laboratory techniques

in biochemistry and molecular biology

general editors: T. S. WORK and E. WORK

determination of sequences in RNA

G. G. BROWNLEE

DETERMINATION OF SEQUENCES IN RNA

G. G. Brownlee

Medical Research Council,
Laboratory of Molecular Biology,
Hills Road, Cambridge, England

1972

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM · LONDON
AMERICAN ELSEVIER PUBLISHING CO., INC. – NEW YORK

Library of Congress Catalog Card Number: 76-157037
ISBN North-Holland – series 0 7204 4200 1
– part 3.I: 0 7204 4209 5
ISBN American Elsevier: 0 444 10102 0

63 illustrations & graphs, 30 tables

Published by:
NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM

Sole distributors for the U.S.A. and Canada:
AMERICAN ELSEVIER PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE, NEW YORK, N.Y. 10017

This book is the pocket-edition of Volume 3, Part I, of the series 'Laboratory Techniques in Biochemistry and Molecular Biology'.
Volume 3 of the series contains the following parts:
Part I Determination of sequences in RNA, by G.G. Brownlee
Part II Techniques of lipidology: isolation, analysis and identification of lipids, by M. Kates

---

Printed in The Netherlands

Contents

Dedicated to my mother

Preface

The aim of this monograph is to survey the methods that have recently been developed for the determination of the sequence of small RNA molecules. I have attempted to give the necessary experimental details for a reader to reproduce these methods in the laboratory and apply them to his own particular sequence problem. The chapter on electrophoresis equipment and technique will, I hope, be especially useful to those envisaging setting up the equipment for sequence determination. In this book, I have covered both the classical methods and the newer radioactive methods. However, my emphasis is naturally towards the latter as it is with these methods that I am most familiar. The recent spate of reports in scientific journals of the sequence of various transfer RNA molecules by radioactive methods have shown that this is a powerful and rapid approach. However, here I have not concerned myself with discussing the results of sequence work and they are only discussed in so far as is necessary to illustrate the methodology of sequence determination. Nor is any reference made to the methods of sequence determination of DNA, in which there has been less progress than with RNA. At the time of writing the longest known RNA sequence is that of the 6S RNA determined by the radioactive approach. It is certain, however, that larger sequences will be established by using these methods. Perhaps we may conjecture that the time will not be too distant when we may know the entire RNA sequence of the ribosome.

List of abbreviations

(excluding those defined in appendix 1 and table 7.2, p. 205)

tRNA	Transfer RNA.
rRNA	Ribosomal RNA.
mRNA	Messenger RNA.
G*	Unknown derivative of guanosine.
Py	Any pyrimidine nucleoside.
Pu	Any purine nucleoside.
R_u	Chromatographic or electrophoretic mobility of an unknown compound relative to uridine 2′(3′)-phosphate = 1.00.
P_i	Orthophosphate.
P-RNase (RNase A)	Pancreatic ribonuclease.
T_1-RNase (RNase T_1)	Takadiastase T_1 ribonuclease.
T_2-RNase (RNase T_2)	Takadiastase T_2 ribonuclease.
U_2-RNase (RNase U_2)	Ustilago U_2 ribonuclease.
DNase	Pancreatic deoxyribonuclease.
mA	Milli-ampères.
C	Curie.
dpm	Radioactive disintegrations per minute.
p.f.u.	Plaque-forming units.
u	Units.
EDTA	Ethylenediaminetetraacetic acid.
Tris	Tris-(hydroxymethyl)-aminomethane.

Subject index p. 261

SDS	Sodium dodecyl sulphate.
SSC	0.15 M sodium chloride–0.015 M sodium citrate, *p*H 7.0.
BBOT	2,5-bis-(2(5-tert-butylbenzoxazolyl))-thiophene.
CMCT	*N*-cyclohexyl-*N'*-(β-morpholinyl-(4)-ethyl)-carbodiimide-methyl-*p*-toluene sulphonate.
TEC	Triethylamine bicarbonate.
BSA	Bovine serum albumin.
M.W.	Molecular weight.
R_t	Chromatographic mobility expressed relative to rate of front.
rpm	Revolutions per minute.

1.1. Historical aspects

Although nucleic acids and the four common bases (fig. 1.1) derived from them have been known to chemists since the last century, there was no precise knowledge of their function until 1944 when Avery et al. showed that DNA was the essential component of the transfor-

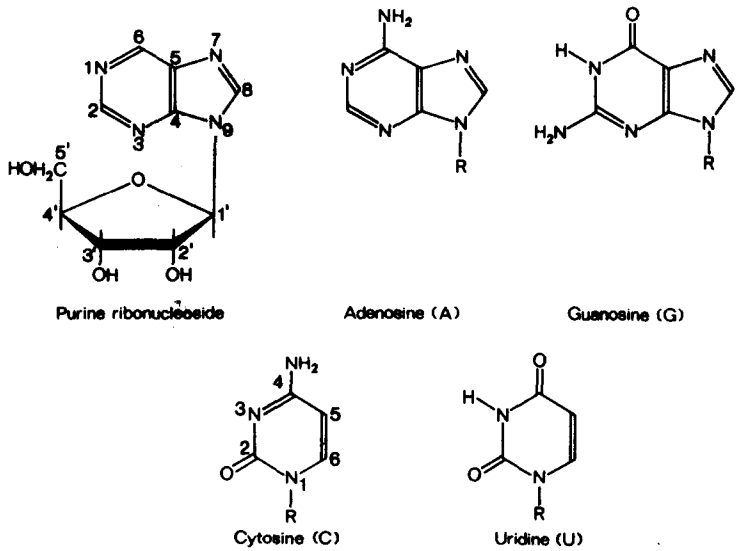


Fig. 1.1. Formula of the four common mononucleosides in RNA, including the numbering of the carbon atoms.

Subject index p. 261

ming factor in pneumococci, and thus carried inheritable information. It required further careful chemical analysis, coupled with advances in paper and ion-exchange chromatography, to prove that the monomeric units of nucleic acids were joined together by a 3′-5′ phosphodiester linkage (fig. 1.2) between adjacent ribose moieties (e.g. see

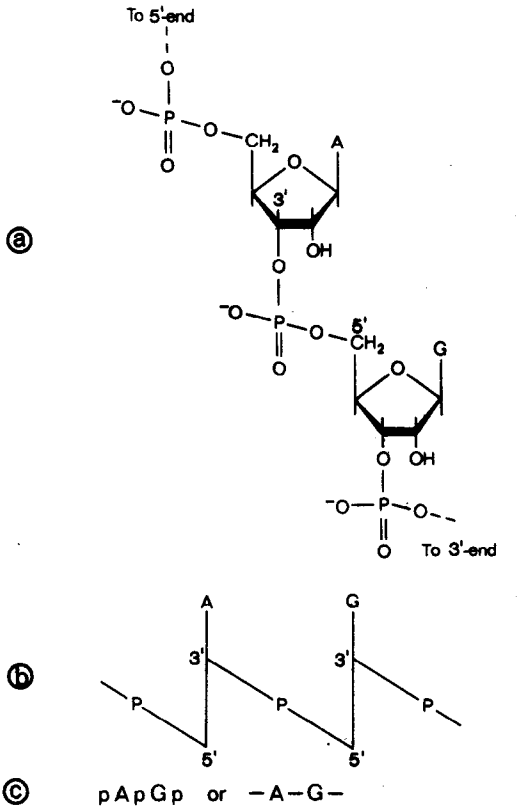
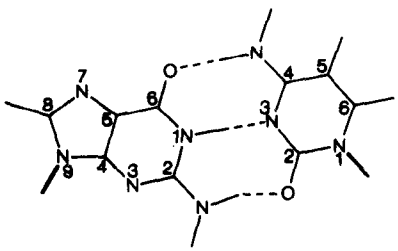
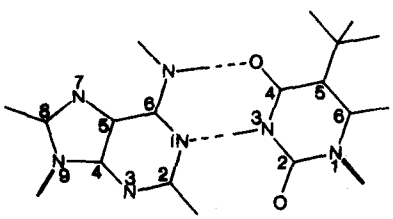


Fig. 1.2. Linkage and nomenclature in RNA. (a) shows the ribose-phosphate backbone which may also be drawn as in (b) or omitted altogether as in (c). Notice that in all cases the phosphodiester bridge is linked to a 3′-C-atom on its left and a 5′-C-atom on its right.

Brown and Todd 1952). However, it was not clear until 1953, when Watson and Crick correctly deduced the fibre structure of DNA, that a precise *sequence* of bases was important in maintaining the double-helical structure of the DNA. In particular, the structure was significantly stabilised by hydrogen bonds between specific complementary bases: adenine with thymine, and guanine with cytosine (fig. 1.3).



G — C pair



A — T pair

Fig. 1.3. Base pairs in DNA drawn approximately to scale. The hydrogen bonds are dotted and the glycosidic linkages are thickened. A–T is replaced by A–U in RNA. G–U ('wobble' pair) occurs in RNA and is rather similar to the G–C pair shown except that the third (uppermost) hydrogen bond is not made.

This specific pairing of bases in the double helix suggested to Watson and Crick (1953) a model for replication involving these base pairs by which the sequence was accurately conserved. Only later did Kornberg and his collaborators (Lehman et al. 1958) isolate a DNA polymerase from *E. coli* which had the correct properties for accurately copying DNA. The implications of these observations were quite clearly that a

nucleotide sequence defined, in an as yet unknown way, the complete phenotype. RNA (which is not usually the gene) must also have a defined sequence as it, also, is synthesised by a copying of (usually) DNA so as to form a complementary sequence to one of the DNA strands in the presence of an RNA polymerase (Hurwitz et al. 1961). Indeed, from our knowledge of the defined amino-acid sequence of any given protein, and a knowledge that it is synthesised on a mRNA template (Volkin and Astrachan 1956) it would be surprising if anything less than a precise sequence existed for mRNA. It is abundantly clear that nucleic acids are exposed to the process of natural selection and that the sequences which exist are those most adapted to function. Thus, in order to define and understand their function we must study their sequence.

1.2. *General approach to sequence determination*

Progress in the determination of nucleotide sequences in nucleic acids has lagged behind progress made in the sequence analysis of proteins. The basic approach is very similar for both polymers and requires the partial degradation of the molecule into smaller fragments of various sizes. After purification, the sequence of these smaller fragments is established by further degradative (usually enzymatic) methods. By analysing a sufficient number of fragments of various sizes, one hopes to accumulate sufficient information to logically deduce the entire sequence. The reason for the slow progress with nucleic acids has been partly due to experimental difficulties in obtaining homogeneous and undegraded low molecular-weight preparations of a given nucleic acid, and partly due to the lack of techniques for fractionation of the extremely complex mixture of related products in a partial digest. In addition, as there are only 4 basic units in nucleic acids compared to twenty amino acids in proteins, it is necessary to isolate, in a pure form, rather longer sequences for them to be unique. As with proteins, only a few of the known nucleic acids are small enough to be susceptible to the present methods of sequence analysis. Unfortunately the smallest known RNA molecule, from satellite tobacco necrosis virus, has over

1100 nucleotides per molecule (Markham 1963) and is not a good candidate for sequence work. However, it has been known for some time that the most promising nucleic acids were the transfer RNA molecules (about 80 residues long) and it is with these that most progress has been made.

As I outline more fully in ch. 2, the particular amino-acyl-specific tRNA in question must first be purified and only then may its sequence be determined. Holley et al. (1965a) in their classic work did this and successfully deduced the sequence of yeast alanine tRNA after separating fragments of the molecule by chromatography on DEAE-cellulose columns. In all this work, the chromatographic fractionation procedure on *columns* was undoubtedly the rate-limiting step and presented the greatest difficulty in sequence determination. However, by making use of uniformly ^{32}P-labelled nucleic acids and using two-dimensional fractionation methods on *modified paper*, with its high resolution, Sanger and his collaborators (Sanger et al. 1965; Sanger and Brownlee 1967) have to a large extent simplified these fractionation problems. Moreover, by using labelled nucleic acids many of the sequence procedures are also considerably simplified because of the sensitivity of the radioactive techniques. This radioactive approach turned out to be a break-through in methodology and allowed the rapid elucidation of sequences of a number of tRNA and 5S ribosomal RNA molecules. The two contrasting approaches are discussed in detail in the rest of this monograph, ch. 2 being devoted to a description of the classical methods of Holley and others, whilst the succeeding chapters describe the radioactive methods of Sanger and his colleagues. The evidence for one sequence, that of 5S RNA of *E. coli*, is presented in detail in ch. 5 so that the reader may be aware of the nature of the evidence required to deduce the structure of a molecule of this length (120 residues). The next chapters describe specialised problems of radioactive sequencing. Ch. 6 is devoted to a discussion of the sequence determination of very long end-products of T_1-ribonuclease digestion. In ch. 7 I discuss some of the difficulties and peculiarities of sequencing tRNA at the radioactive level. Ch. 8 describes two interesting applications of the radioactive methods which are not concerned with de-

Subject index p. 261

fining a total sequence. These are end-group determination and sequences containing methylated bases in high molecular weight ribosomal RNA. The last chapter describes some very recent applications of the radioactive sequence methods for sequencing RNA labelled in vitro. The results of sequence work are not discussed in this monograph.* A useful summary of radioactive sequence methods appears in appendix 7.

1.3. Nomenclature

One of the necessary facts of chemistry is the need to abbreviate the long formal names of compounds in order to simplify their representation. Although this book is concerned with practical techniques rather than nomenclature, the results of applying the techniques will be enumeration of nucleotide sequences. Representation is, therefore, of importance, and the reader will perhaps excuse a digression into a discussion of this problem.

The accepted way of presenting a sequence is a linear order of single capital letters C, A, G and U representing the four common nucleosides cytidine, adenosine, guanosine and uridine. These are joined together by hyphens representing the 3′-5′ phosphodiester linkage if the sequence is known. Where the composition of an oligonucleotide is known, but the residue order is not, this is indicated by enclosing the residues of unknown order within brackets with a comma rather than a hyphen between each residue symbol. Minor nucleosides are represented as derivatives of the nucleosides above, e.g. N^7-methyl guanosine is represented as m^7G. Full rules for representation of minor nucleosides appear in the IUPAC-IUB tentative rules (J. Biol. Chem. (1966) *241*, 527) which have been adopted in this book and examples appear in table 7.2. An extract of these rules appears in appendix 1.

Sometimes, however, it is convenient to abbreviate the accepted nomenclature even further than is suggested in these rules. In sequence

* For a collection of some results of sequence determination see Sober (1968) and Zachau (1969).