (The two technologies can be used in tander

but that is another matter.) There are ethical issues to be considered, but cloning plants and animals not the subject of this book.

Currently, the debate on genetic modification can be envisaged as largely revolving around tw factors: food safety and environmental impact. The first thing to be clear about is that there is n imaginable reason why genetic modification, per se, should make a foodstuff hazardous in any wa There is no reason to suppose that cheese made with rennet from a genetically modified bacterium any more dangerous than similar cheese made with 'natural' rennet. It is possible to imagine a ris associated with some genetically modified foodstuffs, due to unintended stimulation of the productio of natural toxins – remembering, for example, that potatoes are related to deadly nightshade. But th can happen equally well (or perhaps is even more likely) with conventional cross-breeding procedur for developing new strains, which are not always subject to the same degree of rigorous safety testin as GM plants.

The potential environmental impact is more difficult to assess. The main issue here is the use genetic modification to make plants resistant to herbicides or to insect attack. When such plants a grown on a large scale, it is difficult to be certain that the gene in question will not spread to relate wild plants in the vicinity (although measures can be taken to reduce this possibility), or the knock-o effect that such resistance may have on the ecosystem – if all the insects are killed, what will sma birds and animals eat? But these concerns may be exaggerated. As with the bacterial example abov these genes will not spread significantly unless there is an evolutionary pressure favouring them. S we would not expect widespread resistance to weedkillers unless the plants are being sprayed wi those weedkillers. There might be an advantage in becoming resistant to insect attack, but the insec concerned have been around for a long time, so the wild plants have had plenty of time to develo natural resistance anyway. In addition, targeted resistance in a group of plants may arguably have le environmental impact than the less targeted spraying of insecticides. We have to balance the use o genetically modified plants against the use of chemicals. If genetic modification of the plants means reduction in the use of environmentally damaging chemicals, then that is a tangible benefit that cou outweigh any theoretical risk.

The purpose of this book is to provide an introduction to the exciting developments that hav resulted in an explosion of our knowledge of the genetics and molecular biology of all forms of lif from viruses and bacteria to plants and mammals, including of course ourselves – developments th continue as we write. We hope that it will convey some of the wonder and intellectual stimulation tha this science brings to its practitioners.

1.2 Basic molecular biology

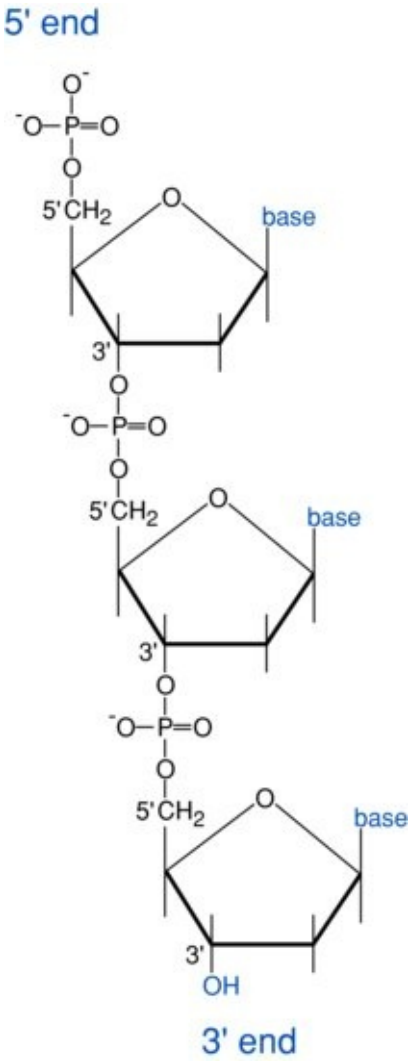
In this book, we assume you already have a working knowledge of the basic concepts of molecular an cellular biology. This section serves as a reminder of the key aspects that are especially relevant t this book.

1.2.1 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 remarkably stabl molecule. Indeed, DNA of sufficiently high quality to be analysed has been recovered from froze mammoths thousands of years old. This stability is provided by the robust phosphate–sugar backbo

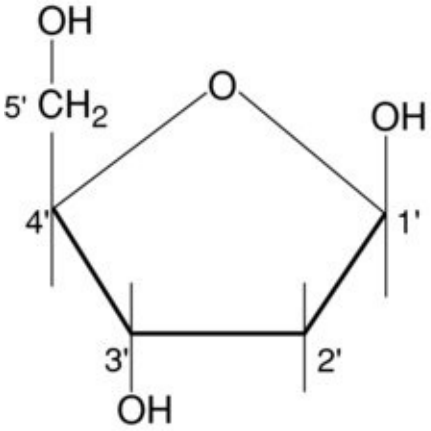
in each DNA strand, in which the phosphate links the 5′ position of one sugar to the 3′ position of th next (Figure 1.1). The bonds between these phosphorus, oxygen and carbon atoms are all *covale bonds*, meaning they are strong interactions that require significant energy to break. Hence, th controlled degradation of DNA requires enzymes (nucleases) that catalyse the breaking of thes covalent bonds. These enzymes are divided into *endonucleases*, which attack internal sites in a DN strand, and *exonucleases*, which nibble away at the ends. (We can for the moment ignore oth enzymes that attack, for example, the bonds linking the bases to the sugar residues.) Some of the enzymes are non-specific, and lead to a generalized destruction of DNA. It was the discovery *restriction endonucleases* (or *restriction enzymes*), which cut DNA strands at specific position coupled with *DNA ligases*, which can join two double-stranded DNA molecules together, that opene up the possibility of *recombinant DNA technology* (‘*genetic engineering*’).

Figure 1.1 DNA backbone.

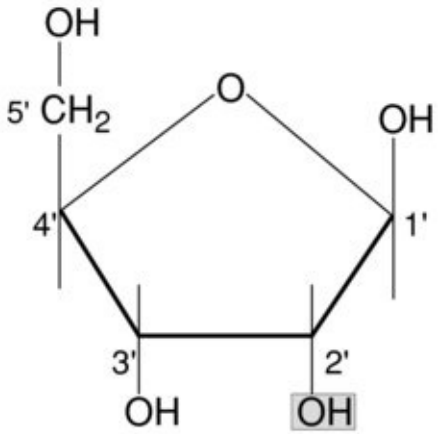


RNA molecules, which contain the sugar ribose (Figure 1.2), rather than the deoxyribose found DNA, are less stable than DNA, often surviving only minutes within the cell. They show great susceptibility to attack by nucleases (*ribonucleases*), and are also more susceptible to chemic degradation, especially by alkaline conditions.

Figure 1.2 Nucleic acid sugars.



2'-Deoxyribose



Ribose

1.2.2 The base pairs

In addition to the sugar (2′-deoxyribose) and phosphate, DNA molecules contain four nitroge containing bases (Figure 1.3): two pyrimidines, thymine (T) and cytosine (C), and two purine guanine (G) and adenine (A). It should be noted that other bases can be incorporated into synthet DNA in the laboratory, and sometimes others occur naturally, but T, C, G and A are the major DN bases. Because the purines are bigger than the pyrimidines, a regular double helix requires a purine one strand to be matched by a pyrimidine in the other. Furthermore, the regularity of the double hel requires specific hydrogen bonding between the bases so that they fit together, with an A opposite a and a G opposite a C (Figure 1.4). We refer to these pairs of bases as *complementary*, and hence each strand as the *complement* of the other.

Figure 1.3 Nucleic acid bases.

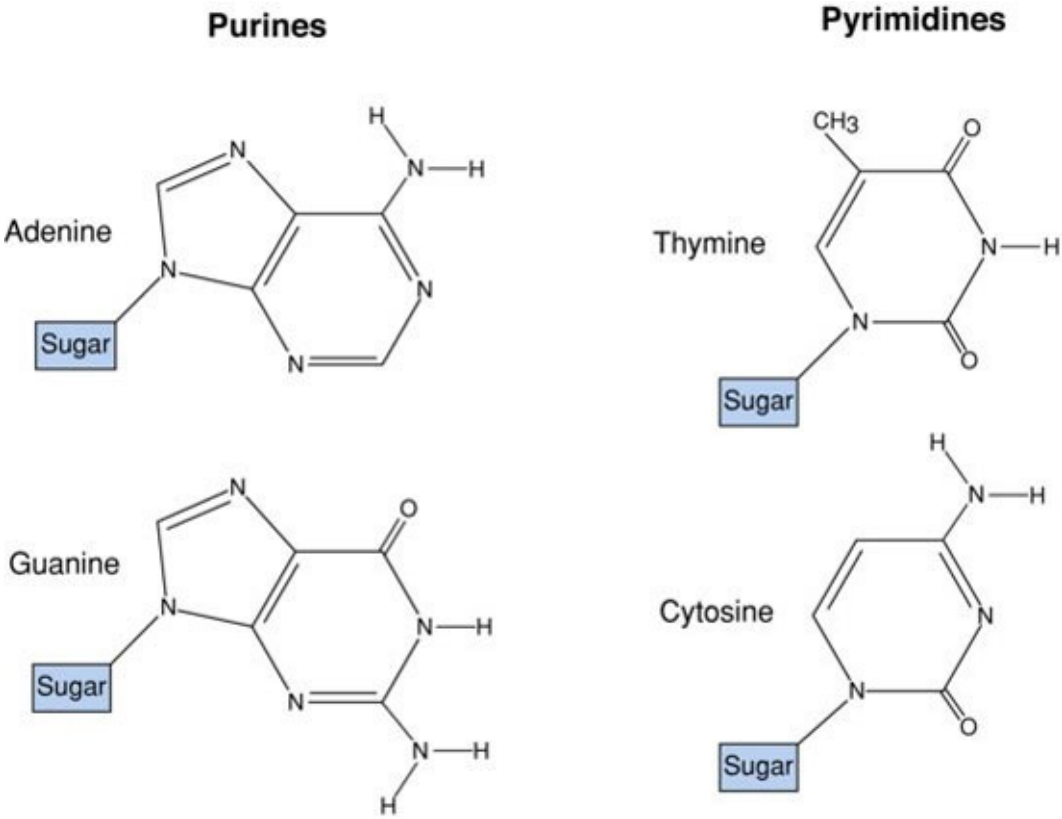
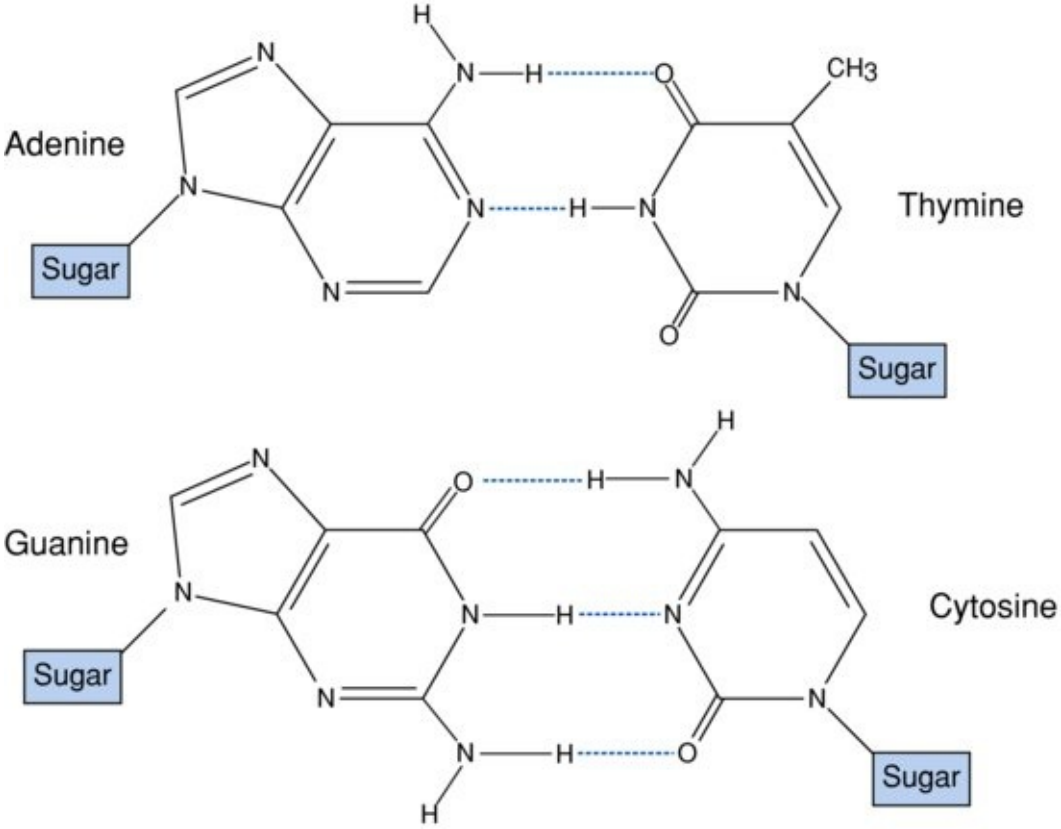


Figure 1.4 Base-pairing in DNA.



Note that the two DNA strands run in opposite directions. In a conventional representation of double-stranded sequence the ‘top’ strand has a 5′ hydroxyl group at the left-hand end (and is said be written in the 5′ to 3′ direction), while the ‘bottom’ strand has its 5′ end at the right-hand en Because the two strands are complementary, there is no information in one strand that cannot b deduced from the other one. Therefore, to save space, the convention is to represent a double-strande DNA sequence by showing the sequence of only one strand, in the 5′ to 3′ direction. The sequence the second strand is inferred from that, and you must remember that the second strand runs in th 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 1.1).

Box 1.1 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

$$5' - \mathrm{AGGCTG} - 3'$$
$$3' - \mathrm{TCCGAC} - 5'$$

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

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).

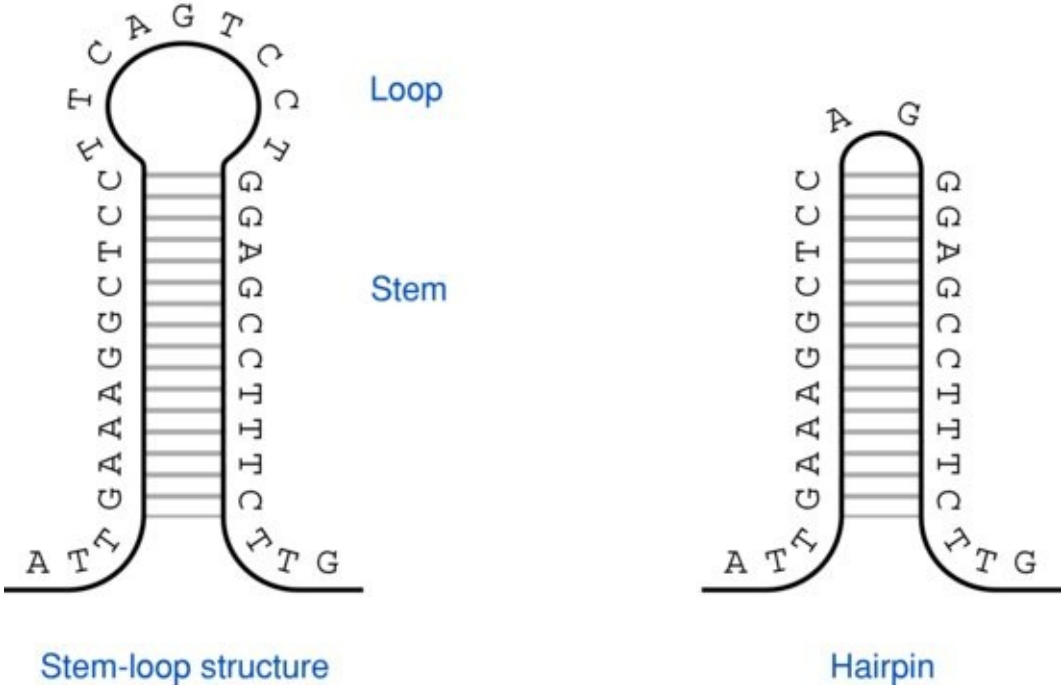
Thanks to this base-pairing arrangement, the two strands can be separated intact – both in the ce and in the test tube – under conditions that, although disrupting the relatively weak hydrogen bon that exist between the bases on complementary strands, are much too mild to pose any threat to th covalent bonds that join nucleotides within a single strand. Such a separation is referred to *denaturation* of DNA, and unlike the denaturation of many proteins it is reversible. Because of th complementarity of the base pairs, the strands will easily join together again and *renature* b

reforming their hydrogen bonds. In the test tube, DNA is readily denatured by heating, and th denaturation process is therefore often referred to as *melting*, even when it is accomplished by mea other than heat (e.g. by NaOH). Denaturation of a double-stranded DNA molecule occurs over a sho specific 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 hav three hydrogen bonds, they are stronger (i.e., melt less easily) than adenine:thymine (AT) pairs, whic have only two. It is therefore possible to estimate the melting temperature of a DNA fragment if yo know the sequence. These considerations are important in understanding the technique known *hybridization*, in which *gene probes* are used to detect specific nucleic acid sequences. We will look hybridization in more detail in Chapter 3.

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

What happens if there is only a single nucleic acid strand? This is normally the case with RNA, b single-stranded forms of DNA (ssDNA) also exist, in some viruses, for example. A single-strande nucleic acid molecule will tend to fold up on itself to form localized double-stranded region producing structures referred to as hairpins or stem-loop structures (Figure 1.5). This has the effect removing the bases from interaction with the surrounding water, again increasing stability. At roo temperature, in the absence of denaturing agents, a single-stranded nucleic acid will normally consi of a complex set of such localized secondary structure elements; this is especially evident with RN molecules.

Figure 1.5 Stem-loop and hairpin structures.



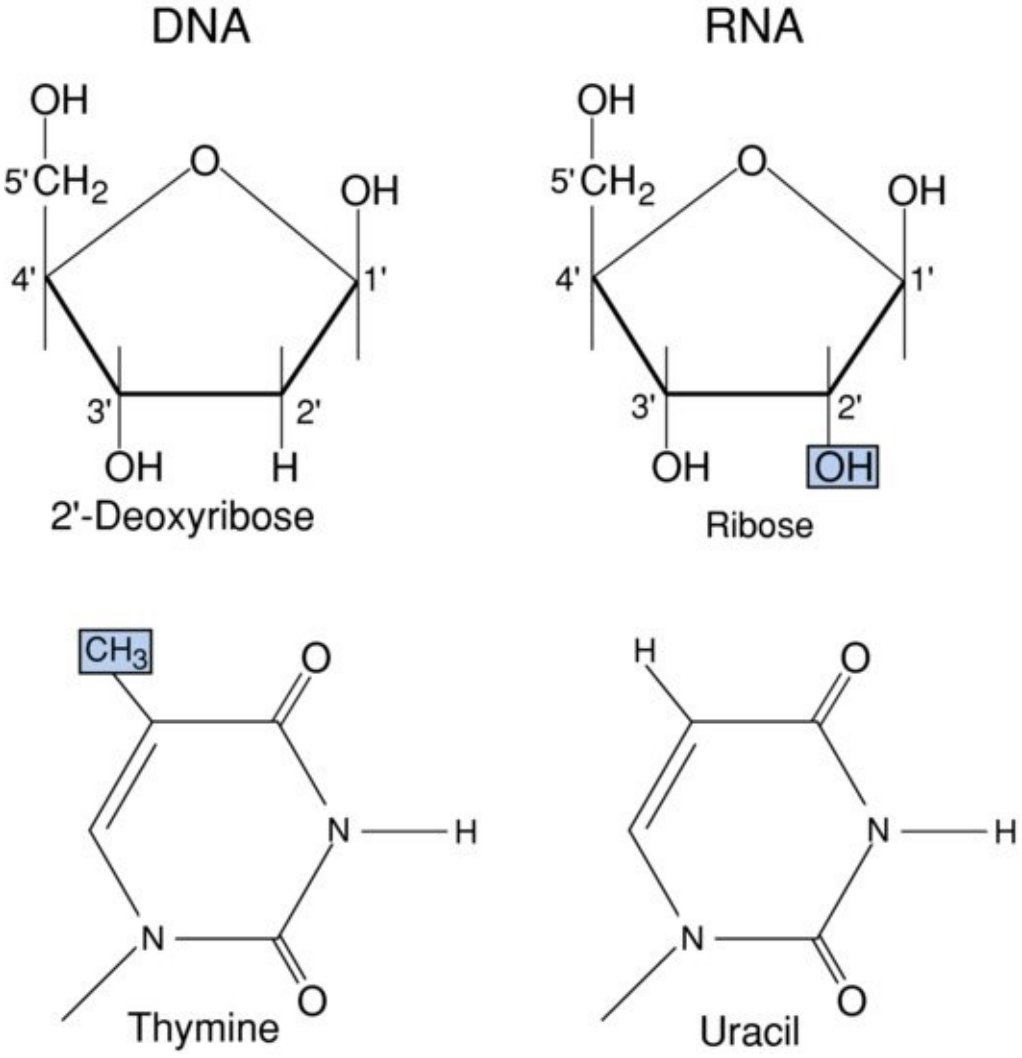
A further factor is the negative charge on the phosphate groups in the nucleic acid backbone. Th

works in the opposite direction to the hydrogen bonds and hydrophobic interactions; the stron negative charge on the DNA strands causes electrostatic repulsion that tends to make the two strand repel each other. In the presence of salt, this effect is counteracted by a cloud of counterion surrounding the molecule, neutralizing the negative charge on the phosphate groups. However, if yo reduce the salt concentration, any weak interactions between the strands will be disrupted b electrostatic repulsion. Hence, at low ionic strength, the strands will only remain together if th hydrogen bonding is strong enough, and therefore we can use low salt conditions to increase th specificity of hybridization (see Chapter 3). Of course, within the cell the salt concentration is suc that the double-stranded DNA is quite stable.

1.2.3 RNA structure

Chemically, RNA is very similar to DNA. The fundamental chemical difference is that whereas DN contains 2′-deoxyribose (i.e., ribose without the hydroxyl group at the 2′ position) in its backbone, th RNA backbone contains ribose (Figure 1.6). This slight difference has a powerful effect on som properties of the RNA molecule, especially on its stability. For example, RNA is destroyed unde alkaline conditions while DNA is stable. Although the DNA strands will separate, they will remai intact and capable of renaturation when the pH is lowered again. However, under such condition RNA will quickly be destroyed. A further difference between RNA and DNA is that the forme contains uracil rather than thymine (Figure 1.6).

Figure 1.6 Differences between DNA and RNA.

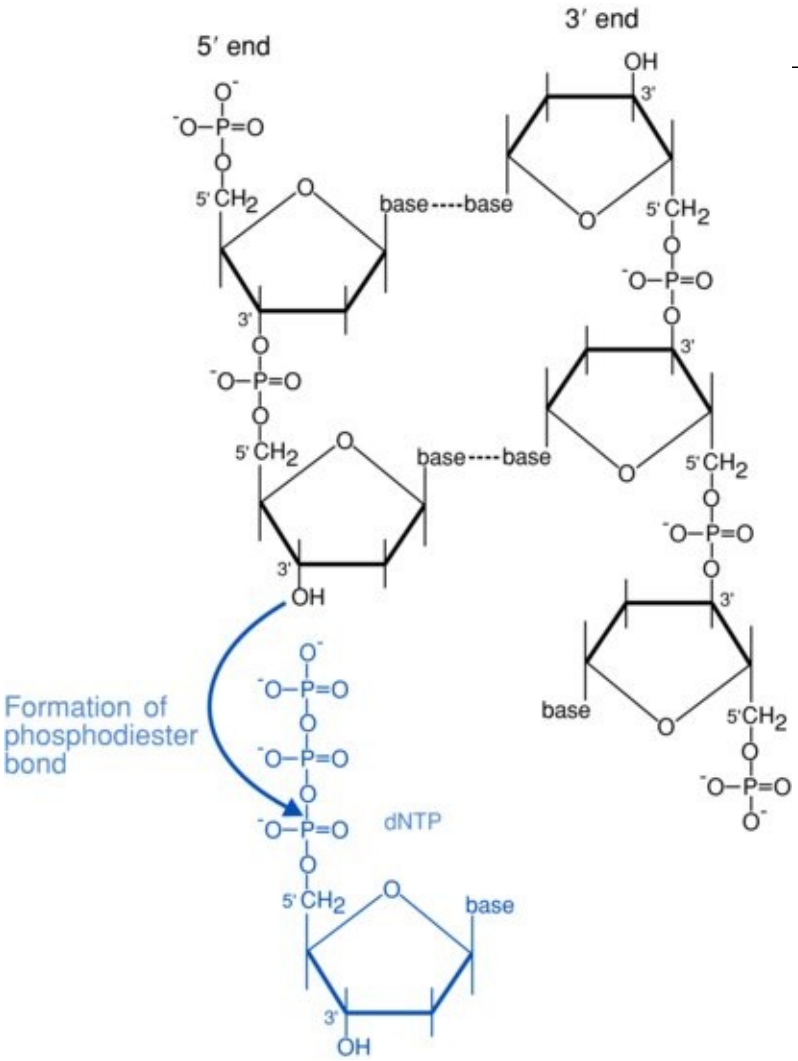


Generally, while most of the DNA we encounter is double-stranded, most of the RNA we me consists of a single polynucleotide strand. However, DNA can also exist as a single-strande molecule, and RNA is able to form double-stranded molecules. Thus, this distinction between RN and DNA is not an inherent property of the nucleic acids themselves, but is a reflection of the natur 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 cel and the replication process that has evolved is based on a double-stranded molecule. By contrast, th roles of RNA in the cell do not require a second strand, and indeed the presence of a secon complementary, strand would preclude its role in protein synthesis. However, there are some viruse that have double-stranded RNA (dsRNA) as their genetic material, as well as some viruses wi single-stranded RNA. In addition, some viruses (as well as some plasmids) replicate via singl stranded DNA forms. Double-stranded RNA is also important in the phenomenon known as RN interference, which we will come to in later chapters.

1.2.4 Nucleic acid synthesis

We do not need to consider here all the details of how nucleic acids are synthesized. The fundament features that we need to remember are summarized in Figure 1.7, which shows the addition of nucleotide to the growing end (3′-OH) of a DNA strand. The substrate for this reaction is the releva deoxynucleotide triphosphate (dNTP), i.e., the one that makes the correct base-pair with th corresponding residue on the template strand. The DNA strand is always extended at the 3′-OH en thus the nucleotide strand grows in the 5′ to 3′ direction. For this reaction to occur it is essential th the existing residue at the 3′-OH end, to which the new nucleotide is to be added, is accurately bas paired with its partner on the other strand.

Figure 1.7 DNA synthesis.



RNA synthesis occurs in much the same way, as far as this simplistic description goes, except th of course the substrates are nucleotide triphosphates (NTPs) rather than the deoxynucleoti triphosphates (dNTPs). There is one very important difference though. DNA synthesis only occurs b extension of an existing strand – it always needs a *primer* to get it started. In contrast, RN polymerases are capable of starting a new RNA strand, complementary to its template, from scratc given the appropriate signals.

1.2.5 Coiling and supercoiling

DNA can be denatured and renatured, deformed and reformed, and still retain unaltered function. Th is a necessary feature, because such a large molecule as DNA will need to be packaged if it is to f within the cell that it controls. The DNA of a human chromosome, if it were stretched into a unpackaged double helix, would be several centimetres long. Thus, cells are dependent on th packaging of DNA into modified configurations for their very existence.

Double-stranded DNA, in its relaxed state, normally exists as a right-handed double helix with on complete turn per 10 base pairs; this is known as the *B form* of DNA. Hydrophobic interactio between consecutive bases on the same strand contribute to this winding of the helix, as the bases a brought closer together enabling a more effective exclusion of water from interaction with th hydrophobic bases.

The DNA double helix can exist in other forms, notably the *A form* (also right-handed but mo compact, with 11 bases per turn) and *Z-DNA*, which is a left-handed double helix with a more irregul appearance (a zigzag structure, hence its designation). However, that is not the complete story. High

orders of conformation are known to exist. The double helix is in turn coiled on itself – an effe
known as *supercoiling*. There is an interaction between the coiling of the helix and the degree
supercoiling. As long as the ends are fixed, changing the degree of coiling will alter the amount
supercoiling, and vice versa. DNA *in vivo* is constrained; the ends are not free to rotate. This is mo
obviously true of circular DNA structures such as (most) bacterial plasmids, but the rotation of line
molecules (other than very short oligonucleotides) is also constrained within the cell. The net effect
coiling and supercoiling (a property known as the *linking number*) is therefore fixed, and cannot
changed without breaking one of the DNA strands. In nature, there are enzymes known
topoisomerases (including DNA gyrase in bacteria) that do just that: they break the DNA strands, ar
then in effect rotate the ends and reseal them. This alters the degree of winding of the helix, and th
affects the supercoiling of the DNA. Topoisomerases also have an ingenious use in the laborator
which we will consider in Chapter 2.

The plasmids that we will be referring to frequently in later pages are naturally supercoiled wh
they are isolated from the cell. However, if one of the strands is broken at any point, the DNA is the
free to rotate at that point and can therefore relax into a non-supercoiled form. This is known as
open circular form (in contrast to the *covalently closed circular* form of the native plasmid).

1.3 What is a gene?

The definition of a ‘gene’ is rather imprecise. Its origins go back to the early days of genetics, when
was used to describe the unit of inheritance of a phenotype. This meaning persists in non-scientif
usage, rather loosely, as the ‘gene for blue eyes’, or ‘the gene for red hair’. As it became realized th
many characteristics were determined by the presence or properties of individual proteins, t
definition became refined to relate to the chromosomal region that carried the information for th
protein, leading to the concept of ‘one gene, one protein’. As the study of genetics and biochemistr
progressed further, it was realized that many proteins consist of several distinct polypeptides, and th
the chromosomal regions coding for the different polypeptides could be distinguished genetically. S
the definition was refined further to mean a piece of DNA containing the information for a sing
specific polypeptide (‘one gene, one polypeptide’). With the advent of DNA sequencing, it becam
possible to consider a gene in molecular terms. So we now often use the term ‘gene’ as bein
synonymous with ‘open reading frame’ (ORF), i.e., the region between the start and stop codons.
bacteria, this is (usually) a simple uninterrupted sequence, but in eukaryotes, the presence of intron
(see below) makes this definition more difficult, since the region of the chromosome that contains t
information for a specific polypeptide may be many times longer than the actual coding sequence. W
also have to be careful as we may want to refer to the whole transcribed region, which will be long
than the translated open reading frame, or indeed we may want to include the control regions that a
necessary for the start of transcription.

Furthermore, this definition, by focusing solely on the regions that code for proteins (
polypeptides), is too limited in its scope. It ignores many regions of DNA that, although not coding f
proteins, are nevertheless important for the viability of the cell, or influence the phenotype in oth
ways. The most obvious of these are DNA sequences that are templates for so-called non-coding RN
molecules. The most well known of these are ribosomal and transfer RNA, although we will encount
other RNA molecules that play significant roles in gene regulation and other activities. Other DN
regions are important in gene regulation because they act as binding sites for regulatory proteins.

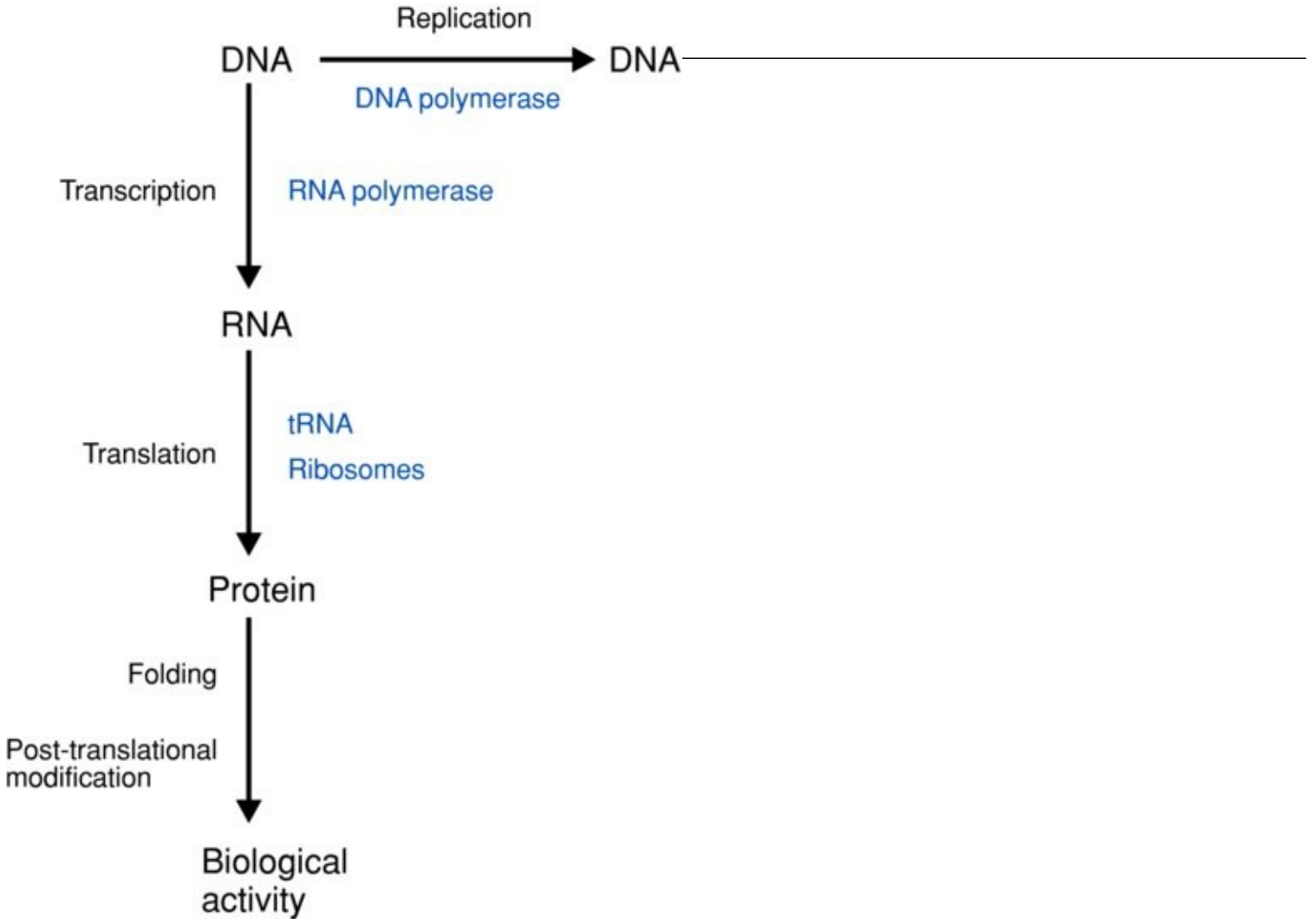
In organisms with small genomes, such as bacteria, a high proportion of the genome is accounted f by these coding and regulatory regions (together with elements that may be described as 'parasiti such as integrated viruses and insertion sequences). There is relatively little DNA to which we ca ascribe no likely function, compared to eukaryotic cells, and especially animal and plant cells wi much larger genomes, where there is a much higher proportion of non-coding DNA. But we should n be too hasty in writing these sequences off as 'junk'. Increasingly, much of it is recognized as havin important functions within the cell. These include enabling the DNA to be folded correctly and ensuring that the coding regions are available for expression under the appropriate conditions, as we as coding for small (non-translated) RNA molecules that play a major role in modulating gen expression.

Thus, we have to accept that it is not possible to produce an entirely satisfactory definition of th word 'gene'. However, this is rarely a serious problem. We just have to be careful how we use depending on whether we are discussing only the coding region (ORF), or the length of sequence th is transcribed into mRNA (including untranslated regions), or whether we wish to include DN regions with regulatory functions as well as coding sequences. In this context, we will also encount the words *allele* and *locus*. A locus is used in the same way as 'gene' in the broad meaning – i.e., could be a coding sequence, a regulatory region, or any other region we wish to consider. An allele one version of that locus. So variation of a genetic characteristic between individuals would be due different alleles at one locus (or several loci).

1.4 Information flow: gene expression

The way in which genes are expressed is so central to the subsequent material in this book that it worth reviewing briefly the salient features. The basic dogma (Figure 1.8) is that while DNA is th fundamental genetic material (ignoring RNA viruses) that carries information from one generation the next, its effect on the characteristics of the cell requires firstly its copying into RN (*transcription*), and then the *translation* of the mRNA into a polypeptide by ribosomes. Furth processes are required before its proper activity can be manifested: these include the folding of th polypeptide, possibly in association with other subunits to form a multisubunit protein, and in som cases modification, for example by glycosylation or phosphorylation. It should be remembered that some cases, RNA rather than protein is the final product of a gene (e.g., ribosomal and transfer RN molecules).

Figure 1.8 Information flow.

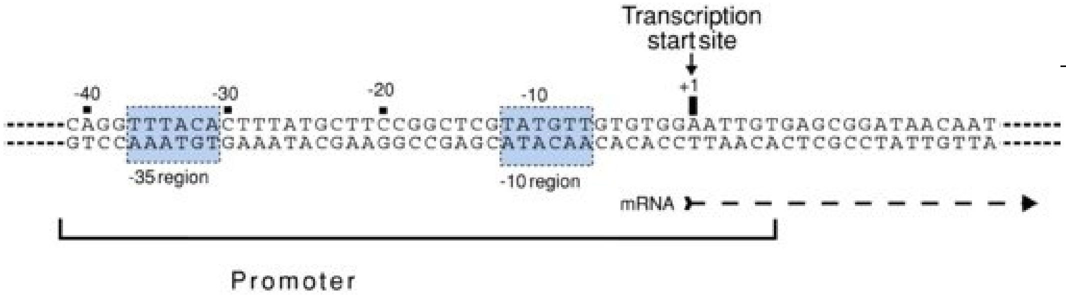


1.4.1 Transcription

Transcription is carried out by RNA polymerase. RNA polymerase recognizes and binds to a specifi sequence (the *promoter*), and initiates the synthesis of mRNA from an adjacent position.

A typical bacterial promoter carries two *consensus* sequences (i.e., sequences that are closely relate in all genes): TTGACA centred at position −35 (i.e., 35 bases before the transcription start site), an TATAAT at −10 (Figure 1.9). It is important to understand the nature of a consensus: few bacteri promoters have exactly the sequences shown, but if you line up a large number of promoters you wi see that at any one position a large number of them have the same base (Box 1.2). The RN polymerase has higher affinity for some promoters than others – depending not only on the exa nature of the two consensus sequences but also, to a lesser extent, on the sequence of a longer regio of DNA. The nature and regulation of bacterial promoters, including the existence of alternative typ of promoters, is considered further in Chapter 5.

Figure 1.9 Structure of the promoter region of the *lac* operon. Note that the −35 and −10 regions of the *lac* promoter do not correspond exactly to the consensus sequences TTGACA and TATAAT respectively.



In eukaryotes, by contrast, the promoter is a considerably larger area around the transcription sta site, where a number of *trans*-acting transcription factors (i.e., DNA-binding proteins encoded b genes in other parts of the genome) bind to various *cis*-acting elements (i.e., elements that affect th expression of the gene next to them) in a considerably more complex scenario. The need for this adde complexity can easily be imagined; if cells carrying the same genome are differentiated into multitude of cell types fulfilling very different functions, a very sophisticated control system needed to provide each cell type with its specific repertoire of proteins, and to fine-tune the degree expression for each one of them. Nonetheless, the promoter region, however simple or complex, giv rise to different levels of transcription of various genes. A further complexity in eukaryotes is th there are commonly additional regulatory elements known as *enhancers*, which may further alter th level of transcription of one, or several, genes. By definition, enhancers are position and orientatio independent, and are often remote from the actual start site of transcription by several thousand ba pairs.

Box 1.2 Examples of *E. coli* promoters

```
                -35                                 -10         1
TGGCGGTGTTGACATAAATA  CCACTGGCGGTGATACTGAGCA  CA   Lambda PL
CGTGCGTGTTGACTATTTTA  CCTCTGGCGGTGATAATGGTTG  CA   Lambda PR
TGCCGAAGTTGAGTATTTTT  GCTGTATTTGTCATAATGACTCCTG    Lambda PO
ATGAGCTGTTGACAATTAAT  CATCGAACTAGTTAACTAGTACGCA    trp
CATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCG     lacI
CCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG  A    lacZ
CGTAACACTTTACAGCGGCG   CGTCATTTGATATGATGCGCC  CG   tyr tRNA
        TTGACA                    TATAAT           consensus
```

Bases matching the −10 and −35 consensus sequences are boxed. Spaces are inserted to optimise the alignment. Note that the consensus is derived from a much larger collection of characterized promoters. Position 1 is the transcription start site.

Eukaryotes have three different RNA polymerases. Only one of these, RNA polymerase II, involved in the transcription of protein-coding transcripts, plus the transcription of a group of sma non-coding RNAs called micro-RNAs, which we will encounter in Chapter 11. RNA polymerase I responsible for the synthesis of large ribosomal RNAs, whilst RNA polymerase III makes small RNA such as transfer RNA (tRNA) and 5S ribosomal RNA.

In eukaryotes, the primary transcript from a protein-coding gene, produced by RNA polymerase I is called a *heterogeneous* or *heteronuclear RNA* (hnRNA). It is very short-lived as such, being rapid processed in a number of steps called *maturation*. A specialized nucleotide *cap* is added to the 5′ en this is the site recognized by the ribosomes in protein synthesis (see below). The precursor mRNA cleaved at a specific site towards the 3′ end and a *poly-A tail*, consisting of a long sequence adenosine residues, is added to the cut end. This is a specific process, governed by polyadenylatio recognition sequences in the 3′ untranslated region. Nature's ‘tagging’ of mRNA molecules comes i