John Howard Parish

Principles and practice of experiments with nucleic acids

John Howard Parish

Principles and practice of experiments with nucleic acids

Longman

LONGMAN GROUP LIMITED
LONDON
Associated companies, branches and representatives throughout the world

First published 1972

ISBN 0582 46285 1

Printed in Great Britain by
William Clowes & Sons Ltd
London, Colchester and Beccles

Preface

This book is largely about chemistry and it is written mainly for biologists. The role of nucleic acids is essential for any overall view of the control of metabolism or genetics. Similarly a precise description of the molecular events involved in such processes as differentiation, carcinogenesis and memory will undoubtedly require an understanding of reactions involving nucleic acids (although we cannot delineate these reactions with certainty at present). Consequently workers in a wide variety of disciplines find their research requires a working knowledge of nucleic acid methodology and the necessary theoretical background. In this book, I have described this background and have exemplified procedures with reference to experiments regarding a variety of biological problems. I hope the choice of examples is balanced rather than random; it is not however, in any sense, comprehensive. The book presents a view of the nucleic acids for those with a real interest in planning experiments or following the literature. It is not a textbook of molecular biology.

Certain topics (such as the isolation and fractionation of nucleic acids) have been described in considerable experimental detail. The reason is that I hope to have presented enough information for the reader to, at least, get started with experiments of his own devising. From such a start, I hope the references are sufficiently comprehensive to cover the more important techniques.

On the other hand, certain topics are almost out-of-bounds to the average biologist. The main example is the technique of X-ray crystallography. For this reason the account of data deduced from diffraction studies is extremely fragmentary. It is important to stress this imbalance as any integrated view of our knowledge of nucleic acid structure would emphasize the enormous (and continuing) contribution from crystallographers.

Any systematic reference to the historical aspects of the subject are also omitted. This omission is regrettable as the elucidation of the basic skeletal structure of RNA and DNA is one of the great achievements of organic chemistry. The chemistry of the nucleic acids that is described here is all relevant to procedures employed in present-day research.

Although nucleic acid research is largely in the province of biochemists, I am aware of the wish of some chemists to change their interest to an aspect of biochemistry—and the study of nucleic acids is a branch of natural product chemistry. Moreover there are several unsolved problems, alluded to in the following chapters, which require the special expertise of physico-organic chemists. It is largely for the reader with an essentially chemical (or physical) academic background, that the first chapter is written. In that chapter, I delineate the outlines of nucleic acid structure and metabolism as a starting point for topics in the rest of the book.

Note

Supplementary notes, signalled in the text by a superscript numeral, are to be found at the end of each chapter.

Acknowledgements

I am extremely grateful to Dr. O. L. Miller, Dr. H. Ris and Dr. D. Dressler for generously supplying the photographs used in the frontispiece and to Dr. J. R. B. Hastings, Dr. I. H. Maxwell, Dr. W. Fuller, Dr. R. M. Malbon, Mr. H. A. Foster and Miss M. Brown for access to unpublished results and for discussions.

Work from the author's laboratory has been supported by the Science Research Council.

Publishers' Acknowledgements

We are grateful to the following for permission to reproduce illustrations:

Academic Press Inc. (London) Ltd: *Journal of Molecular Biology* for Figs. 4.14, 4.19, 4.22, 4.23, 10.4, 10.5, 11.4a, 11.4b, 11.5, 11.6, 12.2; Academic Press Inc. New York: *Advances in Virus Research* for figs. 16.4, 16.7; *Analytical Biochemistry* for fig. 4.7; *Methods in Enzymology* for figs. 4.5c, 4.20; *Molecular Associations in Biology* for fig. 11.10; *The Nucleic Acids* for fig. 2.5; *Progress in Nucleic Acid Research and Molecular Biology* for figs. 1.6, 12.3; American Association for the Advancement of Science: *Science* for Plates 1a, 1c, fig. 12.5; American Chemical Society: *Biochemistry* for fig. 5.12; *Journal of American Chemical Society* for fig. 2.26; American Society for Microbiology: *Journal of Bacteriology* for fig. 5.7; Biochemical Society: *Biochemical Journal* for fig. 4.22; Butterworth & Company Ltd: *Subcellular components* for figs. 4.16, 4.18; Cambridge University Press: *Symposium of the Society for Experimental Biology* for figs. 12.8a, 12.8b, 13.15; J. & A. Churchill: *Cell differentiation–CIBA Foundation Symposium* for fig. 4.17; Cold Spring Harbor Laboratory: *Symposium on Quantitative Biology* for figs. 4.22, 5.5, 12.3, 12.12; Elsevier Publishing Company: *Biochimica et Biophysica Acta* for figs. 2.21, 15.3; *Chemico-Biological Interactions* for fig. 8.9; Genetics Society of America: *Genetics* for fig. 12.9; Macmillan Journals Ltd: *Nature* for figs. 5.8, 11.16, 11.17, 12.10, 16.6, 16.15; National Academy of Sciences: *Proceedings of the National Academy of Sciences* for Plate 1b, figs. 2.27a, 4.22, 4.24, 5.3, 5.9, 5.11, 11.12, 16.2, 16.3.

Contents

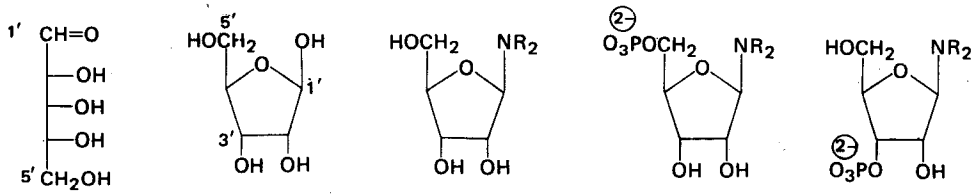
1 Introduction

What are nucleic acids and what are they for? They are polyelectrolytes of high molecular weight, they are the chemicals that genes are made of and they direct the biosynthesis of proteins. I naturally assume that the reader is capable of amplifying these statements in some sort of detail. However, an introductory chapter can serve two useful purposes; it is a polite way of drawing attention to the background material I am assuming and it is the most digestible context for presenting the conventions and abbreviations which must be used throughout the rest of the book. I have omitted any references from the bulk of this chapter as to include them would make it unwieldy. However, an excellent and very readable book which describes the whole background to the subject and its many ramifications is J. D. Watson's *The Molecular Biology of the Gene* (2nd. ed., 1970), Benjamin.

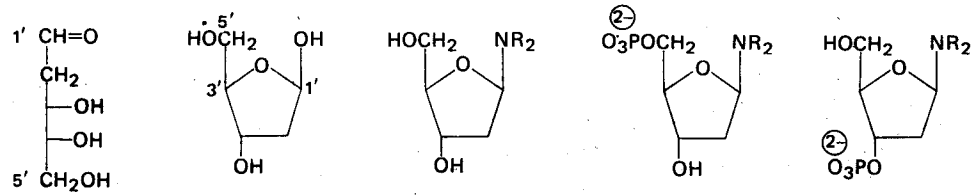
Primary structure

The backbone of a nucleic acid molecule is a chain of sugar moieties linked by phosphodiester groups. As phosphoric acid is tri-basic, phosphodiesters are mono-basic, and each phosphate residue in nucleic acids has one ionizable hydrogen. The pK_a value for this ionization is around 1·0 (or just less). Nucleic acids are not chemically stable at very low pH and consequently they are always handled as their salts. So whenever we say nucleic acid we mean the 'nucleate' anion. The cations present might typically be sodium, potassium, ammonium (or of course a mixture of more than one type). Attached to the sugar residues are heterocyclic bases. Various types of fragment can be obtained from nucleic acids. The base attached to its sugar, but lacking phosphate is called a nucleoside, the base-sugar-phosphate is called a nucleotide so nucleic acids are polynucleotides. The converse of this statement is not true; polynucleotides include in addition to the naturally occurring nucleic acids a number of synthetic polymeric substances. The nucleotides themselves are phosphomonoesters (dibasic acids). Short chains of nucleotides (obtained either synthetically or by partial degradation of polynucleotides) are known as oligonucleotides. The smallest nucleic acids contain about eighty nucleotide residues and the largest of the order of hundreds of millions.

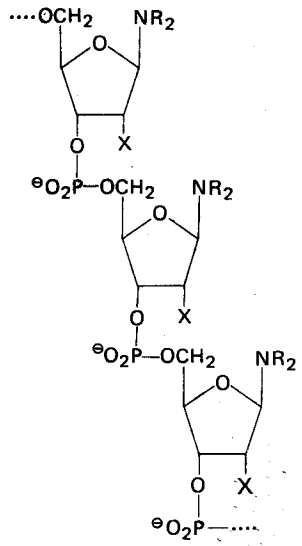
There are two main sugars found in nucleic acids, D-ribose and 2-deoxy-D-ribose.[1] With minor exceptions which are mentioned later, any one nucleic acid molecule contains only one type of sugar. Thus there are two types of nucleic acid (and consequently two families of nucleosides and nucleotides) those in which the sugar is ribose (ribonucleic acid or RNA with ribonucleosides and ribonucleotides) and those in which the sugar is deoxyribose (deoxyribonucleic acid or DNA with deoxyribonucleosides and deoxyribonucleotides).[2] Shortened words for nucleosides and nucleotides are riboside (for ribonucleoside), ribotide (for ribonucleotide) and deoxyriboside and deoxyribotide.



D-Ribose: Straight-chain form | D-Ribose: β-furanoside | Riboside | 5′-Ribotide | 3′-Ribotide



2-Deoxy-D-ribose: Straight-chain form | 2-Deoxy-D-ribose: β-furanoside | Deoxyriboside | 5′-Deoxyribotide | 3′-Deoxyribotide



Fragment of nucleic acid structure. In RNA, X = OH; in DNA, X = H

FIG 1.1 Sugars, nucleosides, nucleotides and nucleic acids. In these formulae, the heterocyclic cyclic amines would be represented as R_2NH.

The remaining details of the outline structure of nucleic acids are as follows. The bases are all secondary amines and are attached to the sugars by *N*-glycoside bonds, the sugars being in the β-furanose configuration. The phosphodiester groups are attached to carbon atoms 5 and 3 of the adjacent sugar residues. These carbons are normally designated 5′ and 3′ (pronounced 'five primed' and 'three primed'). The 'priming' is to distinguish reference to positions in the sugar moieties to ring atoms in the heterocyclic bases. Thus, neglecting for the moment the nature of the bases, we can summarize the structures mentioned so far (Fig. 1.1). Note that a nucleic acid can be regarded as a polymer of either 5′- or 3′-nucleotides.

The details of the chemical degradation of nucleosides, nucleotides and nucleic acids are discussed later in the book, but we may note here that in general *N*-glycosides are relatively stable to alkali but are hydrolysed by acid. For a nucleoside, the sugar and the free base would be the products of such a reaction. Ribose itself is fairly stable in acid but in hot strong acid it is dehydrated (as are all pentoses) to form furfural; this forms the base of the orcinol test for RNA. Deoxyribose, on the other hand is less stable in acid and rearranges to form δ-hydroxylaevulinic acid; this is the basis of the diphenylamine test for DNA.

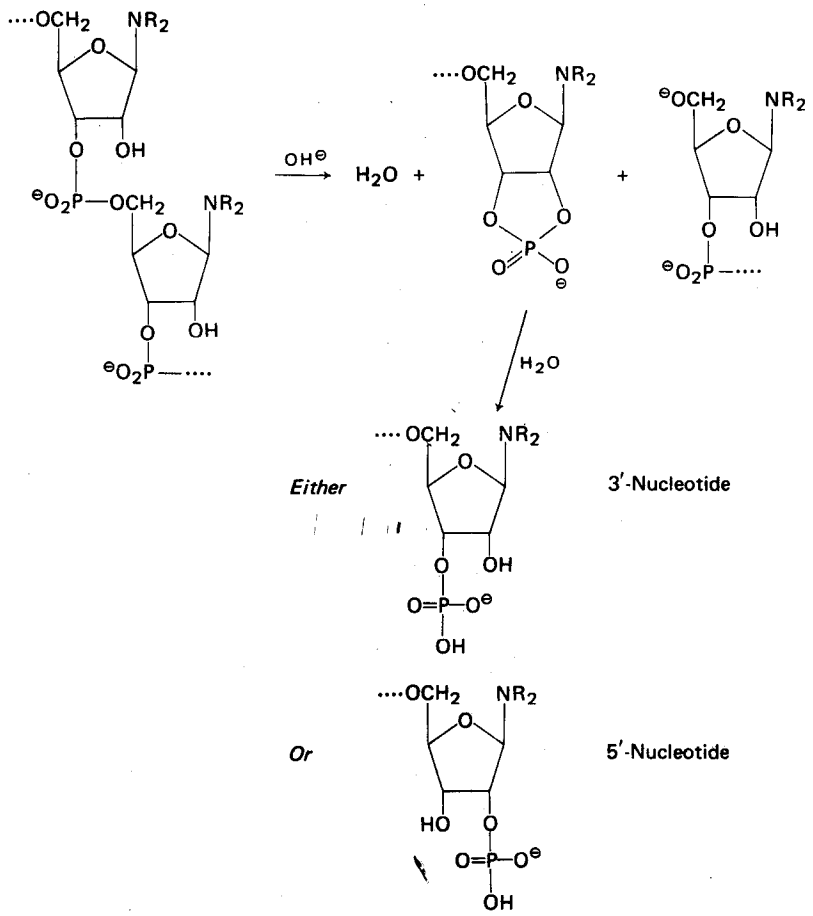


FIG 1.2 Involvement of the 2′-OH groups in the alkaline hydrolysis of a 5′-phosphate ester bond in RNA.

Phosphodiesters are in general fairly stable to hydrolysis. RNA however, is an exception to this general rule. In this case the 5′-phosphate ester bond is relatively easily hydrolysed by alkali. The reason is that the free 2′-hydroxyl group is involved in the reaction to produce a cyclic phosphate intermediate (Fig. 1.2). This type of acceleration of the reaction of one group by the participation of another one in a different part of the same molecule is known as anchimeric assistance. In the present case there are two consequences of this anchimeric effect. One is that when compared to most phosphodiesters (including DNA) RNA is relatively easily hydrolysed by alkali; the other is that the products are mixtures of 3′- and 2′-nucleotides owing

to the alternative modes of hydrolysis of the cyclic phosphate intermediate (see Fig. 1.2). This simple observation (that alkaline hydrolysis of RNA yielded both 2′- and 3′-nucleotides) led to a good deal of confusion during the early chemical work on the elucidation of the outline structure for RNA.

The bases present in nucleic acids are substituted pyrimidines or purines. In Fig. 1.3 are the structures of the parent heterocyclic compounds (pyrimidine and purine themselves) together with the conventional numbering of the atoms in the rings. This is the current convention for numbering and is employed throughout this book. However, many earlier publications use alternative conventions and it is extremely important to establish the convention from the context of the writing in reading any paper dating before about 1965. Obviously pyrimidine itself is not a secondary amine and consequently could not form an *N*-glycoside except as a quaternary ammonium salt. The pyrimidines which occur in nucleic acids are invariably

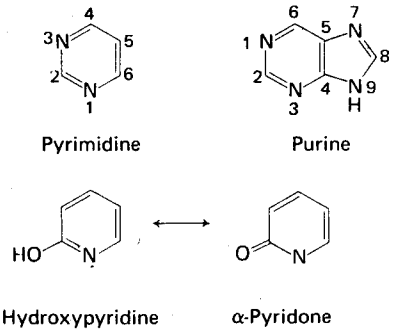


FIG 1.3 Conventions for numbering pyrimidine and purine and an example of a tautomeric equilibrium between a hydroxyimine (lactim) and a lactam.

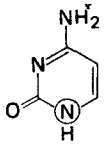
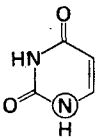
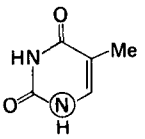
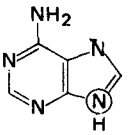
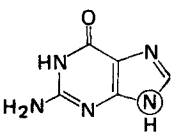
substituted with an oxygen on C-2. A secondary amine group is then generated at N-1 as a consequence of a lactam/lactim equilibrium (sometimes erroneously referred to as a keto/enol equilibrium) of the type illustrated (Fig. 1.3) for the simpler case α-pyridone.

The major bases in RNA are cytosine and uracil (pyrimidines) and adenine and guanine (purines). In DNA, the uracil is replaced by thymine (5-methyluracil); the other three bases are as in RNA. The structures of these bases and the names of the corresponding nucleosides and nucleotides aregiven in Fig. 1.4. In the pyrimidine nucleosides N-1 is involved in the *N*-glycoside bond linking the base to the sugar; in the purines it is N-9. Note that the prefix 'deoxy' is dropped from the names of deoxyribonucleoside and deoxyribonucleotides of thymine. This convention originated at a time when it was assumed that thymine only occurred in DNA. It is now known that some RNA molecules contain thymine as a 'minor base'. In this case, of course, it is attached to ribose and the nucleoside has to be called thymine riboside.

Several, but not all, nucleic acid molecules contain minor components. In RNA the minor components are 2′-methylribose, bases methylated, acetylated, reduced or otherwise modified and a nucleoside in which uracil is attached to the ribose via a carbon-carbon bond (C5-Cl′) rather than an *N*-glycoside. This last nucleoside is known as pseudouridine. In DNA the only minor components identified to date are in the form of substituted bases (certain of which, in some viral DNAs contain oligosaccharide side chains).

The question of shorthand forms of representing these components of nucleic acids is somewhat confused. In laboratories using these compounds (and in many publications) the capital initial letters C, U, T, A and G are used to refer to the bases or nucleosides or

nucleotides depending on the context. The conventional abbreviations currently advocated by the advisory committee of IUPAC-IUB[3] are to use three letter abbreviations for bases and ribosides as follows (base first and then riboside): for C Cyt and Cyd, for U Ura and Urd, for T Thy and Thd, for A Ade and Ado, for G Gua and Guo. Pseudouridine becomes Ψrd. Certain minor bases and their ribosides have similar abbreviations (see chapter 2 for the structures). These include xanthine and xanthosine (Xan and Xao), hypoxanthine and inosine (Hyp and Ino) and orotic acid and oritidine (Oro and Ord). In nucleosides in which the sugar is a pentose other than ribose, the three-letter abbreviation for the nucleoside is prefixed by a small letter: d

BASE	Cytosine	Uracil	Thymine	Adenine	Guanine
RIBONUCLEOSIDE	Cytidine	Uridine		Adenosine	Guanosine
RIBONUCLEOTIDES (2′-, 3′- and 5′-)	Cytidylic acids OR cytidine monophosphates (CMPs)	Uridylic acids OR uridine monophosphates (UMPs)		Adenylic acids OR adenosine monophosphates (AMPs)	Guanylic acids OR guanosine monophosphates (GMPs)
DEOXYRIBONUCLEOSIDE	Deoxycytidine		Thymidine	Deoxyadenosine	Deoxyguanosine
DEOXYRIBONUCLEOTIDE (3′- and 5′-)	Deoxycytidylic acids OR deoxycytidine monophosphates (dCMPs)		Thymidylic acids OR thymidine monophosphates (dTMPs)	Deoxyadenylic acids OR deoxyadenosine monosphosphates (dAMPs)	Deoxyguanylic acids OR deoxyguanosine monophosphates (dGMPs)

FIG 1.4 Structures of common bases and names of their derivatives. A circle has been drawn round the nitrogen atom attached to the sugar (via an *N*-glycoside bond) in the nucleosides.

(deoxyribose), a (arabinose), x (xylose) or l (lyxose). Note that according to this convention, thymidine is dThd and thymidine riboside is Thd. Another example would be the drug known as cytosine arabinoside (the same as cytidine but with arabinose replacing ribose) which means aCyd. The conventions shown in Fig. 1.4 for the nucleotides are allowed by the IUPAC-IUB rules. As an example the two deoxyguanylic acids are 3′-dGMP and 5′-dGMP and the thymidylic acids 3′-dTMP and 5′-dTMP.

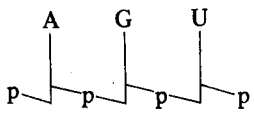
There are several conventions for representing nucleic acid sequences. They all depend on the assumption that each individual nucleoside is 'viewed' from the same vantage point as is implied by the structures in Figs. 1.1 and 1.2, that is to say that 5′ lies to the left of 3′. The shortest convention accords with the following rules. (1) Do not distinguish between RNA and DNA, the nature of the sugar will be apparent from context. (2) Use a single capital letter to identify the major nucleosides. (3) Use a hyphen to indicate a phosphate group. (4) Modifications of bases are shown by small letters preceding the capital initial letter, superscript numbers indicating position and subscripts number or substituents in that position if necessary. (5) Modifications to sugars are represented by small letters following the symbols for nucleosides.

From rules 2 and 3 we could represent 5′-UMP as -U and 3′-UMP as U-. A-U would mean a dinucleoside monophosphate which consists of an adenosine with a free 5′-OH group (no phosphate) a phosphodiester (linking 3′ of the A to 5′ of the U) and a uridine with a free 3′-OH. The dinucleotide -A-U has a phosphomonoester group attached to 5′ of the A and is otherwise the same. The dinucleotide A-U- has 3′-phosphomonoester group on the U. A relatively minor additional point is that A-U! is used to indicate the same dinucleotide in which a 2′,3′-cyclic phosphate is present on the uridine. A more extended oligonucleotide would be represented as A-A-G-A-C-C-U-. In such a molecule the 'left-hand' end is referred to as the 5′-end (in this case an adenosine with a free 5′-OH) and the 'right-hand' end as the 3′-end (in this case a uridine bearing a 3′ phosphate). An alternative to a hyphen for representing phosphate is p. In this book, I only use p when several phosphates are linked by phosphoric anhydride links. Thus pyrophosphate is pp, ATP is pppA and an RNA molecule with a 5′-triphosphate group at one end would be written pppN-N-N-N. . . .

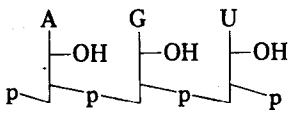
In addition to the letters C, U, T, A and G the following are recognized I (inosine), Ψ (pseudouridine), X (xanthosine), R (unidentified or non-specified purine), Y (unidentified or non-specified pyrimidine), N (completely unidentified or non-specified nucleoside). In this book, all these abbreviations are employed except R and Y. In my view these are extremely confusing as R means 'alkyl' and Y has a quite specific meaning in sequences of tRNA (see p. 301). Therefore I use Py for an unidentified pyrimidine and Pu for an unidentified purine. Examples of prefix letters are m (methyl, ac (acetyl), h (hydrogenated), s (thio). Examples of the use of these abbreviations are contained in the oligonucleotide C-hU-A-m^6_2A-I-ac_4C-Gm-G-, which has a free hydroxyl at the 5′-end and the residues C, 5,6-dihydroU, A, 6-dimethyl A, I, 4-acetyl C, 2′-methyl G and G (with a 3′-phosphomonoester).

Synthetic polynucleotides are represented in the following ways. A homopolymer (one in which every base is the same) is represented simply as, for example, poly A (poly adenylic acid) or poly dA (polydeoxyadenylic acid). A polymer with a short, defined, repeating sequence is represented either as poly d (A-T) or d(A-T)$_n$ (for -A-T-A-T-A-T- . . . , all deoxyribosides). Random polymers are molecules of undefined sequence but known overall composition; these are represented as, for example, $(A_3C_2)_n$ for a random co-polymer of adenylic and cytidylic acids with the molar ratio of A and C as indicated (3:2 in this example).

Another set of conventions used a vertical line to represent the side view of the ring of the sugar molecule. These representations are most easily understood by reference to a particular example. Consider the trinucleotide -A-G-U-. This could be represented as:



or alternatively, if we wished to make it quite clear this was an RNA fragment (and not a DNA fragment) we could include the 2′-OH groups:



so that the equivalent DNA fragment would become: