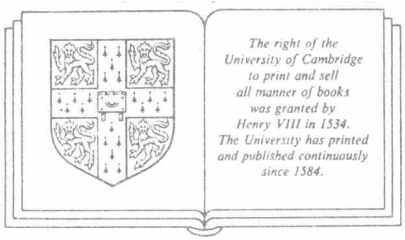
Gene structure and expression

JOHN D. HAWKINS

Gene structure and expression

JOHN D. HAWKINS



CAMBRIDGE UNIVERSITY PRESS
Cambridge
London New York New Rochelle
Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
32 East 57th Street, New York, NY 10022, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1985

Printed in Great Britain by the University Press, Cambridge

Library of Congress catalogue card number: 84–17459

British Library Cataloguing in Publication Data

Hawkins, J. D.
Gene structure and expression.

1. Deoxyribonucleic acid
574.87'3282 QP624

ISBN 0 521 25824 3 hard covers
ISBN 0 521 27726 4 paperback

UP

List of abbreviations

In general, standard biochemical abbreviations are used throughout this book, particularly in figures and tables. The following is a list of those that are used without further explanation.

Amino acids:

Alanine Ala
Arginine Arg
Asparagine Asn
Aspartate Asp
Cysteine Cys
Glutamine Gln
Glutamate Glu
Glycine Gly
Histidine His
Isoleucine Ile
Leucine Leu
Lysine Lys
Methionine Met
Phenylalanine Phe
Proline Pro
Serine Ser
Threonine Thr
Tryptophan Trp
Tyrosine Tyr
Valine Val
Termination Ter
(at end of protein sequences)

Purine and pyrimidine bases and other symbols used in writing DNA or RNA sequences. It should be clear from the context whether bases or nucleotides are intended:

Adenine A
Guanine G
Cytosine C
Uracil U
Pseudo-uridine Ψ
Thymine T
Pyrimidine Y
Purine R
Any nucleotide N
Any base B
Phosphate radical P
Ribose R
Deoxyribose dR
Methyl group m

Introduction

There has been an explosive growth in our detailed knowledge of genetics at the molecular level over the last few years, and it is likely that accretion of new knowledge will occur at an ever increasing rate. It is therefore very difficult even for the specialist to keep abreast of all the latest ideas which rapidly progress from hypothesis to theory to accepted dogma. In the time that it takes to write a comprehensive text book it is inevitable that new ideas will be generated and many problems in the field elucidated so that such a book will certainly be out of date before the writing is finished, let alone published. Even during the writing of this small book, over the course of a little more than a year, much new information has come to light so that were it to be re-written in the next few months, appreciable differences would appear. It does not therefore claim to be a complete guide to the subject under review; nevertheless it attempts to present ideas that are reasonably well established and at the same time to cover a fairly wide field, albeit mostly not in great depth. The selection of topics as examples of our knowledge is somewhat arbitrary and conditioned by the author's own interests and expertise.

I believe that it should be a useful book for medical students who wish to become familiar with recent ideas and techniques in molecular biology to help in understanding further advances when they arrive. It will also be of use to honours and graduate students in genetics, biochemistry and those who would not necessarily regard the topics discussed here as their major interests in these subjects. It assumes a working knowledge of biochemistry that a first or second year university or polytechnic student should have acquired in a fairly elementary course in that subject. This basic material is already excellently covered in such books as Lubert Stryer's *Biochemistry* and Albert Lehninger's *Biochemistry*, as well as a host of others.

Reading Lists for each Chapter are to be found at the back of this book. They are mostly made up of review articles in which references to original work can be found. In several cases parallel reviews covering more or less the same ground have been cited.

I am most grateful to Dr Fay Bendall of Cambridge University Press who encouraged me to write this book in the first place, widening my horizons and giving me a good deal of pleasure in the process: to Dr Audrey O. Smith whose patience and skill in pointing out errors and inconsistencies in the text have helped to clarify it. Needless to say, any errors that do remain are my own responsibility. I am also grateful to many colleagues at St Bartholomew's Hospital Medical College, particularly Dr Clem Lewis for stimulating discussions which I have frequently found helpful in clarifying my ideas. Last, but not least, I am grateful for my wife's forebearance in the face of a neglectful husband who has been preoccupied for many months with the genome rather than with its ramifications as revealed in family life.

June 1984

Contents

1
DNA

1.1 The genetic material

The classic experiments of Avery in 1944 demonstrated that DNA (deoxyribonucleic acid) is the material that can pass genetic information from one bacterium to another. He showed that strain-specific properties of related bacteria could be transferred by DNA that was free of proteins and other substances. DNA is a polymeric molecule built up from only four similar but distinct monomers – nucleotides which are the 5′-phosphates of deoxyguanosine (dGMP), deoxyadenosine (dAMP), deoxycytidine (dCMP), and thymidine (TMP) (Fig. 1.1). In DNA these are joined by phosphodiester linkages between the 3′- and 5′-positions of successive deoxyribose moieties. The initial letters of the bases in the nucleotides are used as abbreviations when writing out sequences in DNA.

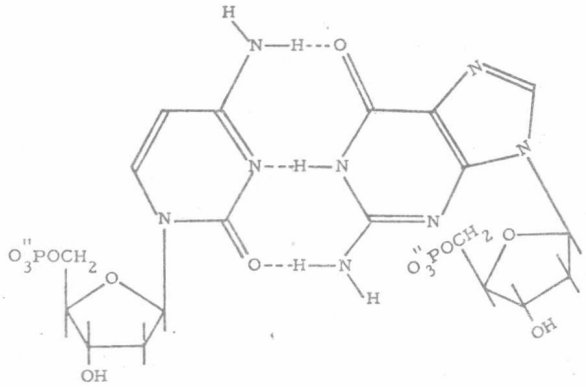
1.2 DNA is a polar molecule

One end of a DNA molecule has a phosphoryl radical on the C-5′ of its terminal nucleotide, while the other end possesses a free -OH on C-3′ of its nucleotide. Thus a polynucleotide possesses *polarity* in an analagous way to the more familiar polarity of proteins with free $-NH_2$ and -COOH groups at each end. This means, for example, that the tetranucleotides TCGA and AGCT are different chemical entities with distinct properties, even though they behave in a very similar way in many respects (Fig. 1.2). By convention, sequences of DNA are written with the nucleotide containing the free phosphoryl radical at the left. Sequences to the left of a given nucleotide are said to be on the 5′-side (often called upstream), and those to the right are said to be on the 3′-side (often called downstream). The symbols N, R and Y are used to denote any nucleotide, a purine nucleotide, and a pyrimidine nucleotide respectively.

1.3 DNA generally exists as a double helix

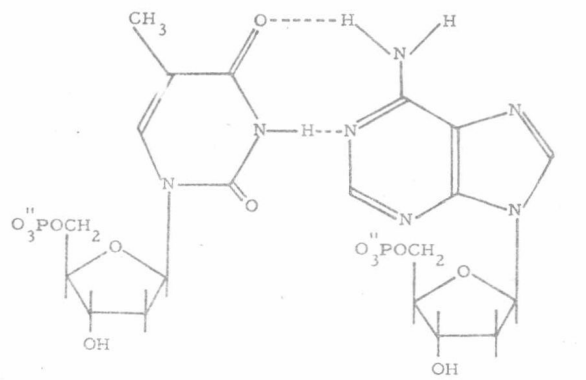
DNA generally exists in double strands because of the propensity of the bases for hydrogen bonding to each other in a highly specific way (Fig. 1.1). A bonds with T, and G with C, though very occasional mismatches or alternative bonding can occur. Thus a double-stranded DNA will always contain equal molar proportions of A and T and of G and C though the content of A (or T) and G (or C) varies widely in DNA from different sources.

Fig. 1.1. The four deoxyribonucleotides that make up DNA, showing how the bases form hydrogen bonds.



DEOXYCYTIDINE-5'-MONOPHOSPHATE
dCMP

DEOXYGUANOSINE-5'-MONOPHOSPHATE
dGMP



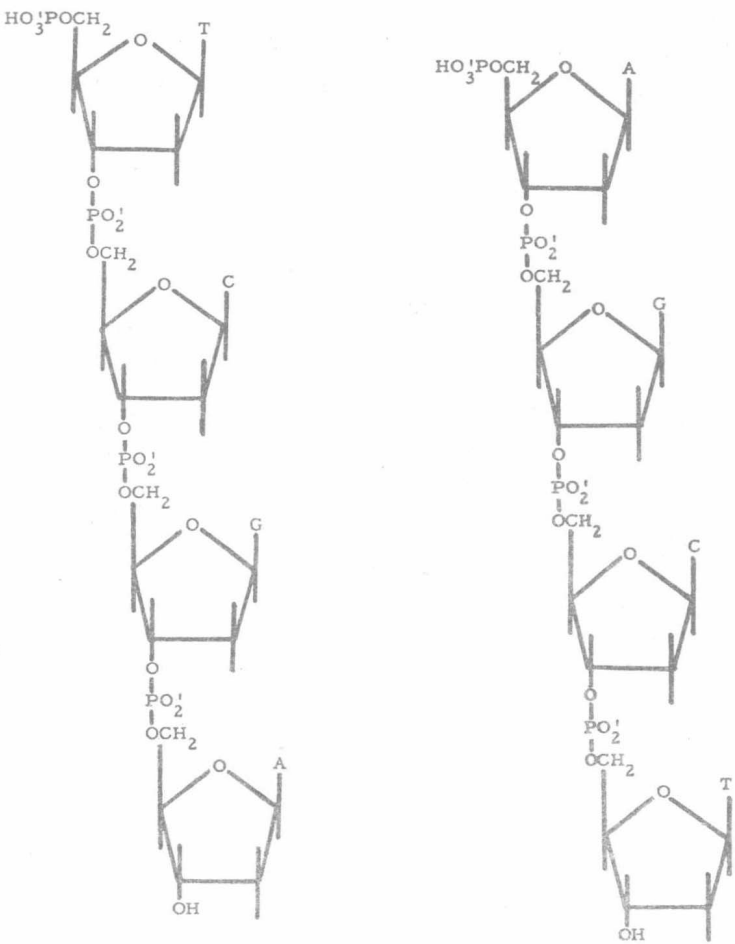
THYMIDINE-5'-MONOPHOSPHATE
dTMP

DEOXYADENOSINE-5'-MONOPHOSPHATE
dAMP

The basic unit in a DNA molecule is the pair of nucleotides hydrogen bonded to each other, which is generally known as a base-pair (bp; with kbp used as an abbreviation for 1000 bp).

This double-stranded molecule takes up a helical conformation in which the continuous deoxyribose-phosphate strands twine round the outside of the *helix* with the base pairs (A and T or G and C) in the interior (Fig. 1.3). Since A and T associate together with only two hydrogen bonds, while the G-C pair has three hydrogen bonds, the former pairing is less stable than the latter. This has important consequences for the stability of different regions of the double helix. In regions that are rich in A and T residues

Fig. 1.2. Polarity in DNA. The two tetranucleotides TCGA (left) and AGCT (right) are different, even though they have the bases in the same order.



the helix can be more easily destabilised and unwound than in G-C rich regions.

The polarity of the two DNA strands is *anti-parallel* – that is to say, one runs in the 5′ to 3′ direction while the complementary strand runs the opposite way. The helix can adopt several conformations. The commonest form (B-DNA) has a pitch of just over 10 residues per turn, and is right-handed when viewed end on. Another form, known as Z-DNA, can arise under certain conditions when there are alternating purine and pyrimidine residues in the sequence. This is a left-handed helix and has a pitch of 11.5 residues per turn (Fig. 1.3). This form is believed to have some biological significance, since short stretches of alternating purine and pyrimidine residues occur at numerous sites in many DNAs. They can be detected by the binding of antibodies which specifically recognise this form of DNA. Other non-antibody proteins have been discovered which also react with Z-DNA in a wide range of cells in higher organisms, insects, bacteria and viruses. These viral sequences occur in regions that are known to be involved in the control of the genome, so it is likely that they have important functions in this process.

Fig. 1.3. A length of double helical DNA, containing 20 bp, showing both B and Z forms. The lines running round the outside represent the backbone of poly deoxyribose phosphate, while the horizontal lines represent the edges of the base pairs in the interior of the molecule. (Reprinted, with permission, from S. B. Zimmerman, *Annual Review of Biochemistry*, Vol **51** © 1982 by Annual Reviews Inc.)

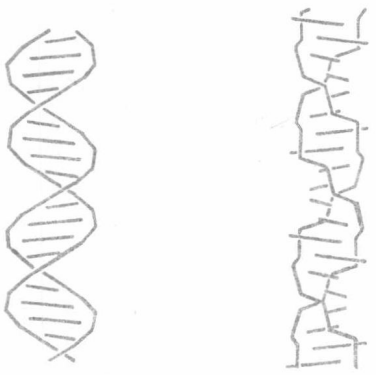


Table 1.1. *Chromosome numbers and DNA content of cells of a representative set of species*

Species	Number of chromosomes	DNA content kilobase pairs
Bacillus subtilis	1	2×10^3
Escherichia coli	1	3.8×10^3
Saccharomyces cerevisiae	34	14×10^3
Drosophila melanogaster	8	2×10^5
Sea urchin	52	1.6×10^6
Frog	26	45×10^6
Chicken	78	2.1×10^6
Mouse	40	4.7×10^6
Human	46	5.6×10^6
Maize	20	30×10^6

All figures are for diploid cells, except for bacteria.

1.4 DNA molecules are very long but can be twisted into compact forms

DNA molecules are extremely long and can be visualised by the electron microscope. A double helix of 10^6 base pairs is 0.34 mm long and only 2 nm in diameter. In prokaryotes the DNA is circular so that there are no free 3′- and 5′-ends, and all the chromosomal DNA is in a single molecule. In eukaryotes the DNA in the chromosomes exists as linear molecules, and different species possess different numbers of chromosomes. The amount of DNA in cells of different species varies very widely, generally with an increase in the DNA content as species become more complex (Table 1.1).

With one exception all eukaryotic chromosomes are paired, with one partner coming from each parent. The exception is the sex chromosome. Females carry two X chromosomes, while males carry an X chromosome inherited from their mother and a Y chromosome from their father. This is the case in mammals and many other orders, but other methods of sex determination do occur. The two chromosomes of a pair are said to be *homologous* since they will nearly always be identical in their organisation and frequently in the genes they carry. However, since there are many mutant genes in a population a pair of homologous chromosomes may carry different genes at particular loci. These are known as *alleles*.

A mutant gene encoding a defective product can generally be *complemented* by a 'good' copy of the gene on the homologous chromosome, but if there is a defect on the single copy of the X chromosome that a male carries it cannot be complemented in this way. Thus there are many sex-linked inherited diseases which are carried by females, but expressed only

in males. These disorders can actually occur in females but the chances of a female inheriting two defective genes are very low.

Since all somatic cells contain a homologous pair of each of the chromosomes they are known as *diploid.* The gametes – sperm and ova – which only contain one member of each pair of chromosomes are known as *haploid* cells. The contribution of one parent to the genetic make-up of the offspring is known as the *haplotype.*

Individual chromosomes are morphologically distinguishable when they are suitably stained. For purposes of identification they have been given numbers in numerical order, starting with the largest one.

In the cell both linear and circular molecules are found in much more compact forms. The helix is coiled on itself several times (like the element in an electric light bulb) so that the overall length is greatly reduced at the expense of an increase in diameter. This conformation is stabilised by proteins in eukaryotic cells (see Chapter 7, sections 2, 3). This kind of structure is very suitable for packaging DNA into a minimum of space, but when the DNA, or strictly speaking, portions of it, becomes functional some uncoiling of this structure must occur accompanied by temporary separation of the two helical strands.

In a circular DNA molecule containing 4000 bp (such as might occur as a bacterial plasmid – Chapter 4.2) the double helix is in the B-form. As this has a pitch of 10 residues per turn there should be 400 turns. In practice such DNA is found to have only about 380 turns because the helix is untwisted to a certain extent. This is known as negative *supercoiling*, and is an important feature of the structure of DNA in bacterial cells. It results in a puckered form of the molecule which gives it a more compact structure, and also places considerable torsional strain on it. An analogy can be made by twisting a rubber band held firmly at two diametrically opposite positions. A molecule in which there is no supercoiling is said to be *relaxed*, and there is a dynamic balance between relaxed and supercoiled forms of DNA as a result of the action of two classes of enzymes called *topoisomerases I* and *II* which catalyse the production of one form from the other.

When a DNA molecule is supercoiled it migrates more rapidly on electrophoresis than when it is relaxed. A family of otherwise identical DNA molecules with different degrees of supercoiling can be made visible as a ladder of bands by this technique (Fig. 1.4). Supercoiled molecules appear more compact than relaxed ones when viewed in the electron microscope.

Formation of the supercoiled form in the cell is brought about by an enzyme called DNA *gyrase* (a class II topoisomerase) and, because it is a strained structure, there is an energy requirement for this reaction, which

Fig. 1.4. The effect of supercoiling on the electrophoretic mobility of Simian virus 40 DNA. The DNA was treated with a class II topoisomerase and then electrophoresed. The thick band at the bottom represents fully relaxed DNA, while molecules with increasing degrees of supercoiling appear as bands of increasing mobility. The arrow shows the direction of electrophoretic migration. (Reproduced from W. Keller, *Proc. Natl. Acad. Sci. USA* (1975), **72**, 2550.)

Fig. 1.5. The action of DNA gyrase in forming a negative supercoil in a circular molecule of DNA. By convention, when the upper strand crosses above the lower strand from left to right the supercoiling is said to be positive. Negative supercoiling is the converse of this.