Lynda Joy Simes

"THE DETERMINATION OF NUCLEOTIDE ARRANGEMENT IN OLIGONUCLEOTIDES DERIVED FROM DEOXYRIBONUCLEIC ACID"

Submitted for the degree of Master of Science in Biochemistry at Massey University, Palmerston North, New Zealand, 1969.

CONTENTS

ABBREVIATIONS AND CONVENTIONS USED IN REPRESENTING NUCLEOTIDES AND OLIGONUCLEOTIDES IN THIS THESIS

For convenience it has been found necessary to adopt a system of abbreviated representations of pyrimidine oligonucleotides, in accordance with the recommendations of the International Union of Pure and Applied Chemistry (1). All the compounds described in this thesis were deoxyribonucleotide derivatives with a regular 3'→5' internucleotide linkage and the recommended prefix "de-3', 5'-". In this thesis however the prefix has been omitted.

C and T are used to represent deoxycytidine and thymidine respectively, while p is used to represent phosphate esterified with the nucleosides. When p is written to the right of the symbol for the nucleoside this indicates the phosphate is esterified with the 3'-hydroxyl group of the nucleoside; when it is to the left, it forms the 5'-ester-.

Further abbreviations have been adopted for sequences that are expected to consist of mixtures of isomer nucleotides in unknown proportions; these are written with nucleoside symbols in parenthesis. Thus CpT represents deoxycytidylyl-(3'→5'-)-thymidine, TpC represents thymidylyl-(3'→5'-)-deoxycytidine and (CT)p represents a mixture of these two

isomers in unspecified proportions.

Several terms describing various types of isomers, used in this thesis, require definition. ISOPLITHS are oligonucleotides of identical chain length. Any group of isopliths, derived from digestion of a naturally occurring nucleic acid, will almost certainly contain sequences that differ from each other in their base composition. These are termed COMPOSITIONAL isomers. An oligonucleotide fraction representing a single compositional isomer is likely to consist of several SEQUENTIAL isomers, which differ from each other only in actual arrangement of bases.

Other abbreviations used in this thesis are:

DNA	deoxyribonucleic acid
RNA	ribonucleic acid
C	deoxycytidine
T	thymidine
Pu	purine
Py	pyrimidine
s.v.	snake venom phosphodiesterase
EDTA	ethylenediamine-tetra-acetate
Tris-HCl	tris(hydroxymethyl)amino methane-HCl

CMCD or CME-carbodiimide	N-cyclohexyl-N'-β-4-methyl-morpholinium ethyl carbodiimide
Pi	inorganic phosphate
UV	ultraviolet light
O.D.	optical density
U	an enzyme unit
E	extinction
$\frac{E_{280}}{E_{260}}$	ratio of extinction at 280 mμ and at 260 mμ

Introduction

(i) The importance of DNA sequences

The genetic material of all animals, plants bacteria, and of many animal and bacterial viruses has long been established as deoxyribonucleic acid or DNA, and recent studies (1a) have shown that its major function is to carry the genetic information required by a cell for the synthesis of species specific proteins.

This information is stored by the nucleic acid macromolecule in the form of a linear code determined by its intrinsic nucleotide sequence or primary structure. The information is carried in such a way that a specific sequence of three nucleotides has the ability to code for one of every type of amino acid found in protein. In recent years the message corresponding to each nucleotide triplet has been established (2).

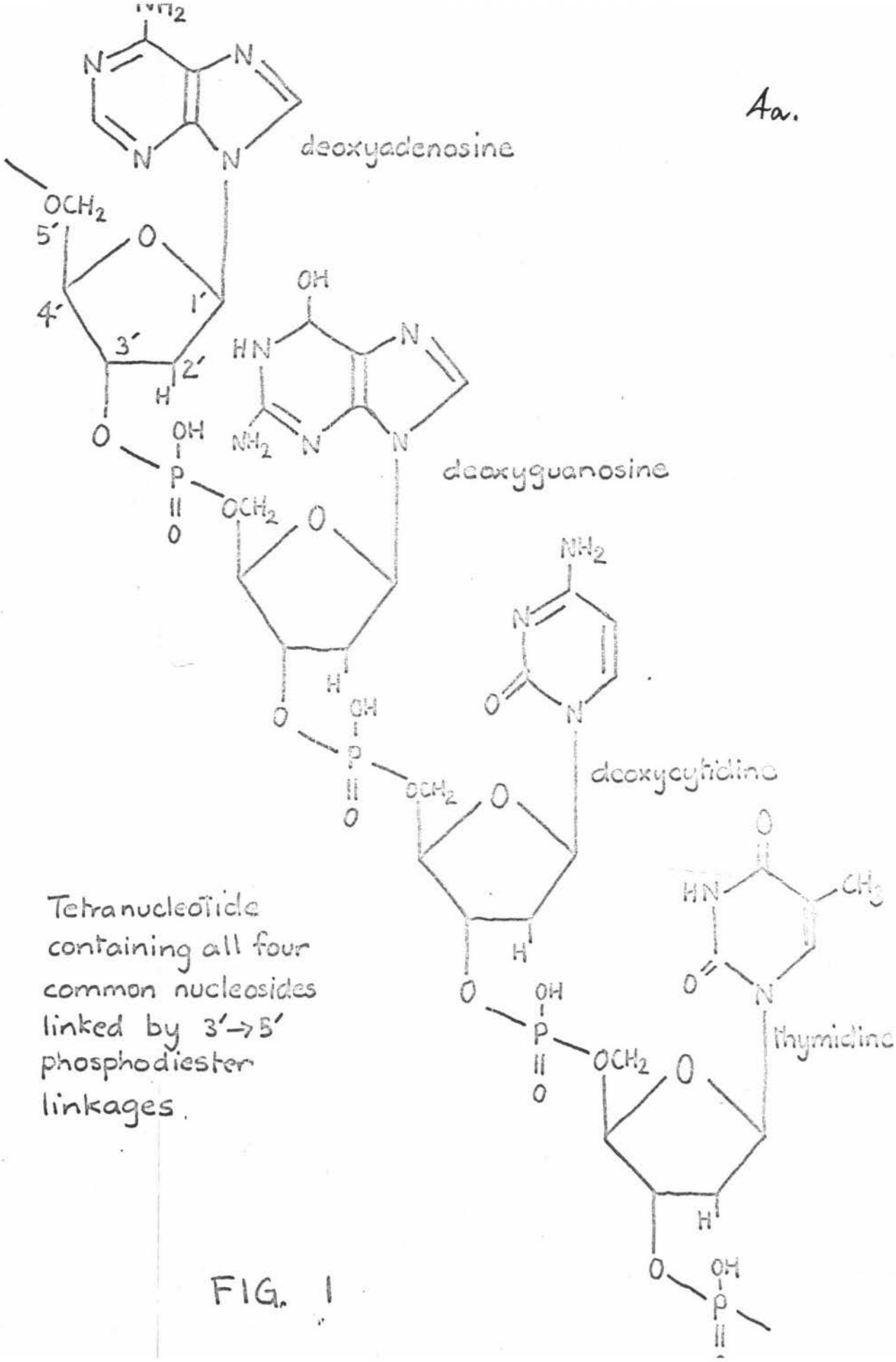
Besides coding for amino acids, nucleotide sequences exist which are known to code for ribosomal transfer RNA's. There are probably other sequences which are involved in a variety of special roles, the more important being

regulating and initiating transcription of the genetic message into functional messenger RNA, and the initiation of DNA replication. Because little information is available about the actual arrangement of bases needed to effect these functions, a knowledge of the complete nucleotide sequence of a biologically active DNA molecule may help to elucidate the nature of these extremely important processes. In addition it is hoped that new approaches can be found towards a better understanding of the actual changes to DNA caused by mutagens and carcinogenic agents, and it is also hoped that some knowledge can be obtained of the extent to which degenerate codons exist in genetic material, and the function of these codons when they do occur.

(ii) The structure of DNA

DNA molecules are linear polymers of deoxyribonucleotides linked through 3'→5' phosphodiester bridges. Four major nucleotides, two containing purine bases (adenine and guanine), and two containing pyrimidine bases (cytosine and thymine), are commonly found. A tetranucleotide containing an arrangement of these four bases is illustrated in figure 1.

In most cases the DNA molecule occurs in the form of a double right-handed helix consisting of two complementary strands of nucleotides held together by hydrogen bonding



Tetranucleotide containing all four common nucleosides linked by 3′→5′ phosphodiester linkages.

FIG. 1

(3). Stereochemical considerations restrict the formation of a stable hydrogen bonded structure to accommodate only pairing between adenine and thymine, or guanine and cytosine. Because of this restriction the two strands have complementary structures, and thus if the nucleotide sequence of one strand is specified then the sequence of the other (complementary) strand can be deduced.

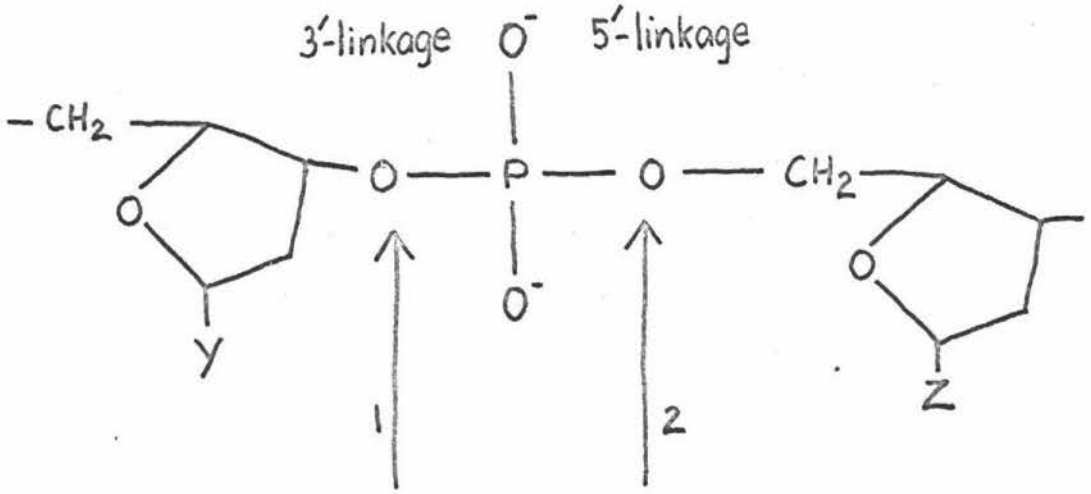
There are some exceptions to this model: the DNA of some small bacteriophages, e.g. the coliphages f1 and ØX 174, is in the form of a single covalently closed circular strand. Some other DNA's have been shown to have a circular double-stranded structure (4, 49).

Polynucleotide strands are polarised in the sense that each has a 3' and a 5' terminus, and in the Watson-Crick DNA model, the two strands of the duplex have opposite polarity.

As a consequence of the 3'→5' phosphodiester bonding of nucleosides, the internucleotide linkage can be broken in two possible ways, as is demonstrated in figure 2. This property has been found very useful in this investigation.

(iii) Determination of nucleotide sequences (15, 40, 50, 54)

Ribonucleic acid, a ribonucleotide polymer, has a similar primary structure to DNA, and is widely



Sites of hydrolysis at the internucleotide linkage

FIG. 2

distributed in cells where it has roles in the translation of the DNA genetic information into active cellular proteins. The primary structures of some small biologically active RNA molecules have been determined by the use of specific hydrolytic enzymes (endonucleases) which catalyse the cleavage of the polynucleotide chain into smaller, overlapping and identifiable sequences (5, 6).

Attempts at determination of nucleotide sequences in DNA have in the main followed along similar lines to these, but the problem is considerably more difficult than for RNA for several reasons:

a) As most cells contain many discrete chromosomes, the cellular DNA content is normally a complex mixture of double-stranded molecules; viruses, especially bacteriophage, provide the only known source of sequentially homogenous DNA. Although the DNA contained in the smallest bacteriophage is a single strand of nucleotides, it still has the high molecular weight of approximately 1.8×10^6, and contains at least 5,500 nucleotides (22).

b) Incomplete knowledge of the specificities of the few purified endonucleases that are capable of hydrolysing DNA prevents the use of these enzymes as analytical tools.

Because of this last restriction, it has been necessary to find suitable chemical aids to degrading

specific internucleotide bonds, and several of these have been described (7 - 12). The most successful of these has been the degradation induced by aromatic amines, particularly diphenylamine, in acid solution (13, 14). When DNA is incubated in 66% (v/v) formic acid containing 2% (w/v) diphenylamine, the acid conditions lead to the quantitative cleavage of the purine-deoxyribose glycosidic linkages, while the stable pyrimidine deoxyribose glycosidic bonds are completely untouched. Subsequent reaction of the resultant "apurinic acid" with the aromatic amine involving a β-elimination mechanism results in the release of inorganic phosphate from positions between adjacent purine nucleosides yielding pyrimidine nucleoside phosphates of general formula Py_nP_{n+1} (13). The reaction is shown in figure 3.

This chemical reaction gives reproducible results and it is quantitative; it produces from any DNA molecule a complex mixture of pyrimidine sequences of varying chain length and base composition. Although detailed methods have been developed for separation of these pyrimidine products into groups on the basis of their cytosine and thymine content, it has proved impossible to resolve mixtures of sequential isomers into their components. For example, the trinucleotide fraction of base composition $(CT_2)p_4$ may contain up to three sequential isomers, TTC,

Diphenylamine Reaction

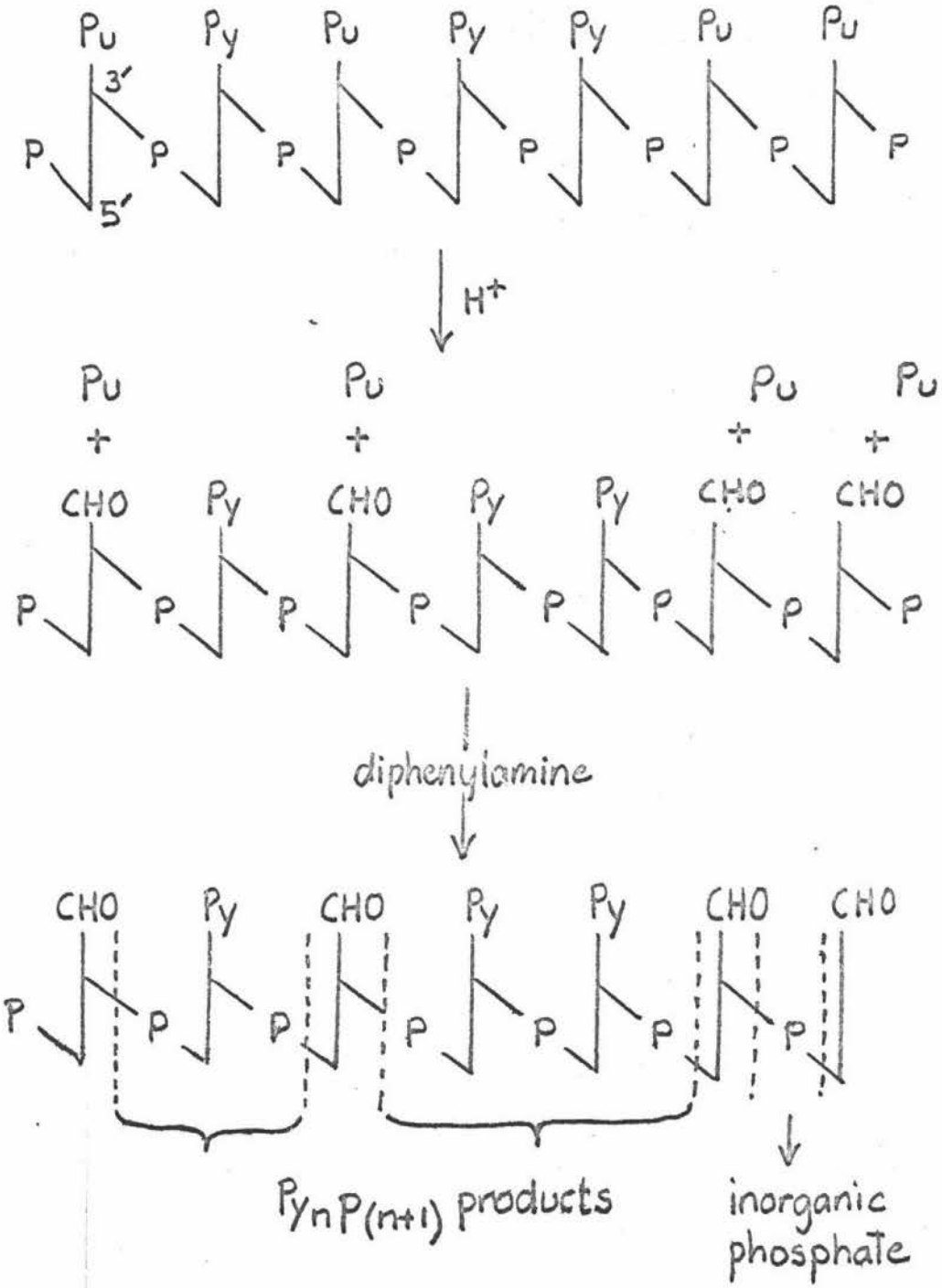


Fig 3

TCT, and CTT, which remain inseparable.

With longer nucleotides the problem becomes more complex, as larger numbers of sequential isomers can theoretically exist according to the general formula which relates numbers of possible sequential isomers to the chain length of the oligonucleotide:

For the isomer $(C_a T_b)p_{a+b+1}$ the number of sequential isomers is given by:

$$N = \frac{(a+b)!}{a! \times b!}$$

(iv) The present study

In the first part of this study, the digestion products from the action of diphenylamine on calf thymus DNA in acid solution have been separated as far as possible by physical methods, characterised, and measured quantitatively.

In addition, an investigation into possible methods of deducing the relative proportions of each sequential isomer in a mixture of isomers of identical base composition was undertaken, with special emphasis on the use of exonucleases (15).

A possible approach to this problem is found from a study of the specificities of snake venom and bovine spleen phosphodiesterases. Although neither of these enzymes will

attack a fully phosphorylated substrate of the type Py_nP_{n+1}, both will catalyse the exonucleolytic cleavage of the Py_nP_{n-1} oligonucleotides obtained by terminal dephosphorylation (16, 17):

$$\underset{(Py_nP_{n+1})}{pCpTpTp} \xrightarrow{\text{phosphomonoesterase}} \underset{(Py_nP_{n-1})}{OH\text{-}CpTpT\text{-}OH} + 2Pi$$

The specificity of snake venom phosphodiesterase is such that it will degrade terminally dephosphorylated oligonucleotides to 5'-mononucleotides, commencing its attack at the 3'-terminus, while the spleen enzyme will catalyse a similar degradation from the 5'-terminus (18).

For example, in the case of the $(CT_2)P_2$ which can exist as the following three sequential isomers (a) - (c):

(5')	(3')	
CpCpT		(a)
CpTpC		(b)
TpCpC		(c)

the two types of enzymic attack will proceed as follows:

$$C + pC + pT \xleftarrow{\text{s.v.}} CpCpT \xrightarrow{\text{spleen}} CpC + Cp + T$$

$$C + pT + pC \xleftarrow{\text{s.v.}} CpTpC \xrightarrow{\text{spleen}} Cp + Tp + C$$

$$T + pC + pC \xleftarrow{\text{s.v.}} TpCpC \xrightarrow{\text{spleen}} Tp + Cp + C$$

It will be seen that mononucleotide thymidine will arise only from the action of the spleen enzyme on isomer (a), or the snake venom enzyme on isomer (c). Measurement of the amount of this product in each type of enzyme digest should therefore allow the relative proportions of these two isomers in the mixture to be determined. Measurement of the other products will help confirm these values.

As shown above, there can be far more sequential isomers in the tetranucleotide and longer fractions than in the trinucleotide fraction. The use of these two nucleases alone therefore proves inadequate for determining the relative proportions of sequential isomers in these cases, although the result of their action does provide useful information. In other experiments therefore, attempts have been made to modify the specificities of the two nucleases, in this way deriving extra tools for the investigation of nucleotide sequences. A chemical blocking agent which attaches to the oligonucleotide was used for the modification (6, 38, 39, 41, 42, 43).

Experiments in which attempts were made to separate trinucleotide sequential isomers by ion exchange chromatography are also described.