NUCLEIC ACID SEQUENCE ANALYSIS

STANLEY MANDELES

Nucleic Acid Sequence Analysis

STANLEY MANDELES

COLUMBIA UNIVERSITY PRESS
NEW YORK AND LONDON 1972

Stanley Mandeles is Professor and Chairman of the Department of Chemistry at Douglass College of Rutgers University, New Brunswick, New Jersey

Library of Congress Catalog Card Number: 79-186389
ISBN: 0-231-03130
Printed in the United States of America

NUCLEIC ACID SEQUENCE ANALYSIS

COLUMBIA SERIES IN MOLECULAR BIOLOGY

Ernest Borek, Advisory Editor

Molecules and Evolution, Thomas Hughes Jukes
Virus-Induced Enzymes, Seymour S. Cohen
The Cancer Problem: A Critical Analysis and Modern Synthesis, Armin C. Braun
The Modified Nucleosides in Nucleic Acids, Ross H. Hall
Of Microbes and Life/Les Microbes et la Vie, Jacques Monod and Ernest Borek, editors
Control Mechanisms and Protein Synthesis, S. D. Wainwright
Nucleic Acid Sequence Analysis, Stanley Mandeles
Information Theory and the Living System, Lila L. Gatlin

To Francine

ACKNOWLEDGMENTS

Any virtue to be found in the following pages is due in large measure to the counsel and encouragement of Professor I. Tinoco, Jr., Department of Chemistry, University of California. I am indebted also to Frances Davis, Roselyn Ferreira, David Garfin, Philip Borer, and Mark Mandeles for their expert technical assistance. I thank the National Institutes of Health for support through Grant No. GM 12158, and the National Aeronautics and Space Administration for support through Grant No. NsG 479.

CONTENTS

1.

INTRODUCTION

WHY DETERMINE SEQUENCES?

It seems certain that the master plan for living, reproducing cells is encoded in deoxyribonucleic acid (DNA). From its molecules comes the information to sustain the homeostatic, adaptive, and replicative mechanisms that are necessary for life to persist. What form, then, does this information take? The evidence is that the biological information resides in the linear sequence of bases on the polynucleotide backbone in almost perfect analogy with the printed material on this page. First, there is a biological alphabet of four letters. These are symbols for the four bases found in DNA: A = adenine, C = cytosine, G = guanine, T = thymine. Words are composed of any three of the four letters and are translated as specific amino acids. Because there are only sixty-four possible arrangements of three letters from an alphabet of four, the dictionary consists of sixty-four words. A set of these words constitutes a sentence that is an explicit instruction for the assembly of specific amino acids into a distinctive protein. The sentence is also an implicit instruction to perform some functional or structural task. A combination of sentences, or a paragraph, often contains the information required for a cell's successful response to an environmental change. Finally, enough of these paragraphs constitute a book (of life, if you will).

The point of this analogy is not its accuracy. It is the hypothesis that all of life's chemical functions and characteristics are ultimately

specified by the sequence of bases in nucleic acids. It should be noted that the role of nucleic acids is not limited to the status of a repository or tape of biological information. Ribonucleic acids (RNA), in particular, participate in the expression of this information. The outlines of this process have become clear. Portions of the DNA are copied to specific messenger RNAs (mRNA) which are then translated into corresponding specific protein structures. Other portions of DNA are copied to form transfer RNA (tRNA), 5S RNA, and ribosomal RNA (rRNA). These molecules play essential roles in the translation of base sequences but do not themselves carry the encoded protein sequence.

In the face of this situation one wonders what there is about mRNA that permits it to be translated into an amino acid sequence while tRNA, 5S RNA, and rRNA serve only as part of the translation mechanism. The answer must be traceable to the differences in the base sequences. The most recent evidence indicates that tRNA, 5S RNA, and rRNA contain sufficient numbers of biochemically modified bases or are so highly structured that they may not be recognizable as mRNA and therefore may not perform as messengers. Furthermore, mRNA may contain specific sequences that signal the initiation of the translation process, and these may be missing from the other varieties of RNA. In any event, it is difficult (if not impossible) to dissociate any property or function of nucleic acids from dependence on base sequence. Knowledge of the base sequence of any nucleic acid is essential for rigorously understanding its function in biological systems.

Recently, answers have been obtained to such fundamental questions as:

1. Can a linear correspondence be demonstrated between the sequence of amino acids in a known protein structure and the sequence of bases in its mRNA? (Yes, first found in the case of coat protein of bacteriophage R17 and a portion of its RNA; Adams *et al.*, 1969.)

2. Are there any base sequences in mRNA that do not code for amino acids? (Apparently so in the R17 RNA initiation areas, Steitz, 1969, and intercistronic areas, Nichols, 1970.)

3. Is the genetic code degenerate? (Yes, at least for bacteriophage RNA; Adams *et al.*, 1969.)

4. In a population of messenger RNA molecules specific for one protein, are there any variations in base sequences? (None determined to date in bacteriophage RNA, one variation in a plant virus.)

5. To what extent do tRNA molecules occur that bear the anticodons for the alternate codons in mRNA? Is the sequence of these other tRNAs the same except for the alteration in the anticodon segment? What are the species differences in the tRNA for any one amino acid? (The data bearing on this last point are rapidly accumulating; see Jukes, 1970.)

6. What are the functions of 5S RNA and rRNA, and to what extent are these functions dependent on specific sequences? (Under study in several laboratories.)

These questions have been directed mainly to RNA structure. Of equal importance and interest are the details of DNA structure. Sequence analysis of DNA should augment the information obtainable from genetic mapping experiments and lead to a better understanding of the actual physical structure of the "taped" genetic instructions. Answers to questions concerning the extent of redundancy, or the mechanism of replication of DNA and the synthesis of RNA copies, undoubtedly depend on sequence information.

Another area of interest is the possibility of the correction of genetic defects through use of available molecular biology techniques. It is plain that detailed information concerning the sequence of bases in DNA will be necessary before any specific action of this sort can be taken. It is hoped that the decisions to utilize the available technology will be made with wisdom in view of the far-reaching effects of such changes.

Use of sequence information in the control of nucleic acid-directed disease processes can be made with a great deal less trepidation. Such diseases include influenza, leukemia, cancer, lupus erythematosis, and some other degenerative types. Knowledge of the base sequences in these errant or intruding nucleic acids should prove important in the prevention, control, and cure of these diseases.

In view of the many compelling reasons, both theoretical and practical, for engaging in nucleic acid sequence analysis, it is unnecessary to invoke the Mount Everest challenge, "because it's

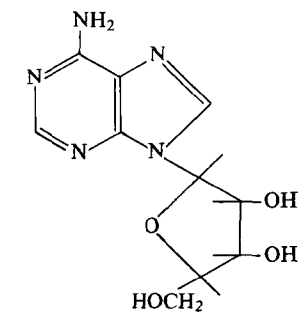
there." There is some attraction, however, in such a seemingly insurmountable problem.

The number of laboratories committed to the determination of base sequences in nucleic acids is rapidly increasing. In the case of RNA, this is partly due to the encouraging successes in the determination of the entire primary structure of 15 tRNAs, two 5S RNAs, and some internal and end sequences in viral and ribosomal RNA. These successes were in turn due to the development of a number of laboratory procedures of great sensitivity and reliability. There are many investigators whose interest in nucleic acid sequences is part of some other problem such as protein and nucleic acid synthesis or biochemical evolution. In order to accommodate a range of concerns, this manual will attempt to provide the working details of the various methods that have produced sequence information, as well as description of some potentially useful procedures that have been proposed but not exploited for one or more reasons. It is hoped that the inclusion of the latter procedures will serve as a stimulant to implementation or to further refinement for greater feasibility.

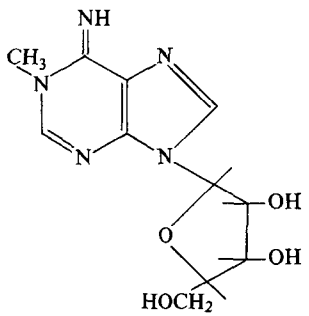
The first portion of the manual presents some basic considerations in nucleic acid sequence analysis, followed by descriptions of the various sequence strategies. In each of these strategies, one or more basic laboratory procedures are deemed important enough to warrant description in greater detail, and they are so indicated and presented in the second portion.

STRUCTURES AND SYMBOLS

In most treatments of chemical structure, some early agreement must be reached on the symbols to be used. For nucleic acids, the conventions recommended by the IUPAC-IUB Commission on Biochemical Nomenclature (1970) will be used to represent structure. As indicated in the names, ribonucleic acid (RNA) and deoxyribonucleic acid (DNA), the distinguishing characteristic is the presence of the sugar ribose in the former and 2-deoxyribose in the latter. The nucleic acids are polymers of the monomer nucleotides, and these consist of a purine or pyrimidine base, a sugar (ribose or deoxyribose), and a phosphate. Figure 1.1 shows the line structural formulas for some of

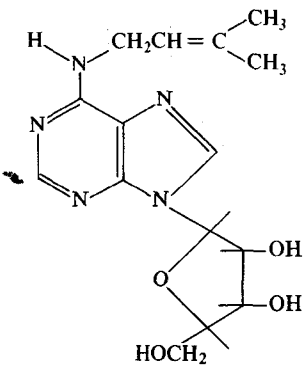


Adenosine

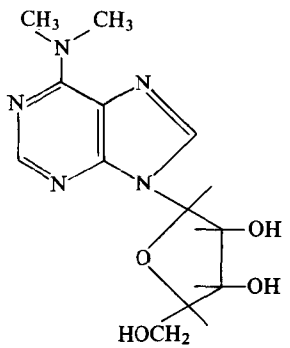


1–Methyladenosine

(a)

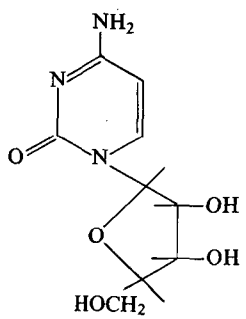


N(6) – Isopentenyladenosine

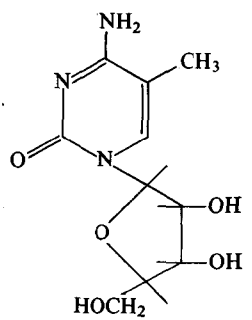


N(6) – Dimethyladenosine

(b)



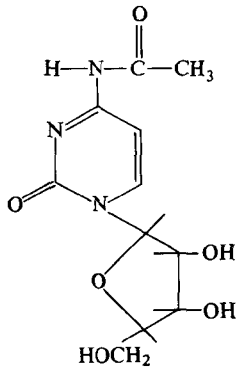
Cytidine



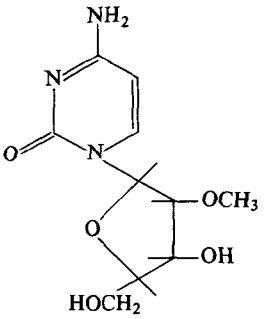
5–Methylcytidine

(c)

Fig. 1.1. Nucleoside structures.

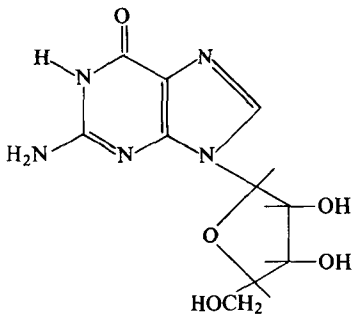


N(4) – Acetylcytidine

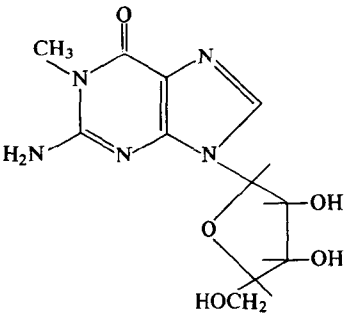


(d)

2′–O–Methylcytidine

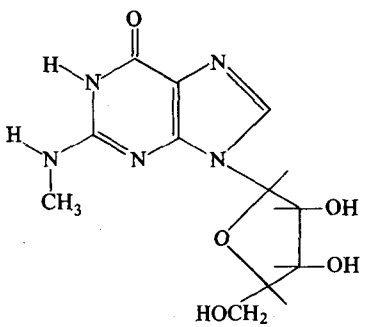


Guanosine

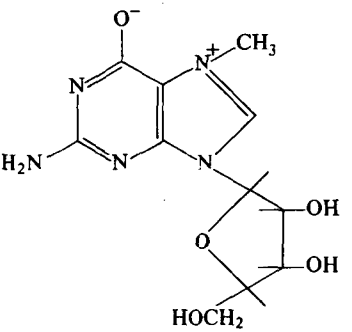


(e)

1–Methylguanosine



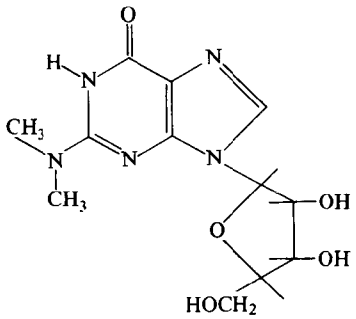
N(2)–Methylguanosine



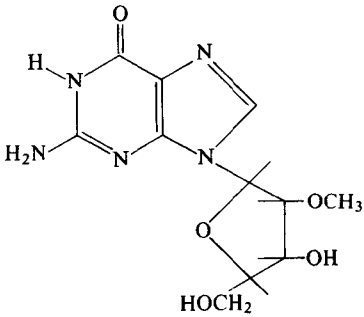
(f)

N(7)–Methylguanosine

Fig. 1.1 (continued)

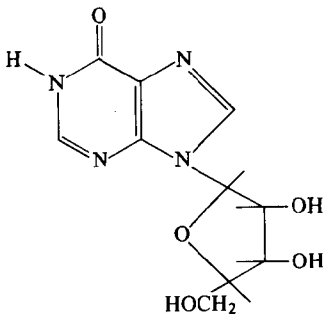


N(2)-Dimethylguanosine

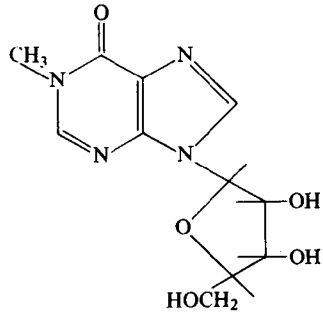


(g)

2'-O-Methylguanosine

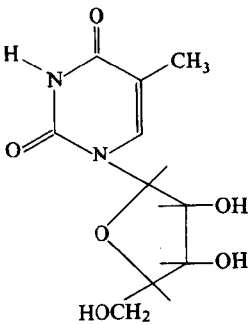


Inosine



(h)

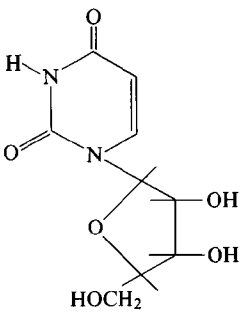
1-Methylinosine



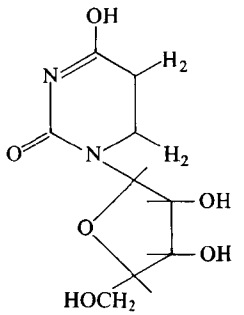
(i)

Ribosylthymine

Fig. 1.1 (continued)

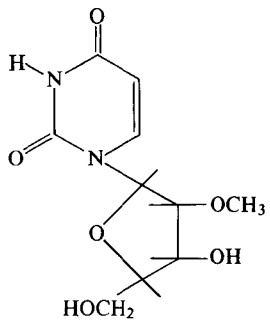


Uridine

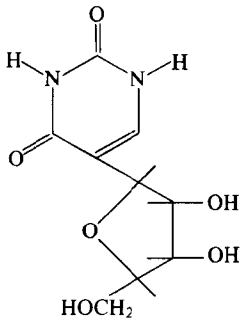


5,6–Dihydrouridine

(j)



2′–O–Methyluridine



Pseudouridine

(k)

Fig. 1.1 (continued)

the purine and pyrimidine bases found in nucleic acids. They are shown here as ribonucleosides, and they include the four major bases found in RNA (adenine, cytosine, guanine, and uracil) and DNA (adenine, cytosine, guanine and thymine), and the minor bases whose positions in nucleic acids have been determined by sequence analysis. The possibilities for keto-enol tautomerism are purposely ignored in favor of presenting single conventionalized structures.

Abbreviations for these structures that have been developed over the years are listed in column 1 of Table 1.1 (see Dayhoff and Eck, 1967). The entries in this column show the biochemists' preference for abbreviations containing no more than three upper-case letters. These abbreviations have been easy to memorize and convenient to use. Up

Table 1.1. Nucleoside symbols

	1	2	3
Adenosine	A	A	A
N(6)-Dimethyladenosine	DMA	m_2^6A	$A^{m_2^6}$
1-Methyladenosine	MA	m^1A	A^{m^1}
N(6)-Isopentenyladenosine	NPA	iA	A^{i}
2-Thiomethyl-6-isopentenyladenosine		(sA)	(sA)
Cytidine	C	C	C
5-Methylcytidine	5MC	m^5C	C^{m^5}
N(6)-Acetylcytidine	NAC	acC	C^{ac}
2′-O-Methylcytidine	OMC	Cm	C^m
Guanosine	G	G	G
N(2)-Dimethylguanosine	DMG	m_2^2G	$G^{m_2^2}$
1-Methylguanosine	1MG	m^1G	G^{m^1}
N(2)-Methylguanosine	2MG	m^2G	G^{m^2}
N(7)-Methylguanosine	7MG	m^7G	G^{m^7}
2′-O-Methylguanosine	OMG	Gm	G^m
Inosine	I	I	I
1-Methylinosine	MI	m^1I	I^{m^1}
Ribosylthymine	T	T	T
Uridine	U	U	U
5,6-Dihydrouridine	DHU	hU	U^h
2′-O-Methyluridine	OMU	Um	U^m
Pseudouridine	PSU	Ψ	Ψ
4-Thiouridine		4S	4S
Uridine derivative	U*	$\overset{*}{U}$	$\overset{*}{U}$
Unknown	Y, N, X	N	N
Purine	Pu	R	R
Pyrimidine	Py	Y	Y