SOME BASE SEQUENCE CHARACTERISTICS OF DEOXYRIBONUCLEIC ACIDS

by

Duncan James McGeoch, BSc

Presented for the degree of Doctor of Philosophy

Institute of Biochemistry

University of Glasgow

December 1970

CONTENTS

CORRECTION

Pages 121 and 123 are consecutive pages; there is no Page 122,

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to the following:-

to Professor J.N. Davidson and Professor R.M.S. Smellie for support and the provision of facilities,

to Dr. J.D. Pitts and Professor H.M. Keir for willing and excellent supervision, advice and discussion,

to the Science Research Council for their financial support in the first two years of this work, and to the Forrest Fund for Medical Research for support in the final year,

to Professor H. Subak-Sharpe, Dr. J.M. Morrison and Dr. J. Hay for valuable discussion,

to Messrs. D.S. Lochhead, P.J. Roach and D.J. Jolly for the gift of RNA polymerase, to Dr. G.G. Brownlee for the gift of U_2 RNase, to Dr. L.V. Crawford for virus preparations, and to Dr. A.M.Campbell, Dr. L.V. Crawford, Dr. M. Green, Professor H.M. Keir, Professor G. Pontecorvo, Dr. F.M. Ritossa and Professor H. Subak-Sharpe for DNA samples.

ABBREVIATIONS AND CONVENTIONS

The abbreviations and symbols recommended by Biochem.J.(1970) 116, 1 are used. However, in indicating the sequences of oligodeoxyribonucleotides, the prefix "d" (for deoxy-) is omitted, for brevity. Occasionally the symbol $\overset{*}{p}$ is used to indicate $[^{32}P]$-phosphate in a nucleotide.

Additional abbreviations used are:-

Ad	Human adenovirus
BSA	Bovine serum albumin
EMC	Encephalomyocarditis virus
MeC	5-methylcytidine
MVM	Minute virus of mice
PPO	2,5-diphenyloxazole
RV	Kilham rat virus
SSC	Standard saline citrate
SV40	Simian virus 40
TEAC	Triethylammonium carbonate
TCA	Trichloroacetic acid

PART 1 : INTRODUCTION

Page

1.1 GENOME STRUCTURE AND FUNCTION

1.1.1 Introduction

This thesis is concerned with some aspects of the structure of genomes, and possible interpretations in terms of function. In this section some general features of genome structure, organisation and function are discussed; only static aspects are dealt with, dynamic and temporal considerations being excluded. The term "genome" is generally used to describe the genetic complement of an organism. Here the term is taken to indicate the physical structure of the complement of genes and associated entities, such as control sites, and also portions of genetic material without any function defined at present.

Genome function in a unicellular organism has two aspects. First, the metabolic activity of the cell must be maintained and controlled. This involves transcription of RNA from the genome to produce specific proteins, and also production of other RNA species - ribosomal RNAs, including 5S RNA, and transfer RNAs. Suitable control systems must exist to govern these activities. Second, the genome must contain elements responsible for its own replication and for cell division. These activities must also be mediated through specific protein production. In the simplest genomes, those of viruses, the first of these functional classes is neglected, and essentially all genome activity is concerned with replication. In complex, multicellular organisms, both aspects are much elaborated: different classes of differentiated cells have distinct metabolic activities and growth characteristics.

Prominent aspects of genome structure and function in bacteria, viruses

and eucaryotes are now discussed in turn. The field covered is so wide that it is difficult to give adequate acknowledgement of sources: references in this section are therefore intended to be illustrative rather than exhaustive.

1.1.2 Bacteria

This topic is discussed at length by Hayes (1968).

Most of the material for bacteria is based on data for Escherichia coli. Although fast-growing cells of this bacterium may contain several chromosomes, the resting "haploid" condition comprises a single, cyclic chromosome of double-stranded DNA with a molecular weight of 2.8×10^9 (Cairns, 1963). Unintegrated episomes are also found. Circumstantial evidence indicates that the chromosomal DNA consists of one continuous double strand; however, it is still possible that "links" are present, e.g. of protein. Unlike eucaryotic DNA, this chromosome probably contains no large repeated sections, except for ribosomal genes (Britten & Kohne, 1968).

An E. coli chromosome contains about 4.5×10^6 base pairs (Cairns, 1963), enough to code for several thousand proteins. Only a fraction of the possible proteins have so far been characterised, but there is no good reason for supposing that most of E.coli DNA is not potentially functional in coding for protein, and for ribosomal and transfer RNAs.

In bacteria, many instances are known where the enzymes of a metabolic pathway are specified by a cluster of genes subject to co-ordinate control. Jacob & Monod (1961) called such a system an "operon".

Epstein & Beckwith (1968) have defined an operon as "a group of contiguous structural genes showing co-ordinate expression and their closely associated control sites". Control sites comprise start and stop signals for transcription, start and stop signals for translation, and sites of action of repressors (i.e. "operator" sites). The structural gene for repressor synthesis is not necessarily closely associated with the operon, and is not considered part of the operon. Additional, intraoperon transcription initiation sites can be present in polycistronic operons (Bauerle & Margolin, 1966; Morse & Yanofsky, 1968).

Control of protein synthesis in bacterial operons appears to act at the level of transcription. There is at present no good evidence for the existence of translational-level controls in bacteria (Epstein & Beckwith, 1968). In the best studied systems, where the repressor molecule has been characterised, the control is negative i.e. the system is repressed when the control element is associated with the DNA (Gilbert & Müller-Hill, 1966). However, there is genetic evidence for positive control in several instances (Epstein & Beckwith, 1968). Polycistronic operons may also possess an additional sequence , of as yet undefined use, at intercistronic boundaries (Rechler & Martin, 1970). The orientation of an operon on the DNA chain does not appear to have functional significance (Beckwith, Signer & Epstein, 1966).

Bacterial DNA must contain other control systems e.g. for replication. Present concepts treat operons as independent functional units, responding directly to the presence of effector molecules. It

is not known whether such piecemeal response constitutes the highest control level in bacteria or whether more general, integrative systems operate. Some possibilities for such systems are discussed below.

Maas & Clarke (1964) have described a system of jointly controlled operons, which they term a "regulon". This particular example does not afford any higher function, since the enzymes produced are all concerned with arginine biosynthesis. Other examples of multiple controls are found, for instance with the production of E.coli alkaline phosphatase (Echols, Garen, Garen & Torriani, 1961; Garen & Otsuji, 1964).

The phenomenon of catabolite repression of enzyme synthesis (Magasanik, 1961) may represent a significant general control system. There is evidence that enzyme repression by metabolites or related small molecules is mediated by the effects of these compounds on the concentration of cyclic 3',5'-AMP in the cell (de Crombrugghe, Perlman, Varmus & Pastan, 1969). The cyclic AMP probably acts, via protein factors, on regions of the genome near the promoters of the operons involved (Pastan & Perlman, 1968; Perlman, de Crombrugghe & Pastan, 1969). This system thus has the properties of a general mechanism whereby the physiological state of the cell affects rate of enzyme synthesis, in addition to the individual actions of classical induction/repression systems.

The discovery of sigma factor, a protein which determines the initiation site specificity of E.coli RNA polymerase (Burgess, Travers, Dunn & Bautz, 1969) suggests possible mechanisms of general metabolic control: if several types of sigma factor, with different specificities, existed then various patterns of operon function, superposed on the

direct induction/repression systems, could result. At present, however, there is no evidence for such sigma-mediated control systems and, indeed, they may not be necessary. (However, a new sigma factor does play a role in phage T4 infection (Travers, 1970)).

1.1.3 Viruses

Viruses have been described as containing either DNA or RNA in the mature virion (Lwoff & Tournier, 1966). This is essentially an empirical definition, since there is at present no strong reason for forbidding a mixture, apart from ideas of economy of function and form. Viral genomes have a variety of structural forms. DNA is found double and single stranded, and in acyclic and cyclic forms. Some viral DNA molecules contain single-strand nicks, some can be extracted as a collection of circularly permuted acyclic species, and some have terminally redundant sequences. The DNA may contain unusual bases (e.g. uracil) and may be modified by methylation or glucosylation. These variations are reviewed by Thomas & MacHattie (1967). RNA viruses may contain single or double-stranded RNA; some animal viruses contain several fragments of RNA (e.g. influenza virus (Duesberg & Robinson, 1967)). Viral genomes vary in size from those of the small RNA phages, containing about 3300 bases and coding for, probably, three proteins (Gussin, 1966) to those of the T-even phages, with about 2×10^5 base pairs, and coding for perhaps two hundred proteins.

Knowledge of control of gene expression in animal viruses is still limited but the understanding of several bacteriophage systems is quite

advanced. Here, some outstanding features of two phage systems, lambda and R17, will be mentioned. The following description of the properties of these virus genomes acting in bacterial cells is in essence an extension of the preceding discussion of bacterial genome function.

Lambda is a DNA phage of E.coli; its double-stranded DNA has a molecular weight of about 3×10^7. The programme of transcription during the lytic infection cycle of this phage has been studied extensively. Lambda DNA contains about fifty genes, with extensive clustering of genes for related functions. There appears to be a complex system of interlocking transcriptional controls, both positive and negative. It is especially interesting that many of these controls are not absolute, but alter transcriptional activity by factors of 5 to 50: the consecutive action of several partial controls then results in effectively complete control. Several reviews on phage lambda have been published. This summary is based on those by Dove (1968) and by Szybalski (1969).

The RNAs of the small phages R17 and Qβ are the only viral nucleic acids in which extensive sequences have been determined (for sequences in Qβ RNA see Billeter, Dahlberg, Goodman, Hindley & Weissman (1969)). The following precis of findings with R17 RNA is based on work by Adams, Jeppesen, Sanger & Barrell (1969), Steitz (1969), Nichols (1970) and Jeppesen, Nichols, Sanger & Barrell (1970).

Known sequences in R17 RNA now total about 500 nucleotides. The following salient features have been found. A sequence at the 5'-terminus of the RNA at least 90 nucleotides long does not specify protein. Its function is unknown - possibly it might serve to increase nuclease

resistance. The sequences of the three ribosome binding sites are known: these are non-identical, each about 35 nucleotides long. One intercistronic region has been sequenced. This is 21 nucleotides long and could code for a hexapeptide. The function of this sequence is also unknown. Sequences corresponding to sections of the coat protein gene are consistent with the known amino acid sequence of the coat protein. The sequence adjacent to the 3'-end of the RNA is known; this presumably contains the replicase recognition site. At this end also a considerable length of chain may not be translated. It is estimated about 70% of the RNA may occur in hairpin loops. These might be involved in increasing the nuclease resistance of the RNA, in packing of the RNA in the complete virus and, possibly, in control of translation.

This small genome is thus seen to possess several sequences not obviously relevant to protein specification. Some of these regions may have functions peculiar to a viral RNA e.g. for nuclease resistance and packing into a virion, which have no direct analogues in larger, DNA genomes. Also, since control of protein synthesis in an RNA phage must perforce be at the level of translation, any translational control found in such a phage may not be used in other systems. Nevertheless, these studies indicate the kind and extent of sequences, other than those specifying protein, which occur in the simplest genomes.

1.1.4 Eucaryotes

Eucaryotic cells are defined as possessing a discrete nucleus enclosed by a membrane (Dougherty, 1957). Organisms composed of such

cells vary greatly in complexity. The emphasis of the following discussion is on the genomes of vertebrates; much may not be applicable to, say, yeast.

Eucaryote genomes are very large compared with those of procaryotes: for instance, the bovine genome contains about 3×10^9 base pairs i.e. it is 10^3 times larger than that of E.coli (McCarthy, 1965). The genome size also varies greatly within the class of eucaryotes. There can be detected a rough correlation of genome sizes with qualitative ideas of the "complexity" of the organisms (Britten & Kohne, 1968). However, within this trend, genome size also varies considerably even between similar species (King & Jukes, 1969).

The nuclear DNA of eucaryotes is organised into a number of chromosomes, where it is associated with large amounts of other macromolecules - histones and other basic polypeptides, "acidic" proteins and RNA. These associations have a formidable structural complexity, most impressively in the condensed metaphase state; this topic is discussed at length by Du Praw (1968). The size of the DNA "molecules" within chromosomes is unknown. Each chromosome apparently contains many replication units (Huberman & Riggs, 1968). Mitochondria and chloroplasts also contain DNA, which is cyclic with molecular weight about 10^7 (Kroon, 1969).

Kinetic studies on reassociation of fragmented, denatured DNA indicate that most eucaryote DNAs contain large amounts of repeated sequences (Britten & Kohne, 1968). DNA sequences of many degrees of repetition occur, from "single-copy" DNA to sequences, such as those of

mouse satellite DNA, repeated at least 10^6 times. The "families" of DNA defined in this way consist of similar rather than identical sequences. Surprisingly large differences in such families are found even between closely related animal species (Hennig & Walker, 1970). Fractions of widely varying (G + C) content can be isolated from vertebrate DNAs (Walker & McLaren, 1966; Martin & Hoyer, 1967) : this is another example of the complex structure of these DNAs.

King & Jukes (1969) estimate that if the mammalian genome contains 4×10^4 genes each 10^3 base pairs long, then this accounts for only 1% of the genome. It is difficult to envisage any much larger number of structural genes. Also, the genetic assignment of a unique locus to each gene appears to imply that each gene is not present in many copies. This discrepancy could be taken as meaning that most eucaryote DNA is superfluous, an interpretation that is supported by the finding of a wide range of cellular DNA content among vertebrates, which might be thought to require comparable genetic complements.

The occurrence of repetitive DNA may be relevant to this problem; there are several possibilities. Britten & Davidson (1969) have postulated that repeated sequences are involved in extensive control and integration systems. They argue that the great phenotypic complexity of higher eucaryotes demands very widespread and subtle responses in different cellular states during differentiation and development. Several models can then be constructed implicating families of DNA sequences in regulatory roles. Next, the master-slave hypothesis of Callan (1967) proposes that many non-hereditable copies of each gene are

created at each generation. Some theories of the immune response demand the existence of large numbers of related genes (e.g. Edelman & Gall, 1969). Repeated sequences might play a role in the structural organisation of chromosomes. Finally, much DNA might indeed be devoid of direct function: it might then be useless or might act as a "reservoir" for evolutionary change.

Little is known of the control of eucaryote gene expression in the detail with which some procaryote systems are being examined. The operon concept, with extensive clustering of functionally related genes, seems not to apply. The lack of clustering of related genes might allow a more flexible response from the cell but would require much greater elaboration of elements analogous to promoter and operator regions. The stable patterns of enzyme activities in differentiated cells and the profound effects of hormones on target cells imply the existence of control systems which have stable states yet are capable of swift and extensive change. Cyclic AMP acts as an integrating intracellular "second messenger" in many such systems (Robison, Butcher & Sutherland, 1967). The roles of different DNA-associated molecules in such control are not known in any detail; it is suggested that histones provide a gross "masking" effect while the detailed patterns of gene action are determined by other proteins or by RNA (see Britten & Davidson, 1969).

1.1.5 The Genetic Code.

This introduction has not dealt with the mechanisms of protein synthesis, so that a description here of the nature of the mRNA triplet

TABLE 1. THE GENETIC CODE

1st Position	2nd Position U	2nd Position C	2nd Position A	2nd Position G	3rd Position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	CT-1	Trp	A
	Leu	Ser	CT-2	CT-3?	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The table shows the complete set of 64 triplets, and their corresponding amino acids. Three triplets do not code for amino acids; two of these triplets (CT-1, CT-2) terminate the polypeptide chain and the third (CT-3?) probably does so.

From Woese (1970).

code must be rather superficial. The following brief outline is included because later parts of this thesis are concerned with some sequence characteristics of DNAs, and their interpretation in terms of codon usage.

The mRNA triplet code, from Woese (1970), is shown in Table 1. All amino acids, except methionine and tryptophan, have more than one codon. On present evidence, the set of codons' assignments appears to be universal (Marshall, Caskey & Nirenberg, 1967) i.e. in any system studied any given codon specifies the same amino acid. Several cases are known in which the synonymous codons for an amino acid are not used with equal frequency. For instance, in E. coli the codons AAA and GAA are used in preference to their synonyms AAG and GAG (Weigert, Galluci, Lanka & Garen, 1966).

1.2 NEAREST-NEIGHBOUR ANALYSIS

1.2.1 Introduction

The preceding section describes some aspects of the structure of genomes. The study of genome structure in terms of base sequences is at this time not technically possible, except for the very small bacteriophage RNAs, as described in Section 1.1.3. Therefore, in a general study of sequence characteristics of different genome types, the technique of nearest-neighbour analysis was used. This method measures the frequency of different dinucleotides in the nucleic

acid but does not give data on longer sequences. However, even this limited information can give some insight into the sequence characteristics of the nucleic acid.

In this section the principles of nearest-neighbour analysis are outlined, the kind of results obtained are described, and the information given by the results and the scope of the technique are discussed.

1.2.2 Principles of the Method

The technique of nearest-neighbour analysis was introduced by Josse, Kaiser & Kornberg (1961); it is a method of measuring the frequencies of occurrence in a given DNA of the sixteen possible dinucleotide sequences.

The strategy of the method is as follows. The DNA under study is used as a template for DNA polymerase. Four samples of DNA are made *in vitro*, each labelled with, in turn, one [α ^{32}P]- deoxynucleoside 5'-triphosphate. The unincorporated nucleotides are removed and the DNA is degraded to deoxynucleoside 3'-monophosphates, by digestion first with micrococcal nuclease and then with spleen phosphodiesterase. The monophosphates are separated by electrophoresis and the radioactivity in each is measured. Thus, the DNA is synthesised from a precursor which has ^{32}P in the 5'-position; the DNA is then degraded to 3'-monophosphates, and so the ^{32}P is transferred from the 5'-position of the original nucleotide to the 3'-position of the nucleotide adjacent to (on the 5'-side of) the original labelled nucleotide. Generally, therefore, all

four 3'-monophosphates will contain ^{32}P.

The fraction of the total radioactivity which is found in a given deoxynucleoside 3'-monophosphate then indicates the frequency with which this species occurred on the 5'-side of the original labelled nucleotide in the synthesised DNA i.e. it is a measure of the dinucleotide frequency. For example, if $[\alpha\ ^{32}P]$- dATP is used, then the ^{32}P in the four 3'-monophosphates is a measure of the relative frequencies of CpA, ApA, GpA and TpA.

The relative amounts of each of the four bases incorporated into the DNA are dependent on the template DNA used. It is therefore necessary, to obtain the absolute values of dinucleotide frequencies in the synthesised DNA, to multiply the fractional values for each reaction by a factor equal to the frequency of occurrence of the original $[\alpha\ ^{32}P]$ - nucleotide in the labelled DNA. Such incorporation factors could be obtained in three ways: first, from the base composition of the template DNA, if this is known; second, from the total radioactivity incorporated in each reaction (assuming the specific activities are known and the extent of reaction and the recovery are the same in each case); and, third, by calculation from the electrophoresis results. This third method, which is generally used, is described below.

Call the molar proportions of adenine, thymine, guanine and cytosine, in the synthesised DNA, a, t, g and c respectively. Let ApA represent the absolute frequency, in the synthesised DNA, of the dinucleotide ApA, and so on. If the DNA is labelled with $[\alpha\ ^{32}P]$-dNTP, then call the fraction of radioactivity in Ap, $\overline{\text{ApN}}$, and so on.

The proportion of adenine incorporated as a 5'-nucleotide must be equal to the proportion recovered as a 3'-nucleotide.

i.e. $ApA + TpA + GpA + CpA = ApA + ApT + ApG + ApC.$

i.e. $a = (\overline{ApA} \times a) + (\overline{ApT} \times t) + (\overline{ApG} \times g) + (\overline{ApC} \times c)$

Similar equations can be written for g, t and c. A set of three simultaneous equations in four unknowns (and one redundant equation derivable from the other three) is thus obtained. Together with the initial condition, $a + t + g + c = 1$, these can be solved to give the base ratios of the DNA.

Therefore, from the nearest-neighbour analysis of a DNA, sixteen dinucleotide frequencies are obtained and, in addition, the base composition of the synthesised DNA. In their original study, Josse et al. (1961) found that, for a number of double-stranded DNAs used as templates, the base compositions obtained by nearest-neighbour analysis agreed well with those found by chemical analysis of the template DNA. Each DNA analysed gave a characteristic and non-random pattern of dinucleotide frequencies: this pattern was stable under different conditions of synthesis. In each case the pattern obtained implied that the DNA was synthesised with pairing of adenine to thymine, and guanine to cytosine, between the two strands of DNA, and with opposite polarity of the strands, as proposed by Watson & Crick (1953).

Some of the data of Josse et al. (1961) are shown in Table 2 to illustrate their findings. Each analysis gives six pairs of dinucleotides which are complementary in a double-strand opposite polarity model of DNA:

the frequencies of such pairs are in each case very similar. There are, in addition, four dinucleotides which with this scheme are self-complementary (CpG, GpC, ApT and TpA). On comparing results for different DNAs, it is evident that even where the (G + C) contents of two DNAs are identical, as with calf thymus and Bacillus subtilis in Table 2, the nearest-neighbour patterns can still be quite distinct.

Swartz, Trautner & Kornberg (1962) extended this system to include single-stranded DNA as template. By restricting the amount of synthesis they were able to obtain frequencies corresponding only to the complementary strand. In this case complementary dinucleotides do not necessarily occur with equal frequency.

The technique of nearest-neighbour analysis was developed using DNA and DNA polymerase. However, the idea of examining a nucleic acid in this way can be generalised, depending on the versatility of polymerising enzymes available. Template DNA can be transcribed using an RNA polymerase: the synthesised RNA is then hydrolysed to 2', 3'-monophosphates with alkali. This approach has been used by Weiss & Nakamoto (1961) with RNA polymerase from Micrococcus lysodeikticus. They found that the nearest-neighbour patterns obtained in this way for several DNAs were identical to those obtained by the DNA polymerase method.

In a further extension of the method, RNA is used as the template. This method has been used with several RNA viruses. There are two kinds of RNA-RNA polymerase. First, such enzymes are formed in cells after infection with RNA viruses, and have been purified in some cases (e.g. August, Shapiro & Eoyang, 1965; Shapiro & August, 1965). Second,

TABLE 2. EXAMPLES OF NEAREST-NEIGHBOUR FREQUENCIES

	Calf Thymus	Bacillus subtilis	Escherichia coli	Micrococcus lysodeikticus
ApA TpT	89 87	92 95	71 76	19 17
CpA TpG	80 76	67 68	71 71	52 54
GpA TpC	64 67	67 65	55 56	65 63
CpT ApG	67 72	57 58	55 55	50 49
GpT ApC	56 52	48 48	55 54	56 57
GpG CpC	50 54	46 46	56 56	112 113
TpA	53	52	51	11
ApT	73	80	68	22
CpG	16	50	67	139
GpC	44	61	83	121
(G+C) %	42	44	50	71

Frequencies of dinucleotides are in parts per 10^3.
These data are from Josse et al. (1961).

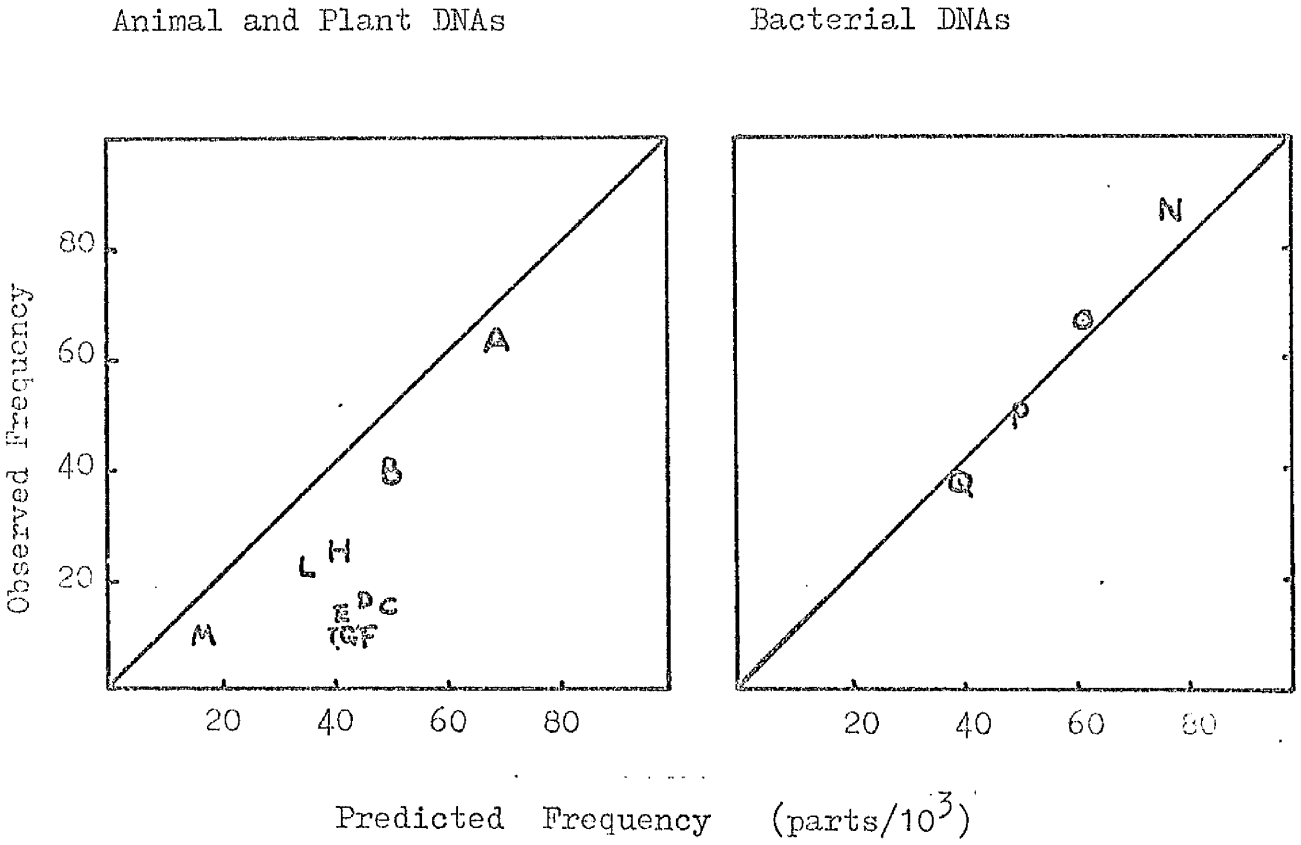
M. lysodeikticus DNA-RNA polymerase can, in the absence of DNA, utilise RNA as a template. (Fox, Robinson, Haselkorn & Weiss, 1964; Hay & Subak-Sharpe, 1968).

1.2.3 Presentation of Results

Several schemes have been devised for visual presentation of nearest-neighbour data, which are difficult to compare in the tabular form of Table 2. Kaiser & Baldwin (1962) introduced the idea of comparing graphically the measured frequency of a given dinucleotide with the frequency expected in a random sequence chain having the base composition of the DNA under study (i.e. the random frequency equals the product of the frequencies of the constituent nucleotides of the dinucleotide). This is illustrated in Fig.1. This method of presentation shows quite well the variations from random frequency of various dinucleotides.

Two presentations of value were introduced by Subak-Sharpe et al. (1966). In the first, the frequencies are presented as a series of histograms. Superposition of the data for two DNAs then allows a direct visual comparison (Fig.2). For the second method, the idea of deviation from randomness is again used. The dinucleotide frequencies are "normalised" to the value they would have in a DNA of 50% (G + C) content. Thus for a dinucleotide XpY of frequency z, the normalised frequency is $\frac{z \times 0.0625}{x \times y}$, where x and y are the respective frequencies of occurrence of X and Y. The factor of 0.0625 allows for the fact that in a 50% (G + C) DNA all dinucleotides have a random expectation of occurrence of 0.0625. A histogram is then drawn, with a base-line of

FIGURE 1. FREQUENCIES OF CpG IN EUCARYOTE AND PROCARYOTE DNAs

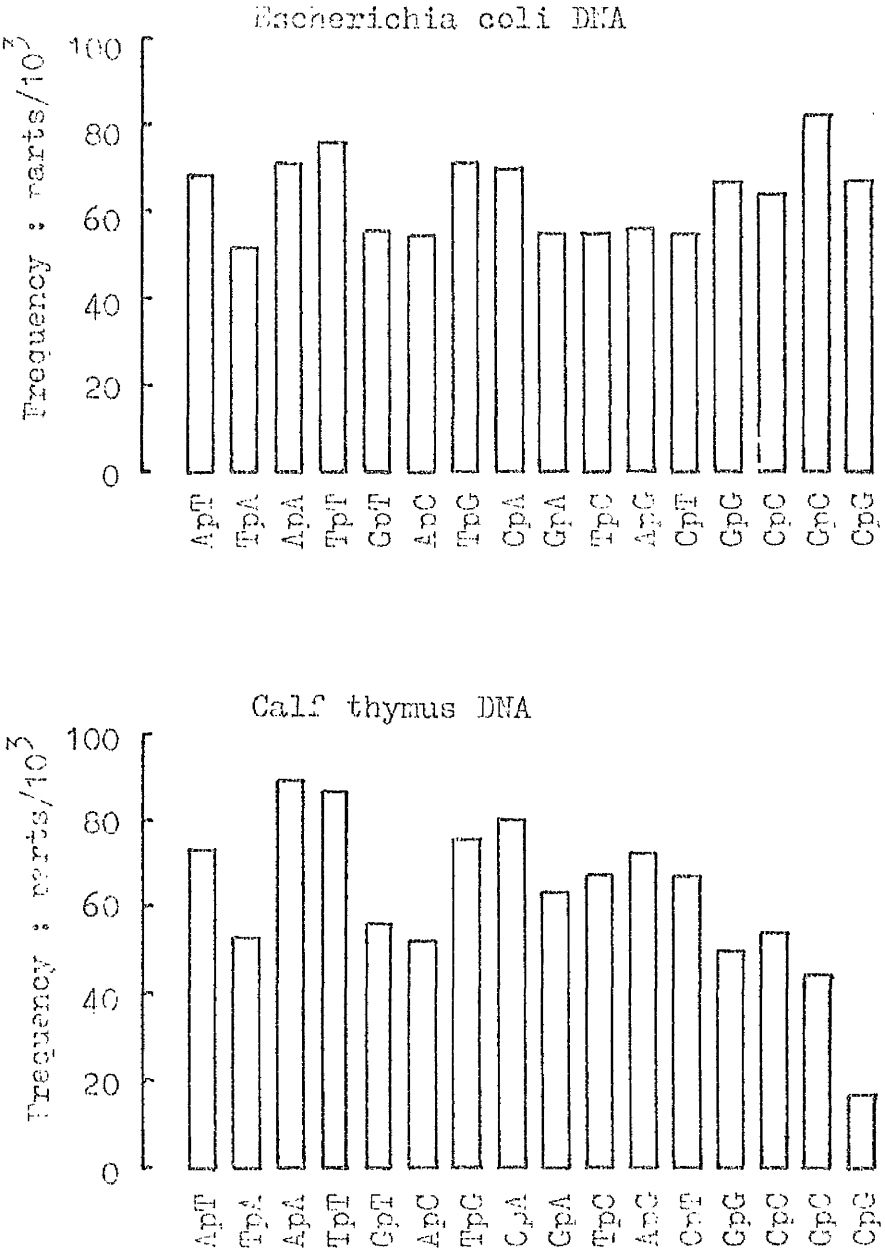


The frequencies of the dinucleotide CpG in animal, plant and bacterial DNAs are compared with values predicted from random association. The line through each origin represents random expectation, and the upper case letters represent measured frequencies of CpG in the DNAs from the following organisms :-

A Chlamydomonas	F Chicken	M T. pyriformis
B Wheat	G Mouse	N A. aerogenes
C Calf	H Starfish	O E. coli
D Salmon	I Human	P B. subtilis
E Rabbit	L Sea urchin	Q H. influenzae

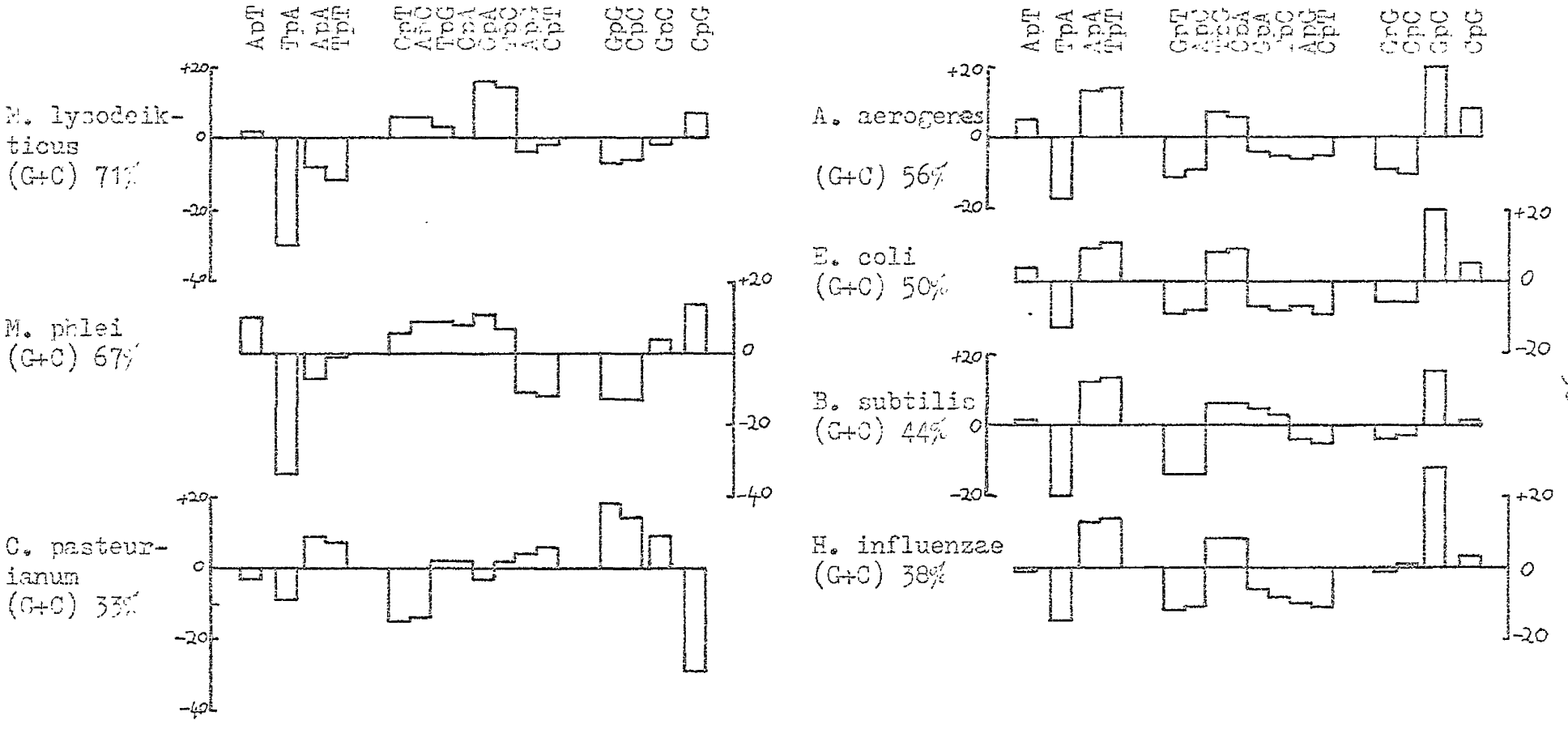
Based on similar figures in Swartz et al. (1962).

FIGURE 2. PLANAR-2 DINUCLEOTIDE FREQUENCY HISTOGRAMS



Based on data from Josse et al. (1961).

FIGURE 3. NEAREST-NEIGHBOUR PATTERNS OF BACTERIAL DNAs



Frequencies are as parts per 10^3 deviation from random expectation. All these diagrams are based on data from Josse et al.(1961), except that for C. pasteurianum, which is based on data from Skalka et al. (1966). This last analysis was performed with M. lysodeikticus RNA polymerase.

62.5, showing for each dinucleotide the deviation (+ or -) in parts/ 10^3 from random, as shown in Fig. 3. In effect, this procedure calculates the specific frequency of each dinucleotide i.e. the frequency per unit of each constituent nucleotide. The deviation histogram then represents the difference of each specific frequency from unity. The introduction of the concepts of "normalisation to 50% (G + C)" and "deviation from $62.5/10^3$" merely multiplies throughout by a factor of 62.5. This presentation is therefore derived conceptually from that of Kaiser & Baldwin (1962). This is a useful method of comparing different DNAs: its limitations are discussed later.

1.2.4 Description of Results

Josse et al. (1961) determined the nearest-neighbour frequencies of a number of bacterial DNAs. When these frequencies are plotted as deviation histograms, as described above, they fall into several classes (Fig.3). First, the patterns for Escherichia coli, Aerobacter aerogenes, Bacillus subtilis and Haemophilus influenzae are all similar, while Micrococcus lysodeikticus and Mycobacterium phlei give another pattern. The frequencies for Clostridium pasteurianum DNA were determined by Skalka, Fowler & Hurwitz (1966) and give another distinct pattern. Note that the DNAs in the E.coli group give similar patterns of specific frequencies in spite of large differences in (G + C) content. Bacteriophages of E.coli in general give patterns very similar to their host. (Josse et al., 1961; Swartz et al., 1962).

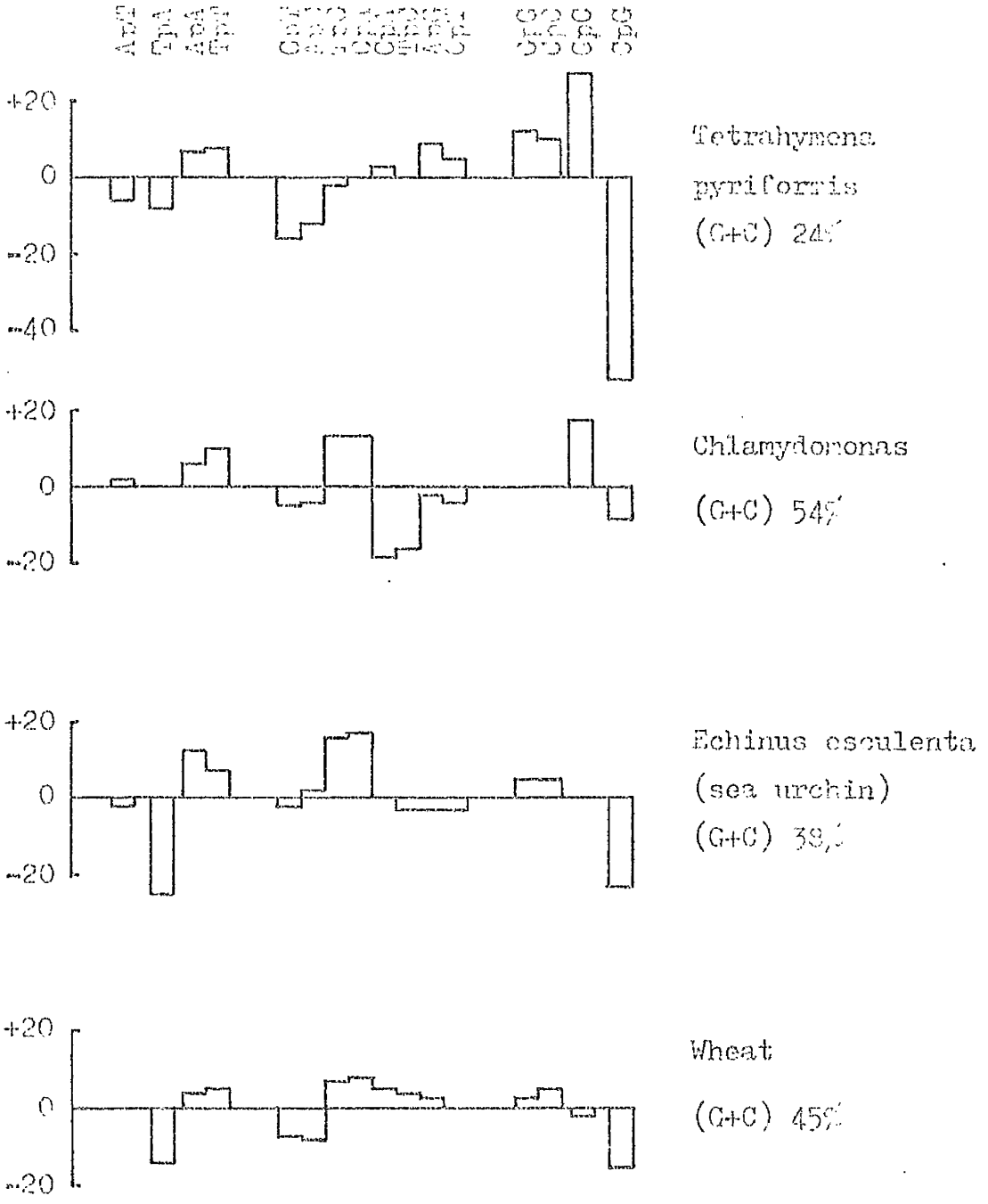
Patterns for DNAs from eucaryotic organisms other than vertebrates

are shown in Fig. 4. These give highly variable patterns, some with resemblances to the vertebrate patterns discussed below.

Frequency patterns for several vertebrate DNAs are shown in Fig. 5; it can be seen that these patterns are all very similar. The patterns for DNAs from different tissues of the same species, and from related tumours, are identical (Swartz et al., 1962). As with the bacterial DNAs, a low relative frequency of TpA is found. The most outstanding feature of this vertebrate pattern is the very low frequency of occurrence of CpG (Figs 2 and 5).

The patterns for several DNA animal viruses and for one RNA virus have been determined (Subak-Sharpe et al., 1966; Morrison, Keir, Subak-Sharpe & Crawford, 1967; Hay & Subak-Sharpe, 1968). As shown in Fig. 6 these fall into two distinct groups. First, large viruses, with DNAs of widely varying (G+C) content, all give patterns quite close to random, but without any particular features common to all. This class comprises pseudorabies, vaccinia, herpes simplex, equine rhinopneumonitis and, possibly, adenovirus 2. The second class, of the small viruses polyoma, SV 40, Shope papilloma and human papilloma, gives highly non-random patterns similar to those found with vertebrate DNAs. The RNA virus, EMC, gives a similar pattern. In this case the pattern represents the single, complementary strand of RNA. The nucleic acids of these viruses have (G+C) contents in the range 39-48% i.e. close to the values found for vertebrate DNAs.

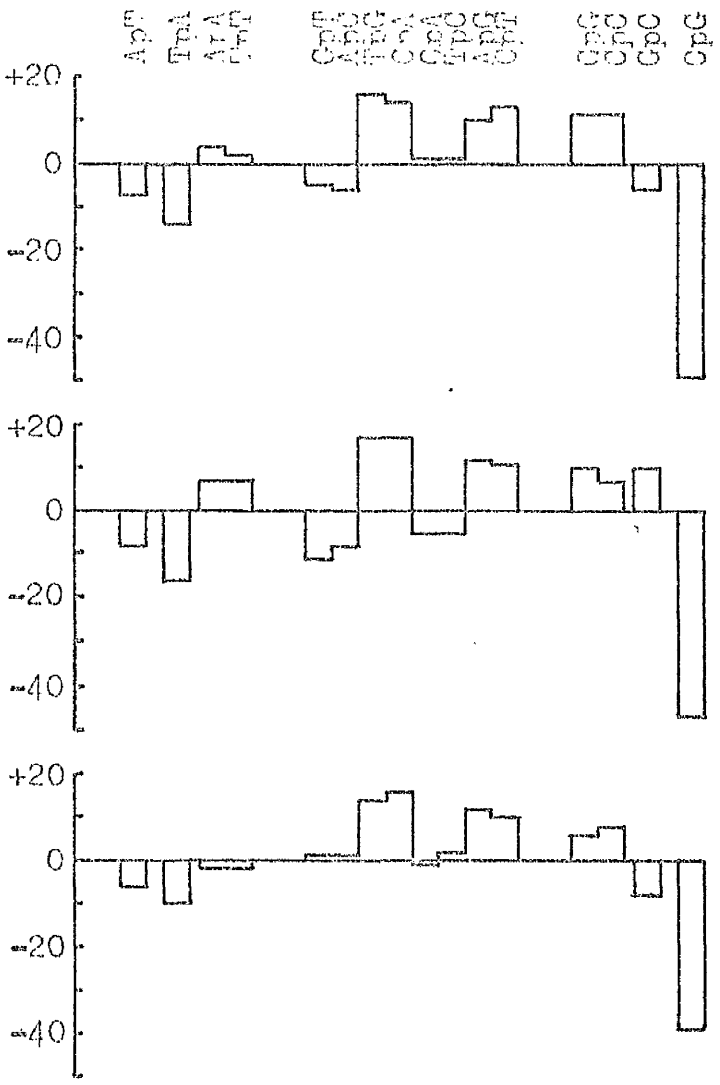
FIGURE 4. [illegible]



Frequencies are as parts per 10^3 deviation from random.

Based on data from Swartz et al.(1962).

FIGURE 5. NEAREST-NEIGHBOUR PATTERNS OF VERTEBRATE DNAs



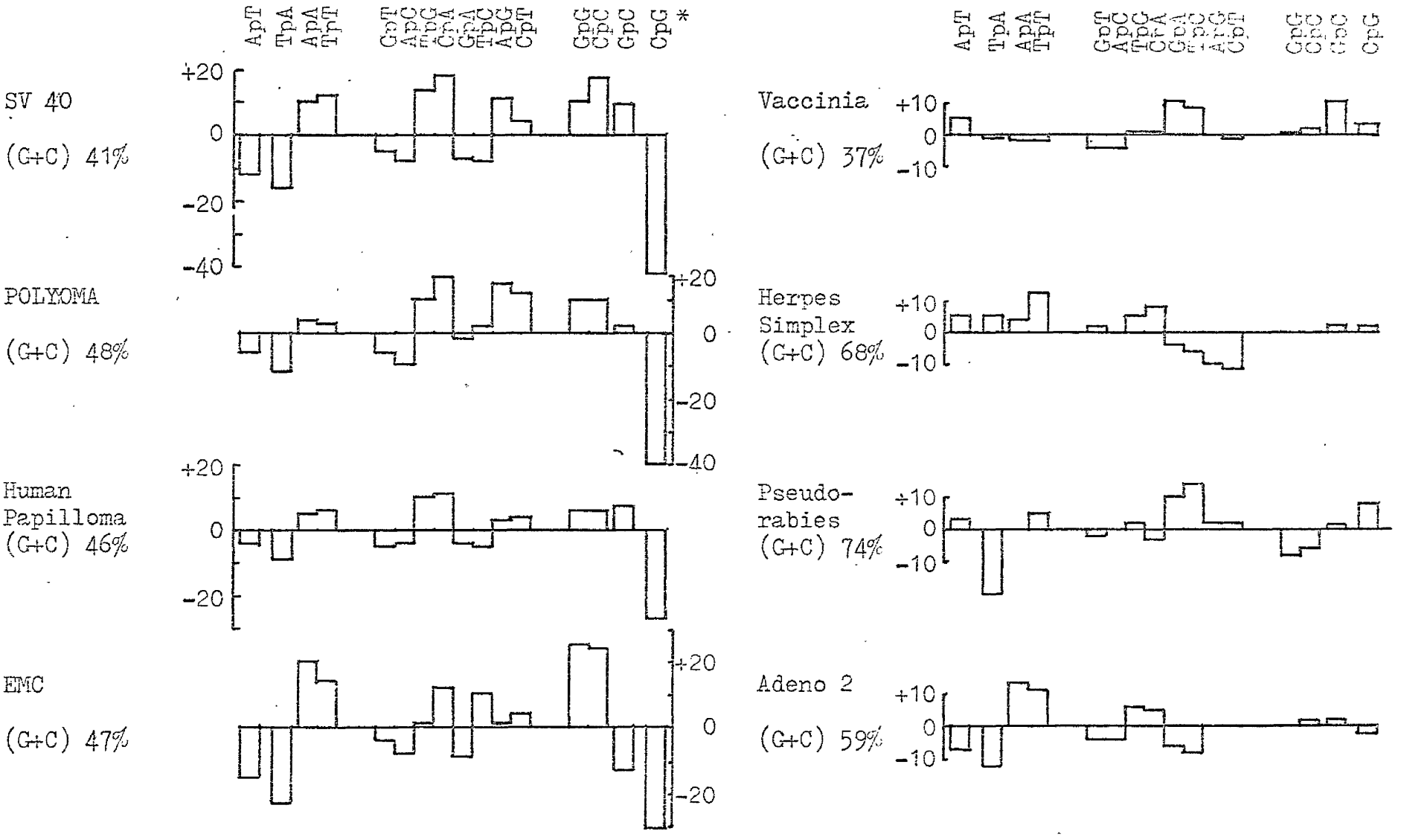
Mouse DNA
(G+C) 42%

Chicken DNA
(G+C) 42%

Salmon DNA
(G+C) 43%

Frequencies are as parts per 10^3 deviation from random.
Based on data from Swartz et al.(1962).

FIGURE 6. NEAREST-NEIGHBOUR PATTERNS OF VIRUS DNAs



Frequencies are as parts per 10^3 deviation from random expectation.

* For EMC, U replaces T.

From Subak-Sharpe et al. (1966); Morrison et al. (1967); Hay & Subak-Sharpe (1968).

1.2.5 Implications and Comments

Much material relevant here will be dealt with more fully in later sections.

Bacteria are seen to give distinct patterns, non-random but with few large deviations from random. By the criteria of the deviation histograms they appear to fall into perhaps three classes. Most bacterial DNA must specify proteins, and the nearest-neighbour pattern must therefore depend largely on the frequencies with which different codons are used. In all of the bacterial DNAs the frequency of TpA is unusually low: this can be rationalised as resulting from the low content of codons UAA and UAG, the stop-codons. (In addition, UAU and UAC code for tyrosine, which occurs with low frequency in proteins). The low frequency of TpA is visible because TpA is self-complementary. Many other large variations in frequency of doublets in one DNA strand could be disguised because of averaging between complementary strands. This restriction does not apply to the "limited replication" results for the complementary strand of phage ϕX 174. In this case the frequencies of complementary dinucleotides are not generally equal; however, there are still no very large deviations from random. Averaging of complementary doublets gives a pattern resembling that of the host, E.coli.

CpG is another self-complementary doublet showing, in vertebrate DNAs, large deviations from random: this shortage of CpG is the outstanding feature of these DNAs' patterns. However, again the reservation must be made that large deviations in other dinucleotides in

a single strand could be concealed by averaging over both strands. The CpG shortage was originally found in vertebrate DNA, i.e. in genomes which are unlike those of bacteria in many respects, as discussed in Section 1.1. The most relevant conclusion from discussion of the complexity of vertebrate DNA is that much of vertebrate DNA may not specify protein. Therefore, it is not at once necessary or obvious that the large systematic deviations found in vertebrate DNA (in particular the CpG shortage) are related to, or interpretable in terms of, specification of polypeptides. They could represent other DNA functions: for instance, highly non-random DNA sequences could be involved in the structural organisation of the chromosomes, or in a complex functional and control organisation. They could be present in "inert" DNA, e.g. as heterochromatin.

However, the finding of the same nearest-neighbour pattern in the DNAs of small viruses quite alters the nature of the problem. For example, polyoma DNA is about 5000 base pairs long (Crawford, 1964). This is enough to code for up to ten proteins: it therefore seems unlikely that there can be much DNA devoid of a protein specifying function. If the virus proteins were specified by sequences containing near-random CpG levels, then only about one third of the DNA would be available for protein specification, if the rest were completely lacking in CpG sequences. Thus, it seems that in this case the non-random nearest-neighbour pattern must be relevant to protein specification.

If CpG is found as an intracodon dinucleotide, the amino acids involved are serine, proline, threonine and alanine, each of which has one codon of

the form NCG, and arginine, which has the four codons CGN. All of these amino acids have other codons. The level of CpG in vertebrate-pattern DNA is consistent with, first, the doublet being completely excluded from intracodon occurrence and, second, being lower than expected even in the intercodon position. Subak-Sharpe et al. (1966) have calculated that a DNA of random sequence, except that the stop-codons TAA and TAG, and the codons NCG and CGN are banned, would have a nearest-neighbour pattern similar to that of vertebrate DNA. All this, therefore, suggests that CpG-containing codons are little used in vertebrate systems.

The situation is quite different with large animal viruses, which contain "random expectation" amounts of CpG. However, these viruses multiply in cells whose DNA does have a low CpG level. Subak-Sharpe et al. (1966) proposed that such vertebrate cells would possess a population of tRNAs optimally adapted to the translation needs of the cell: this was interpreted as meaning that these cells would possess low levels of tRNAs for CpG-containing codons. Then, when the CpG-rich mRNA of the virus came to be translated, a "bottleneck" might result from lack of necessary tRNAs. There is some evidence that herpes simplex virus may code for new arginyl tRNAs, which are produced on infection (Subak-Sharpe & Hay, 1965; Subak-Sharpe, Shepherd & Hay, 1966).

The concept of an "optimal population" of tRNAs containing low amounts of tRNAs for CpG-containing codons presents some difficulties. There is evidence, from mutation studies, that mammalian DNA does use arginine CGN codons (Perutz & Lehmann, 1968; King & Jukes, 1969); Therefore, there must be some arginyl tRNA for CGN. One could argue that an

"optimal population" of tRNAs should be one which allows protein synthesis to proceed at an even rate, with the same step-time per codon i.e. the concentrations of different species of activated tRNAs should be comparable. Each tRNA species occurs in four forms: activated, bound to ribosome, free, and bound to activating enzyme. A tRNA for a much used codon will presumably have a larger total amount ribosome-bound than a little used tRNA. However, the amounts of free and enzyme-bound species necessary to maintain the level of activated tRNA will depend on the concentration of activating enzyme as well as that of tRNA. Thus, it does not appear at all necessary that a tRNA which is little used should be present in the cell in much lower amount than other tRNAs.

The methylation pattern of mammalian DNA may be relevant to an understanding of the CpG levels. 5-methylcytosine is found in small amounts (about 1% of total bases) in mammalian DNA, where it is the only methylated base (Wyatt, 1951). Most or all of the methylcytosine is found in the sequence MeCpG, and it is estimated that all or most of the cytosine in CpG becomes methylated (Doskocil & Sorm, 1962). This is quite unlike polyoma virus DNA, which has a similar nearest-neighbour pattern but is not methylated (Kaye & Winocour, 1967). Any possible function of the low CpG phenomenon is discussed later.

Subak-Sharpe et al. (1966) used these data to argue that the nucleic acids of small animal viruses must have arisen from the DNA of a host-type cell at some stage, since the number of directed base changes required to alter significantly the nearest-neighbour pattern of a virus

after its nucleic acid became functional is very large. This argument could also be applied to the small phages of E.coli. On the other hand, it appeared that the nucleic acids of the large animal viruses must have foreign origins.

These ideas can be extended to use nearest-neighbour analysis as a method of classification and an indicator of possible (evolutionary) relations between organisms. This should be a more valid approach than attempts to use base ratios for this purpose. Classification can be performed in two ways. First, qualitatively, the patterns given by different DNAs can be compared visually. Next, a quantitative approach has been used by Bellett (1967), who did not use absolute frequencies but instead their deviations from random: this gave ten parameters for each DNA (complementary doublets were averaged). By computer methods, the "distance", squared, in a 10-dimensional space for each DNA from every other DNA was calculated, and results classified by a flexible sorting method. DNAs were then grouped by a principle-axis technique. Results agreed generally with those from visual interpretation. Interestingly, the papilloma viruses were placed close to invertebrate DNAs, and adenovirus 2 was a borderline case, not easily classified.

The detail of this method seems to have several defects. The use of deviations from random instead of total frequencies appears unjustified, since thereby the (G + C) content is ignored. Bellett (1967) states that inclusion of the (G + C) content as another parameter made little difference; this is inevitable since, presented in this way, it is one parameter out of eleven. Finally, the use of all ten doublet and

doublet-pair frequencies can be criticised since only seven of these parameters specifies the set (Kaiser & Baldwin, 1962).

The nearest-neighbour pattern appears to be a very stable characteristic of a DNA e.g. mammalian DNAs give near-identical patterns. These data therefore seem useful for comparison of remotely related organisms, unlike the DNA hybridisation approach (McCarthy, 1965). However, the possibility of spurious relations appearing, through convergence, or the same codon usage, must be remembered.

Some attempts have been made to use nearest-neighbour analysis to study variations in transcription of DNA under different conditions e.g. Skalka et al. (1966) studied the variation of nearest-neighbour pattern with addition of histones.

1.2.6 Evaluation of the Method

To summarise, nearest-neighbour analysis of nucleic acids gives stable patterns characteristic of a given nucleic acid. It can give some limited information about codon usage. The CpG shortage in vertebrate DNA is revealed as a very interesting, large and consistent deviation. Patterns are very insensitive to change, since a large number of directed base changes is required before there is a visible alteration in frequencies; this is limited by the accuracy of estimation. The system is therefore useless for comparison of DNAs from closely related species. The method can be used to obtain base ratios from a small amount of DNA (down to 20 μg). Interpretation of the results assumes that a representative sample of labelled product has been made. This assumption

appears to be well justified with large, double-stranded DNAs, but is less well grounded with viral nucleic acids, especially single-stranded species. Here, results could possibly depend on the amount of product made, on the integrity of the template material, and on the enzyme to template ratio. Care must therefore be taken. The method is quite demanding technically, since it requires quantitative conversion of the product to mononucleotides.

PART 2 : MATERIALS AND METHODS

Page

Page

2.5 RNA SEQUENCE STUDIES

2.1 MATERIALS

2.1.1 Enzymes.

E.coli DNA polymerase (DNA nucleotidyltransferase, E.C. 2.7.7.7), fraction p4, was bought from Worthington. This preparation was further purified by DEAE-cellulose column chromatography according to Richardson, Schildkraut, Aposhian & Kornberg (1964). The polymerase obtained by this method was totally dependent on added DNA for activity.

Micrococcal nuclease (E.C. 3.1.4.7) was bought from Worthington, Schwarz Bioresearch and Mann. The protein was dissolved in water at 15,000 units/ml (enzyme units as defined by Cunningham, Catlin & de Garilhe (1956)). It was then heated at 100°C for 1 min to destroy any contaminating phosphatase (Ohsaka, Mukai & Laskowski, 1964). Spleen phosphodiesterase (E.C. 3.1.4.1) was bought from Worthington, Schwarz and Mann, and was dissolved in water at 20 units/ml (enzyme units as defined by Hilmoe (1960)). Some preparations of this enzyme were contaminated with phosphatase. Pancreatic DNase (E.C. 3.1.4.5) was bought from Worthington. E. coli alkaline phosphatase (E.C. 3.1.3.1) was bought from Nutritional Biochemicals.

E.coli RNA polymerase (RNA nucleotidyltransferase, E.C. 2.7.7.6) was prepared by the method of Burgess (1969), and contained sigma factor. The enzyme was stored at 5mg/ml, -10°C, in a solution containing 0.01 M-tris-Cl, pH 7.9, 0.01 M-$MgCl_2$, 0.1M-KCl, 0.1mM-dithiothreitol, 0.1mM-EDTA and 50% (v /v) glycerol. Preparations used had specific activities of 400-500 units/mg (enzyme units of

Burgess (1969) and were free of nuclease and polynucleotide phosphorylase activities. This enzyme was prepared by D.S. Lochhead, P.J. Roach and D.J. Jolly.

T_1 ribonuclease was bought from Worthington and Sankyo, and was dissolved in water at 1 mg/ml. Pancreatic ribonuclease was bought from Worthington and was dissolved in water at 1 mg/ml. U_2 ribonuclease was a gift from G.G. Brownlee, and was dissolved in 0.05 M-sodium acetate, pH 4.5, containing 0.002 M-EDTA and 0.1 mg/ml crystalline BSA, at a concentration of 10 units/ml (Arima, Uchida & Egami, 1968 a,b).

2.1.2 Nucleotides and Nucleic Acids

$[\alpha^{32}P]$- deoxynucleoside triphosphates were purchased from International Chemical & Nuclear Corp. $[\alpha^{32}P]$- ribonucleoside triphosphates were bought from Sigma. $[^{14}C]$- ribonucleoside triphosphates were bought from New England Nuclear. The purity of samples was checked by chromatographing overnight on Whatman No.1 paper with isobutyric acid-water-14M-ammonia (62.5: 35.3: 2.2, by vol.).

E.coli tRNA was purchased from Calbiochem. High molecular weight yeast RNA was bought from British Drug Houses.

DNAs of the parvoviruses MVM, H1 and RV were prepared by L.V. Crawford. MVM DNA was also prepared, as follows, from virus supplied by L.V. Crawford.

200 µg of virus was suspended in 0.6ml 0.2 M-KOH. After 5 min at 20°C the suspension was layered on to a step gradient of CsCl

consisting of 1.4ml of CsCl (1.40g/ml) in 0.05 M-tris-Cl, pH 8.0, (bottom layer) then 1.4ml of unbuffered CsCl (1.35g/ml) and 1.4ml of unbuffered CsCl (1.30g/ml). The gradient was centrifuged at 48,000 rev./min for 20 h in the SW50 head of a Spinco Model L centrifuge. The supernatant was removed, except for a small volume containing the DNA pellet, which was then suspended in 0.02 M-tris-Cl, pH 8.0, recentrifuged at 48,000 rev./min for 20 h and dissolved in 0.1ml of 0.02 M-tris-Cl, pH 8.0.

DNAs from the following organisms were prepared by H.M. Keir, by the method of Marmur (1961): Bacillus megaterium, Proteus vulgaris, Rhodospirillum rubrum and Serratia marcescens. Aspergillus nidulans DNA was obtained from G. Pontecorvo, and phage α DNA from H. Subak-Sharpe. DNAs of human adenoviruses 4, 7, 11, 12, 18, 21 and 27 were obtained from M. Green. Drosophila melanogaster DNA was obtained from F.M. Ritossa, Rana catesbeiana DNA from A.M. Campbell, and mouse (C3H) satellite and main band DNAs from P.M.B. Walker. DNA from calf thymus was prepared by the method of Kay, Simmons & Dounce (1952). The buoyant densities of these DNAs were measured by equilibrium centrifugation in a Spinco Model E ultracentrifuge (Meselson, Stahl & Vinograd, 1957).

Calf thymus DNA, for use as a carrier in nearest-neighbour analyses, was denatured before use by heating, at 1 mg/ml, for 10 min at 100°C and then cooling on ice.

2.2 MEASUREMENT OF RADIOACTIVITY

Radioactivity (^{32}P and ^{14}C) was measured by several methods:-

1. Routine measurements of incorporation of ^{32}P and ^{14}C into DNA or RNA, and of radioactivity in small aliquots of labelled fractions, were made in a low-background gas-flow counter, or by drying the aliquot on paper and measuring total count rate in toluene, 0.5%(w/v) PPO, in a liquid scintillation counter.

2. ^{32}P effluent fractions from column chromotography were collected directly into scintillation vials. The ^{32}P was then determined by measuring the whole of each sample, without addition of scintillator, for Cerenkov radiation in a liquid scintillation counter (Clausen, 1968). ^{32}P counted at about 20% efficiency in this system, and was not affected by the presence of paper, or by high salt concentrations. ^{14}C did not register at all: 20,000 dpm gave background count rate. By this means the whole of a column effluent could be recovered for further use, and inaccuracy due to the withdrawal of aliquots, for measurement, from fractions of slightly variable volume was avoided.

3. ^{14}C and ^{32}P were determined simultaneously by double label counting in toluene, 0.5% (w/v) PPO, in a liquid scintillation spectrometer, with efficiencies determined by an external standard channels ratio method.

Radioactivity on paper was solubilised by incubation at 60°C for 10 min in the presence of 0.5ml of 1.0 M-hyamine hydroxide. The

presence of paper did not, in practice, affect the external standard.

4. Radioactivity on paper chromatograms was recorded with a Nuclear Chicago Actigraph strip scanner.

2.3 NEAREST-NEIGHBOUR ANALYSIS

The principles of this technique were described in the Introduction. The method used was based on that introduced by Josse *et al.* (1961) and extended and modified by Swartz *et al.* (1962), Josse & Swartz (1963), Subak-Sharpe *et al.* (1966) and Morrison *et al.* (1967).

2.3.1 Activation of DNAs and Measurement of Template Activities

Many DNA preparations have quite low template activity with *E.coli* DNA polymerase. Most DNAs were, therefore, "activated" by very limited digestion with pancreatic DNase, as described by Aposhian & Kornberg (1962). Single-stranded (parvovirus) DNAs gave satisfactory results without DNase treatment.

DNAs were dissolved in 0.05 M-tris-Cl, 0.005 M-EDTA, pH 7.5, or in 1/10 SSC. The activated DNAs were prepared by incubating 50µg DNA for 15 min at 37°C in 1.0ml of solution containing 5×10^{-5}µg of crystalline pancreatic DNase, 0.5mg of crystalline BSA, 5 µmol $MgCl_2$ and 50 µmol tris-Cl, pH 7.5. DNase activity was then destroyed by heating at 77°C for 5 min.

The template activity of each DNA preparation was measured in a trial incubation with polymerase. 5 µg of DNA was incubated at 37°C

for 30 min in 0.3ml of solution containing 5 nmol each dATP, dGTP, dCTP and TTP, with one of these [α^{32}P]- labelled, 20 µmol tris-Cl, pH 7.5, 2 µmol $MgCl_2$, 30 nmol 2-mercaptoethanol, and DNA polymerase. A 0.05ml sample was then withdrawn and pipetted on to a Whatman No.1 paper disc previously treated with 0.05ml BSA, 2mg/ml. This disc was immersed for 10 min in cold 5% (w/v) TCA, containing 50 mM-sodium pyrophosphate, and was then washed twice in cold 5% (w/v) TCA, twice in ethanol, and twice in ether, and counted in a gas-flow counter. This assay was sometimes run at half scale. Sufficient polymerase was used to give about 5% incorporation of radioactivity with activated calf thymus DNA. Usually [α^{32}P]- TTP was used as label. The template activity of activated calf thymus DNA was measured at the same time to allow later comparisons of activity. Activity measurements on DNA polymerase and checks on labelled triphosphates were made in the same way.

2.3.2 Preparation of DNA in vitro.

Each reaction mixture contained, in 0.3ml, 5µg of the DNA under study, 20 µg of E.coli tRNA (to inhibit any contaminating endonuclease) 5 nmol each of dATP, dGTP, dCTP and TTP, 20 µmol tris-Cl, pH 7.5, 2 µmol $MgCl_2$, 30 nmol 2 - mercaptoethanol, and DNA polymerase. Each triphosphate in turn was [α^{32}P]- labelled, with specific activity between 10 and 100 µCi/µmol. Sufficient DNA polymerase was added to each tube to give 20-30% replication of the DNA in 30 min at 37°C. However, rather than make up exact dilutions of polymerase for each DNA, the template activities of the DNAs were approximated into three

or four classes and corresponding amounts of enzyme added. Before use the polymerase was dialysed against 0.05 M-tris-Cl, 0.01 M-2-mercaptoethanol, pH 7.5. All dilutions of DNA polymerase were made with a solution containing 0.05 M-tris-Cl, pH 7.5, 0.1 M-ammonium sulphate, 0.01 M-2-mercaptoethanol, and crystalline BSA, 1mg/ml (Richardson et al, 1964). Round-bottomed glass tubes (10 x 1.5cm) were used for the incubations since these were suitable for the single-electrode pH meter system used later in the procedure. Tubes were incubated at 37°C for 30 min, then cooled on ice. All determinations were carried through in duplicate.

Reaction mixtures were kept in ice during the purification procedure. 0.2ml of denatured calf thymus DNA (1 mg/ml) was added to the incubation mixture and the mixture vortexed. At once, 0.5ml of cold 7% (v/v) perchloric acid was added, and the tube vortexed again. After standing for 5-10 min, 2.5ml of cold water was added, and the precipitate collected by centrifuging at 1000g for 10 min, at 0°C. The supernatant was discarded, and the precipitate dissolved in 0.3ml of cold 0.2 M-NaOH. Two more cycles of acid precipitation were carried out in the same way. Finally, the tube was drained and dried carefully, and the DNA dissolved in 0.1ml of cold 0.05M-NaOH, 0.2M-tris-base. 0.3ml cold water was added and the pH adjusted to 8.6 ($\pm$ 0.2) with 0.1 M-HCl, using a pH meter equipped with a single electrode system.

2.3.3 Digestion of DNA to 3'-mononucleotides.

0.01ml of 0.1 M-$CaCl_2$ and 0.01ml of micrococcal nuclease solution

(15,000 units/ml) were added, and the mixture incubated for 2 h at 37°C. 0.01ml of 0.1 M-sodium arsenate, pH 7.0, was then added (to inhibit any phosphomonoesterase in the spleen phosphodiesterase) and the pH adjusted to 6.8 ± 0.2 with 0.1 M-HCl. 0.01ml of spleen phosphodiesterase solution (20 units/ml) was then added and the mixture incubated at 37°C; further 0.01ml quantities of phosphodiesterase were added each hour for 4 h.

Completeness of digestion to deoxynucleoside 3' - monophosphates can be ascertained in three ways. The original method of Josse et al. (1961) measures the proportion of ^{32}P which is sensitive to alkaline phosphatase i.e. is in a terminal position in a nucleotide chain. The following method is simpler and more reliable and was regularly used. A small aliquot of the digest, usually 0.01ml, was applied to DEAE paper. About 20 µg of dGMP were added as u.v. marker. The paper was eluted with 0.2 M-ammonium acetate containing 7 M-urea. Chromatograms were run ascending for 20cm. This procedure separates oligonucleotides according to chain length (Ohsaka et al., 1964). The mononucleotides run fastest and form a double peak with dCMP and TMP slightly ahead of dAMP and dGMP. Any radioactivity in the oligonucleotide region indicated that further digestion was required. This method was used with about one fifth of the tubes in each batch. In addition, a critical examination of the results of electrophoresis afforded a good indication of the state of digestion. This is discussed below.

2.3.4 Electrophoretic Separation of 3'- mononucleotides.

Each digest was centrifuged at 1000g for 10 min to sediment any protein which had precipitated during the digestion period. The supernatant was then carefully removed. Often, the precipitate was very fragile and some was removed with the supernatant. However this did not seem to have any effect on the electrophoresis. The supernatants were dried with an air stream at room temperature, and then dissolved in 0.1ml water.

The deoxynucleoside 3'-monophosphates were separated by high voltage electrophoresis, on paper, at pH 3.5 (Markham & Smith, 1952). The inclusion of arsenate in the digestion mixture had an adverse effect on the electrophoretic separation, causing bad streaking. This was overcome by spotting only 0.04ml of each digest on to Whatman 3MM paper: however, if counts were low, the total sample was applied as a 3cm streak. The buffer used for electrophoresis was 0.05 M-ammonium formate adjusted to pH 3.5 with 100% formic acid. Samples were electrophoresed for 7 kV-h in a water-cooled flat-plate apparatus.

After the electrophoresis run the dried paper was examined under u.v. light. Nucleotides should show up clearly as discrete spots. The order is: origin, dCMP, dAMP, dGMP, TMP. Often there was some u.v. absorbing material near the origin: since this was non-radioactive, it was probably protein. Any streaking of the spots, inadequate separation, or presence of u.v. absorption, either continuous or as discrete spots outside the main spots, was noted. All spots were marked and the entire paper strip corresponding to each digest was cut up and ^{32}P

measured in a scintillation counter, using toluene - PPO scintillator. (Since the nucleotides and inorganic phosphate are insoluble in toluene scintillator, this can be recovered and re-used). The radioactivity in each of the nucleotide spots was expressed as a fraction of the total nucleotide counts.

Each electrophoresis result was examined by the following criteria. Apart from the irregular non-radioactive absorption sometimes found near the origin there should be no u.v. absorption between the mononucleotides. The nucleotides should appear as well-separated non-streaked spots. Over 98% of the total radioactivity should be in the nucleotide spots: u.v. absorption or more than 2% of the total counts outside the nucleotide areas generally indicated that digestion was incomplete. There should be close agreement (within 1.5% units) between corresponding nucleotides in duplicate runs. Larger differences usually also corresponded to incomplete digestion. (Incomplete digestion gives products, e.g. dinucleotides, which move faster than dCMP or dAMP so that the fraction of radioactivity in the dGMP and TMP areas is elevated). If any of these signs were seen, the remaining portion of the digestion mixture was made up to 0.3ml with water, the pH adjusted to 6.8 and phosphodiesterase treatment continued. The electrophoresis was then repeated.

An important auxiliary function of electrophoresis was to separate inorganic phosphate from the nucleotides and thus allow the detection of any ^{32}P in inorganic phosphate, resulting from phosphomonoesterase action. However, it was found necessary, to achieve good separation, to run TMP almost to the limit of the paper, and since inorganic phosphate

moves faster than TMP, it was therefore run off the paper. Accordingly, for some of the assay tubes in each batch, a portion was electrophoresed for 3kV-h. Marker phosphate was included and detected with an ammonium molybdate spray reagent. Radioactivity in the inorganic phosphate was usually about 1% of the total, and should not exceed 2%.

When the $[\alpha^{32}P]$-triphosphate supply situation allowed, it was useful, for a given DNA, to perform the $[\alpha^{32}P]$-dCTP and $[\alpha^{32}P]$-dGTP analyses at the same time. This is because, for a double-stranded DNA, the fraction of radioactivity in 3'-dCMP, from $[\alpha^{32}P]$-dCTP, should equal that in 3'-dGMP from $[\alpha^{32}P]$-dGTP, without final frequency calculation; and similarly for dAMP and TMP. This is a useful check that the system is functioning properly, but it does not hold for single-stranded DNAs.

The principles of obtaining dinucleotide frequencies from the data have been outlined in the Introduction. In practice an Olivetti Programma desk computer was used to solve the simultaneous equations involved; this machine was also used for other routine calculations.

2.4 PYRIMIDINE SEQUENCE STUDIES.

The design of these experiments is described in Section 4.1.

2.4.1 Early Experiments.

DNA labelled with $[\alpha^{32}P]$-dGTP was synthesised *in vitro* as described for nearest-neighbour analysis. MVM and calf thymus DNAs

were used as templates. Such preparations were digested to pyrimidine runs by the methods of Shapiro & Chargaff (1957) and of Burton & Petersen (1960). Samples of these digests were fractionated by two-dimensional paper chromatography, according to Shapiro & Chargaff (1963). Whatman no.1 paper was used, and was developed, descending, first with isopropanol-water (7:3, v / v) in an atmosphere of concentrated ammonia, for 72h at 30°C. The paper was then dried and developed at right angles to the first dimension with isobutyric acid - 0.3M-NH_4OH (5:3, v/v), for 36h at 20°C. Radioactivity was detected by autoradiography with Kodak Industrex x-ray film. Generally, 10-12 spots were resolved.

A detailed investigation of pyrimidine sequences was made as follows.

2.4.2 Preparation of Pyrimidine Runs from MVM DNA.

DNA labelled with ^{32}P and ^{14}C, replicated *in vitro* from MVM DNA, was prepared as follows. 40 µg MVM DNA was incubated for 150 min at 37°C in a volume of 2.35ml containing 160 µg *E.coli* tRNA, 100 nmol [α ^{32}P]- dGTP (S.A. 400 µCi/µmol), 100 nmol [^{14}C]- dCTP (S.A. 29 µCi/µmol), 150 nmol each of dATP and TTP, 160 µmol tris-Cl, pH 7.5, 16 µmol $MgCl_2$, 240 nmol 2-mercaptoethanol and 11 units of *E.coli* DNA polymerase (enzyme units of Richardson *et al.* (1964)). The progress of the reaction was followed by withdrawing 0.005ml samples at 0, 40, 80, 120 and 150 min. The acid-precipitable radioactivity in these was measured in a gas-flow counter as described for the DNA polymerase assay. After 150 min the reaction mixture was chilled in ice. 4ml of denatured calf thymus DNA (1.8mg/ml) was added as carrier to the cold reaction

mixture and vortexed. At once, 4 ml of 7% (v/v) perchloric acid was added, and the mixture vortexed. After standing for 5 min, 20ml of cold water was added. The precipitate was collected by centrifuging at 1000 g for 15 min. The supernatant was discarded and the precipitate dissolved in 2ml of 0.2M-NaOH. Acid precipitation was repeated twice, and the DNA finally dissolved in 2.3ml of 0.05M-NaOH. The recovery of acid-precipitable radioactivity through the washing procedure was greater than 90%. A small aliquot was withdrawn for nearest-neighbour analysis. This was performed as described before, except that radioactivity was measured by double label counting for ^{32}P and ^{14}C in a scintillation spectrometer.

After addition of 10ml of calf thymus DNA, 1.8mg/ml, the labelled DNA was digested to pyrimidine runs by the method of Burton & Petersen (1960). 24ml of 3% (w/v) diphenylamine in 100% formic acid was added to the DNA solution, and incubated for 18 h at 37°C in the dark. 20ml of water was added and the reagents were then removed by repeated extraction with six volumes of ether. Traces of formic acid were then removed by evaporation at 60-65°C *in vacuo*. Lithium acetate was added to the remaining solution to a final concentration of 0.01M and the pH adjusted to 5.3 with 0.1M-NH_4OH. The recovery of radioactivity over acid digestion and ether extraction was greater than 70%.

2.4.3 Separation of Pyrimidine Runs by Length.

The pyrimidine runs were fractionated by column chromatography on DEAE-cellulose, according to Spencer, Cape, Marks & Mushynski (1969).

Microgranular, preswollen DEAE-cellulose (Whatman DE52) was freed from fines by repeated decantation from 0.5 M-NaCl, and packed into a column, 30 x 1.0cm. The column was washed with 500ml of 2.0 M-NaCl and then with 300ml of 0.01 M-lithium acetate, pH 5.3. All solutions were pumped through the column at 25ml/h. The solution of digested DNA was next loaded on to the column. Fractions were collected every 30 min directly into scintillation counter vials, and the u.v. absorbance of the column effluent was monitored at 254nm with a Uvicord. The column was washed with 300ml of 0.01 M-lithium acetate, pH 5.3, to remove purine bases. Pyrimidine runs were eluted with a linear gradient of LiCl in 0.01 M-lithium acetate, pH 5.3, from 0 to 0.4 M-LiCl. The total volume of the gradient was 1100ml. After starting the gradient, fractions were collected at 20 min intervals until isostich V was eluted, and then at 15 min intervals. After completion of the gradient, the column was washed with 1.0M-LiCl. ^{32}P eluted from the column was detected by measuring the whole of each sample for Cerenkov radiation. Recovery of ^{32}P from the column was greater than 99%.

As described by Spencer & Chargaff (1963b) and by Spencer et al. (1969), the purine bases were washed off the column immediately. However, a large peak of ^{32}P was eluted closely after the purine bases. This was tentatively identified as inorganic phosphate; its position contrasts with the findings of Cerny, Mushynski & Spencer (1968), who detected an inorganic phosphate peak only after the start of the gradient. In the present case the peak described had a long trailing

edge, and a small, residual peak was finally eluted after starting the gradient, in the position described by Cerny et al. (1968) for inorganic phosphate. This elution pattern was presumably due to some small differences in ion concentrations. The appearance of u.v. peaks in the gradient corresponded closely with the results of Spencer et al. (1969). Minor u.v. peaks, not containing ^{32}P, were observed in front of isostichs I and II; possibly these contained some dephosphorylated material. They were not further examined. ^{32}P peaks coincident with the u.v. peaks were found for isostichs I - XIII. ^{32}P was also found in the final peak, designated "IX". The size of the ^{32}P peaks decreased, with increase of isostich length, more rapidly than the u.v. peaks, since the ^{32}P was present only at the 3'-ends of isostichs.

Fractions corresponding to each isostich were pooled and their pH adjusted to 5.0 with 0.1 M-acetic acid. Their volumes were noted and the absorbance at 270 nm measured. (This is the isosbestic wavelength for pyrimidine nucleotides at pH 5.0, determined by Spencer, Cape, Marks & Mushynski (1968)). From these data the distribution of pyrimidine runs in calf thymus DNA was calculated.

2.4.4 Separation of Pyrimidine Runs by Base Composition.

Samples of isostichs I to V were desalted by gel-filtration on Biogel p2 (Uziel, 1967). 10ml samples were loaded on to a column of Biogel p2 (100 x 2.5cm) and washed through with water at a flow rate of 10ml/min. 10ml fractions were collected into scintillation vials.

The peak ^{32}P fractions were pooled and concentrated by rotary evaporation at 40-50°C. Aliquots were applied to 3MM paper and electrophoresed for 4 kV-h in 0.05 M-ammonium formate adjusted to pH 3.0 with 100% formic acid. Marker nucleotides were run in parallel. Although markers were separated cleanly, the isostich material streaked badly, presumably because some salt had not been removed. Later separations were therefore carried out on DEAE-cellulose columns at low pH, by the method of Cerny et al. (1968), as follows.

DEAE-cellulose (Whatman DE52) was freed of fines by decantation, and packed into a column, 30 x 1.0cm, in 0.5 M-formic acid. The column was washed with 100ml of 1.0M-formic acid and then with water till the pH of the effluent was greater than 3.5. All solutions were pumped through the column at 25ml/h. An aliquot of the isostich fraction under study was diluted with two volumes of water and pumped on to the column without any pH adjustment. The u.v. absorbance of the column effluent was monitored at 254 nm with a Uvicord. Fractions were collected at 20 min intervals into scintillation vials. The column was washed with 0.1M-formic acid until pH and absorbance at 254 nm of the effluent were the same as the eluent. The column was then developed with a linear gradient of ammonium formate. The first gradient vessel contained 0.1 M-formic acid (pH 2.7), and the second 0.5 M-ammonium formate adjusted to pH 3.1 with 100% formic acid. The total volume of the gradient was 400-500ml. ^{32}P was detected by Cerenkov counting.

With this technique, pyrimidine runs containing only cytidine are eluted first, and runs with higher proportions of thymidine appear in succession. ^{32}P peaks were coincident with u.v. peaks detected on the Uvicord. Isostichs I to IV were fractionated in this way. In each case, between 4 and 8% of the ^{32}P was found in the 0.1 M- formic acid wash.

Fractions corresponding to each peak were pooled, their pH adjusted to 3.0 with 0.1 M-acetic acid, and their volume measured. Absorbance at 267.5 nm (the isosbestic wavelength at pH 3.0, according to Cerny et al. (1968)) was measured. Aliquots from each peak were measured for ^{32}P by Cerenkov counting. Samples of each peak were concentrated by rotary evaporation in vacuo at 40-50°C. Ammonium formate was then removed by sublimation in vacuo at 60-65°C. Nucleotide material was dissolved in 1.0ml of 4mM-tris-Cl, pH 8.9. Aliquots were measured in a scintillation spectrometer to determine ^{32}P and ^{14}C dpm.

2.4.5 Determination of 3'-end Groups.

0.3ml of each sample, in 5mM-tris-Cl, pH 8.9, was incubated with 0.01ml of 0.1M-$CaCl_2$ and 0.03ml of micrococcal nuclease (15,000 units/ml) for 20h at 37°C. 0.01ml of 0.1M-sodium arsenate, pH 7.0, was then added to each sample and the pH adjusted to 6.8 $\pm$ 0.2 with 0.1M-HCl. 0.01ml of spleen phosphodiesterase (20 units/ml) was added and the sample incubated for 2 h at 37°C. Each sample was centrifuged (1000g, 10min) to remove any protein precipitate; the supernatants were dried with an air stream and redissolved in 0.1ml water.

DEAE-paper was prewashed with 1.0M-formic acid and then with water,

and dried. Digestion mixtures streaked on to the paper on 2cm base lines with marker TMP, dCMP and d(pTp) (20 µg each). The paper was developed for 1h, descending, with 0.05M-formic acid, and was dried at room temperature. The paper was then developed in the same direction with 0.2 M-ammonium formate. The final order of components on the paper was then: origin, undigested material, d(pTp) and d(pCp), TMP, dCMP, and dC. Markers were located under u.v. light, and the whole paper strip for each sample cut into fractions and measured for ^{32}P and ^{14}C by double label scintillation counting.

2.4.6 Pyrimidine Runs from Calf Thymus DNA.

A similar procedure was followed with activated calf thymus DNA as template for DNA polymerase. Only [$\alpha^{32}P$]- dGTP was used as label. After digestion to pyrimidine runs as described above, the nucleotide material was isolated by a modified method (Spencer et al., 1969). The incubation mixture (30ml) was diluted to 250ml with water and cooled on ice. The resulting precipitate of diphenylamine was removed by vacuum filtration through a fine grade sintered glass funnel, which was then washed with 50ml water. The filtrate and the washings were evaporated to about 20ml in a rotary evaporator. 200ml of water was added, and the solution again evaporated. This cycle was repeated until the pH was about 3.5 and the remaining diphenylamine precipitated out. The precipitate was removed by filtration. Lithium acetate was added to the filtrate to a final concentration of 0.01M and the pH adjusted to 5.3 with 0.1M-NH_4OH. Separation by length and by base composition and

determination of 3'-end groups were performed as described above.

2.5 RNA SEQUENCE STUDIES.

2.5 RNA SEQUENCE STUDIES.

2.5.1 Preparation of Labelled RNA.

RNA was synthesised *in vitro* using conditions based on those of Burgess (1969). Typically, MVM, H1 or calf thymus DNA was incubated with *E.coli* RNA polymerase in 0.25ml of a solution containing 10 µmol tris-Cl, pH 7.9, 40 µmol KCl, 2.5 µmol $MgCl_2$, 0.1 µmol K_2HPO_4, 50 nmol 2-mercaptoethanol, 0.1mg crystalline BSA, 50 nmol each of three unlabelled nucleoside triphosphates, and 10-20 nmol of radioactive nucleoside triphosphate. Incubations were at 37°C, generally for 2h (the rate of incorporation of label into RNA was linear at least up to 3h). 5-10 µg of MVM or H1 DNA, or 20-40 µg calf thymus DNA, was used as template. Generally 5-10 µg RNA polymerase was used. Specific activities of $[\alpha^{32}P]$- triphosphates were between 10 and 400 µCi/µmol. $[^{14}C]$- triphosphates had specific activities about 300 µCi/µmol. Sometimes, incubations were carried out in a total volume of 0.12ml.

Estimates of the extent of incorporation of radioactivity into RNA were made by withdrawing an aliquot of 0.005ml from the incubation mixture, spotting this on to DEAE-paper, and developing with 0.3M-ammonium formate. This elutes unreacted nucleoside triphosphates away from the origin, while RNA remains in the origin area. Radioactivity in the origin area was then measured on a gas-flow counter.

Except as noted, RNAs were isolated free of nucleoside triphosphates as follows. RNA preparations were kept on ice during

the isolation procedure. 0.02ml of yeast RNA, 20mg/ml, was mixed with the incubation mixture. 0.5ml of cold 5% (w/v) TCA was added, and the mixture vortexed. After standing for 5-10 min 3.0ml of cold water was added, and the precipitate collected by centrifuging at 1000 g for 10 min. The supernatant was poured off, the tube dried, and the precipitate dissolved in 0.2ml of cold 0.05 M-NaOH. This acid precipitation cycle was repeated twice, then the precipitate was washed with 4ml ethanol/ether (3:1 v/v), dried at 20°C with an air stream, and dissolved in 0.1ml of 0.05 M-tris-base. 0.3ml of 0.001 M-EDTA was added, and the pH adjusted to 7.5-8.0 with 0.1 M-HCl. Sometimes the RNA was then heated to 95°C for 5 min and cooled in ice. Any precipitate formed at this stage was non-radioactive and was discarded.

2.5.2 Enzymatic Hydrolysis of RNA.

Pancreatic and T_1RNase digests were generally made with an enzyme: substrate ratio of 1:20 (w/w) with the RNA prepared as described above. Digestion was for 1h at 37°C. Combined T_1 and U_2 digests were made by first digesting with T_1 RNase, as above, then adding sodium acetate, EDTA and crystalline BSA to final concentrations of 0.05M, 0.002M and 0.1mg/ml, respectively. The sodium acetate was added from a stock of 0.55M, pH 4.2; the final pH of the solution was about 4.5. 0.5 unit of U_2 RNase was added per mg RNA, and the mixture incubated at 37°C for 2-4 h. Digests of volume greater than 0.1ml were freeze-dried and dissolved in 0.1ml water before applying to paper.

2.5.3 Fractionation of Pancreatic RNase Digests.

Digests were applied to 100 x 7cm lengths of DEAE-paper, which were developed for 10-12 h, descending, with 0.20-0.22 M-ammonium formate, containing 7 M-urea; this gave good separation of oligonucleotides into length classes, up to hexanucleotides. DEAE-paper was prewashed with 1.0 M-formic acid and then with water, and dried. Paper prepared in this way gave a much faster flow rate than normal DEAE-paper. Separations were checked with known mono-, di- and trinucleotides. However, higher order peaks were not characterised: their identity was assumed by analogy with other work on DEAE-paper (e.g. Ohsaka et al., 1964; de Wachter & Fiers, 1969), and with the results of Tomlinson & Tener (1963) using DEAE-cellulose columns.

After such fractionation, nucleotides were identified under u.v. light, and radioactivity located with an Actigraph scanner. The DEAE-paper was then washed with ethanol for at least 50 h to remove urea. Radioactivity in different fractions was measured by immersing the appropriate sections of paper in toluene - PPO and counting in a scintillation counter. Paper sections were then washed with toluene, ethanol and ether, and dried. Nucleotide fractions were eluted with 30% TEAC (v/v), pH10 (Sanger, Brownlee & Barrell, 1965). TEAC was removed by repeated cycles of freeze-drying and addition of water. Alkaline hydrolysis was performed as described below (section 2.5.5). Dinucleotides were fractionated by electrophoresis at pH 1.9 as described in section 2.5.4. T_1 digests were made of part of each fraction longer than dinucleotide. These were fractionated by descending

chromatography on DEAE-paper (prewashed), developing with 0.05M-formic acid for 2 h. This separates CMP from other components (Jacobson, 1964).

2.5.4 Fractionation of T_1 and U_2 RNase Digests.

T_1 and U_2 digests were fractionated by electrophoresis on DEAE-paper at pH 1.9 (Sanger et al., 1965). By this means the dinucleotide CpGp can be isolated directly. Digests were applied as 1cm streaks. Electrophoresis of T_1 digests was carried out for 15-16 kV-h on a water-cooled flat-plate apparatus. Combined T_1 and U_2 digests were electrophoresed for 16-18 kV-h. The buffer consisted of 8.7% (v/v) acetic acid, 2.5% (v/v) formic acid. The potential drop across the paper was kept below 2kV. The paper was then dried, and the carrier nucleotides marked under u.v. light.

Sometimes the whole paper strip from each digest was scanned with the Actigraph. In every case, radioactivity was measured by counting sections of the paper in a low-background gas-flow counter. Generally, the first 20cm of paper was divided into ten 2cm sections and the rest of the paper, which contained the nucleotide species of particular interest, was divided into 0.5cm sections. Oligonucleotides were removed from fractions of interest by elution with 30% (v/v) TEAC, pH 9.5, into glass capillaries (Sanger et al., 1965). TEAC was removed as described above.

2.5.5 Alkaline Hydrolysis of Oligonucleotides and RNA.

The procedure was based on that of Sanger et al.(1965).

Oligonucleotide samples were freeze-dried and then taken up in 0.02ml of 0.3 M-NaOH, containing carrier RNA, 4mg/ml, if necessary for subsequent location of electrophoresis products. Each sample was sealed into a glass capillary and incubated at 37°C for 16-18 h. Digests were applied as 1.5cm streaks to Whatman 3MM paper (without prior neutralisation) and the mononucleotides separated by electrophoresis at pH 3.5 for 7 kV-h (Markham & Smith, 1952). The buffer used was 0.1 M-ammonium formate adjusted to pH 3.5 with 100% formic acid. Nucleotides were identified under u.v. light and corresponding areas of paper measured for ^{32}P in a scintillation counter. GMP was observed to fractionate into two adjacent areas with this method; presumably these were 2^1- and 3^1- GMP. Samples of the original RNA preparations for nearest-neighbour analysis were treated in this way.

PART 3 : NEAREST - NEIGHBOUR ANALYSES

3.1 PARVOVIRUSES

3.1.1 Introduction

Parvoviruses are small, DNA animal viruses. They have been isolated from tumours and irradiated tissues, and as contaminants of adenovirus stocks. Their pathology is not clearly defined: they cause acute disease when inoculated into newborn hamsters (Kilham, 1961), but may generally exist in a latent state of low pathogenicity.

Some physical properties of the DNA of the minute virus of mice (MVM) have been studied (Crawford, 1966); these were consistent with the DNA being single stranded. Kilham rat virus (RV) (Kilham & Olivier, 1959) is in many ways similar to MVM. May, Niveleau, Berger & Brailovsky (1967) reported that RV DNA had a double-stranded structure. However, more recently RV DNA was examined by Robinson & Hetrick (1969) and their results indicated that the DNA was single stranded. H-1 virus (Toolan, 1960) is similar to MVM and RV. Cheong, Fogh & Barclay (1965) reported briefly that H-1 DNA was unusual in that it was not denatured by heating at 100°C for 7 min. Recently, Usategui-Gomez, Toolan, Ledinko, Al-Lami & Hopkins (1969) reported that H-1 DNA was single stranded.

The DNAs of MVM, RV and H-1 were thus of interest in several ways. They came from a poorly characterised group of very small animal viruses, and were perhaps single stranded. Nearest-neighbour analyses were therefore made of MVM, RV and H-1 DNAs in an attempt to define their base compositions, structures and relations with other DNAs. Nearest-neighbour analyses were also carried out on MVM and H-1 DNAs using *E.coli*

RNA polymerase. The results of these analyses are shown in Table 5 alongside the DNA polymerase results, but interpretation of the RNA polymerase results is deferred till after discussion of the DNA polymerase results.

3.1.2 DNA Polymerase Nearest-Neighbour Analyses

As mentioned in Section 1.2.6 it is possible that the extent of replication achieved *in vitro* with such small DNA templates may affect the nearest-neighbour frequencies obtained. In this work DNA was produced *in vitro* equal to 20 to 30% of the input template DNA. Thus, if the DNAs are single stranded, only the complementary strand should be produced. It is assumed that this amount of DNA constitutes a representative sample of total virus DNA. Subsequent comments depend on this assumption, so a brief study was made of the influence of the extent of replication using MVM DNA and $[\alpha\ ^{32}P]$-dGTP in three separate incubations for different times. The extent of replication was estimated from the amount of acid-precipitable radioactivity in aliquots of the incubation mixtures (Table 3). In this case the dinucleotide frequencies are essentially independent of the extent of replication, and it seems reasonable to expect a similar situation with the other labelled triphosphates, and with the other virus DNAs. (It is of interest that attempts to obtain more than 100% replication of template, using MVM and H-1 DNAs, were not successful. However, this matter was not pursued).

With double-stranded DNAs, the accuracy of a nearest-neighbour analysis can be judged by the agreement between complementary

dinucleotides' frequencies, and by the agreement between the (G+C) value found by nearest-neighbour analysis and that given by buoyant-density centrifugation or by chemical methods. Neither of these criteria were applicable to work with the parvovirus DNAs. Since it was suspected that the DNAs were single stranded, it was not necessary that complementary doublets' frequencies should be equal. Also, the base compositions were unknown. The only criteria available were that duplicate analyses agreed closely, and that results were stable to additional nuclease treatment of the DNA digests.

In Table 4 are shown the A/T and G/C ratios found by nearest-neighbour analysis for the DNAs of MVM, RV and H-1, and also the mean values for fifteen double-stranded DNAs (viral and cellular) obtained concurrently. As found by Swartz et al. (1962), all the double-stranded DNAs yielded A/T ratios less than 1.00. The A/T ratios for MVM, RV and H-1 DNAs, and the G/C ratios for MVM and RV DNAs, are outside the range of values obtained for known double-stranded DNAs and thus are not compatible with these virus DNAs being double stranded. The nearest-neighbour frequencies of the three DNAs also are consistent only with a single-stranded structure (Table 5 and Fig.7) since several of the complementary pairs, particularly GpA:TpC and GpT:ApC show large differences.

The conclusion that MVM DNA is single-stranded confirms the results of Crawford (1966) and Crawford, Follett, Burdon & McGeoch (1969) with other techniques. RV DNA has been studied by May et al. (1967) and by Robinson & Hetrick (1969), with contradictory results. The present

results provide strong support for those of Robinson & Hetrick (1969) and raise the question of the identity of the double-stranded DNA studied by May et al. (1967). The conclusion that H-1 DNA is single stranded is supported by the work of Usategui-Gomez et al.(1969).

The nearest-neighbour analyses also provide information on the similarity of these DNAs. The overall patterns of the DNAs are very similar, and differences between complementary pairs are in the same sense: GpA > TpC and GpT < ApC. Estimated by the differences between duplicate analyses, the precision of the values for individual frequencies is $\pm 5\%$. On this basis each of the analyses contains several values which differ significantly from the corresponding values in the other analyses. RV and H-1 DNAs show only small differences but MVM DNA is distinct.

If the frequencies of each complementary pair of dinucleotides are averaged, then it can be seen that these DNAs possess the typical vertebrate nearest-neighbour pattern (Fig. 8); in particular, they all show a low frequency of the dinucleotide CpG. Apart from EMC RNA, which does show some resemblances to the vertebrate pattern, the parvovirus DNAs are thus the first single-stranded nucleic acids known with this sequence structure. This topic is discussed further in Section 4.1. According to the arguments of Subak-Sharpe et al. (1966) this similarity in pattern to vertebrate DNA suggests that the DNAs of these parvoviruses may have been derived from DNA closely related to that of the host organism.

TABLE 3. INFLUENCE OF EXTENT OF REPLICATION OF MVM-DNA ON NEAREST-NEIGHBOUR FREQUENCIES

Extent of Replication	20-30%	50-60%	90-100%
3'-dAMP	32.4 *	32.0	33.3
3'-TMP	36.4	36.5	37.6
3'-dGMP	24.0	24.6	23.4
3'-dCMP	7.1	7.1	5.7

* Percentage of total ^{32}P (from $[\alpha^{32}P]$-dGTP) associated with each 3'-mononucleotide.

---ooOoo---

TABLE 4. A/T AND G/C RATIOS OF PARVOVIRUS (-) STRAND DNAs

DNA	DNA Polymerase		RNA Polymerase	
	A/T	G/C	A/T	G/C
MVM	1.23	1.10	1.56	1.14
H-1	1.15	1.00	1.48	1.15
RV	1.10	1.10	–	–
Double-stranded DNAs	0.95±0.03 (0.91-0.99)	1.00±0.04 (0.93-1.07)	0.99	1.10

The values for DNA polymerase analyses of 15 viral and cellular double-stranded DNAs are shown as mean ± standard deviation with range in brackets. Only one RNA polymerase analysis of a double-stranded DNA (calf thymus DNA) is available.

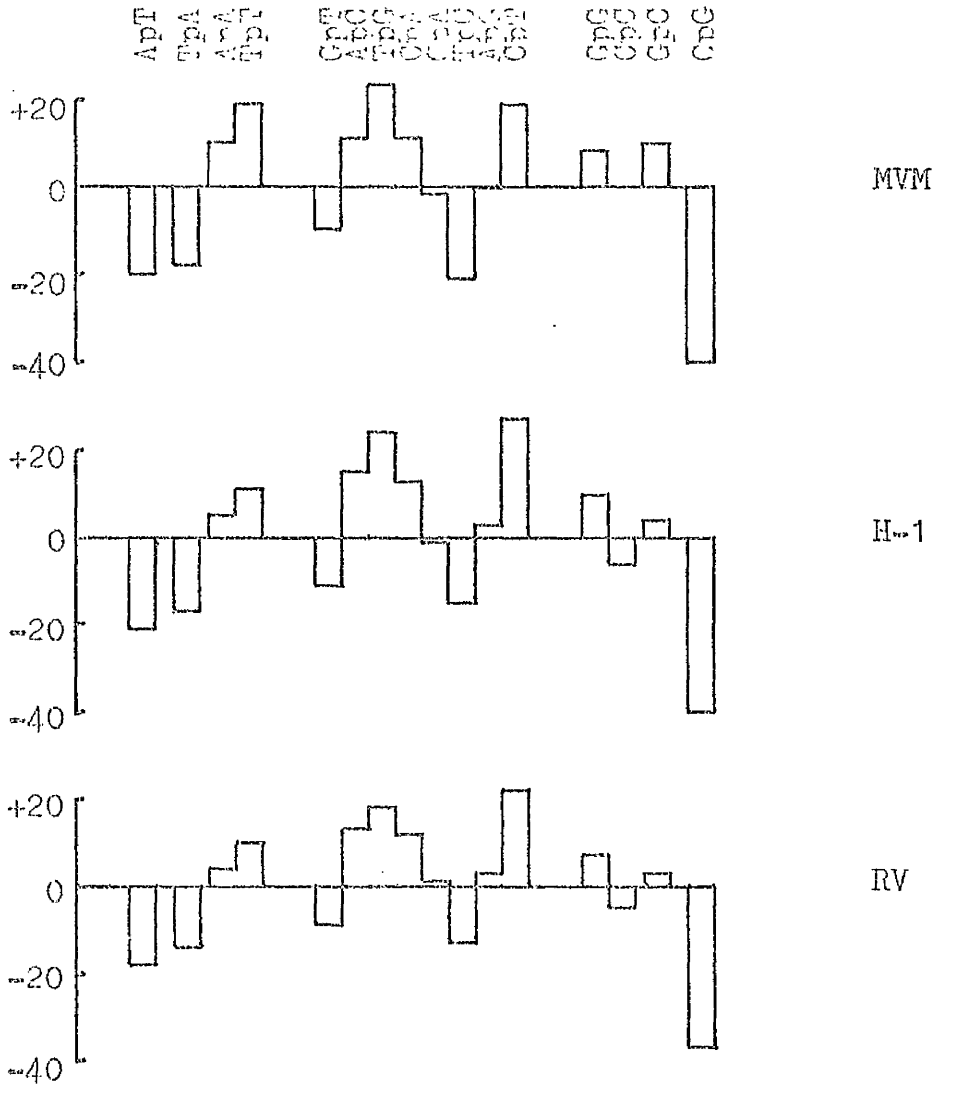
TABLE 5. NEAREST-NEIGHBOUR FREQUENCIES OF PARVOVIRUS DNAs

		MVM		H-1		RV	
Analyses with DNA Polymerase	ApA	TpT	123	91	93	76	94	84
	CpA	TpG	75	78	80	80	72	79
	GpA	TpC	67	34	65	44	69	44
	CpT	ApG	67	69	82	69	75	72
	GpT	ApC	47	75	47	82	53	74
	GpG	CpC	51	38	60	46	59	39
	TpA		62		55		61	
	ApT		59		49		57	
	CpG		15		18		20	
	GpC		48		34		49	

		MVM		H-1	
Analyses with RNA Polymerase	ApA	UpU	119	74	104	62
	CpA	UpG	86	59	84	65
	GpA	UpC	81	34	85	39
	CpU	ApG	57	100	63	96
	GpU	ApC	45	76	46	78
	GpG	CpC	49	37	56	39
	UpA		62		57	
	ApU		53		52	
	CpG		17		23	
	GpC		51		53	

Frequencies are expressed as parts per 10^3.

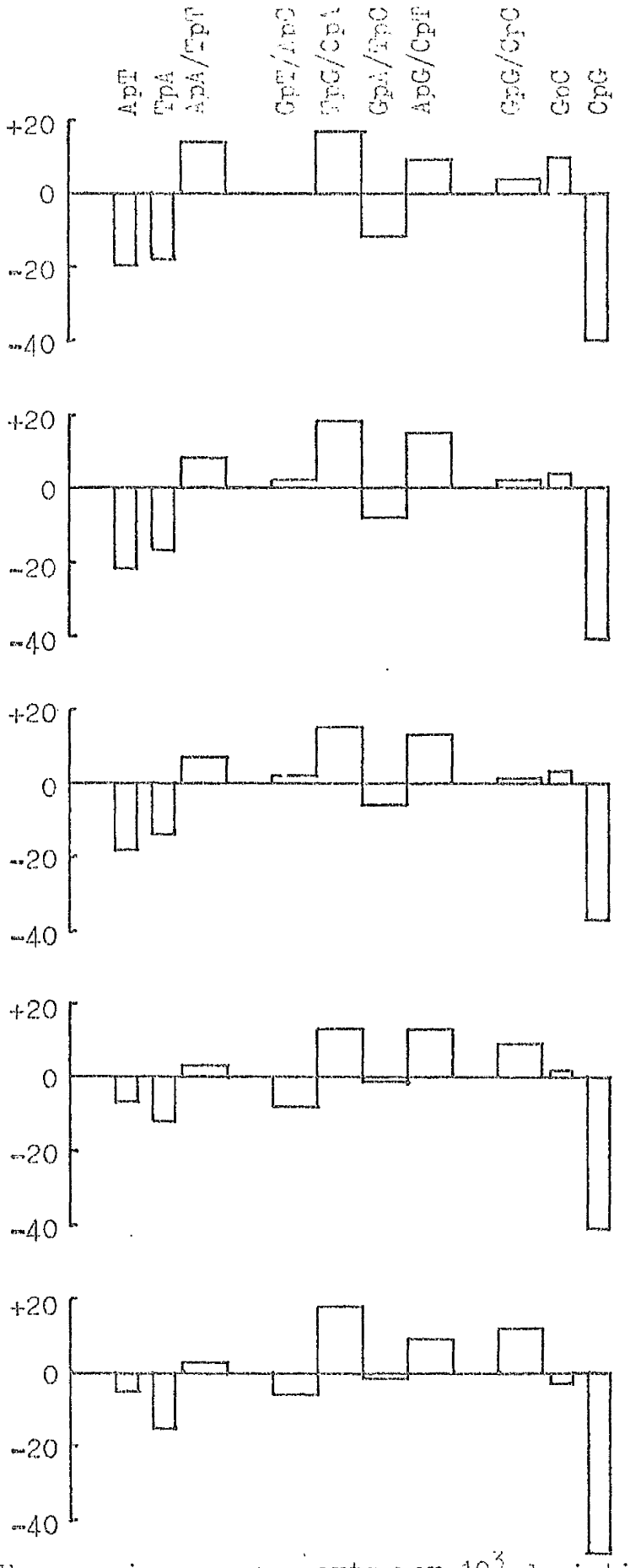
FIGURE 7. [illegible] POLYPURINE-PYRIMIDINE-RICH [illegible]
PATTERNS OF PARVOVIRUS DNAs



Frequencies are as parts per 10^3 deviation from random expectation.

FIGURE 8. [illegible] PARVOVIRUS DNA H.A EXT-

[illegible]



Frequencies are as parts per 10^3 deviation from random.
Frequencies of complementary doublets are averaged.

3.1.3 RNA Polymerase Nearest-Neighbour Analyses

In the course of studies on the CpG shortage in vertebrate pattern DNAs (Section 4.3), nearest-neighbour analyses were made of MVM, H-1 and calf thymus DNAs using E.coli RNA polymerase. The results for MVM and H-1 are shown in Table 5 and Fig. 9. These show some differences from the DNA polymerase analyses. In this case it seems better to analyse the differences between the two methods for the same DNA by comparing the fractional incorporation data (Table 6) and not the final calculated frequencies. This is because the absolute frequencies are obtained by first calculating base compositions from the fractional incorporation data, and then multiplying the fractional incorporations by the appropriate base frequencies. Thus, any localised doublet frequency change will alter the base composition and therefore could affect all the calculated doublet frequencies to some extent. This method of comparing results is suitable only for the special case of results obtained by different methods for the same DNA. In the following discussion, as a matter of convenience, the DNA polymerase analyses are taken as standards, to which the changes in the RNA polymerase analyses are referred. The abbreviation "T" for thymidine is used to represent both T and U in this section.

In Table 6 it is shown that the frequencies of dinucleotides with A, T and C at the 3'-end are, with both viral DNAs, not changed very much, and that changes in frequencies are generally in the same directions in both cases. However, the doublet set with G at the 3'-end shows radical change: ApG frequency shows a large increase, and

TpG a large decrease in the RNA polymerase results. These changes are not found with the calf thymus DNA analysis, where all the frequencies remain constant with the two methods.

This large change is not due to analytical errors. All results were duplicated at different times, and the relevant MVM DNA-polymerase frequencies were constant for different amounts of replication (Table3). The same H-1 DNA preparation was used for all analyses, but different MVM DNA preparations were used for the DNA and RNA polymerase systems.

It is conceivable that the two polymerase systems might preferentially copy different sequences of the virus DNAs. The average chain lengths synthesised, especially with the RNA polymerase, may be of importance: in this work large amounts of RNA polymerase were used (several polymerase molecules per virus DNA molecule). If the RNA chains synthesised are short, it is possible that transcribed initiation sequences could constitute a measurable proportion of the chains.

It has been found that DNA polymerase nearest-neighbour analysis often over-estimates T content. This does not appear to apply to RNA polymerase. Therefore, this phenomenon is probably responsible for the small decrease, in the RNA polymerase analyses, in the frequencies of TpN doublets, where N represents C, A and T. However, it does not account wholly for the large differences found with the NpG sequences, and observed only with the single-stranded virus DNAs.

The RNA polymerase NpG results do not at all imply that TpG sequences are replaced directly by ApG sequences. The results mean that the relative amounts of ApG and TpG copied are changed. It is obvious

that a number of explanations are possible, taking either the DNA or RNA polymerase analyses as standard, and considering the changes in terms of TpG or ApG. For several reasons, preference is given here to the hypothesis that, with the RNA polymerase, ApG sequences are elevated from the "normal" DNA polymerase level. First, the virus DNA NpG levels found with DNA polymerase resemble the invariant calf thymus DNA levels more than the RNA polymerase values do. Second, it seems likely that DNA polymerase would act as a less discriminating copier than RNA polymerase. Third, as mentioned above, DNA polymerase gives results for MVM DNA which are constant with varying extents of replication. (However, analogous experiments with RNA polymerase have not been performed). It is known that RNA polymerase finds more initiation sites on single-stranded DNA than on double-stranded DNA, and that with single-stranded DNA as template many short chains are synthesised (Maitra, Cohen & Hurwitz, 1966; Geiduschek & Haselkorn, 1969). It is therefore possible that the proportions of transcribed initiation sequences are exaggerated. The notable depression of TpG levels is then due to, first, removal of the DNA polymerase high-T error and, second, a proportional lowering, together with GpG and CpG, as ApG is raised. Since more TpG than GpG or CpG is found, the reduction is largest with TpG. This scheme can account semi-quantitatively for the changes.

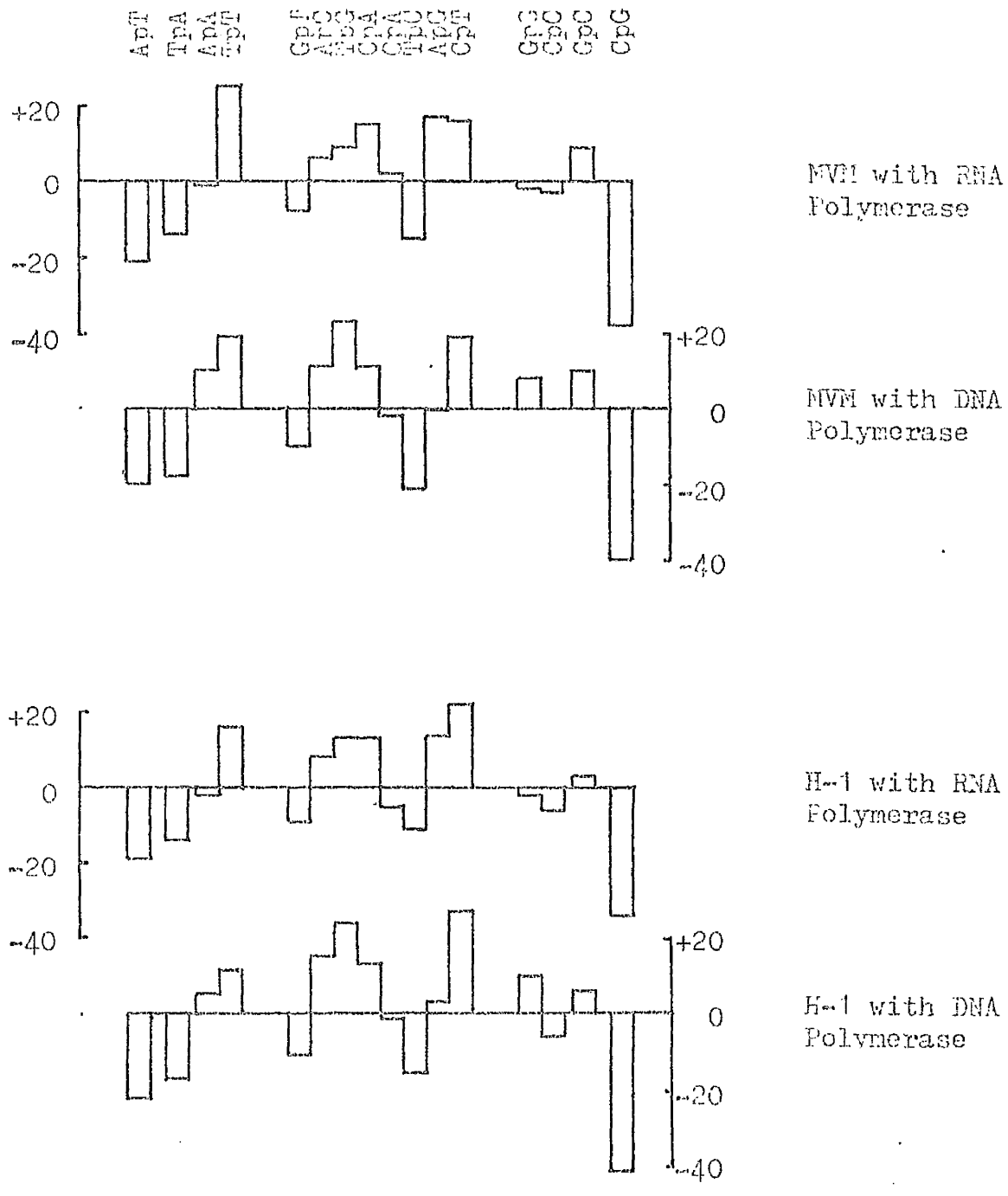
To summarise, the two nearest-neighbour methods give identical results for native calf thymus DNA, but give similar but not identical results for the single-stranded parvovirus DNAs. (It would be of interest to carry out an RNA polymerase analysis on denatured calf thymus

TABLE 6. COMPARISON OF NEAREST-NEIGHBOUR ANALYSES WITH DNA POLYMERASE AND WITH RNA POLYMERASE

	MVM		H-1		CALF THYMUS	
	DNA	RNA	DNA	RNA	DNA	RNA
ApA	37.6	34.2	31.7	31.5	31.1	31.1
TpA	19.0	18.0	18.8	17.2	18.5	21.9
GpA	20.5	23.2	22.3	25.9	22.4	22.9
CpA	22.9	24.8	27.3	25.4	27.9	24.1
ApT	22.4	23.3	19.2	23.3	25.8	26.9
TpT	34.3	32.1	29.9	28.0	30.7	32.1
GpT	17.8	19.7	18.6	20.6	19.8	18.4
CpT	25.4	25.0	32.3	28.1	23.7	22.7
ApG	32.4	44.3	30.6	40.3	33.7	34.6
TpG	36.4	26.4	35.1	27.0	35.5	35.7
GpG	24.0	21.6	26.3	23.3	23.4	24.5
CpG	7.1	7.7	8.0	9.5	7.5	5.3
ApC	38.4	38.2	36.2	37.3	25.0	25.2
TpC	17.6	17.2	19.5	18.5	30.9	32.1
GpC	24.6	25.7	24.0	25.3	20.3	20.1
CpC	19.4	18.8	20.3	18.8	24.9	22.7

The frequency of each dinucleotide XpY is shown as a percentage of the total frequencies of all four dinucleotides having Y their 3'-nucleoside. The underlined values are those showing a major difference between the two methods of analysis.

FIGURE 9. [illegible] PATTERNS OF PARVOVIRUS DNAs OBTAINED WITH RNA AND DNA POLYMERASES



Frequencies are as parts per 10^3 deviation from random.

DNA). These differences do not vitiate the conclusions drawn earlier from the DNA polymerase results. As shown in Table 4, the RNA polymerase results also indicate a single-stranded structure for the virus DNAs, and most of the features of the analyses do not show large changes.

3.1.4 Base Compositions of Virus DNAs

The base compositions of the virus DNAs obtained by the two nearest-neighbour methods are shown in Table 7. Since nearest-neighbour analyses refer to the complementary strand, the experimental values for A and T, and for G and C, have been interchanged. The compositions shown are therefore those of the virus strands. Data obtained by Crawford et al. (1969) for MVM DNA and by Usategui-Gomez et al. (1969) for H-1 DNA are also shown. These data were obtained by enzymatic hydrolysis of [^{32}P]-DNA to 5' or 3'-nucleotides.

DNA polymerase nearest-neighbour analysis of double-stranded DNAs often overestimates the thymine content of the DNA. If this error applies to single-stranded DNAs, then the results obtained by DNA polymerase nearest-neighbour analysis may underestimate the thymine contents of the virus DNAs. Inspection of the MVM DNA results indicates that this error probably is operating, since higher thymine contents are obtained by the other methods. Except for this reservation, the results obtained for MVM DNA by different methods agree quite well. The values for H-1 DNA obtained by the two nearest-neighbour methods also agree,

with the same reservation. These values are similar also to those found for RV DNA.

This general agreement gives confidence in the set of results discussed so far. However, the results of Usategui-Gomez et al. (1969) for H-1 DNA differ considerably from the others in their values for G and C contents. This determination was made by digesting [^{32}P]- DNA to 5'-mononucleotides, which were separated by low pH electrophoresis. There are several possible causes of the discrepancy. First, if the specific activities of the different deoxynucleoside 5'-triphosphates in the cell were different, these differences would be seen in the final results, since each phosphate group remains attached to the same nucleoside throughout. Second, the result could be due to a digestion or electrophoresis artifact. As described in Sections 2.3.3 and 2.3.4, such degradation procedures require considerable care. It is of interest that the result of Crawford et al. (1969) for MVM DNA, using hydrolysis to 5'-nucleotides, also exhibits a relatively low G value.

All these DNAs have a high thymine content. This seems to be quite a general characteristic of single-stranded virus DNAs: it is also shown by the single-stranded DNAs of phages ØX 174, fd and M13 (Sinsheimer, 1959; Hoffmann-Berling, Marvin & Durwald, 1963; Salivar, Tzagoloff & Pratt, 1964). The parvovirus DNAs also have (G + C) values in the range 40-45% i.e. close to the values for vertebrate DNAs.

Some of the results described here have been published (Crawford

TABLE 7. BASE COMPOSITIONS OF PARVOVIRUS DNAs

VIRUS	METHOD OF ANALYSIS	A %	T %	G %	C %
MVM	Hydrolysis of [^{32}P]-DNA to 5'-nucleotides*	24.8	35.8	17.8	21.8
	Hydrolysis of [^{32}P]-DNA to 3'-nucleotides*	23.6	33.3	20.7	22.4
	DNA polymerase nearest-neighbour analysis	26.5	32.7	19.5	21.4
	RNA polymerase nearest-neighbour analysis	23.0	34.8	19.8	22.5
H-1	Hydrolysis of [^{32}P]-DNA to 5'-nucleotides‡	25.2	33.1	14.3	27.4
	DNA polymerase nearest-neighbour analysis	25.5	29.3	22.6	22.6
	RNA polymerase nearest-neighbour analysis	22.2	33.0	20.8	24.0
RV	DNA polymerase nearest-neighbour analysis	26.8	29.6	20.6	22.9

*From Crawford et al. (1969).

‡From Usategui-Gomez et al. (1969).

et al., 1969; McGeoch, Crawford & Follett, 1970).

3.2 ADENOVIRUSES

3.2.1 Introduction

This group of viruses has been described by Pereira, Huebner, Ginsberg & Van der Veen (1963). They contain double-stranded, acyclic DNA. Most of them are associated with respiratory infections. Some adenoviruses cause malignant tumours in newborn hamsters. The work described here was concerned with human adenoviruses, of which there are 31 serological types (Lacy & Green, 1967). The human adenoviruses have been divided into three groups according to the base compositions of their DNAs (Pina & Green, 1965): first, those with (G + C) contents of 48-49%, which include the highly oncogenic Ad 12, 18 and 31; second, those with (G + C) contents of 50-52%, which include Ad 3, 7, 11, 14, 16 and 21, and are typically weakly oncogenic; and, third, those with high (G + C) contents (55-61%), which include the non-oncogenic Ad 1, 2, 4, 5, 6, 8, 9, 10, 13, 15, 17, 19 and 22-30.

The Ad DNAs have molecular weights of 20-25 million (Green et al., 1967). The DNAs of the oncogenic viruses are slightly smaller than those of the non-oncogenic viruses (Green et al., 1967; Van der Eb, Van Kesteren & Van Bruggen, 1969). Lacy & Green (1964, 1965, 1967) have examined relations between the groups of adenoviruses by DNA-DNA hybridisation. The results indicate that close sequence relations exist within the highly oncogenic group, and within the weakly oncogenic group; lesser homologies were detected between DNAs of non-oncogenic

viruses. Sequences in the highly oncogenic group are quite distinct from those in the weakly oncogenic group.

The nearest-neighbour pattern of Ad 2 DNA was determined by Morrison et al. (1967) (Fig.6). As described above, the adenoviruses can be divided into three groups by several criteria. It was, therefore, of interest to determine the nearest-neighbour patterns of the DNAs from several viruses in each group, to see if these correlations could be extended to the nearest-neighbour frequencies. The nearest-neighbour patterns found for the oncogenic papova viruses resemble those of mammalian DNA, while the patterns found for large non-oncogenic viruses are different: it was of interest to ascertain whether this trend existed also within the adenovirus group.

3.2.2 Results and Discussion.

The results of nearest-neighbour analyses for eight Ad DNAs are shown in Table 8 and Fig.10. The base compositions of the Ad DNAs are shown in Table 9. The results for Ad 2 are from Morrison et al. (1967). Agreement between the frequencies of complementary dinucleotides is, in many cases, not good. This variability is most marked in the results for Ad 12, 18 and 27. The A/T ratios given by nearest-neighbour analysis are all less than 1.00. Again, the results for Ad 12, 18 and 27 show the largest discrepancies. The (G + C) contents obtained were all lower than those found by Pina & Green (1965).

Difficulties have been experienced in determining Ad DNA base compositions by other methods: Pina & Green (1965) found that the

buoyant densities of the intermediate (G + C) group were consistently higher than expected. In the present study it was found that the Ad DNA preparations used had low template activities for E.coli DNA polymerase. The DNAs were therefore activated as described in Section 2.3.1. Since the distribution of A-T rich regions in Ad DNAs is non-uniform (Doerfler & Kleinschmidt, 1970; Kimes & Green, 1970), it is possible that the activation procedure led to some preferential copying of A-T rich regions. This hypothesis could account for the low (G + C) values obtained but not for the low A/T ratios or the discrepancies between complementary doublets' frequencies, and it is evident that some other error must be operating. Many of these analyses had to be performed on half the normal scale because of the small quantities of DNAs available.

The errors operating in these analyses seem to be similar for all the DNAs, and the results obtained are sufficiently precise for comparisons to be made. All the DNAs give similar deviation histograms, which are also similar to the pattern obtained by Morrison et al.(1967) for Ad 2. However, the Ad 2 DNA results show a higher specific CpG content than found for the other DNAs in the present study. To facilitate detailed comparisons, the complementary doublet frequencies for each DNA were averaged: this reduced data (Table 10) demonstrates the close similarity between the frequencies of the members of each group. This near-identity of pattern within each group encouraged the calculation of typical frequency patterns for each group by averaging

all the results for each group. These data (Table 10) show that the frequencies of dinucleotides containing both a purine and a pyrimidine are very similar in all three groups, while those dinucleotides containing either two pyrimidines or two purines all vary between the groups in accordance with (G + C) changes. This is also illustrated by Fig.11. The nearest-neighbour analyses therefore suggest, together with other data on (G + C) contents and sizes of the DNAs and on the biology of the adenoviruses, that the human adenoviruses are closely related to each other, in terms of the sequence structures of their DNAs, and that they probably have arisen by divergence from a common ancestor. Apart from the changes in base composition, which are applied proportionally to all dinucleotides, there are no distinctive features which can be correlated with oncogenic potential.

The nearest-neighbour data thus indicate unequivocally that the Ad DNAs are closely related. However, the relation of the whole group of Ad DNAs to other DNAs is less clear. Ad DNAs do not exhibit the very low frequency of CpG found with vertebrate DNAs. However, the CpG frequency in Ad DNA is still less than random. In other respects (Fig.11) the deviation histograms of the Ad DNAs are qualitatively quite similar to the vertebrate pattern. However, similarities can also be seen with the E.coli pattern. In conclusion, it seems that the Ad DNA pattern does not show similarities to other DNAs great enough to formulate any firm theories on the origin of the virus DNAs. It is of interest that, from computer comparisons of nearest-neighbour data,

TABLE 8. NEAREST-NEIGHBOUR FREQUENCIES OF ADENOVIRUS DNAs

	Ad 2*		Ad 4		Ad 7		Ad 11	
ApA	TpT	64	68	59	66	69	77	73	83
CpA	TpG	66	71	67	70	73	74	70	72
GpA	TpC	55	55	53	65	54	60	59	64
CpT	ApG	64	62	70	58	66	61	67	61
GpT	ApC	59	56	55	52	55	52	58	54
GpG	CpC	72	72	73	84	65	68	59	58
TpA		44		42		50		53	
ApT		48		53		64		65	
CpG		62		56		45		42	
GpC		82		76		71		62	

	Ad 12		Ad 18		Ad 21		Ad 27	
ApA	TpT	74	106	75	96	73	75	53	63
CpA	TpG	67	69	68	68	70	79	70	73
GpA	TpC	44	68	47	61	63	54	53	70
CpT	ApG	65	54	67	54	64	63	84	61
GpT	ApC	57	42	56	54	57	59	53	58
GpG	CpC	53	65	56	66	68	58	71	69
TpA		60		66		48		43	
ApT		75		72		59		48	
CpG		39		37		44		50	
GpC		61		56		65		77	

Frequencies are expressed as parts per 10^3.

*The data for Ad 2 are from Morrison et al. (1967).

TABLE 9. BASE COMPOSITIONS OF ADENOVIRUS DNAs

VIRUS	A%	T%	G%	C%	A/T	G/C	(G+C)%	(G+C)% *
Ad 2 ‡	22.9	23.9	26.7	26.4	0.96	1.01	53	57
Ad 4	22.1	24.3	25.8	27.8	0.91	0.93	54	57
Ad 7	24.5	26.1	24.4	25.1	0.94	0.97	50	51
Ad 11	25.4	27.3	23.4	23.9	0.93	0.98	47	50
Ad 12	24.6	30.3	21.5	23.6	0.81	0.91	45	49
Ad 18	25.5	29.2	21.6	23.7	0.88	0.91	45	47
Ad 21	25.4	25.6	25.4	23.7	0.99	1.07	49	52
Ad 27	22.1	25.2	25.4	27.4	0.88	0.93	53	60

These data are from nearest- neighbour analyses. The (G+C) values by this method are compared with values obtained by direct chemical methods.

* From Pina & Green (1965).

‡ Data for Ad 2 are from Morrison et al. (1967).

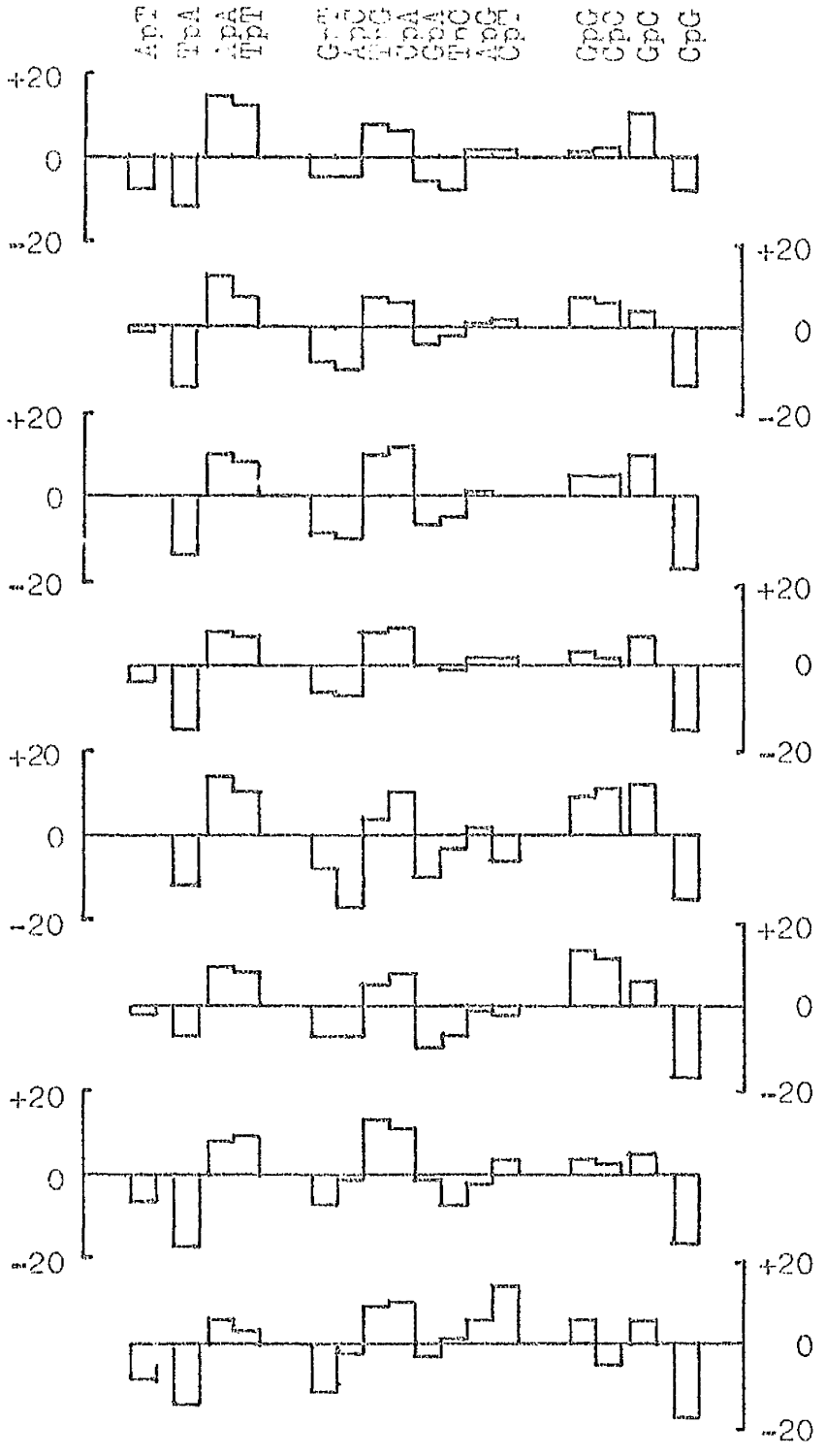
TABLE 10. REDUCTION OF ADENOVIRUS DNA NEAREST-NEIGHBOUR DATA

	LOW (G+C) DNAs		MEDIUM (G+C) DNAs			HIGH (G+C) DNAs			MEANS		
	12	18	7	11	21	2*	4	27	LOW	MEDIUM	HIGH
ApA/TpT	90	86	73	78	74	66	63	58	88	75	62
CpA/TpG	68	68	74	71	75	69	69	72	68	73	70
GpA/TpC	56	54	57	62	59	55	59	62	55	59	59
CpT/ApG	60	61	63	69	64	63	64	73	60	64	67
GpT/ApC	50	55	54	61	58	58	54	56	52	56	56
GpG/CpC	59	61	67	59	63	72	79	70	60	63	74
TpA	60	66	50	53	48	44	42	43	63	50	43
ApT	75	72	64	65	59	48	53	48	73	63	50
CpG	39	37	45	42	44	62	56	50	38	44	56
GpC	61	56	71	62	65	82	76	77	58	66	78

The mean frequency of complementary doublets is given. Frequencies are in parts per 10^3.

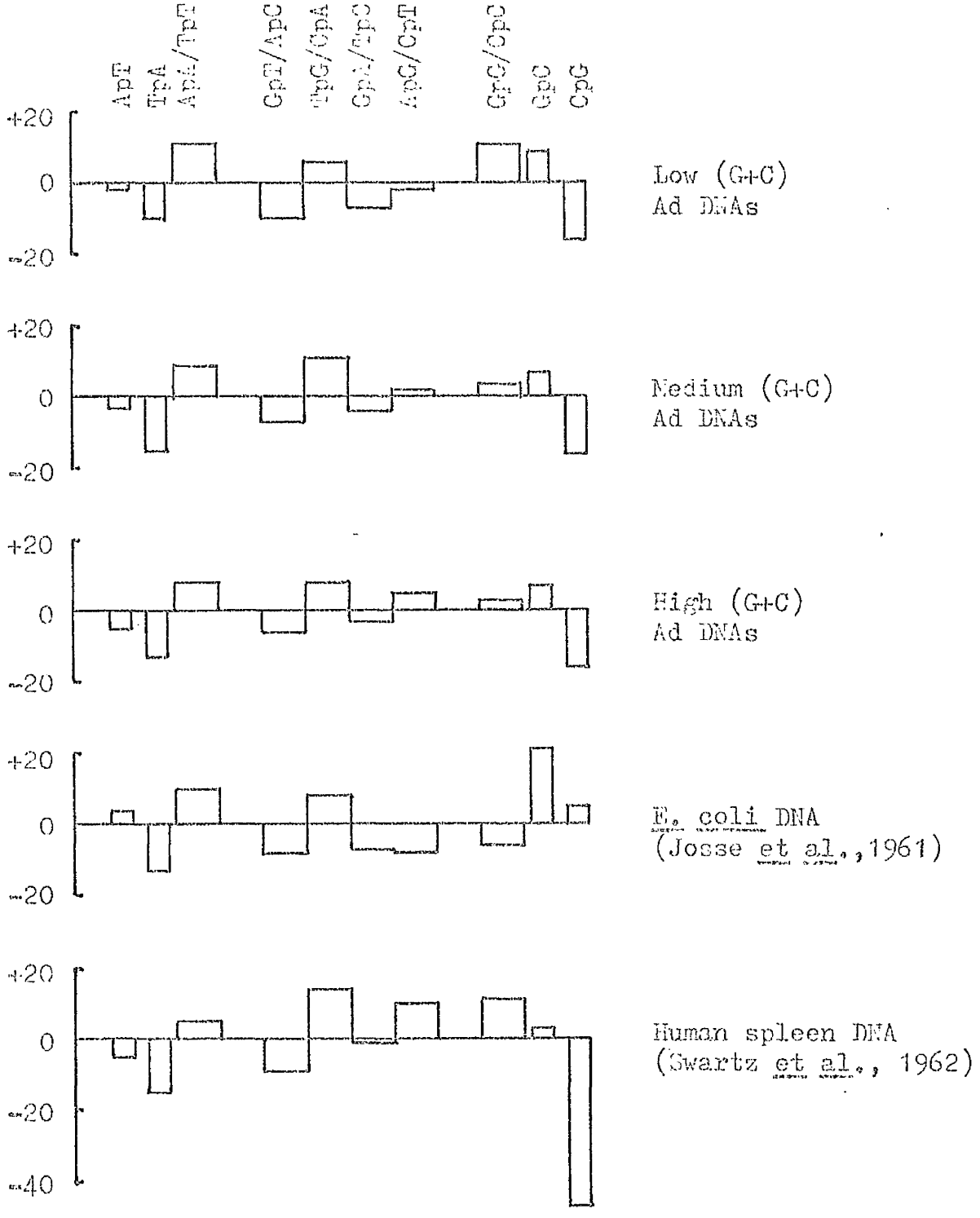
* Data for Ad 2 are from Morrison et al. (1967).

FIGURE 10. NEAREST-NEIGHBOUR PATTERNS OF ADENOVIRUS DNAs



Frequencies are as parts per 10^3 deviation from random.

FIGURE 11. DINUCLEOTIDE FREQUENCIES OF ADENOVIRUS CLASSES



Frequencies are as parts per 10^3 deviation from random. The frequencies of complementary dinucleotides are averaged.

Bellett (1967) classified Ad 2 DNA as an intermediate type between mammalian and bacterial DNAs (See Section 1.2.5). Those similarities in frequency pattern which can be detected with other DNA types could be due not to any common history but to a similar use of the genetic code. This is discussed in Section 3.6.

3.3 BACTERIA

As described in Section 1.2.4 , the known nearest-neighbour patterns of bacterial DNAs can be divided into only a few classes. Further nearest-neighbour analyses of bacterial DNAs were undertaken to extend knowledge of possible patterns. In Table 11 are shown the nearest-neighbour frequencies and derived base compositions for the DNAs of B.megaterium, P.vulgaris, S.marcescens and R.rubrum, and also for phage α DNA. All the results give good agreement between frequency values of complementary doublets. The base compositions obtained by nearest-neighbour analysis agree closely with those estimated from buoyant densities of the DNAs, and also with values tabulated by Hill (1966). However, with these DNAs also, the A/T ratios are all less than 1.00 (0.92 to 0.99).

Fig. 12 shows the deviation histograms of these DNAs, and of E.coli DNA. DNAs from the Eubacteria B.megaterium, P.vulgaris and S.marcescens all give near-random patterns qualitatively similar to that for E.coli DNA, in spite of a variation in (G + C) content from 34% to 57%. The basis of the E.coli pattern, in terms of codon use, is discussed in

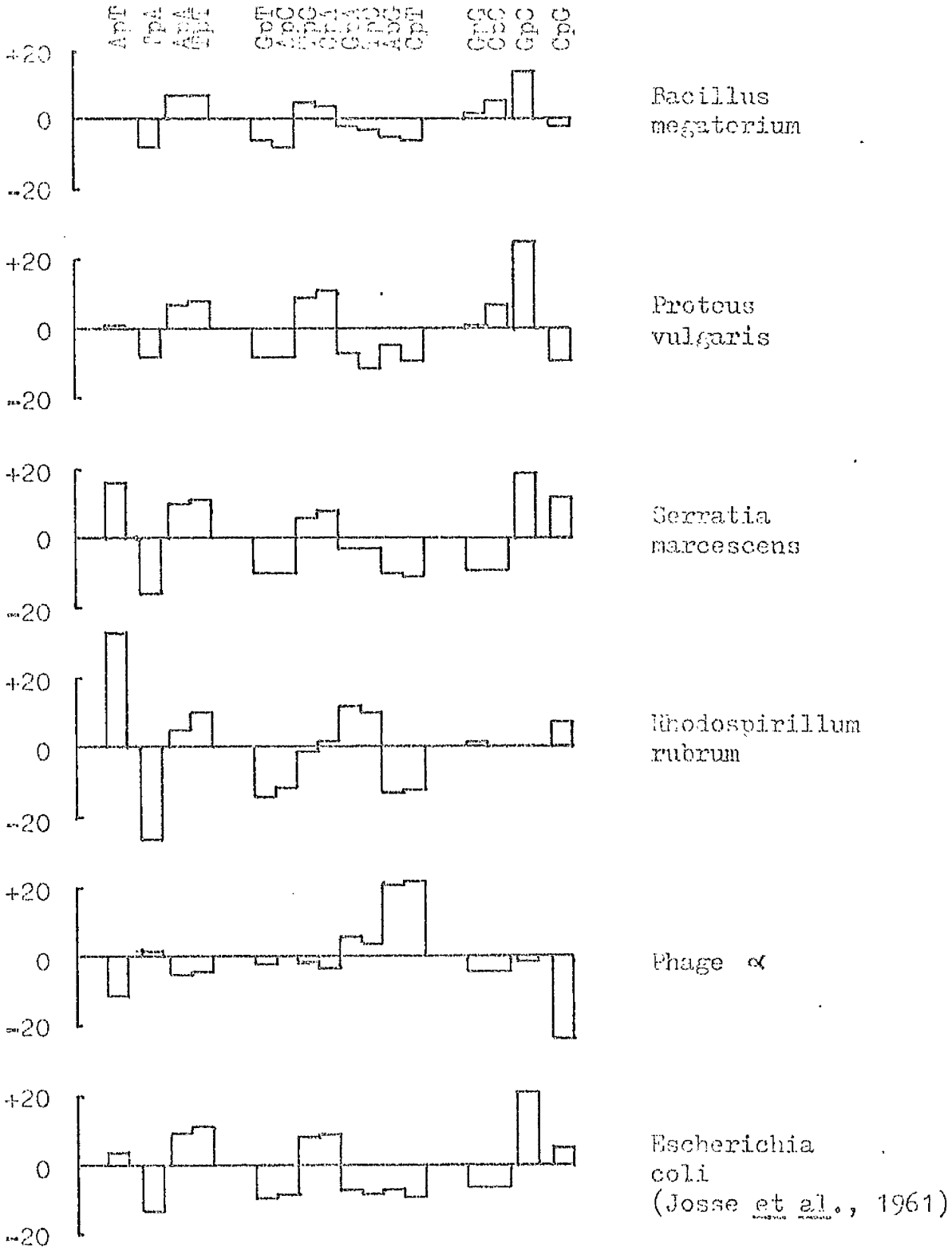
TABLE 11. NEAREST-NEIGHBOUR FREQUENCIES OF BACTERIAL DNAs

	Bacillus Megaterium	Proteus Vulgaris	Serratia marcescens	Rhodosp. rubrum	Phage ∝
ApA TpT	120 122	105 120	50 59	37 47	73 81
CpA TpG	58 62	64 70	67 69	59 59	58 61
GpA TpC	55 53	51 48	56 61	66 73	65 67
CpT ApG	51 52	50 53	53 49	51 45	85 79
GpT ApC	51 50	54 48	54 50	47 47	60 60
GpG CpC	31 31	36 35	69 69	95 98	41 43
TpA	95	88	35	22	85
ApT	108	102	59	56	69
CpG	28	29	96	105	29
GpC	35	47	105	95	45
(G+C) %	24	37	57	62	43
Quoted (G+C)	34	37	55	63	44

Frequencies are in parts per 10^3.

The "quoted" (G+C) values were obtained by buoyant density centrifugation, and agree with the values given by Hill (1966).

FIGURE 12. [illegible] PATTERNS OF BACTERIAL DNAs



Frequencies are as parts per 10^3 deviation from random.

Section 3.6. DNA from the photosynthetic bacterium R.rubrum gives a distinct and unusual pattern. All bacteria, other than R.rubrum, whose nearest-neighbour patterns have been measured to date belong to the order Eubacteriales; R.rubrum belongs to the Rhodobacteriales (Thimann, 1963). Its nearest-neighbour pattern presumably then is a result of its distinct evolution. Phage ∝ DNA also gives an unusual pattern, quite unlike that of its host, B.megaterium.

3.4 EUCARYOTES

Three eucaryote DNAs, from A.nidulans, D.melanogaster and R. catesbeiana, were analysed. Their nearest-neighbour frequency patterns and base compositions are shown in Table 12 and Fig.13. All three DNAs gave good agreement between the frequencies of complementary dinucleotides. A/T ratios are again less than 1.00. The base compositions obtained by nearest-neighbour analysis for D.melanogaster and R.catesbeiana DNAs agreed with buoyant density determinations and literature values. Nearest-neighbour analysis and buoyant density measurements both indicated a (G + C) content of 44-45% for A.nidulans DNA. However, previous estimates for A.nidulans DNA indicated a (G + C) content of 50% (G. Pontecorvo, personal communication). This discrepancy has not been resolved, and results for A.nidulans must therefore be treated with caution.

The deviation histogram for A.nidulans DNA shows frequencies very close to random, and quite similar to those found by Jackson et al. (1965) for baker's yeast DNA (Fig.13). D.melanogaster DNA also gives

a near-random pattern, similar to the E.coli DNA type. This is the first example of an insect DNA nearest-neighbour pattern. It differs markedly from the other analysis of an Arthropod DNA:- main band DNA of Cancer borealis (Swartz et al., 1962). The D.melanogaster DNA pattern is unique among patterns known for multicellular organisms, since it shows no marked deviations from random and, in particular, no CpG shortage. The haploid amount of DNA for D.melanogaster is only about ten times the amount of DNA found in an E.coli chromosome (Kurnick & Herskowitz, 1952). This is an exceptionally low, and perhaps minimum, value for a multicellular organism (Shapiro, 1968). Using the arguments of Section 1.1, it appears likely that much of D.melanogaster DNA is used for protein specification. The amount of DNA available for control and integration functions (and for other uses) must be much less than in higher organisms. It therefore seems reasonable that the sequence structure of D.melanogaster DNA, as summarised by the nearest-neighbour analysis, should be simpler than the structures of DNAs from higher organisms with much larger cellular DNA complements.

DNA from R.catesbeiana (North American bullfrog) exhibits the typical vertebrate pattern, and is the first DNA from an amphibian analysed. This pattern has now been observed in all orders of the Chordata, except for reptiles, for which no results are available. The basis of this pattern is discussed in Section 3.6.

TABLE 12. NEAREST-NEIGHBOUR FREQUENCIES OF EUCARYOTE DNAs

	Aspergillus nidulans	Drosophila melanogaster	Rana catesbeiana
ApA TpT	79 81	99 111	80 92
CpA TpG	58 65	61 68	76 77
GpA TpC	63 59	58 53	51 66
CpT ApG	66 65	58 58	73 64
GpT ApC	52 51	53 51	58 53
GpG CpC	52 . 49	44 39	55 55
TpA	77	76	58
ApT	82	86	73
CpG	43	37	17
GpC	58	51	48
(G+C) %	44	40	44
Quoted (G+C)	50*	41 ‡	43 †

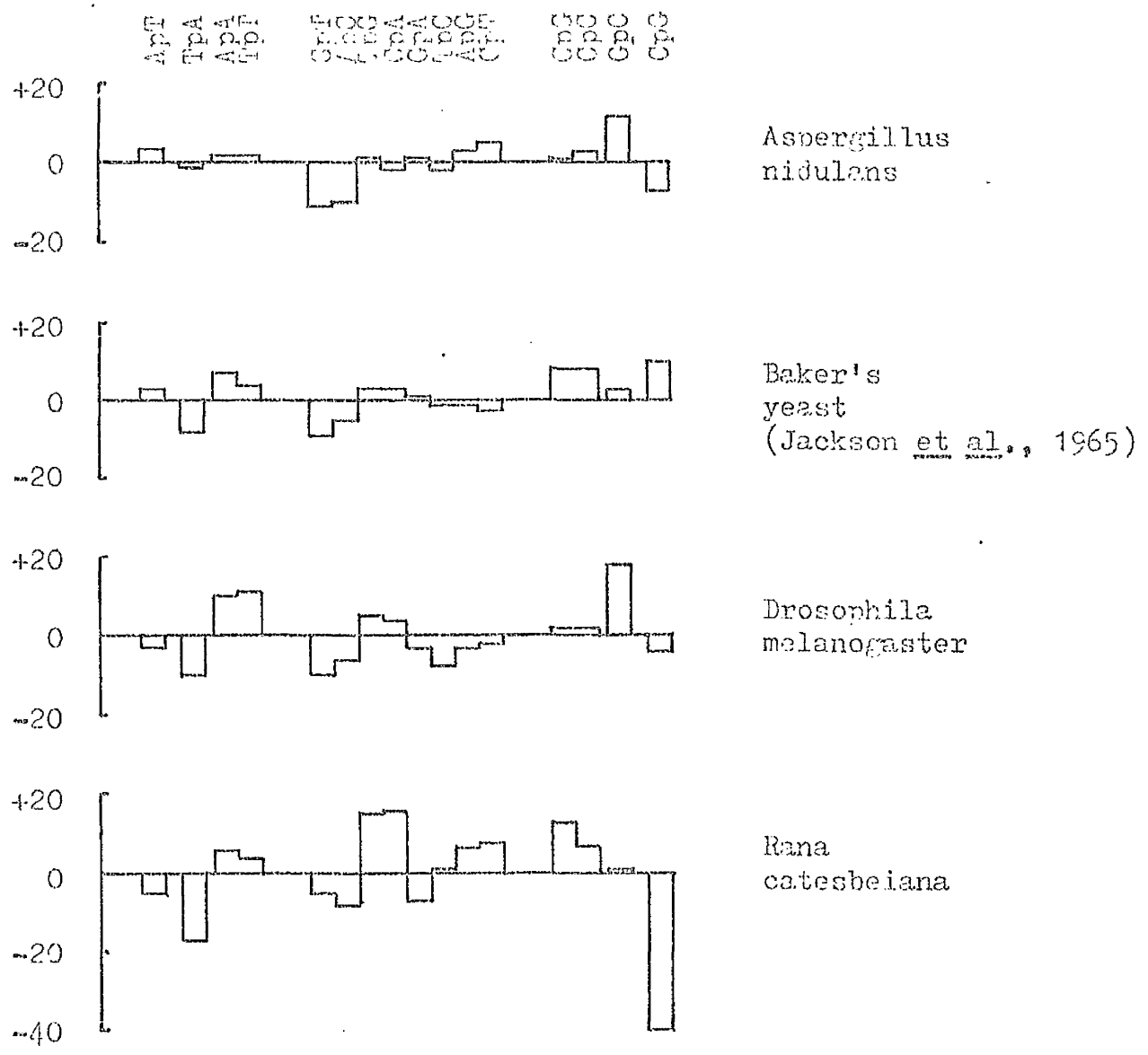
Frequencies are in parts per 10^3.

* G.Pontecorvo, personal communication.

‡ Hastings & Kirby (1966).

† Finamore & Volkin (1958).

FIGURE 13. AVERAGE FOUR BASE RUNS OF EUCARYOTIC DNAs



Frequencies are as parts per 10^{-3} deviation from random.

3.5 MOUSE SATELLITE AND MAIN BAND DNAs.

3.5.1 Introduction

Satellite DNAs are fractions of nuclear DNA which are found in most higher organisms. The work described here was performed on the light satellite of mouse DNA. This material forms about 10% of total mouse DNA and is distinguished from the bulk DNA by, first, its lower buoyant density (Kit, 1961) and, second, its high reassociation rate after denaturation (Waring & Britten, 1966). The kinetics of reassociation suggest that the satellite DNA is composed of repeats of similar sequences each containing 300-600 base pairs (Waring & Britten, 1966). The two strands of mouse satellite DNA can be separated by centrifugation in alkaline caesium chloride (Flamm, McCallum & Walker, 1967). This satellite DNA is also distinguished by its intracellular behaviour. In metaphase chromosomes it is situated near the centromeres (Jones, 1970). In polyoma virus infected cells the satellite DNA replicates earlier in the cell cycle than the bulk of the cellular DNA (Smith, 1970). However, it appears that in uninfected cells the satellite DNA may replicate later than most of the DNA (Helleiner & Cohen, 1970). Mouse satellite DNA is not transcribed *in vivo* (Flamm, Walker & McCallum, 1969).

This material is thus an extremely interesting fraction of mouse DNA, of unknown function. A nearest-neighbour analysis of mouse satellite DNA was undertaken to define its relation to main band DNA in terms of sequence structure, and with the thought that, since this DNA appeared to have a short repetition length, quite extensive insight into

its sequences might be obtained.

3.5.2 Results and Discussion.

The nearest-neighbour frequencies and derived base compositions of mouse satellite and main band DNAs are shown in Table 13 and Fig.14. These are the means of results obtained with two separate preparations of the DNAs. The main band DNA gives frequencies very similar to those of unfractionated mouse DNA (Fig.14), as might be expected of a fraction representing 90% of the total DNA. Beside having a lower (G + C) content than main band DNA, mouse satellite DNA also has a distinct nearest-neighbour deviation histogram i.e. the difference in base composition from main band DNA is not due to a uniform replacement of G and C by A and T. This implies that the sequences present in the satellite were not derived from a DNA of main band pattern by random mutations. Much of the difference in base composition can be ascribed to the high frequencies of ApA and TpT in satellite DNA. The CpG frequency of the satellite is twice that of main band DNA. This correlates with the results of Salomon, Kaye & Herzberg (1969), who found that the level of 5-methylcytosine in mouse satellite DNA was about twice the level found in the bulk DNA. As mentioned in Section 1.2.5, mammalian DNA appears to be methylated only in the sequence MeCpG. The satellite DNA is apparently not composed of very simple sequences, since all dinucleotides are present in appreciable amounts. The nearest-neighbour analysis of the satellite does not preclude the

possibility that the satellite sequences could have polypeptide-specifying potential.

The base sequence of guinea-pig α-satellite DNA has been examined by Southern (1970) who considers that the fundamental repeating unit is very short (perhaps a hexanucleotide) and that this basic sequence is much modified by accumulation of mutations. No sequence relationship was detectable between the guinea-pig α-satellite and mouse satellite. Southern (1970) also considers that mouse satellite DNA is derived from the repetition and mutation of a sequence of 8 to 13 base pairs, which includes the pyrimidine sequence T-T-T-T-T-C. The concept that mouse satellite DNA might be derived from a short sequence suggests a new method for examining the nearest-neighbour results:- classify the doublets into high and low frequency groups; consider the high frequency sequences as part of the basic sequence, and the low frequency members as resulting from accumulation of mutations. The finding of a high frequency species T-T-T-T-T-C by Southern (1970) is consistent with the present finding of very high ApA/TpT frequencies, and moderately high TpC/GpA frequencies. Southern (1970) obtained the pyrimidine hexanucleotide by diphenylamine - formic acid degradation of the DNA (Burton & Petersen, 1960). This implies that purine nucleosides should lie on both sides thus: Pu-T-T-T-T-T-C-Pu. The nearest-neighbour data indicate that the nucleoside on the 3'side is probably A: the mean frequency of the dinucleotide pair CpA/TpG is 3.7 times that of the alternative CpG, and is very close to the TpC frequency. Similarly, the more likely

candidate for the 5'-side is A. This gives a basic sequence containing A-T-T-T-T-T-C-A. In conclusion, the nearest-neighbour analysis is qualitatively consistent with the sequence studies, and indicates possible extensions of the basic sequence. From the nearest-neighbour data it is apparent that, if mouse satellite DNA is indeed composed of very short repeating sequences, then a large number of mutations must have accumulated.

The sequence studies of Southern (1970) on satellite DNAs differ from studies on RNAs in that they do not examine unique sequences, and are really a form of oligonucleotide frequency analysis. This suggests two ways of studying satellite sequences further. First, nearest-neighbour analysis of the separated strands of a satellite DNA should give a valuable extension to the pyrimidine run approach. Next, a single strand of satellite DNA could again be used as template for DNA polymerase, and the strand produced labelled using one $[\alpha^{32}P]$-dPuTP species at a time. Diphenylamine-formic acid degradation and fractionation of the pyrimidine runs would then extend the sequences found by Southern (1970) to give the predominant purine on the 3'-side, so allowing better integration of the data to give probable longer sequences. A similar method was used in this thesis to study other problems (Section 4.2).

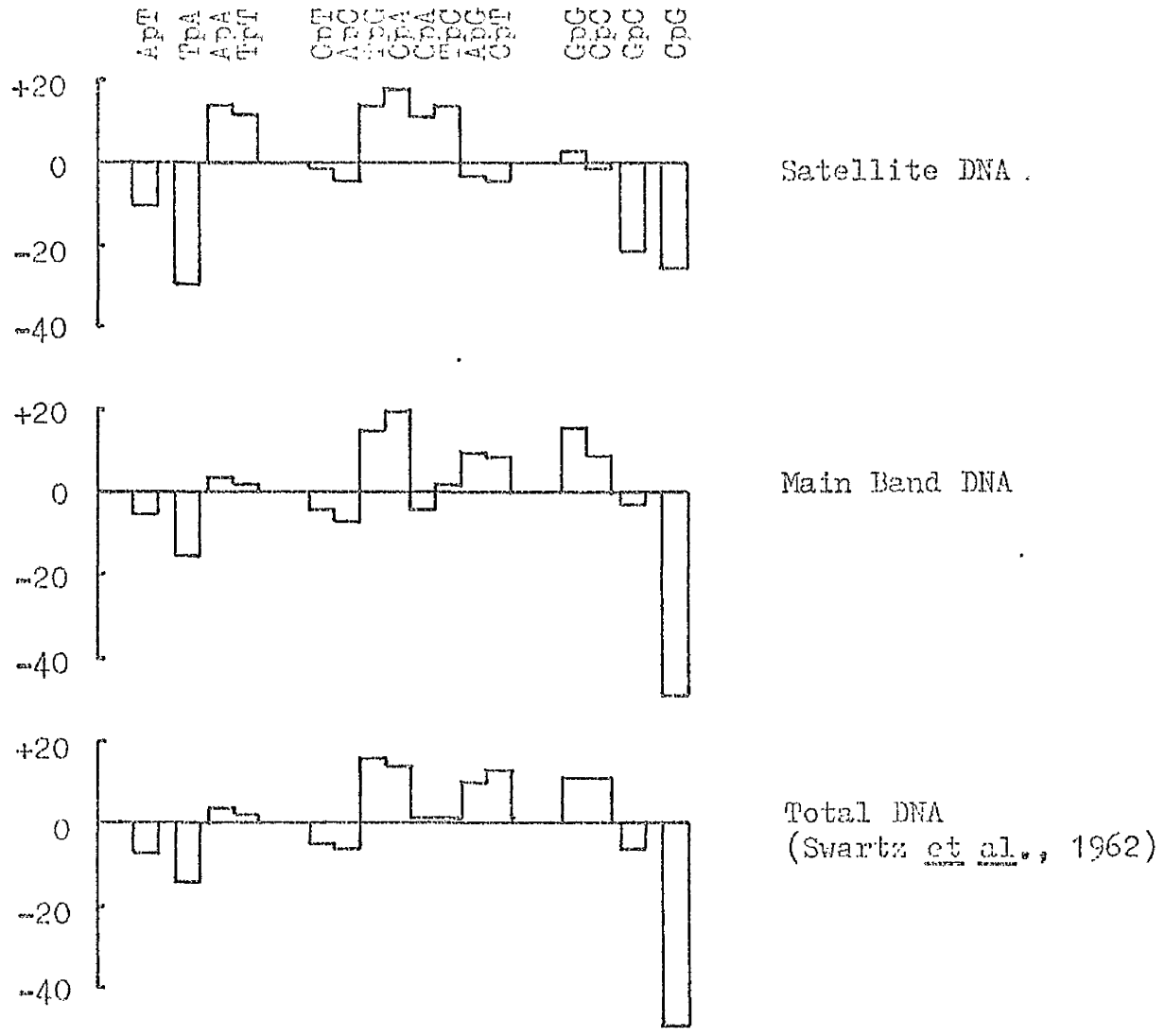
TABLE 13. NEAREST-NEIGHBOUR FREQUENCIES OF MOUSE DNA FRACTIONS

	Satellite DNA	Main Band DNA	Total DNA *
ApA TpT	119 133	83 91	88 93
CpA TpG	72 71	80 77	72 78
GpA TpC	64 74	54 66	61 63
CpT ApG	57 52	73 67	75 70
GpT ApC	57 53	57 54	57 54
GpG CpC	32 33	53 53	51 50
TpA	55	64	67
ApT	87	76	75
CpG	19	9	9
GpC	21	42	39
(G+C) %	35	42	41

Frequencies are as parts per 10^3.

* From Swartz et al. (1962).

FIGURE 14. [illegible] PATTERNS OF MOUSE DNA FRACTIONS



Frequencies are as parts per 10^3 deviation from random.

3.6 THE BASIS OF NEAREST-NEIGHBOUR PATTERNS

In this section some attempts are made at explaining nearest-neighbour patterns of DNAs in terms of the sequences present in the DNAs. As remarked in Section 1.2.5, it appears reasonable that the nearest-neighbour pattern of a DNA should depend on the codons used in the DNA, at least with genomes where most of the DNA is used to specify polypeptides. The frequency of occurrence of codons in a given DNA should depend on two factors. First, it must be related to the proportions of different amino acids in the proteins specified by the DNA. Second, it must depend on the relative usage of the members of sets of synonymous codons. The question of what forces might be involved in the evolutionary construction of such a codon set is not elaborated at length here. Selection might be for a given amino acid composition in proteins, or for a given total base composition of the DNA. Both of these trends could be driven by the need to maximise response to various physico-chemical demands of the organism's structure and functioning, or by the nutritional or metabolic availability of different anabolites; all these types of force could operate at once. In addition, if sequences other than codons (e.g. control sequences) were frequent enough and unusual enough in structure, then they too would contribute to the observed pattern.

Several attempts have been made to construct models to rationalise nearest-neighbour patterns using simple hypotheses.

Subak-Sharpe *et al.* (1966) calculated the nearest-neighbour pattern for a DNA having a random base sequence except that the codons TAA, TAG, CGN and NCG were forbidden: this pattern was shown to be similar to that of vertebrate DNA. Although similar calculations, banning TAA and TAG codons, appeared to give a good model for some bacterial DNAs, these were later found to be in error (H. Subak-Sharpe, personal communication) and there are no satisfactory models of bacterial DNA patterns based on this approach. The basic assumption of this method is that it takes sequences to be random except for special constraints. Another approach is to start with the relative amounts of different codons used by the DNA and calculate a nearest-neighbour pattern from these. The only way to obtain such a codon set at present is from the amino acid compositions of proteins. King & Jukes (1969) have compiled a set of relative frequencies of amino acids in vertebrate proteins, based on the known compositions of 53 proteins. Insufficient data for bacterial proteins is available to allow a similar approach in this case. An alternative source of data is amino acid composition determinations performed on bulk bacterial protein (Sueoka, 1961; Fitch, 1964). These values represent weight averages of the amino acid compositions of the proteins in the cell. However, results were constant under different growth conditions, which is some indication that they approximate the required number average. The sets of amino acid frequencies obtained by these methods are shown

in Table 14.

Some work on model nearest-neighbour patterns using these data has been published by Woese (1967). These results were used to deomonstrate that the amino acid compositions of proteins in organisms with extreme (G + C) contents are consistent with the known genetic code. By this method Woese (1967) showed that the nearest-neighbour pattern for M.lysodeikticus DNA could be closely approximated by a model pattern.

It has now proved possible to employ an even simpler approach with DNAs of intermediate (G + C) content. E.coli DNA, with a (G + C) content of 50%, was chosen for model-building. It was assumed that all codons for a given amino acid were used in equal amounts. Since the available data did not distinguish between glutamic acid and glutamine, and between aspartic acid and asparagine, it was assumed that both members of each pair occurred with equal frequency. The frequencies of intracodon dinucleotides were calculated on this basis and the final nearest-neighbour pattern for the model mRNA was obtained by allowing for the frequencies of intercodon dinucleotides, estimated as the products of the base frequencies for the third and first positions in codons. This assumed that no constraints existed on the ordering of codons. The nearest-neighbour pattern for a double-stranded version of this model is illustrated in Fig.15 as Model I. The pattern obtained is qualitatively similar to that determined experimentally for E.coli

DNA. One major discrepancy is the relatively high ApG/CpT content of the model. It is interesting that Weigert *et al.* (1966) found, by genetic methods, that in *E.coli* the codons AAA and GAA are preferred to their respective synonyms AAG and GAG. Incorporation of such a preference into the model would certainly lower the ApG/CpT frequencies. The nearest-neighbour pattern for such a modified model is shown in Fig.15 as Model II. Since the available amino acid frequency data did not distinguish between glutamic acid (codons GAG and GAA) and glutamine (codons CAG and CAA), the use of CAG was banned in this model, as well as GAG and AAG. This *ad hoc* elaboration of one aspect probably over-extends the model.

It is considered that, in view of the crudity of the available data and the simplicity of the assumptions, these models for *E.coli* DNA give remarkably close approximations to the experimentally determined nearest-neighbour patterns, and indicate that such a basis for the nearest-neighbour pattern's determination is adequate for this DNA.

These *E.coli* DNA models gave (G + C) contents of 48% and 46%. reasonably close to the 50% (G + C) value found in practice. Nearest-neighbour patterns similar to that of *E.coli*, as judged by their deviation histograms, are found in DNAs with (G + C) contents ranging from 34% to 57%. Two DNAs in this class have similar *specific* frequencies for corresponding doublets i.e. to transform one set of frequencies to the other, all doublet absolute frequencies

are changed in proportion to the overall base change.

These observations can be rationalised quite simply, as follows. Suppose that a DNA of E.coli pattern is descended from a DNA of (G + C) content 46-48%, also having this pattern. The base composition of the evolving DNA then changes over a period of time: this could be for a variety of reasons. It is supposed that this base composition change is applied uniformly over all codons, so that as the (G + C) content changes, the codon set of the DNA also changes. However, as long as the new codon set arrived at in this way allows the formation of adequately functional proteins, there should be no pressure to change the relative frequencies of the codons in this new set i.e. the specific frequencies of the doublets in the codons should also remain constant and the DNA should thus still show the E.coli pattern.

In this scheme the different deviation histogram patterns observed with DNAs of extreme (G + C) values then result from the changing of the codon set away from that which would be obtained by uniform base change to a given (G + C) value, since this latter codon set would not allow the proper functioning of proteins. In other words, as the base composition becomes extreme, base changes can no longer occur with equal frequency at all codon sites.

This idea of a nearest-neighbour pattern being determined by codon usage should also account adequately for the patterns found for A.nidulans and D.melanogaster DNAs, and also for the adenovirus DNAs. As discussed earlier, all these DNAs have intermediate (G + C)

contents with doublet patterns which are close to random and similar to the E.coli pattern in many respects.

Models for vertebrate DNA were made using the data of King & Jukes (1969) (Table 14) and the assumptions made for the first E.coli model. As shown in Fig.16 (Model III), allowing all codons to be represented gives a pattern quite distinct from the vertebrate pattern. Banning the CGN arginine codons gives little improvement (Model IV). However, if all CpG containing codons are banned, but CpG is still allowed as an intracodon dinucleotide, quite good duplication of the vertebrate pattern is obtained (Model V), although the CpG level is still rather high. The pattern obtained by Subak-Sharpe et al. (1966) by forbidding stop codons and CpG containing codons in a random sequence DNA is also shown in Fig.16. These models of vertebrate DNA are based on the assumption that most of the DNA specifies protein. As discussed earlier, this probably is applicable to bacterial and virus DNAs but may not be to vertebrate DNAs. Therefore, although these methods of model construction give reasonable agreement with the experimentally observed pattern, it should not be concluded that the pattern necessarily arises from this particular type of CpG restriction. What these models do indicate is, that the restriction of the CpG sequences is a predominant factor in determining the overall nearest-neighbour pattern in vertebrate type DNAs.

Fig.17 shows the single-strand version of Model V and compares it with patterns for the two strands of MVM DNA. It is evident that,

if MVM DNA is really a single-stranded version of the vertebrate type, then the model building is not good enough to predict accurately frequencies for single-stranded DNA.

All the model systems discussed above predict high A contents for the mRNAs (Table 15). As noted in Section 3.1.4, this is also a characteristic of the (-) strands for single-stranded DNA viruses. However, in the case of the bacterial viruses of this type, it appears that that viral mRNA is equivalent to the (+) strand of the DNA (e.g. Hayashi, Hayashi & Spiegelman, 1963). Nothing is known about the mRNA of parvoviruses. Table 15 also shows two conflicting values, from the literature, for the base composition of mammalian RNA fractions believed to be mRNA. These are not consistent and it is therefore not possible to use such data to test the models.

King & Jukes (1969), using the data shown in Table 14 for the amino acid contents of vertebrate proteins, calculated the expected base composition of mRNA for these proteins, on a basis similar to that used above. They then calculated the expected frequency of occurrence of codons for each amino acid in a random sequence chain of the base composition previously estimated. The observed frequency of each amino acid was then plotted against this estimated frequency, as shown in Fig.18. The points are clustered around the line of unit slope through the origin; this was interpreted as evidence for the existence of non-Darwinian evolutionary trends. However, the point of interest to the present discussion is the anomalous behaviour

of arginine in this plot: the observed frequency is much lower than the random prediction. This low arginine frequency was correlated with the low CpG frequency in vertebrate DNA, and it was suggested that perhaps "the amount of arginine that can be tolerated in animal proteins is less than the amount which would result from the occurrence of all six arginine codons at a random rate, so that the CpG content of animal DNA has been lowered by natural selection" (King & Jukes, 1969). If this argument were valid it would have profound implications for ideas on the origin and function of the CpG shortage.

However, there are two objections to this viewpoint. First, as shown by Model IV for vertebrate DNA, the banning of CpG containing codons for arginine does not necessitate a very low CpG frequency. Second, it can be shown that in bacteria, whose DNAs do not have low CpG frequencies, arginine also occurs at low levels. For instance, B.subtilis has about the same (G + C) content as mammalian DNA but does not show a CpG shortage. Using the relevant data of Table 14, a calculation similar to that of King & Jukes (1969) was made for B.subtilis DNA. It can be seen (Fig.18) that in this case also the observed arginine frequency is conspicuously lower than predicted. Inspection of the data in Table 14 indicates a similar situation with other bacteria. A low CpG frequency cannot, therefore, be regarded as a necessary concomitant of a low arginine frequency.

In summary, this section has shown, first, that the nearest-neighbour patterns of DNAs with near-random doublet frequencies can

be quite well accounted for in terms of the codons used in the DNA. Next, models for vertebrate DNA showed that the CpG shortage plays a large part in the determination of the total nearest-neighbour pattern. Last, it was shown that there is no real evidence for the CpG shortage in vertebrate or animal virus DNAs being correlated exclusively with low arginine levels in the proteins specified.

TABLE 14. AMINO ACID COMPOSITIONS OF PROTEINS IN DIFFERENT ORGANISMS

	Bacillus cereus	Bacillus subtilis	Vertebrate	Escherich. coli	Serratia marcescens	Micrococc. lysodeik.
(G+C) of DNA %	35	42	42	50	58	72
Ala	10.1 *	9.0	7.4	10.6	11.2	15.0
Arg	4.3	4.3	4.2	5.4	4.8	5.4
Asn	9.8	10.3	4.4	10.4	11.3	9.1
Asp			5.9			
Cys	0.3	0.2	3.3	0.4	0.4	-
Gln	12.4	11.9	3.7	11.1	10.3	12.6
Glu			5.8			
Gly	9.6	8.7	7.4	8.6	9.0	11.2
His	2.1	2.4	2.9	2.1	1.8	1.7
Ile	6.8	5.8	3.8	5.4	4.5	3.9
Leu	8.8	8.5	7.6	8.8	8.9	8.5
Lys	5.7	7.5	7.2	6.4	5.7	4.5
Met	1.9	2.2	1.8	2.9	3.6	0.8
Phe	3.8	3.9	4.0	3.5	3.8	2.4
Pro	3.7	4.6	5.0	4.1	3.8	4.7
Ser	4.3	4.7	8.1	4.6	5.1	4.0
Thr	5.7	5.5	6.2	5.7	5.1	5.9
Trp	-	-	1.3	-	-	-
Tyr	3.0	2.8	3.3	2.8	3.2	1.4
Val	7.7	7.7	6.8	7.2	7.2	8.5

* Percentage of total amino acids in protein.

Based on tables in Woese (1967) and King & Jukes (1969).

TABLE 15. BASE COMPOSITIONS OF mRNAs

	E. coli Models		Vertebrate Models			Experimental Results for Vertebrates	
	I	II	III	IV	V	A	B
A	29.1 *	30.9	27.9	29.1	29.9	31.0	26
U	22.9	22.9	22.6	22.4	22.9	19.9	28
G	26.5	24.5	26.4	26.6	24.8	22.8	21
C	21.2	21.2	23.2	22.0	22.5	25.4	24

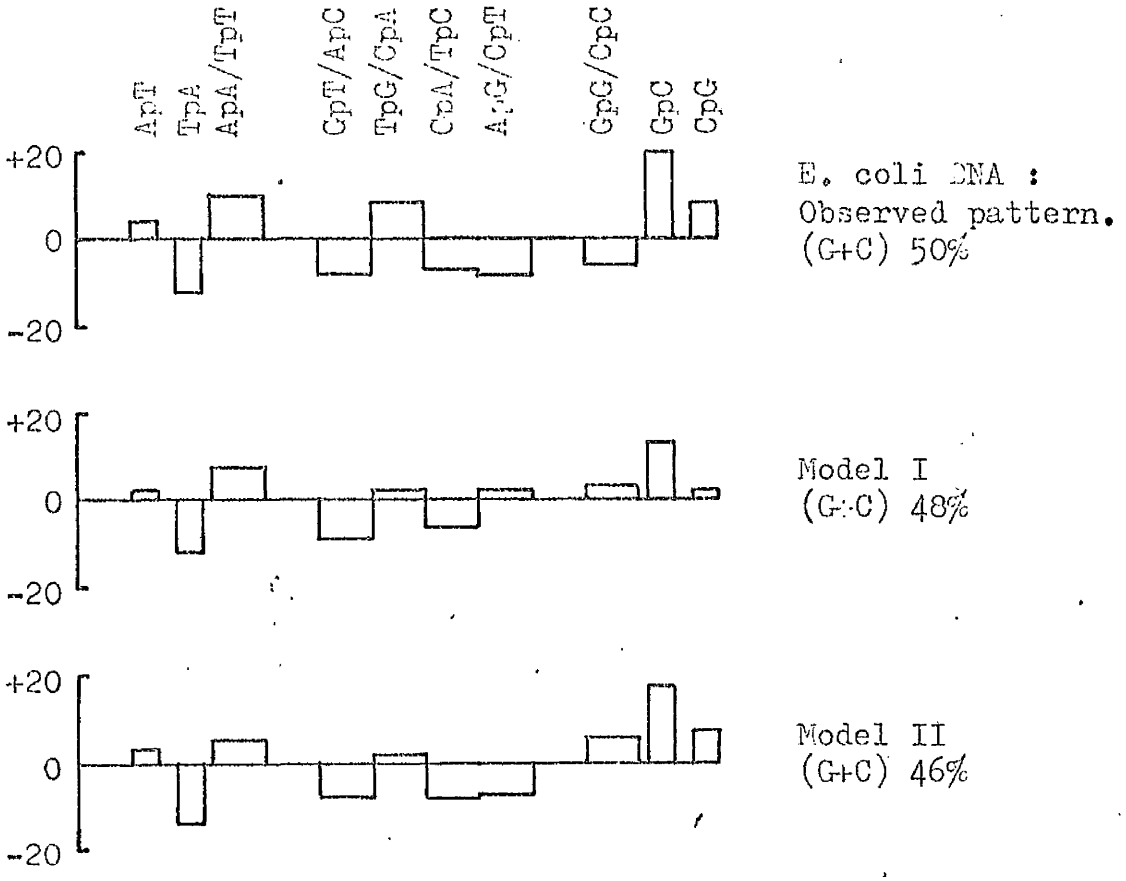
* Percentage of total bases.

Experimental results for vertebrates:-

A : Henshaw (1968).

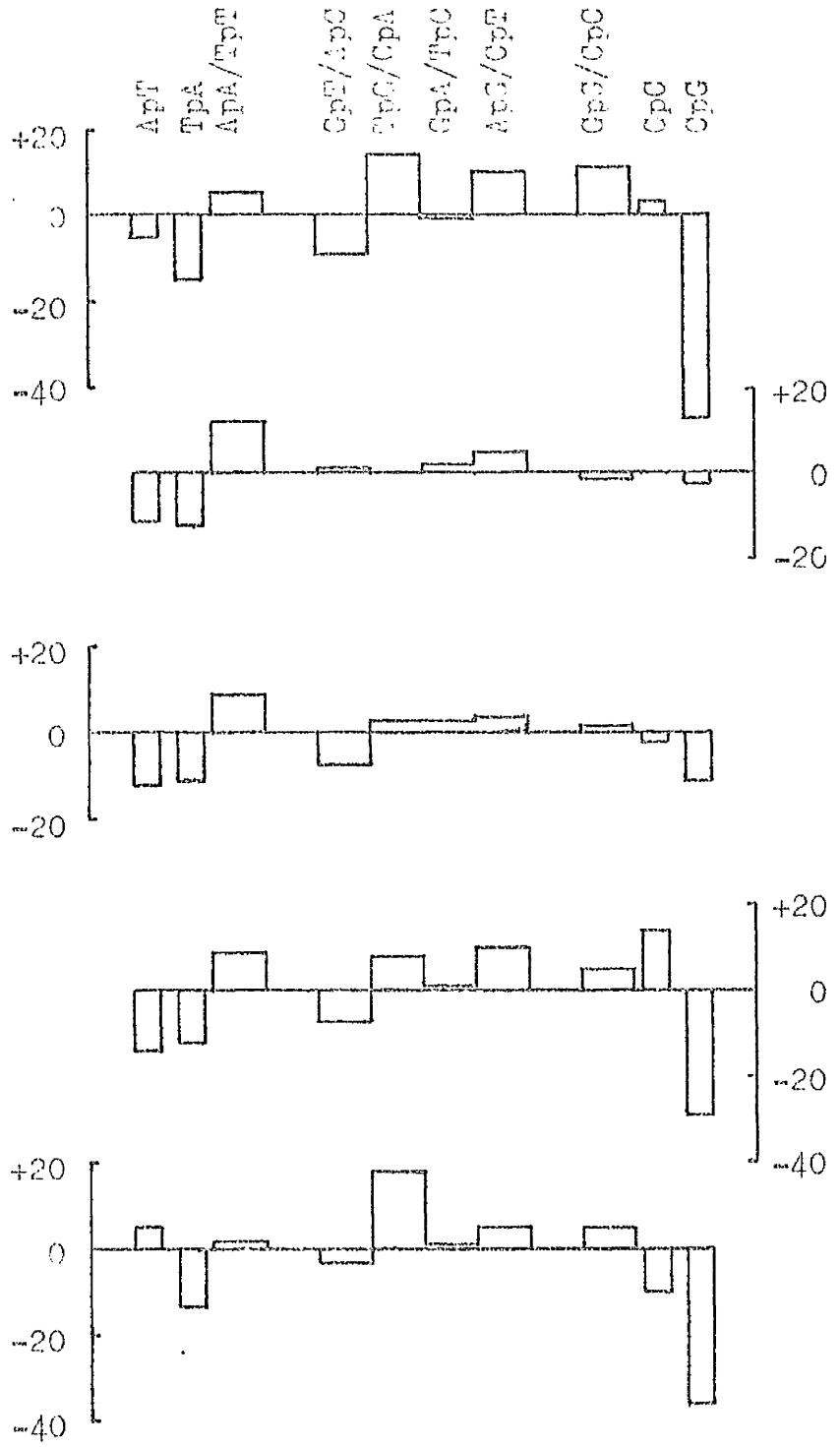
B : Darnell (1968).

FIGURE 15. MODELS OF THE E. COLI DNA NEAREST-NEIGHBOUR PATTERN



Frequencies are as parts per 10^3 deviation from random expectation. The data for the observed *E. coli* DNA pattern are from Josse *et al.* (1961), and the frequencies of complementary dinucleotides have been averaged.

FIGURE 16. MODEL OF THE V PROTEIN OF A NEAREST-NEIGHBOUR FREQUENCY



Human Spleen DNA: Observed pattern (Josse et al., 1961)

Model III : All codons allowed

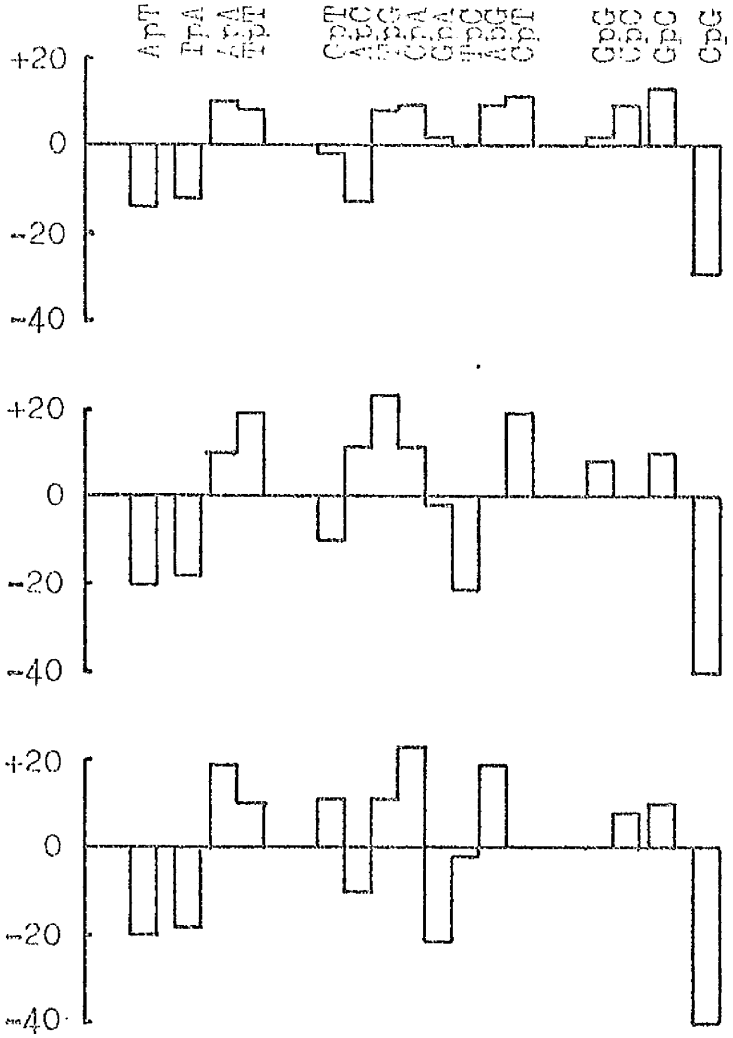
Model IV : No CGN codons

Model V : No CGN or NCG codons

Model by Subak-Sharpe et al. (1966)

Frequencies are as parts per 10^3 deviation from random. Frequencies of complementary dinucleotides are averaged.

FIGURE 17. COMPARISON OF VERTEBRATE DNA MODEL WITH MVH DNA NEAREST-NEIGHBOUR PATTERNS



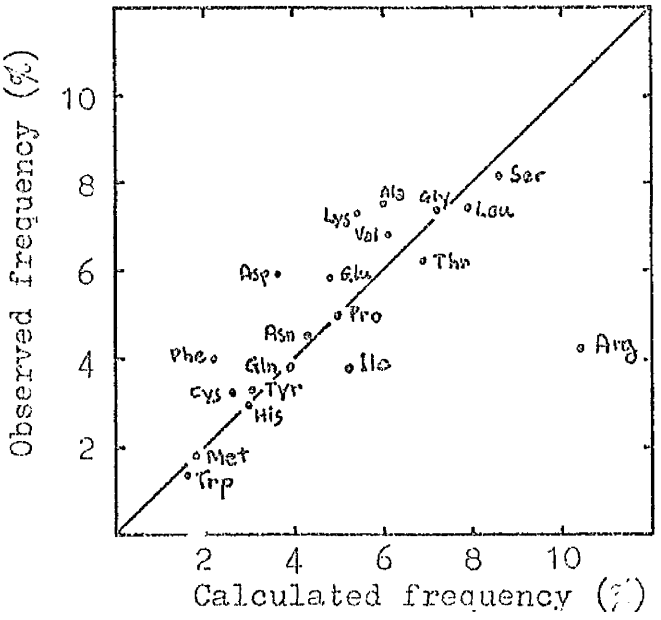
Vertebrate DNA : Model V, mRNA version.

MVH (-) strand DNA : experimental results.

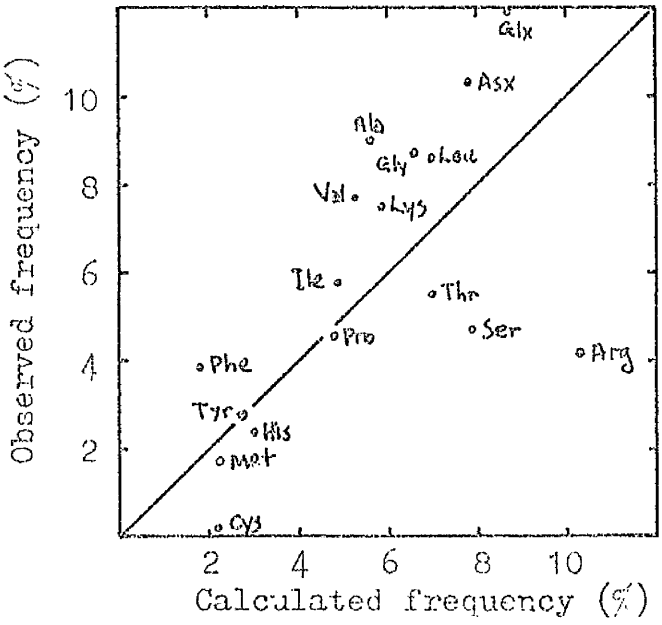
MVH (+) strand DNA : calculated from results for (-) strand.

Frequencies are as parts per 10^3 deviation from random.

FIGURE 18. AMINO ACID COMPOSITIONS OF VERTEBRATE AND BACTERIAL PROTEINS



Vertebrates : from King & Jukes (1969).



Bacillus subtilis

The observed frequency of occurrence of each amino acid in proteins of the organism is expressed as a percentage of the total amino acids in proteins, and is plotted against the frequency predicted from a random sequence model.

PART 4 : THE CpG SHORTAGE IN VERTEBRATE AND VIRUS DNAs

Page

Page

4.4 DISCUSSION OF RESULTS

4.1. DEFINITION OF THE PROBLEM AND EXPERIMENTAL APPROACHES

4.1.1 Introduction

As discussed in Sections 1.2 and 3.6, vertebrate DNAs and the nucleic acids of small, animal viruses contain very low levels of the dinucleotide CpG. Work described in Section 3.1 showed that this low CpG level is also found in single-stranded parvovirus DNAs. Although models presented in Section 3.6 gave some correlations between the CpG levels and other features of the nearest-neighbour patterns, these models gave no indication of *why* the CpG frequency should be low. A suggestion by King & Jukes (1969) that the low CpG level might be closely related to low arginine levels in vertebrate proteins was discounted. It is known that in mammalian DNA the C in CpG sequences is methylated while C in other sequences is not (Doskocil & Sorm, 1962). The recognition sites on the DNA for methylating enzymes most probably, therefore, contain the CpG doublet. However, the CpG sequences of polyoma virus are not methylated *in vivo* (Kaye & Winocour, 1967), but this may be due to some aspect of the virus function in the infected cell, rather than to the absence of potential methylation sequences.

It is conceivable that CpG sequences may play some role in control of expression of genes. Possibly, the CpG sequences might occur within only one or a few types of longer sequence; the CpG shortage could be the only clear sign of such sequence specificity detectable by nearest-neighbour analyses. It can be calculated that on *one strand* of polyoma DNA, which specifies at most ten

proteins, there are about 80-90 CpG dinucleotides i.e. there are perhaps 8 CpG sequences per strand per gene. This appears to be rather a high number for the sequences to have a direct and unique role in, for instance, start or stop signals for transcription or translation, unless several CpG sequences in tandem were involved.

4.1.2 General Plans of Attack

To study the low CpG phenomenon further, attempts were made to define the longer sequences within which the CpG sequences occurred. This was a difficult problem to approach experimentally, since it demanded the characterisation of a few particular kinds of sequences out of a large number. Further, because of the low levels of CpG, it was expected that the sequences of interest would be present as only a small proportion of the total.

If all tetranucleotides of the form X-C-G-Y, where X and Y are A,T,G or C, were present in equal amounts in vertebrate pattern DNA, then each species would contain 0.3 - 0.4% of all the bases in the DNA. Since it seemed possible that one or a few of the possible sequences might be preferred, while others were absent or present in much lower amounts, it was considered desirable to have an analytical method capable of detecting tetranucleotides, or their derivatives, containing perhaps less than 0.1% of the bases in the DNA under study.

One approach to the problem would be to make partial nuclease

digests of the DNA under study and then fractionate and identify nucleotides containing CpG sequences (or MeCpG sequences). This approach did not seem to offer the resolving power required. Another approach would be to equate the presence of 5-methylcytosine with the presence of the CpG sequence, and examine the distribution of 5-methylcytosine in fractions of DNA. Some work of this type has been reported by Doskocil & Sorm (1962), who concluded that the 5-methylcytosine could occur in sequences of the types Pu-MeC-G, C-MeC-G and T-MeC-G in double stranded DNA.

The general approach finally adopted was as follows. First, a low CpG DNA was used as a template for a polymerase *in vitro* to produce a radioactively labelled copy. In particular, much use was made of labelling with a specific $[\alpha\,^{32}P]$-nucleoside triphosphate. Specific degradation and fractionation methods were then applied to the labelled product to examine various aspects of the sequences. Degradation schemes producing oligonucleotide species with 3'-phosphate groups allowed use of the nearest-neighbour analysis technique of ^{32}P transfer between adjacent residues.

In double-stranded DNA, CpG sequences must be found in complementary positions on both strands. It was considered that this would probably lead to complications in interpretation of the results, so the single-stranded DNA from the parvovirus MVM was used as a model for vertebrate DNA. In this case the material produced *in vitro* should correspond to the (-) strand only. The

use of MVM DNA as a model for vertebrate DNA assumes that, because both exhibit low CpG frequences, the kinds of sequences within which CpG is found are similar in the two DNA types: this assumption may not be valid. Most experiments were therefore performed in parallel on MVM DNA and on calf thymus DNA.

It was not possible to devise a general scheme to examine all aspects of sequences containing CpG dinucleotides. Several approaches were necessary, with each approach examining one aspect of the problem. Briefly, a scheme using DNA polymerase with $[\alpha\,^{32}P]$-dGTP, and subsequent acid digestion of the product, was used to examine sequences, mainly of pyrimidines, to the 5'-side of CpG sequences. Purine sequences to the 5'-side of CpG were examined by a scheme using RNA polymerase with $[\alpha\,^{32}P]$-GTP, and digestion with pancreatic RNase. Nucleotides to the 3'-side of CpG were examined by producing RNAs *in vitro* labelled with each $[\alpha\,^{32}P]$-NTP species in turn, and then digesting with T_1 and U_2 RNases.

In the following description of the methods used and the results obtained, all descriptions of CpG sequences refer to the material produced *in vitro*: the relations of these findings to sequences in the template DNAs are then discussed in Section 4.4. The rationales of the different experimental approaches are now discussed in turn.

4.1.3 Pyrimidine Run Experiments

Treatment of DNA with diphenylamine and formic acid removes purine bases, degrades the exposed deoxyribose residues to which the purines were attached, and leaves oligonucleotide species of the general formula $Py_n p_{n+1}$ i.e. pyrimidine runs with both 3'- and 5'-phosphate groups (Burton & Petersen, 1960). Phosphate groups originally found between two purine nucleosides are liberated as inorganic phosphate. The reaction has been shown to be specific and complete, and no exchange of components can be detected (Burton & Petersen, 1960; Jones, Tittensor & Walker, 1966). This procedure was used as the basis of a method to examine pyrimidine sequences to the 5'-side of CpG groups.

The DNAs under study (calf thymus and MVM DNAs) were used as templates for E.coli DNA polymerase to produce, in vitro, DNA labelled with $[\alpha\ ^{32}P]$-dGTP. The synthesised DNA was isolated and digested to pyrimidine runs. Where the ^{32}P was originally in sequences of the type $Pup\overset{*}{p}G$ it should now be found as inorganic phosphate. In the cases where the ^{32}P occurred in sequences of the type $Py\overset{*}{p}G$, it should now be attached to the 3'-ends of various pyrimidine run species. By reasoning similar to that used in nearest-neighbour analysis (Section 1.2.2), the presence of ^{32}P on the end of a given pyrimidine isostich $(Py)_n$ indicates the presence in the DNA chain of the sequence $(Py)_n$ - G before degradation. If the mononucleotides at the 3'-ends of the various isostichs can be isolated with the 3'-phosphates attached, then the distribution of

$C\overset{*}{p}$ should indicate the kinds of sequences which occur to the 5'-side of CpG.

In early experiments, the pyrimidine runs were fractionated by two-dimensional paper chromatography. This indicated that the ^{32}P was distributed in a number of molecular species, but the separation achieved was not good enough for quantitative determinations. Later experiments therefore used anion exchange column chromatography to fractionate the pyrimidine runs first by length (i.e. into isostichs) and then by base composition.

The 3'-terminal nucleotides in runs containing both C and T were measured as follows. Although oligonucleotides bearing both 3'- and 5'- phosphates are resistant to attack by most nucleases, they can be digested with large amounts of micrococcal nuclease (Mikulski, Sulkowski, Stasiuk & Laskowski, 1969). Mixed C and T species were therefore digested with micrococcal nuclease, and then with spleen phosphodiesterase to ensure production of mononucleotides. The method used to separate the digestion products deserves some mention. It seemed desirable to have a method which would separate clearly any undigested material from dCMP and TMP, which rules out low pH electrophoresis. A system based on a method described by Jacobson (1964) was devised. The digest was spotted on to DEAE - paper which had been prewashed with formic acid and water and dried. The paper was developed first with 0.05M - formic acid. This elutes 3'-dCMP away from the

origin; all other phosphorylated species remain near the origin. The paper was dried and then developed in the same direction with 0.2 M-ammonium formate, which washed 3'- TMP away from any oligonucleotides. Once the solvent front reaches the dCMP, the dCMP moves at about the same rate as the TMP and the two species remain separated.

Further information can be obtained if [^{14}C]-dCTP is incorporated into DNA along with [α ^{32}P]- dGTP in this type of experiment. The [^{14}C]- label gives some information about pyrimidine run species other than those terminated at the 3'- end by G, and allows a check on the identity of fractions separated by base composition. This double labelling system was used with MVM DNA as template.

4.1.4 RNA Digestion Experiments

Two general approaches for examining sequences to the 3'- side of CpG dinucleotides were considered. First, the same techniques of acid digestion of DNA, as described above, could be employed if the two strands of a low CpG DNA were fractionated and each used separately as a template. This approach would examine sequences to the 5'- side of CpG of each strand: sequences to the 3'-side of CpG on one strand could then be deduced from the results for the other strand. For this approach it would be necessary to fractionate the strands of a suitable, defined, double-stranded DNA, such as polyoma DNA or SV40 DNA. At the time of starting this

work, attempts to isolate strands of these virus DNAs, by techniques used with bacteriophage DNAs, had proved unsuccessful (L.V.Crawford, personal communication). This approach was therefore dependent on a technical advance and has not been pursued. Recently, Westphal (1970) described a method for fractionating the strands of SV40 DNA. Some implications of Westphal's work will be discussed in Section 4.4.

The following approach was finally employed. MVM DNA was used as a template for E.coli RNA polymerase to produce RNA samples each labelled with one [α ^{32}P]-NTP species. T_1 RNase digestion of these labelled RNAs then gave a mixture of oligonucleotides with 3'-GMP as their 3'-end group. The digest was then fractionated by electrophoresis on DEAE-paper at pH 1.9 (Sanger et al., 1965). By this means the dinucleotide CpGp can be isolated directly. Any ^{32}P at the 3'-end of CpGp should then indicate the presence in the original RNA chain of the sequence G-C-G-N (where N was the [^{32}P]-labelled nucleotide). In addition, U_2 RNase was used to extend the analysis. This enzyme specifically cleaves RNA to the 3'-side of purine nucleoside 3'-phosphates, with a preference for cleavage next to adenine residues (Arima et al., 1968b). ^{32}P from the 3'-end of CpGp isolated from a U_2 RNase digest should indicate the occurrence of the species Pu-C-G-N. By subtraction, estimates for A-C-G-N should then be obtained. Since only small amounts of U_2 RNase were available and since the enzyme exhibits a preference for cleavage next to adenine residues, U_2 RNase digests were made

by first digesting with T_1 RNase and then with T_2 RNase.

This scheme was incomplete since no information on sequences of the type Py-C-G-N was obtainable. Also, no information about sequences further to the 3'-side of CpG could be obtained. The interpretation of results depended on the assumption that the RNA produced *in vitro* would contain sequences similar to those found in the DNA made using DNA polymerase. This could be checked at least in part by comparing nearest-neighbour analyses made with the two systems.

The production of RNA labelled with $[\alpha\ ^{32}P]$- GTP offered the possibility of obtaining information about the purine sequences to the 5'-side of CpGp i.e. information complementary to that given by the pyrimidine run experiments. RNA labelled with $[\alpha\ ^{32}P]$-GTP was digested with pancreatic RNase, and the digest fractionated into length classes, giving a series of oligonucleotides of general formula $(Pup)_nPyp$. Alkaline hydrolysis of each length class, separation of the resulting 3'-mononucleotides, and estimation of the ^{32}P in 3'-CMP in each case then gives a measure of the occurrence of $(Pu)_n$-C-G species.

The oligonucleotides can be separated into length fractions on DEAE-cellulose columns in 7M-urea (Tomlinson & Tener, 1963). However, a method of fractionating pancreatic RNase digests on DEAE - paper was developed. Several systems have been devised for length fractionation of oligonucleotides on DEAE-paper developed with a salt solution (e.g. see Laskowski (1967)), but the

resolution obtainable by such methods did not seem adequate. De Wachter (1968) described a method for applying a salt gradient to paper by which oligonucleotides up to 7 or 8 units long could be separated (de Wachter & Fiers, 1969). This system was tested but did not give reproducible results. Finally, it was found that adequate separation of pancreatic RNase digests into length classes up to hexanucleotide could be achieved by developing long (100 cm) strips of DEAE - paper with salt in 7M-urea, and this system was used.

It seemed possible also to examine the occurrence in the various fractions of the species $(Pu)_{n-1}$-G-C-G by digesting samples of each fraction with T_1 RNase and measuring the ^{32}P content of the 3'-CMP liberated. The 3'-CMP was isolated by chromatography on DEAE - paper with 0.05 M-formic acid (Jacobson, 1964).

The specificity of U_2 RNase allowed the design of an experiment completely analogous to the DNA acid digestion experiments, using the RNA system. A complete T_1 and U_2 RNase digest consists of oligonucleotide species of the type $(Py)_n$.Pu. These should be separable into length classes using the system described above for pancreatic RNase digests. If the RNA is labelled with [α ^{32}P]-GTP, the label should be found on the 5'-sides of terminal G residues: alkaline hydrolysis of fractions should then locate CpG sequences. A preliminary experiment showed that the fractionation procedure was feasible, but adequate data have not been obtained.

4.2 DNA ACID DIGESTION EXPERIMENTS

4.2.1 10-12 fractions could be isolated from acid digests of [α^{32}P]-dGTP labelled DNA preparations, using two dimensional paper chromatography. These experiments showed that the ^{32}P was present in a number of molecular species, as expected. However, the separations obtained, and the reproducibility of the system, were not good enough for detailed characterisation of different pyrimidine run species.

4.2.2 Preparation of DNAs for Major Experiments

The extent of DNA polymerase action on MVM DNA was estimated to be equivalent to 90-100% replication of the input template DNA. The course of the reaction is shown in Fig.19. It is of interest that in this experiment, and in several experiments with H-1 DNA, the reaction stopped after about 100% replication. The nearest-neighbour analysis of this material is shown in Table 16 in the 90-100% column: these figures agree quite closely with results for lesser extents of replication. An extent of copying of about 100% was chosen so that, hopefully, all sequences would be represented equally. Although the nearest-neighbour analysis agrees with previous work, this does not show that only the (-) strand is being produced. Also shown in Table 16 are the nearest-neighbour percentages expected for production of the (+) strand only, calculated on the assumption that the original,

TABLE 16. NEAREST-NEIGHBOUR ANALYSES OF [^{32}P]-G DNAs

	MVM-DNA			CALF THYMUS DNA	
	Extent of copying		(+) strand		
	20-30%	90-100%		A	B
3'-dAMP	32.4 *	33.3	34.5	33.2	33.7
3'-TMP	36.4	37.6	38.5	36.8	35.5
3'-dGMP	24.0	23.4	19.0	23.4	23.4
3'-dCMP	7.1	5.7	7,7	6.6	7.5

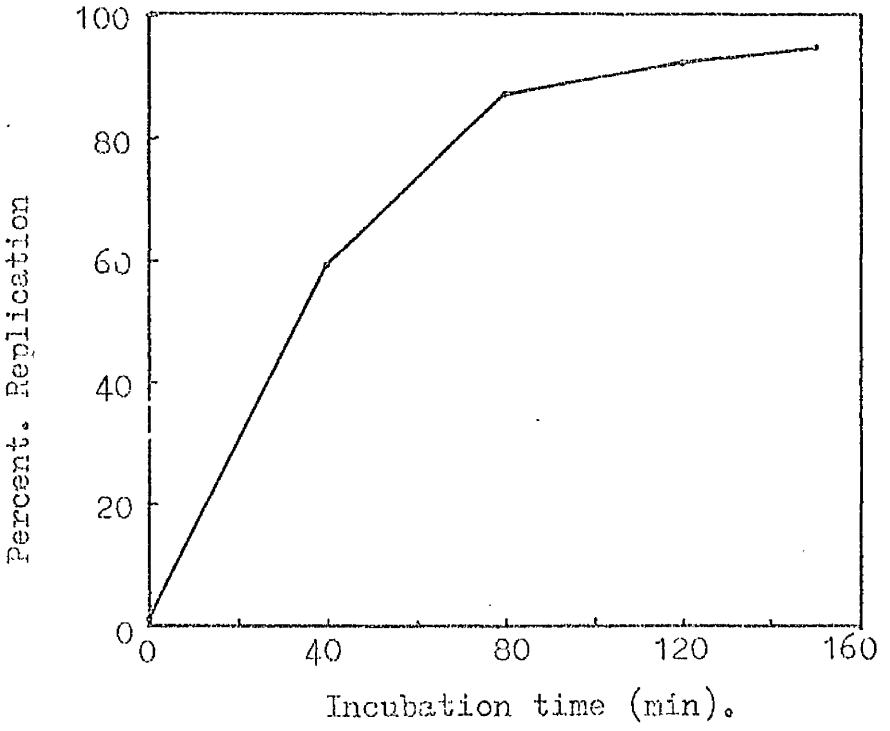
* Percentage of total ^{32}P (from [α^{32}P]-dGTP) associated with each 3'-mononucleotide.

Calf thymus DNA results :-

A : results of present study

B : from Josse et al. (1961).

FIGURE 19. PRODUCTION OF EVM DNA IN VITRO



Aliquots were withdrawn from the incubation mixture at 0, 40, 80, 120 and 150 min, and assayed for acid-insoluble ^{32}P. The amount of DNA synthesised is expressed as a percentage of the input template DNA.

limited-copy nearest-neighbour analysis was indeed made on (-) strands only. It is evident from the similarity of the results expected for both strands that a considerable proportion of (+) strand made *in vitro* could escape detection by these criteria. However, it does seem highly probable that most of the DNA made *in vitro* is (-) strand. The observation that the polymerisation stops after producing about 100% of the input DNA could also be taken to support this conclusion.

Nearest-neighbour analysis of DNA made on a calf thymus DNA template gave results shown in Table 16: these are in good agreement with the results of Josse *et al.* (1961).

4.2.3 Fractionation of Pyrimidine Runs by Length

The separations into isostichs are illustrated by Fig.20. With both experiments isostichs I to VIII were obtained, together with a final purged fraction, designated IX. In both cases the appearance of the u.v. peaks was closely similar to that found by Spencer *et al.* (1969). Except for the initial inorganic phosphate peak, the ^{32}P peaks coincided with the u.v. peaks. In the MVM DNA experiment, minor u.v. peaks, not containing ^{32}P, were found before isostichs I and II: these did not appear in the calf thymus DNA experiment, and possibly resulted from some degradation of pyrimidine runs during the isolation procedure. These peaks were not further investigated.

Over 99% of the ^{32}P in the MVM DNA digest applied to the column

was recovered. 58.3% of the ^{32}P recovered was found as inorganic phosphate. Since the DNA was labelled with $[\alpha\ ^{32}P]$-dGTP, the ^{32}P found as P_i should represent the radioactivity present before digestion as $Pup^{*}G$. The nearest-neighbour analysis gave a value of 56.7% of total radioactivity for PupG, agreeing well with that indicated by the acid digestion.

In the calf thymus DNA experiment, part of the P_i peak was lost and it was therefore not possible to apply directly the quantitation used above. Since recovery of ^{32}P from the column was complete in the MVM DNA experiment, it seemed reasonable to expect this with the calf thymus DNA digest also. 45% of the input ^{32}P was recovered in pyrimidine runs; the P_i peak should therefore contain 55%. This agrees well with the nearest-neighbour estimate of 56.6%, and so lends weight to the assumption of complete recovery.

The percentages of ^{32}P found in various fractions are shown in Table 17. It is evident that results for the two DNAs agree quite closely. Table 17 also shows the u.v. contents of the isostichs in the two experiments, expressed as mol pyrimidines per 100 mol bases of input DNA (i.e. the total for all isostichs is adjusted to 50.0). Both sets of results represent calf thymus DNA, and agree well with values published by Hall & Sinsheimer (1963) and by Spencer *et al.* (1969), also shown in Table 17. Some implications of the u.v. results are discussed in Section 4.4.

TABLE 17. PROPORTIONS OF ^{32}P AND U.V. IN ISOSTICHS

Isostich	Percent of total ^{32}P		u.v. absorbance : mol Py per 100 mol bases *			
	MVM	Calf thymus	A	B	C	D
Pi	58.28	(55)	-	-	-	-
I	19.30	22.17	11.5	11.7	10.6	11.2
II	10.93	10.63	10.9	11.0	10.0	10.5
III	5.70	5.14	7.3	8.2	7.7	8.3
IV	2.29	3.01	5.8	6.5	6.4	6.1
V	2.14	1.82	4.4	4.7	4.1	4.1
VI	0.64	1.13	2.8	3.1	3.4	3.2
VII	0.22	0.56	1.8	1.8	2.4	2.1
VIII	0.33	0.26	1.6	1.1	1.6	1.5
"IX"	0.19	0.30	2.6	1.9	3.0	3.0

* All u.v. results represent calf thymus DNA. The sources of these data are :-

A : Experiment with [^{32}P]-labelled MVM DNA.

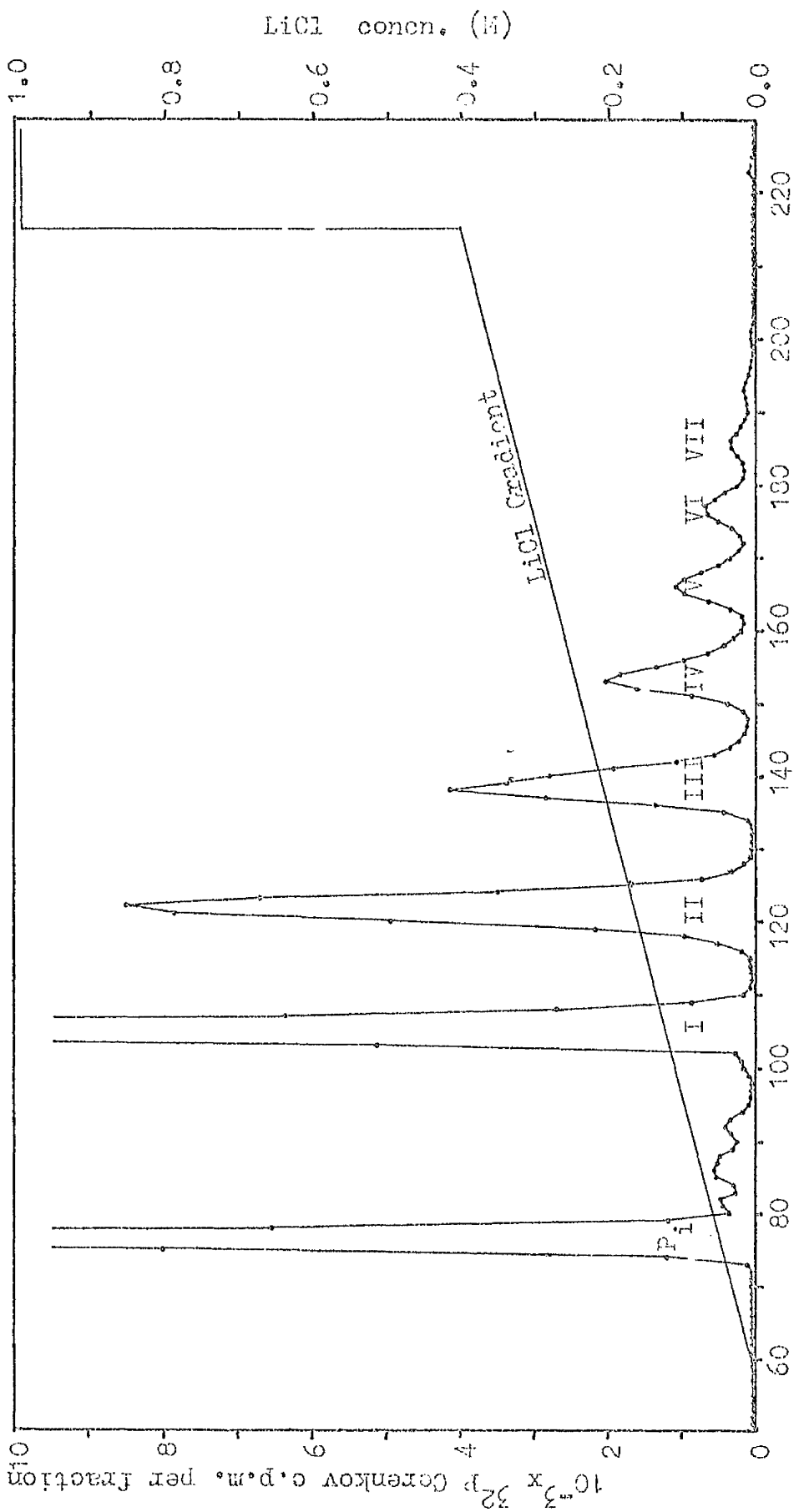
B : Experiment with [^{32}P]-labelled calf thymus DNA.

C : Hall & Sinsheimer (1963).

D : Spencer *et al.* (1968).

C and D are included for comparison.

FIGURE 20. FRACTIONATION OF PYRIMIDINE RUNS INTO ISOSTICHS



The acid-digest of DNA was loaded on to a DEAE-cellulose column at pH 5.3. Purine bases were first removed by washing with buffer, and the pyrimidine runs were then eluted with a salt gradient: only this latter part of the procedure is shown here. The figure shows results for a calf thymus DNA digest; similar results were obtained for MVM DNA. The u.v. peaks (not shown) coincided with the ^{32}P peaks.

4.2.4 Fractionation of Isostichs by Base Composition

In both experiments, isostichs I to IV were fractionated by base composition as shown in Figs.21 and 22. In all cases the ^{32}P peaks coincided with u.v. peaks. The percentages of ^{32}P and u.v. absorbing material in the various fractions are shown in Table 18. Peaks were identified as follows. First, the positions and numbers of peaks in the salt gradient agreed with the results of Cerny et al. (1968). Next, the u.v. spectra of the peaks were measured at pH 3.0. The ratio of absorbance at 290 nm to that at 267.5 nm was calculated. 267.5 nm is the isosbestic wavelength for T and C at pH 3.0 (Cerny et al., 1968); at 290 nm the molar absorbance of C is much higher than that of T. The ratio therefore gives a measure of the proportion of C in each fraction. In a similar way the $^{14}C/^{32}P$ ratios of each fraction, in the MVM DNA experiment, gave some confirmation of the other methods: the ^{14}C value represents the total cytosine content of the fraction, while the ^{32}P represents, not actually the total frequency of occurrence of the fraction, but the frequency of occurrence of runs in the fraction which ended with G.

The first peak to be eluted in the gradient using this system has the formula C_nP_{n+1} where n is the length of the isostich under study. Since this fraction contains only C, any ^{32}P found must represent CpG. All runs of type C_nP_{n+1}, for n equals 1 to 4, did contain ^{32}P, but only pCp contained large amounts: in the other

TABLE 18. PROPORTIONS OF ^{32}P AND U.V. IN FRACTIONS SEPARATED BY BASE COMPOSITION

Fraction	Percent of total ^{32}P in isostich		Percent of total u.v. absorbance in isostich *		
	MVM	Calf thymus	A	B	C
C	22.9	16.3	42.9	39.3	41
T	77.1	83.8	57.1	60.7	59
C_2	6.4	5.1	19.4	15.9	19
CT	55.2	57.6	50.0	56.6	51
T_2	38.4	37.4	30.6	27.5	30
C_3	2.3	1.8	14.5	9.0	16
C_2T	25.1	29.9	29.9	34.9	33
CT_2	37.3	44.0	31.3	37.6	31
T_3	35.4	24,3	24.5	18.5	21
C_4	1.0	2.1	-	7.8	7
C_3T	7.5	16.9	-	22.5	20
C_2T_2	37.3	33.7	-	30.8	33
CT_3	38.8	33.6	-	26.4	27
T_4	15.4	13.7	-	12.5	12

Results are expressed as percentages of the material recovered from the column in the gradient i.e excluding material eluted in the prewashing procedure.

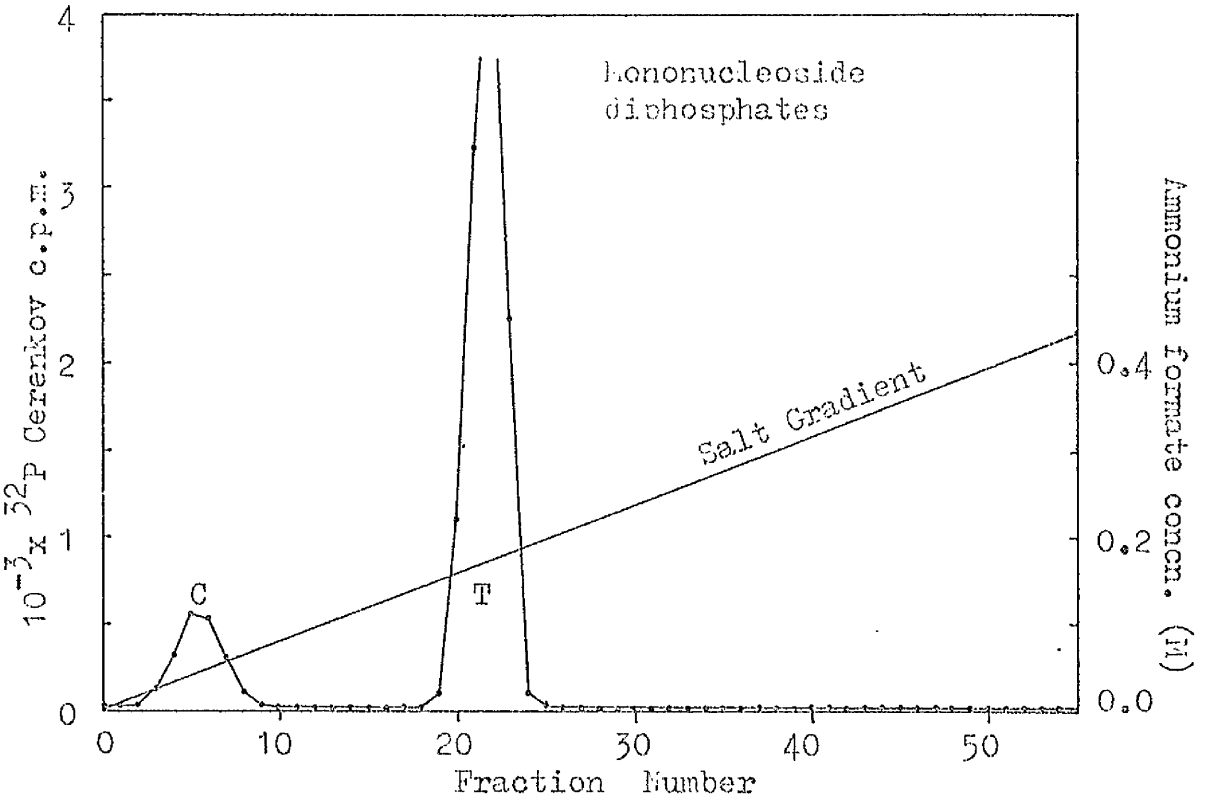
* All u.v. results represent calf thymus DNA. The sources of these data are :-

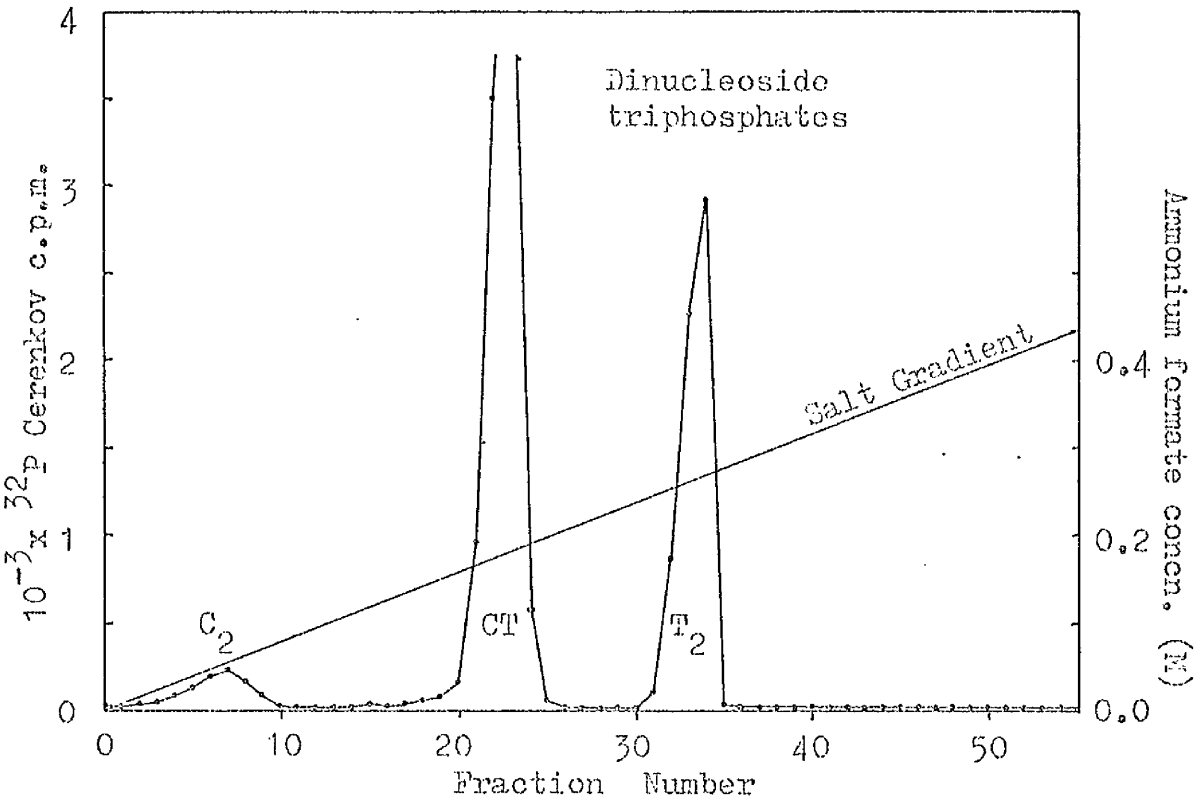
A : Experiment with [^{32}P]-labelled MVM DNA.

B : Experiment with [^{32}P]-labelled calf thymus DNA.

C : Hall & Sinsheimer (1963).

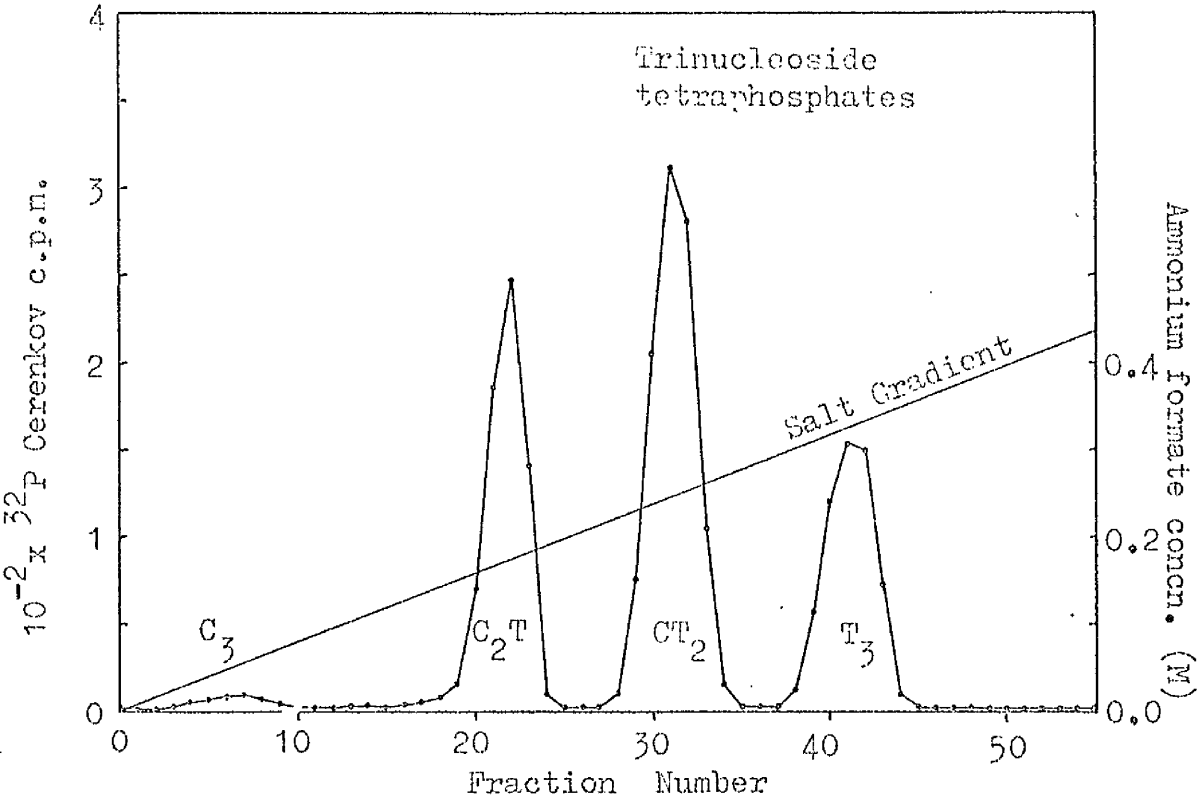
FIGURE 21. FRACTIONATION OF ISOMERICS BY BASE COMPOSITION

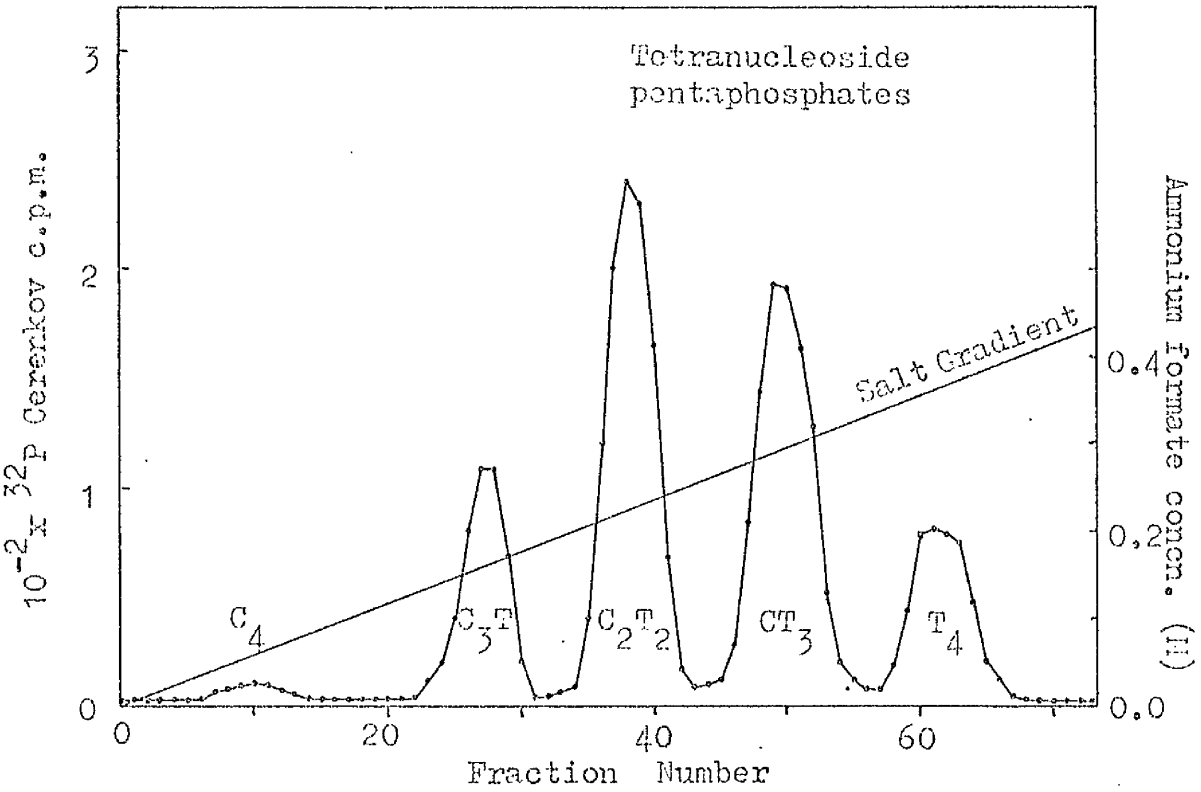




These results are for calf thymus DNA; MVM DNA gave similar results.

FIGURE 22. FRACTIONATION OF ISOMERS BY BASE COMPOSITION





See also Fig. 21.

cases the amounts were very low (Table 18). It was therefore of particular concern that the prewashing of the column with 0.1 M-formic acid eluted 4-8% of the total applied ^{32}P: since C_nP_{n+1} was the first peak eluted after starting the gradient, and since levels of ^{32}P in these fractions were low, it seemed possible that this easily eluted material was being lost preferentially in the prewash.

This possibility was excluded by the following considerations. First, the ratio of the absorbance at 290 nm to that at 267.5 nm, and the ratio of ^{14}C to ^{32}P contents, of the prewash fractions indicated the presence of both C and T. Second, the proportions of u.v. absorbing material in various fractions, particularly the all cytosine fractions, obtained by this method agreed well with those found by Hall & Sinsheimer (1963) using other methods. It is probable that the prewash fractions represent some degraded pyrimidine run material.

4.2.5 Determination of 3'-end Groups

The results of 3'-end group estimates for the separated isostichs from MVM DNA are shown in Table 19. All fractions examined contained between 5 and 16% of their ^{32}P in 3'-dCMP. The system gave satisfactory digestion and separation, although in a few cases undigested material was detected. In some cases ^{14}C was detectable in the forward area where deoxycytidine runs, indicating some dephosphorylation. Inorganic phosphate runs with TMP in this system, so any dephosphorylation will

TABLE 19. DETERMINATION OF 3'-END GROUPS IN MVM DNA PYRIMIDINE FRACTIONS

	CTp_3		C_2Tp_4		CT_2p_4		$C_2T_2p_5$		CT_3p_5	
	^{14}C %	^{32}P %	^{14}C %	^{32}P %	^{14}C %	^{32}P %	^{14}C %	^{32}P %	^{14}C %	^{32}P %
Undigested Material, + pCp & pTp	68	6	36	10	42	4	40	13	41	14
3'-TMP	0	86	0	78	0	76	1	68	2	76
Area between TMP & dCMP	0	2	1	3	0	5	1	3	1	3
3'-dCMP	30	6	53	9	53	14	53	13	48	5
Area forward of dCMP	2	0	10	0	4	1	5	2	8	2

These results are for fractions from MVM DNA (-) strand labelled with [^{14}C]-dCTP and [α^{32}P]-dCTP. Each fraction was digested with nuclease and chromatographed on DEAE-paper; the portions of the paper indicated were then cut out and their ^{14}C and ^{32}P contents determined.

cause the over-estimation of 3'-TMP end groups at the expense of 3'-dCMP. The mononucleoside diphosphate fraction should contain any [^{14}C]-3',5'-dCDP from the 5'-ends of runs. Duplicate determinations on several fractions agreed well.

In the calf thymus DNA experiment, satisfactory results were obtained only for the CTp$_3$ fraction, where 84% of the ^{32}P was found in TMP and 16% in dCMP. The amounts of ^{32}P in other fractions were too small to obtain good estimates of ^{32}P in dCMP.

4.2.6 Conclusions

Table 20 lists the frequency of occurrence of sequences ending with -C-G and with -T-G. It is evident that in both MVM (-) strand DNA and in calf thymus DNA the CpG sequences can occur in a variety of positions with respect to sequences found on the 5'-side. The immediate 5'-neighbour can be Pu or C or T. This agrees qualitatively with the conclusions of Doskocil & Sorm (1962) for the distribution of 5-methylcytosine in calf thymus DNA. Table 20 shows that the amounts of CpG and TpG found in various fractions add up to values close to the totals indicated by nearest-neighbour analysis. Table 21 indicates the relative amounts of CpG found as Pu-C-G, C-C-G, T-C-G and unassigned Py-C-G. TpG is treated similarly.

If it is assumed that the DNA made *in vitro* from MVM DNA template is completely representative of (-) strand then the number of copies of each sequence per (-) strand DNA molecule can be calculated, using base composition values obtained from nearest-neighbour analysis and

a (+) strand molecular weight of 1.7×10^6 (Crawford et al., 1969). These estimates are also shown in Table 20. All the values are greater than unity except for the hexanucleotide Pu-C-C-C-C-G. In a random chain about 5000 units long a given hexanucleotide might, on a random basis, be expected to occur 1-2 times, so the values fall into the expected order of magnitude. The MVM DNA molecule is expected, from its size, to contain about ten genes i.e. there are, per gene, five copies of Pu-C-G sequences and one each of C-C-G and T-C-G in the (-) strand.

From the $^{14}C/^{32}P$ ratios in the fractions separated by base composition, the total frequency of occurrence of cytosine-containing fractions in MVM DNA was estimated. By subtraction this also gave estimates for cytosine-containing runs ended with A. The estimates obtained in this way are shown in Table 22. Other information on sequences deduced from the nuclease digest results is shown in Table 23. These estimates may not be very accurate because of the indirect methods used to obtain them.

In retrospect, some improvements might be made in the detail of the experiments described above. First, although the micrococcal nuclease system worked well, it did use large quantities of enzyme, and this might give rise to artifacts. Richards & Laskowski (1969) have reported that, at low pH, oligonucleotides with 3'-phosphate end groups are relatively easily hydrolysed by snake venom phosphodiesterase. This might provide an alternative method of analysis. In this case the 3'-end groups would appear as mononucleoside 3', 5'-diphosphates.

Secondly, the use of ion exchange colums for separation was time-consuming although accurate. Southern (1970) has recently published a method for the separation of fully phosphorylated pyrimidine runs by two dimensional electrophoresis on a gel and on DEAE-paper: this would seem to be well suited to investigations of the type described here.

TABLE 20. FREQUENCIES OF CpG AND TpG CONTAINING SPECIES IN MVM AND CALF THYMUS DNAs

Species containing CpG	MVM (-) Strand % of G	MVM (-) Strand Times/ DNA Strand	Calf Thymus % of G	Species containing TpG	MVM (-) Strand % of G	MVM (-) Strand Times/ DNA Strand	Calf Thymus % of G
Pu-C-G	4.43	52	3.61	Pu-T-G	14.87	176	18.56
Pu-C-C-G	0.70	8	0.54	Pu-C-T-G	5.64	67	5.13
Pu-T-C-G	0.39	6	1.00	Pu-T-T-G	4.20	50	3.97
Pu-C-C-C-G	0.13	2	0.09	Pu-C-C-T-G	1.28	15	-
Pu-(C,T)-C-G	0.15	2	-	Pu-(C,T)-T-G	1.79	21	-
Pu-T-T-C-G	0.34	4	-	Pu-T-T-T-G	2.02	24	1.25
Pu-C-C-C-C-G	0.02	0.3	0.06	Pu-(C_2,T)-T-G	0.73	9	-
Pu-(C,T_2)-C-G	0.13	2	-	Pu-(C,T_2)-T-G	0.84	10	-
Pu-T-T-T-C-G	0.05	1	-	Pu-T-T-T-T-G	0.35	4	0.41
Sum of Frequencies of above Species	6.34		4.30	Sum of Frequencies of above Species	31.72		29.32
Total CpG *	5.7-7.1		6.6-7.5	Total TpG *	36.4-37.6		35.5-36.8

Frequencies of species in both DNAs are expressed as percentages of the total ^{32}P incorporated into DNA from [$\alpha^{32}P$]-dGTP. For MVM DNA, frequencies are also shown as estimated number of occurrences per DNA (-) strand.

* Total CpG and TpG estimates are from nearest-neighbour analyses. See Table 16.

TABLE 21. RELATIVE FREQUENCIES IN MVM AND CALF THYMUS DNAs OF THE 5'-NEIGHBOURS OF CpG AND TpG SEQUENCES

Species	MVM (-) Strand %	Calf Thymus %	Species	MVM (-) Strand %	Calf Thymus %
Pu-C-G	63	51	Pu-T-G	41	51
C-C-G	12	10	C-T-G	19	14
T-C-G	11	14	T-T-G	18	16
Py-C-G *	15	25	Py-T-G *	22	19

* Py-C-G and Py-T-G represent the differences between the summed frequencies of fractions examined and the values for total CpG and TpG, respectively, obtained by nearest-neighbour analysis. These estimates are thus dependent on the accuracy of the nearest-neighbour analysis.

---ooOoo---

TABLE 22. FREQUENCIES IN MVM DNA OF PYRIMIDINE SEQUENCES WITH A AS 3'-NEIGHBOUR

Species	Times/DNA	Species	Times/DNA
Pu-C-A	110	Pu-(C,T)-A	39
Pu-C-C-A	52	Pu-(C_2,T)-A	12
Pu-C-C-C-A	3	Pu-(C,T_2)-A	16
Pu-C-C-C-C-A	1	Pu-(C_3,T)-A	2
		Pu-(C_2,T_2)-A	4
		Pu-(C,T_3)-A	2

These data are for MVM DNA (-) strand, and were deduced from the $^{14}C/^{32}P$ data for each fraction.

---ooOoo---

TABLE 23. ADDITIONAL FREQUENCY INFORMATION ON MVI DNA

Species	Times/DNA
Pu-T-C-A	48
Pu-C-T-A	62
Pu-T-C-C-Pu	6
Pu-C-T-T-Pu	17
Pu-C-T-T-T-Pu	5

These data are for MVM DNA (-) strand, and were deduced from the $^{14}C/^{32}P$ nuclease digest results.

4.3 RNA DIGESTION EXPERIMENTS

4.3.1 Preparation of RNAs

Sinsheimer & Lawrence (1964) and Chamberlin & Berg (1964) showed that the single-stranded DNA of phage ϕX174 could act *in vitro* as a template for *E.coli* RNA polymerase. It has now been found that the single-stranded parvovirus DNAs are also active as templates. In different experiments the RNA produced *in vitro* was equivalent to between 10 and 30% of the template DNA. Nearest-neighbour analyses were made of samples of the [^{32}P]-RNAs: these results have already been discussed in Section 3.1.3. Calf thymus DNA gave closely similar nearest-neighbour patterns with the DNA and RNA polymerase methods, but the results for the DNAs of parvoviruses MVM and H-1 showed differences between the methods. However the RNA preparations made on parvovirus DNA templates did exhibit low CpG levels, and were used for studies on the CpG phenomenon.

4.3.2 Electrophoresis at low pH

The separation obtained with T_1RNase digests using the electrophoretic system of Sanger *et al.*(1965) is illustrated in Figures 23 and 24. Species of interest were identified by comparison with the results of Sanger *et al.* (1965) and, with [^{32}P]-labelled digests, by alkaline digestion of material eluted from the paper. In addition, the displacements of species from the origin were expressed as percentages of the distance moved by 3'-GMP, which could readily be

TABLE 24. SEPARATION OF OLIGONUCLEOTIDES BY ELECTROPHORESIS AT LOW pH

Fractionation of T_1 RNase Digests

Compound	Mobility *
ApGp	69
CpCpGp	77
CpGp	83
Gp	100

Fractionation of T_1 and U_2 RNase Digests

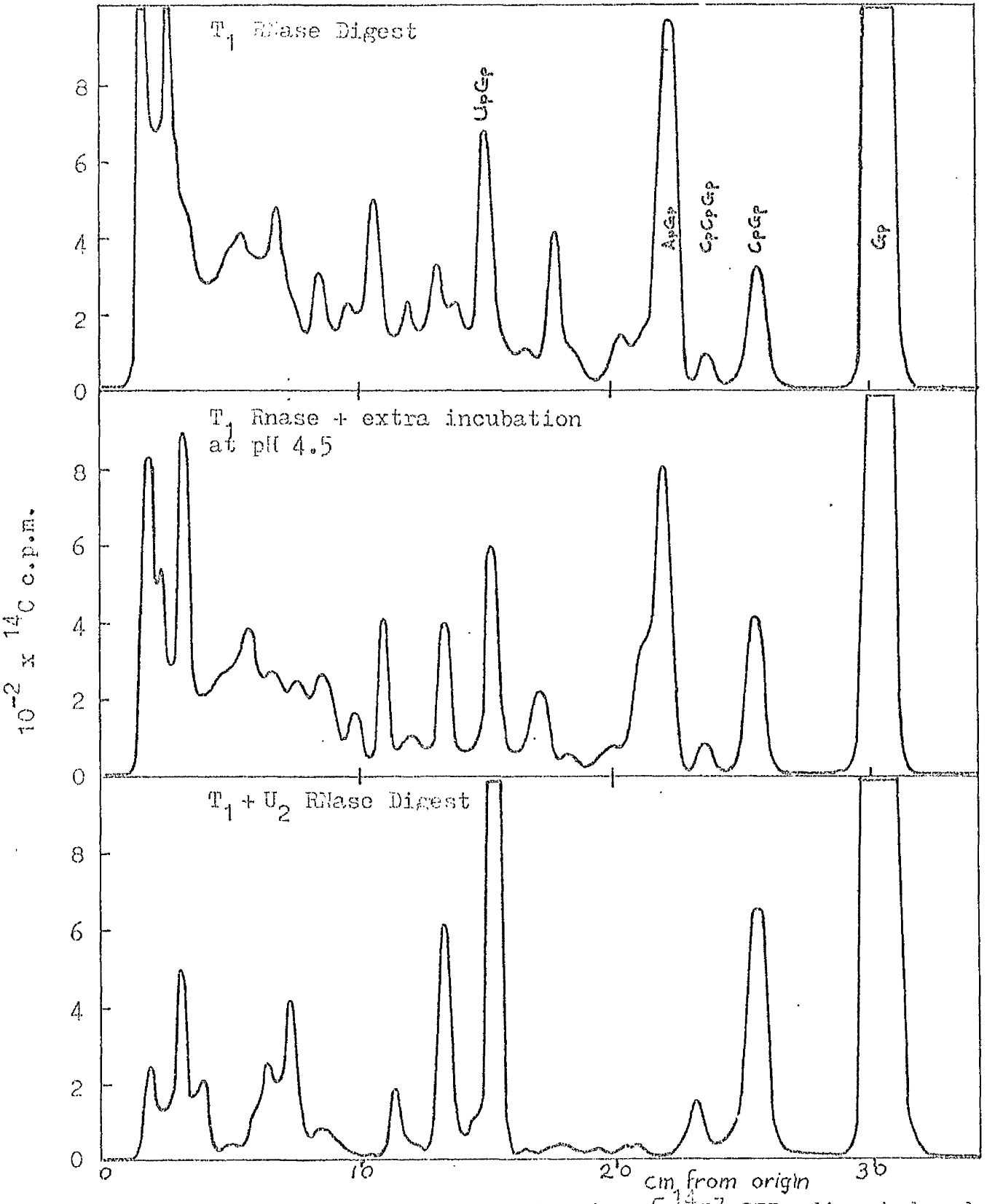
Compound	Mobility *
CpCpGp	77
CpGp	83
Unknown ‡	87
Unknown ‡	91
Gp	100
CpAp	107
Ap	113

Nuclease digests of RNA were fractioned by electrophoresis on DEAE-paper at pH 1.9.

* The distance from the origin moved by each species is expressed as a percentage of the distance moved by 3'-GMP.

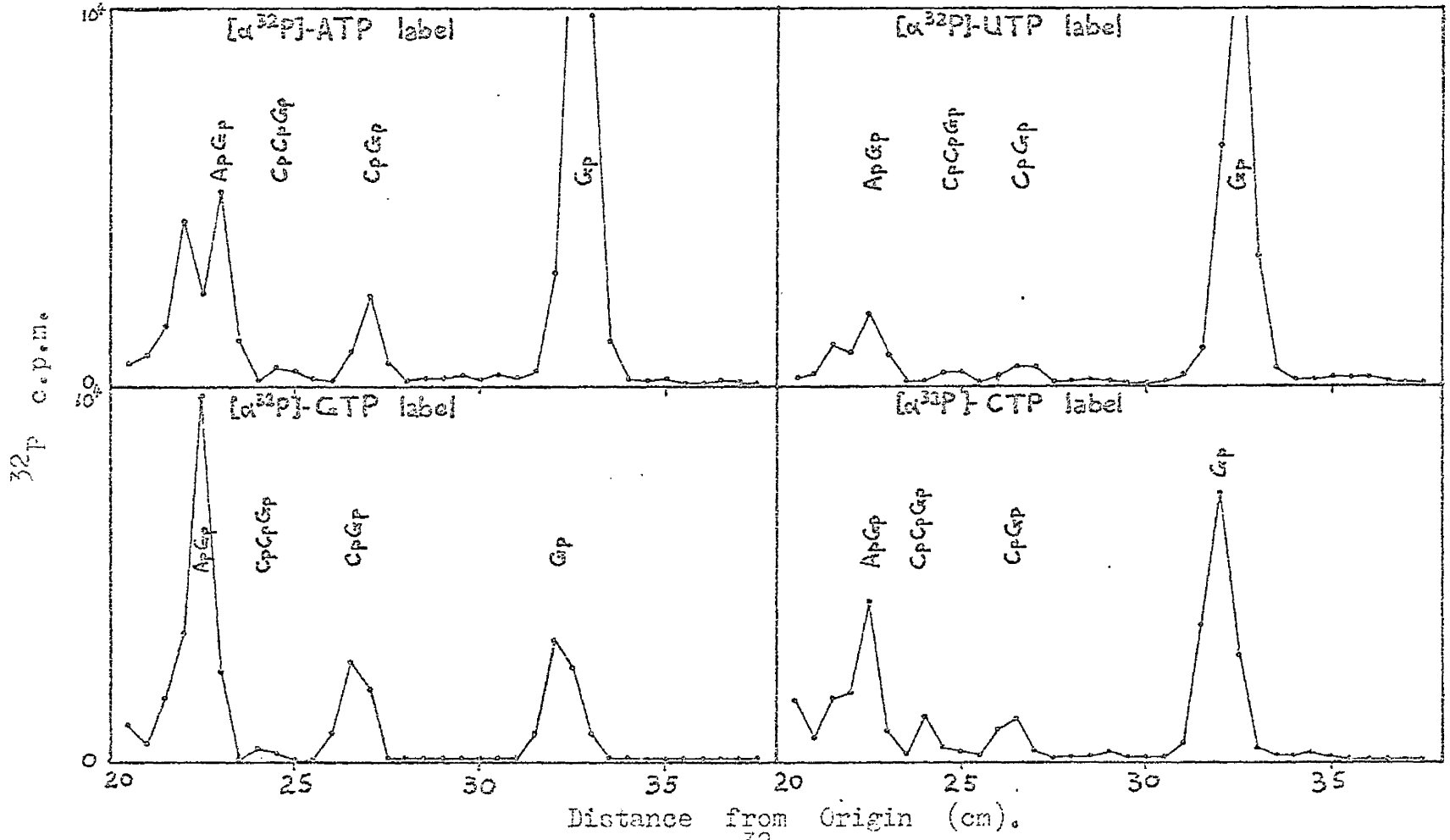
‡ The species marked "Unknown" contained C and A.

FIGURE 23. LOW pH ELECTROPHORESIS OF [^{14}C]-RNA DIGESTS



RNA was copied in vitro from MVM DNA using [^{14}C]-GTP, digested and electrophoresed. These diagrams are from Actigraph scans of the electrophoretograms, taken with a slit width of 1.5 mm and scan rate 30cm/h. Similar results were obtained with RNA copied from calf thymus DNA.

FIGURE 24. LOW pH ELECTROPHORESIS OF [^{32}P]-RNA DIGESTED WITH T_1 RNASE



RFA was copied from MVM DNA using each [α^{32}P]-NTP species in turn, digested with T_1 RNase, and fractionated by low pH electrophoresis. These diagrams show details of the forward areas of the electrophoretograms. The paper was cut into 0.5 cm strips, which were counted in a gas-flow counter. The ^{32}P c.p.m. scales have been adjusted so that all represent about the same percentages of total ^{32}P label. Full scale represents about 10^4 cpm. Similar results were obtained with RNAs copied from calf thymus and H-1 DNAs.

identified under u.v. light, by comparison with marker GMP. As shown in Table 24, this gave consistent results for the compounds of interest.

This system gave good separation of CpGp from other digestion products for T_1 RNase digests of RNA labelled with [^{14}C]-GTP or [α ^{32}P]- GTP. For digests of RNA labelled with other [α ^{32}P]-NTP compounds the separation was adequate but sometimes some background radioactivity was detectable between CpGp and Gp, and forward of Gp. This was ascribed to non-specific digestion and is discussed later. The separation obtained with combined T_1 and U_2 RNase digests is illustrated in Figures 23 and 25. Again, the separation is excellent for digests of RNA labelled with [^{14}C]-GTP. However, with [^{32}P]- label, radioactivity was found in other digestion products near CpGp, and in many cases the separation was not good. These digests were therefore electrophoresed for 16-18 kV-h instead of 15-16 kV-h. Nevertheless, resolution in the critical forward area was variable.

4.3.3 Experiments with T_1 RNase

In early experiments it became apparent that in some cases the Worthington T_1 RNase preparation used was not completely specific. Amounts of label in Gp were much higher than expected from nearest-neighbour analysis, and oligonucleotide species not containing G could be detected. The contaminating nuclease activity was not destroyed by heating at 100°C for 10 min. Later experiments used

Sankyo T_1RNase, which was more satisfactory. However, even here care was necessary, since trace amounts of species not expected in T_1 RNase digests could be detected with $[^{32}P]$- RNA.

The following criteria were used to judge the quality of T_1 RNase digests. First, the proportion of radioactivity in GMP should give some indication of the state of digestion. GMP should be liberated wherever the sequence -G-G- is found. Therefore, for T_1 RNase digests of RNA labelled with $[^{14}C]$- GTP, the amount of ^{14}C in GMP should be a direct measure of the frequency of occurrence of GpG ie. the proportion of ^{14}C in GMP should be equal to the proportion of ^{32}P in GMP with a nearest-neighbour analysis using $[\alpha^{32}P]$- GTP. The situation with $[^{32}P]$- RNAs is less simple. In this case, if the RNA was labelled with $[\alpha^{32}P]$- NTP, the amount of $[^{32}P]$- GMP in a T_1 RNase digest should represent the occurrence of the sequence G-G-N. The results can only finally be checked by summation of the calculated total frequencies for all four G-G-N species, which gives an estimate for absolute frequency of CpG.

Two other criteria were therefore used to check $[^{32}P]$- RNA digests. Some of the possible non-specific breakdown products (i.e. oligonucleotides with 3'-end groups other than 3'-GMP) run between CpGp and Gp, and others run ahead of Gp, in the electrophoresis system. The amount of ^{32}P in these areas should be minimal. Next, the reproducibility of the distribution of ^{32}P with a given $[^{32}P]$- RNA was considered. In particular, the amounts of ^{32}P in (C,A)pGp and ApGp, in CpCpGp, in CpGp and in Gp were compared for different

digests.

The results obtained with T_1 RNase digests of RNA transcribed from MVM DNA and labelled with $[^{14}C]$- GTP are illustrated in Fig. 23. With these digests the ^{14}C in CpGp should give a measure of the occurrence of G-C-G. As shown in Table 25, in six determinations values between 2 and 3% of the total radioactivity were obtained. The less extensive data for calf thymus DNA give similar results (Table 26).

Results obtained with digests of $[^{32}P]$- RNAs are illustrated in Fig.24. Table 27 shows the percentages of ^{32}P found in various fractions with MVM and H-1 DNA experiments. Columns 1 and 2 of Table 27 give sets of results for different digests of the same RNA preparations from MVM DNA. Digestion conditions for the column 1 results were standard:- 1h at 37°C with an enzyme: RNA ratio of 1:20. No steps were taken to remove the template DNA in these experiments. The digests whose results are shown in column 2 were prepared by treatment with T_1 RNase for 30 min. at 37°C (enzyme: RNA ratio 1:20), then heating at 95°C for 5 min and cooling on ice, and incubating at 37°C for 30 min. more. This procedure was intended to dissociate any DNA/RNA complexes which may have inhibited complete digestion.

Except for $[\alpha\ ^{32}P]$- UTP, the results for these two sets of digests were quite close and were averaged. The result for $[^{32}P]$- UTP in column 1 showed a very high percentage of ^{32}P in Gp, and the total distribution of ^{32}P was also atypical. This result was,

therefore, discarded.

In these digests, ^{32}P can occur in oligonucleotides internally or at the 3'-ends. The CpGp fraction should contain internal ^{32}P only when $[\alpha\ ^{32}P]$- GTP has been used as label. This was verified by alkaline hydrolysis of CpGp fractions from the various digests. When the RNA had been prepared with $[\alpha\ ^{32}P]$- GTP, ^{32}P was found in both Cp and Gp from the alkaline hydrolysate, while with other labels only Gp contained ^{32}P. Labelling with $[\alpha\ ^{32}P]$- GTP, for MVM 19.9% of the ^{32}P in CpGp was external, and for H-1 this figure was 20.6.

Further treatments of T_1 RNase results for MVM and H-1 are shown in Tables 28 and 29 respectively. Part A of Table 28 shows the absolute frequency values derived for G-C-G-N and G-G-N species by multiplying the percentage values by the appropriate incorporation factor, as in nearest-neighbour analysis. Part B shows similar calculations for $[^{14}C]$- RNA. Part C of Table 28 compares the GpG frequencies obtained by different methods. The values obtained with T_1 RNase are higher than the nearest-neighbour value; this is probably caused by some non-specific digestion. The sequence G-C-G-N is found for all four species of N (Table 28, part A). However, the total amounts involved are small, and part D shows the poor agreement between different ways of estimating total G-C-G. It is evident that any over-digestion which produces extra CpGp can raise the estimated value of G-C-G, and the minimum estimate of 24% of total CpG as G-C-G seems most likely. The primary aim of these

experiments was to ascertain the nature of sequences on the 3'-side of CpG. Thus the results indicate quite clearly that all four nucleosides are found as C-G-N. The results for H-1 (Table 29) are similar and warrant the same conclusions.

The results for T_1 RNase digests of $[^{32}P]$-RNA transcribed from calf thymus DNA are shown in Table 30. The various digests were made as follows. The digests represented in columns 1 and 2 of Table 30 were made as described for columns 1 and 2 respectively of Table 27. These results show percentages of ^{32}P in CpGp and Gp from calf thymus DNA generally lower than those found for the parvovirus DNAs. 5μg of parvovirus DNA was used as template in each RNA synthesis mixture, while 20μg of calf thymus DNA was used. It therefore seemed possible that the low recoveries of CpGp and Gp indicated interference by the larger quantities of DNA with complete digestion, rather than a genuine difference.

The digests represented in column 3 of Table 30 were therefore made as follows. The DNA was first removed by incubating 200 μg RNA with 5μg crystalline pancreatic DNase for 30 min. at 37°C in 0.3ml of a solution containing 0.05 M-tris-Cl, pH 7.5, and 5mM-MgSO4. This was followed by one cycle of acid precipitation as described in Section 2.5.1, and by digestion with T_1 RNase for 1 h at 37°C (enzyme: RNA ratio 1:20). This procedure gave results similar to those found before. As a final check, new preparations of RNA were made. In this case, after incubation with RNA polymerase, the template DNA was removed by treatment with 5μg DNase for 30 min. at

TABLE 25. FRACTIONATION OF [^{14}C]-RNA MADE FROM MVM DNA

Experiment	Fraction	T_1 RNase	T_1 RNase + extra time at pH 4.5	T_1 + U_2 RNases
1	CpGp	2.14	-	-
	Gp	25.17	-	-
2	CpGp	2.03	-	4.23
	Gp	22.86	-	55.70
3	CpGp	2.99	-	7.11
	Gp	23.21	-	67.20
4	CpGp	2.44	5.84	8.32
	Gp	28.01	35.16	61.79
5	CpGp	2.19	3.75	6.36
	Gp	20.33	21.23	48.75
6	CpGp	2.84	3.13	6.88
	Gp	21.85	23.31	47.59
Expected *	Gp	21.6	-	65.9

RNA was copied from MVM DNA, labelled with [^{14}C]-GTP, and digested as shown. The amounts of ^{14}C found in CpGp and Gp are shown as percentages of total ^{14}C for six determinations.

* These estimates are from RNA polymerase nearest-neighbour analyses.

---ooOoo---

TABLE 26. FRACTIONATION OF [^{14}C]-RNA MADE FROM CALF THYMUS DNA

Experiment	Fraction	T_1 RNase
1	CpGp	1.22
	Gp	20.55
2	CpGp	2.26
	Gp	26.45
Expected *	Gp	24.5

This table is constructed in the same way as the previous table.

TABLE 27. FRACTIONATION OF T_1 RNASE DIGESTS OF [^{32}P]-RNAs MADE FROM MVM AND H-1 DNAs

Label	Fraction	MVM 1	MVM 2	H-1
[α^{32}P]-ATP	ApGp + (Ap,Cp)Gp	4.38	4.13	4.55
	CpCpGp	0.28	0.47	0.31
	CpGp	0.85	0.77	1.14
	Between CpGp & Gp	0.42	0.68	0.40
	Gp	6.57	6.28	7.86
	Forward of Gp	0.06	0.14	0.14
[α^{32}P]-UTP	ApGp + (Ap,Cp)Gp	3.99	1.89	1.60
	CpCpGp	0.39	0.19	0.30
	CpGp	1.14	0.49	0.39
	Between CpGp & Gp	0.42	0.23	0.20
	Gp	16.98	6.31	6.86
	Forward of Gp	0.03	0.24	0.26
[α^{32}P]-GTP	ApGp + (Ap,Cp)Gp	10.38	10.00	9.53
	CpCpGp	0.17	0.31	0.27
	CpGp	2.11	2.47	2.63
	Between CpGp & Gp	0.05	0.10	0.08
	Gp	3.22	3.33	3.68
	Forward of Gp	0.01	0.02	0.06
[α^{32}P]-CTP	ApGp + (Ap,Cp)Gp	4.79	3.94	3.70
	CpCpGp	0.63	0.81	0.79
	CpGp	1.18	1.19	0.99
	Between CpGp & Gp	0.36	0.45	0.24
	Gp	4.72	6.15	4.78
	Forward of Gp	0.02	0.29	0.22

RNAs were copied from MVM and H-1 DNAs using [α^{32}P]-NTP labels, digested with T_1 RNase and fractionated by electrophoresis on DEAE-paper at pH 1.9. The amounts of ^{32}P in the compounds shown are expressed as percentages of the total ^{32}P in each digest.

TABLE 28. POOLED DATA FOR T_1 RNASE DIGESTS OF RNA MADE FROM MVM DNA

PART A : T_1 RNase Digests of [^{32}P]-RNAs

Sequence	N=A	N=U	N=G	N=C
	Freq./10^3 bases	Freq./10^3 bases	Freq./10^3 bases	Freq./10^3 bases
G-C-G-N	2.82	1.13	1.03 *	2.36
G-G-N	22.4	14.5	7.4	10.8

These frequencies are calculated from the data of Table 27.

* Calculated from ^{32}P at external position in CpGp.

PART B : T_1 RNase Digests of [^{14}C]-RNA

Sequence	Frequency /10^3 bases
G-C-G	5.75
G-G	54

RNA was copied from MVM DNA using [^{14}C]-GTP. These frequencies are calculated from the data of Table 25.

PART C : Estimates of G-G Frequency

Method	Freq./10^3 bases
RNA nearest-neighbour analysis	49
T_1 RNase digests of [^{32}P]-RNAs (Part A)	55
T_1 RNase digests of [^{14}C]-RNA (Part B)	54

PART D : Estimates of G-C-G Frequency

Method	Freq./10^3 bases	% of total C-G
T_1 RNase digests of [^{32}P]-RNAs (Part A)	7.3	43
T_1 RNase digests of [^{14}C]-RNA (Part B)	5.75	34
Ratio of internal to external ^{32}P in CpGp from RNA made with [α^{32}P]-GTP	4.12	24

All these data refer to MVM DNA (-) strand.

TABLE 29. POOLED DATA FOR T_1 RNASE DIGESTS OF RNA MADE FROM H-1 DNA

PART A : T_1 RNase Digests of [^{32}P]-RNAs

Sequence	N=A	N=U	N=G	N=C
	Freq./10^3 bases	Freq./10^3 bases	Freq./10^3 bases	Freq./10^3 bases
G-C-G-N	3.77	0.87	1.30 *	2.06
G-G-N	26.0	15.2	8.8	10.0

These frequencies are calculated from the data of Table 27.

* Calculated from ^{32}P at external position in CpGp.

PART B : Estimates of G-G Frequency

Method	Freq./10^3 bases
RNA nearest-neighbour analysis	56
T_1 RNase digests of [^{32}P]-RNAs (Part A)	60

PART C : Estimates of G-C-G Frequency

Method	Freq./10^3 bases	% of total C-G
T_1 RNase digests of [^{32}P]-RNAs (Part A)	8.00	35
Ratio of internal to external ^{32}P in CpGp from RNA made with [α^{32}P]-GTP	5.01	22

All these data refer to H-1 DNA (-) strand.

TABLE 30. FRACTIONATION OF T_1 RNASE DIGESTS OF [^{32}P]-RNAs MADE FROM CALF THYMUS DNA

Label	Fraction	1	2	3	4
[α^{32}P]-ATP	ApGp + (Ap,Cp)Gp	3.83	-	4.42	5.43
	CpCpGp	-	-	0.17	0.38
	CpGp	0.34	0.52	0.55	0.69
	Between CpGp & Gp	0.29	0.29	0.30	0.73
	Gp	5.72	6.21	6.30	8.11
	Forward of Gp	0.15	0.19	0.26	0.17
[α^{32}P]-UTP	ApGp + (Ap,Cp)Gp	1.55	-	1.98	2.40
	CpCpGp	-	-	0.12	0.30
	CpGp	0.25	0.29	0.34	0.62
	Between CpGp & Gp	0.14	0.20	0.08	0.46
	Gp	3.79	4.19	2.29	4.17
	Forward of Gp	0.07	0.10	0.07	0.35
[α^{32}P]-GTP	ApGp + (Ap,Cp)Gp	12.04	-	13.67	13.17
	CpCpGp	-	-	0.33	0.59
	CpGp	1.11	1.01	1.24	1.73
	Between CpGp & Gp	0.06	0.19	0.08	0.10
	Gp	5.43	5.66	5.92	7.49
	Forward of Gp	0.01	0.08	0.04	0.05
[α^{32}P]-CTP	ApGp + (Ap,Cp)Gp	2.80	-	2.60	3.15
	CpCpGp	0.37	-	0.38	0.97
	CpGp	0.32	0.41	0.54	0.85
	Between CpGp & Gp	0.17	0.41	0.29	1.04
	Gp	4.69	4.92	4.90	7.37
	Forward of Gp	-	0.11	0.10	0.43

RNAs were copied from calf thymus DNA using [α^{32}P]-NTP labels, digested with T_1 RNase and fractionated by electrophoresis on DEAE-paper at pH 1.9. The amounts of ^{32}P in the compounds shown are expressed as percentages of the total ^{32}P in each digest.

TABLE 31. POOLED DATA FOR T_1 RNASE DIGESTS OF RNA MADE FROM CALF THYMUS DNA

PART A : T_1 RNase Digests of [^{32}P]-RNAs

Sequence	N=A	N=U	N=G	N=C
	Freq./10^3 bases	Freq./10^3 bases	Freq./10^3 bases	Freq./10^3 bases
G-C-G-N	1.38	0.81	0.58 *	0.82
G-G-N	17.9	10.2	12.2	9.4

These frequencies are calculated from the data of Table 30.

* Calculated from ^{32}P at external position in CpGp.

PART B : T_1 RNase Digests of [^{14}C]-RNA

Sequence	Frequency /10^3 bases
G-C-G	3.72
G-G	50

RNA was copied from calf thymus DNA using [^{14}C]-GTP. These frequencies are calculated from the data of Table 26.

PART C : Estimates of G-G Frequency

Method	Freq./10^3 bases
RNA nearest-neighbour analysis	52
T_1 RNase digests of [^{32}P]-RNAs (Part A)	50
T_1 RNase digests of [^{14}C]-RNA (Part B)	50

PART D : Estimates of G-C-G Frequency

Method	Freq./10^3 bases	% of total C-G
T_1 RNase digests of [^{32}P]-RNAs (Part A)	3.59	32
T_1 RNase digests of [^{14}C]-RNA (Part B)	3.72	33
Ratio of internal to external ^{32}P in CpGp from RNA made with [α^{32}]-GTP	3.90	35

37°C. The RNA was then isolated as usual, and digested for 1 h at 37°C. (enzyme: RNA ratio 1:10). This treatment gave rather higher percentages of ^{32}P in CpGp and Gp (column 4) but the levels of non-specific background ^{32}P indicated that these were probably due to non-specific digestion. The results for the first three digestion sets were averaged.

These mean results are treated in Table 31. Estimates of GpG frequencies from the T_1 RNase digests and from nearest-neighbour analysis agree well, so the higher estimates in column 4 of Table 30 were not considered further. The estimates for G-C-G also agree well (although that based on the $\angle^{14}C\_7$- G data of Table 26 is suspect since it was obtained by averaging two substantially different estimates). Again, all four species of G-C-G-N are found.

All T_1 RNase digests gave some radioactivity in the CpCpGp fraction, indicating the occurrence of sequences of type G-C-C-G-N. This fraction was not investigated further. It is apparent from Figures 23 and 24 that a large variety of sequences is present in the RNAs made *in vitro* and that this system could be used to investigate other aspects of the RNA sequences.

4.3.4 Experiments with U_2 RNase

Since only a small amount of U_2 RNase was available, U_2 RNase digests were made by digesting first with T_1 RNase at pH 7.5, then lowering the pH to 4.5 and digesting with U_2 RNase. Control incubations at pH 4.5 without U_2 RNase also gave some further

breakdown of the T_1 RNase products (see Fig.23). The contaminant of the T_1 RNase preparations might be T_2 RNase, which is maximally active at pH 4.5 (Egami, Takahashi & Uchida, 1964). T_2 RNase is not base specific, but does exhibit some preference for cleavage next to 3'-AMP. If the contaminating activity is T_2 RNase, this preference should minimise the extent of undesired digestion.

The resolution obtained with electrophoresis of U_2 RNase digests of $[^{32}P]$-RNA is illustrated by Fig.25. All fractions near CpGp were investigated by alkaline hydrolysis. Two peaks are present between CpGp and Gp, which were not found with T_1 RNase digests (or were detected only in trace amounts). These peaks contain C and A but their sequences have not been defined.

The results of U_2 RNase digests can be checked by methods similar to those outlined for T_1 RNase. Radioactivity in Gp gives a measure of the frequency of PupG. Radioactivity in CpGp measures the frequency of Pu-C-G: values for the frequency of this sequence are also obtainable from pyrimidine run data, and from pancreatic digest experiments. The amount of label in CpGp and Gp from control incubations at pH 4.5 should give some indication of the extent of non-specific digestion.

The results of U_2 RNase digests of RNA copied from MVM DNA and labelled with $[^{14}C]$-GTP are shown in Fig.23 and Table 25. The results of U_2 RNase digests of $[^{32}P]$-RNAs transcribed from MVM DNA are given in Table 32. The control digests for MVM show that in all

TABLE 32. U_2 RNASE DIGESTS OF RNA MADE FROM MVM DNA

Label	Fraction	Control Digest	$T_1 + U_2$ RNases	Sequence indicated	Freq./10^3 bases
[α^{32}P]-ATP	CpGp	1.73	2.40	Pu-C-G-A	8.4
	Gp	8.42	18.31	Pu-G-A	64.0
[α^{32}P]-UTP	CpGp	0.64	0.89	Pu-C-G-U	2.1
	Gp	9.23	12.44	Pu-G-U	28.5
[α^{32}P]-GTP	CpGp	0.76 *	1.29 *	Pu-C-G-G	2.9
	Gp	4.57	11.03	Pu-G-G	25.0
[α^{32}P]-CTP	CpGp	2.10	2.93	Pu-C-G-C	5.8
	Gp	8.87	21.12	PU-G-C	42.0

RNA was copied from MVM DNA using [α^{32}P]-NTP labels, as shown. The amounts of ^{32}P in CpGp and Gp in different digests are given as percentages of the total ^{32}P in the digest. The results in the column headed "Control Digest" were obtained by digesting with T_1 RNase and then incubating, without U_2 RNase, at pH 4.5. The frequencies of occurrence of sequences (for MVM (-) strand) deduced from the U_2 RNase results are given on the right.
* Calculated from ^{32}P in the external position of CpGp.

Estimates of Pu-G Frequency

Method	Freq./10^3 bases
RNA nearest-neighbour analysis	149
U_2 RNase digests of [^{32}P]-RNAs	160
U_2 RNase digests of [^{14}C]-RNA	127

Estimates of Pu-C-G Frequency

Method	Freq./10^3 bases
U_2 RNase digests of [^{32}P]-RNAs	19.1
U_2 RNase digests of [^{14}C]-RNA	14.8
Pancreatic RNase digest of [^{32}P]-RNA	15.0

RNA nearest-neighbour analysis gives a total C-G frequency of 17 times per 10^3 bases.

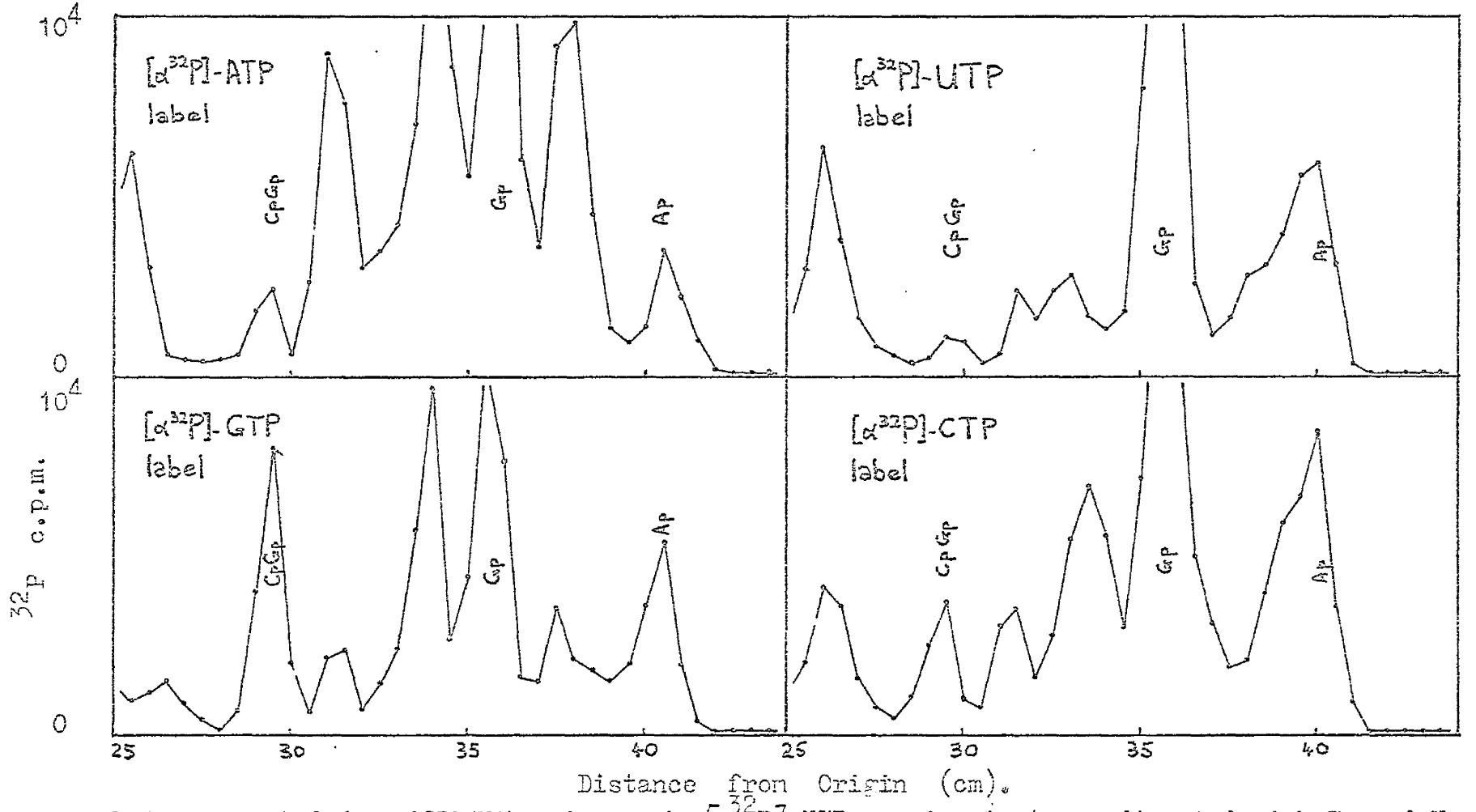
TABLE 33. U_2 RNASE DIGESTS OF RNA MADE FROM CALF THYMUS DNA

Label	% of ^{32}P in CpGp	Sequence indicated	Freq./10^3 bases
[α^{32}P]-ATP	1.00	Pu-C-G-A	2.9
[α^{32}P]-UTP	1.25	Pu-C-G-U	3.7
[α^{32}P]-GTP	0.50 *	Pu-C-G-G	1.1
[α^{32}P]-CTP	1.23	Pu-C-G-C	2.5

RNA was copied from calf thymus DNA using [α^{32}P]-NTP labels, and digested with T_1 and U_2 RNases. The amounts of ^{32}P in CpGp in different digests are shown as percentages of the total ^{32}P in the digest. The frequencies of occurrence of the sequences deduced from these results are given on the right.

* Calculated from external ^{32}P in CpGp.

FIGURE 25. LOW pH ELECTROPHORESIS OF [^{32}P]-RNA DIGESTED WITH T_1 AND U_2 RNASES



RNA was copied from MVM DNA using each [α^{32}P]-NTP species in turn, digested with T_1 and U_2 RNases, and fractioned by low pH electrophoresis. These diagrams were prepared as described for Fig. 24. Similar results were obtained with RNAs copied from calf thymus DNA.

cases some extra digestion does occur on incubation at pH 4.5. Table 32 also shows absolute frequencies of the various Pu-C-G-N and Pu-G-N species in the MVM experiment. The value for the total frequency of Pu-C-G is high compared with the frequency derived from the pancreatic RNase experiment. It therefore appears that some non-specific digestion is occurring with U_2 RNase. As with the T_1 RNase results, all four nucleosides are found on the 3'-side of CpG. The results for calf thymus DNA (Table 33) are similar. The increases in ^{32}P found in CpGp over the corresponding T_1 RNase results indicate that all species of A-C-G-N occur.

4.3.5 Experiments with Pancreatic RNase

RNA transcribed from MVM DNA was labelled with $[\alpha\,^{32}P]$-GTP and digested with pancreatic RNase. Three such digests were fractionated by chromatography on DEAE-paper. The first fractionation used 0.2 M, the second 0.02 M and the third 0.22 M-ammonium formate in 7 M-urea. The latter two digests were samples of the same RNA preparation. The percentages of ^{32}P found in the various fractions agreed well for the three separations (Table 34). The separation obtained with the 0.22 M-ammonium formate is shown in Fig.26. The fractions of the second and third runs were further investigated after elution from the paper. The proportion of radioactivity in each 3'-mononucleotide after alkaline hydrolysis of each length class was measured. Table 35 shows the amount of radioactivity in each mononucleotide in each fraction as a percentage of the total radioactivity.

The distribution of ^{32}P estimated in this way gives quite good agreement with the direct nearest-neighbour analysis results, although the percentage of ^{32}P in AMP is rather low. ^{32}P is found in 3'-CMP in all the fractions examined, showing that the sequence CpG can occur in a variety of positions with respect to purine tracts found on the 5'-side. Sequences of the form G-C-G for each length class were estimated by hydrolysing samples of each length class with T_1 RNase and measuring the $[^{32}P]$- CMP released. Radioactivity was detectable in CMP in all the digests. However, only for the fraction of oligonucleotides longer than six were ^{32}P levels high enough to quantitate the $[^{32}P]$- CMP. In this case 2.2% of the ^{32}P in the fraction was found in CMP, as against 7.1% in CMP from alkaline hydrolysis. Therefore, all the fractions investigated contained sequences representing G-C-G, and the "large oligonucleotide" fraction contained about 30% of the CpG as G-C-G.

TABLE 34. SEPARATION BY LENGTH OF OLIGONUCLEOTIDES IN PANCREATIC RNASE DIGESTS

Length of Oligonucleotide Class	Separation with 0.22M Salt	Separation with 0.20M Salt	Separation with 0.02M Salt
1	15.16	15.64	15.58
2	6.29	4.88	6.12
3	16.73	14.31	78.30
4	12.86	9.82	
5	11.86	11.17	
6	7.93	41.56	
> 6	29.16		

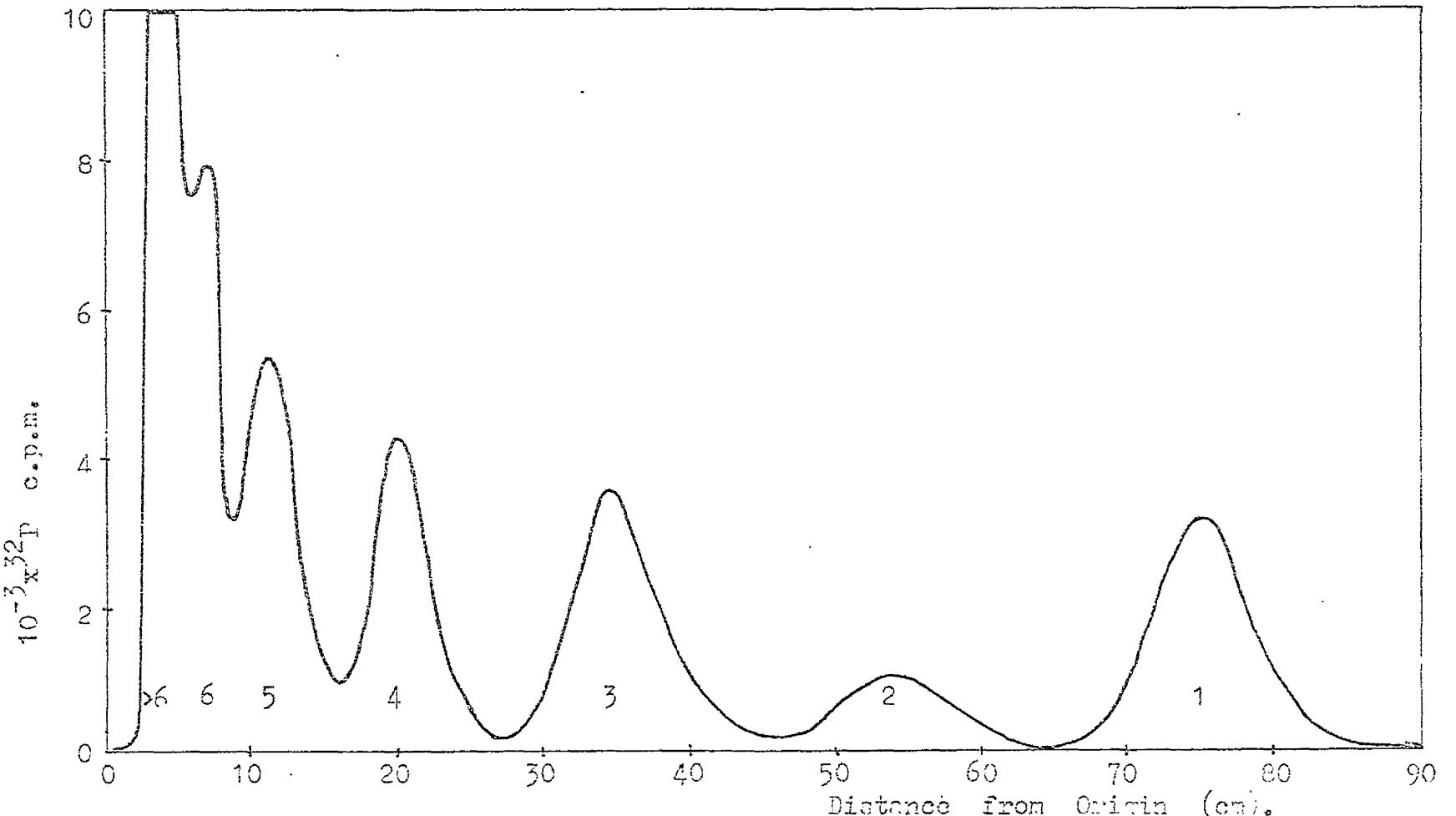
RNA was copied from MVM DNA using [α^{32}P]-GTP, digested with pancreatic RNase, and fractionated into length classes by chromatography on DEAE-paper. The amounts of ^{32}P in each length class are expressed as percentages of the total, for each of three experiments using different eluting salt concentrations.

TABLE 35. ALKALINE HYDROLYSIS OF PANCREATIC RNASE DIGEST FRACTIONS

	Length of Oligonucleotides							Total	Direct Estimate
	1	2	3	4	5	6	>6		
3'-AMP	-	-	4.26	6.51	7.16	4.77	17.32	40.0	44.3
3'-UMP	12.90	4.51	4.85	2.64	1.37	0.86	2.36	29.5	26.4
3'-GMP	-	-	6.57	2.93	2.71	1.86	7.40	21.5	21.6
3'-CMP	2.26	1.78	1.20	0.78	0.62	0.43	2.08	9.2	7.7

RNA was copied from MVM DNA using [α^{32}P]-GTP, digested with pancreatic RNase and fractionated into length classes (Table 34). Each fraction was then digested with alkali and the 3'-mononucleotides separated. This table shows the proportion of ^{32}P in each 3'-NMP species as a percentage of the total ^{32}P in all length classes. The column headed "Total" sums all these percentages. The column headed "Direct Estimate" gives the corresponding result expected from RNA nearest-neighbour analysis.

FIGURE 26. FRACTIONATION BY LENGTH OF OLIGONUCLEOTIDES IN A PANCREATIC RNASE DIGEST



RNA was copied from MVM DNA using $[\alpha^{32}P]$-GTP, digested with pancreatic RNase, and chromatographed on DEAE-paper with 0.22 M-ammonium formate, 7 M-urea. This diagram is from an Actigraph scan of the chromatogram, taken with a slit width of 6 mm and a scan rate 60 cm/h. The numeral at each peak indicates the nucleotide chain length of the peak material.

4.4 DISCUSSION OF RESULTS

4.4.1 The Immediate Neighbours of CpG Sequences

In Table 36 are given estimates for the relative frequencies of occurrence of G-C-G-N species, in MVM, H-1 and calf thymus DNAs, and of Pu-C-G-N species in MVM and calf thymus DNAs. The values for G-C-G-C and Pu-C-G-G are both 25% by the most direct method available:- measuring the ratio of internal to external ^{32}P in CpGp from RNA labelled with $[\alpha\ ^{32}P]$- GTP. The frequencies of the other species were then obtained by proportion from the data of Tables 28, 29 and 31. It is considered that direct use of all the data in these latter Tables under-estimates the relative frequency of the G-C-G-G and Pu-C-G-G species.

The results for G-C-G-N species in the two parvovirus DNAs are qualitatively similar: in each case G-C-G-A is the most frequently occurring species (33-42%) and G-C-G-T the least frequent (10-14%). Similar results were also obtained for Pu-C-G-N species in MVM DNA. The results for calf thymus DNA are less extreme: with the G-C-G-N frequencies, A is the most common 3'-nucleoside, but with the Pu-C-G-N determination, T occurs most frequently. For both MVM and calf thymus DNAs, the Pu-C-G-N species represent most of the total C-G-N sequences (at least 50-70%) and are probably similar to the relative frequencies of C-G-N species.

Also shown in Table 36 are the relative abundances of the GpN doublets, in each DNA, as measured by RNA nearest-neighbour

TABLE 36. ESTIMATES OF C-C-G-N AND Pu-C-G-N SEQUENCES IN MVM, H-1 AND CALF THYMUS DNAs

Estimates of G-C-G-N Frequencies

Sequence	MVM	H-1	Calf Thymus
G-C-G-A	33	42	33
G-C-G-T	14	10	21
G-C-G-G	25	25	25

Relative frequencies are as percentages of total G-C-G.
The data for MVM and H-1 DNAs represent the (-) strands.

Estimates of Pu-C-G-N Frequencies

Sequence	MVM	Calf Thymus
Pu-C-G-A	38	24
Pu-C-G-T	11	31
Pu-C-G-G	25	25
Pu-C-G-C	26	20

Relative frequencies are as percentages of total Pu-C-G.
The data for MVM DNA represents the (-) strand.

Relative Frequencies of G-N Sequences

Sequence	MVM	H-1	Calf Thymus
G-A	36	35	31
G-T	20	19	26
G-G	22	23	24
G-C	22	22	18

Relative frequencies are as percentages of total G.
These data are derived from RNA nearest-neighbour analysis.

---ooOoo---

TABLE 37. POOLED DATA FOR RELATIVE FREQUENCIES OF N-C-C SEQUENCES IN MVM DNA

Sequence	% of total	Nearest-Neighbour :	RNA	DNA
A-C-G	45	A-C	38	38
T-C-G	15	T-C	18	17
G-C-G	25	G-C	25	26
C-C-G	15	C-C	20	19

The relative frequencies of N-C-G species (for MVM (-) strand) are estimated from results with pyrimidine runs, pancreatic RNase digests, and T_1 and U_2 RNase digests, and are compared with estimates for total N-C from nearest-neighbour analyses.

analysis. Comparison with these data shows that the GpN sequences found as G-C-G-N and Pu-C-G-N are quite similar in relative frequency to the total G-N species, although the low C-G-T frequency in the parvovirus DNAs is not matched by a low total GpT frequency.

Good quantitative estimates of the relative frequencies of N-C-G species cannot be made from the available data. However, by pooling results for MVM DNA from pyrimidine run experiments, T_1 and U_2 RNase experiments, and pancreatic RNase experiments, approximate estimates can be made: for MVM DNA, the relative frequencies of A-C-G, T-C-G, G-C-G and C-C-G are, respectively, 45%, 15%, 25% and 15% of the total. As shown in Table 37, these data give a similar distribution of NpC species to estimates for total NpC frequencies.

Since calf-thymus DNA is double-stranded, estimates of N'-C-G species can be made from the data for the complementary C-G-N species (Table 36). In summary, the relative frequencies of the immediate neighbours of CpG in MVM (-) strand DNA and in calf thymus DNA are close to those expected from nearest-neighbour analysis. It is clear from these experiments that in neither of these DNAs is the CpG doublet contained in an unique longer sequence.

4.4.2 Oligonucleotide Neighbours of CpG Sequences

The pyrimidine run experiments give information on sequences, up to three residues long, on the 5'-side of CpG sequences. The quantitative examination of these data requires some background of theory, and in this section some aspects of the frequencies of various

species expected in random sequence chains are developed and then used as references for the consideration of the experimental data.

Consider the set of pyrimidine runs from a DNA chain of random sequence. As in Section 1.2.2, the frequencies of A, T, G and C are represented by a, t, g and c. Isostich I represents sequences of the type Pu-Py-Pu and its frequency of occurrence is therefore $(a+g)(c+t)(a+g)$. Similarly the isostich II frequency is $(a+g)(c+t)(c+t)(a+g)$ and, in general, the frequency of occurrence F_n of an isostich n units long is given by $F_n = (a+g)^2(c+t)^n$. This equation can be converted to $\log F_n = 2 \log (a+g) + n \log (c+t)$ i.e. plotting $\log F_n$ against n gives a straight line of slope $\log (c+t)$. Since in practice $(c+t)$ is fractional, $\log (c+t)$ is negative.

Similar models can be developed for other fractions of the DNA. In each case the $\log^{-1}$(slope) gives a measure of the frequency of occurrence, in the DNA, of the bases found in the runs, or, in other words, a measure of the probability of a run $(n+1)$ units long being formed from a run n units long. If, in a real, non-random DNA, such a plot gives a straight line, then the slope of the line gives a measure of the frequency of the internal base constituents; the intercept on the F_n-axis is of less value as it measures the total frequency of the end constituents of the runs, and in general only gives information obtainable by more direct means.

Fig.27 shows the MVM DNA experimental data for pyrimidine runs ended with -C-G and with -T-G, and for purine runs ended with -C-G and with -T-G, plotted as $\log F_n$ against n. In each case the points

fall close to a straight line, so the model developed above is relevant to the data in this respect. Using data from nearest-neighbour analyses, $\log^{-1}$(slope) should be 0.43-0.46 for the pyrimidines and 0.54-0.57 for the purine runs. The values obtained from the plots are as follows.

Considering the runs ended with -C-G, the addition of pyrimidines occurs with a frequency of 0.37, against the expected range of 0.43-0.46. The addition of purines occurs with a frequency of 0.67, against the expected 0.54-0.57. These two sets of data complement each other well, giving a total base value of 1.04 (0.37+0.67) close to the ideal value of 1.00. It is therefore evident that in MVM DNA (-) strand, the region to the 5'-side of CpG sequences has a higher purine content than the bulk of the DNA; this phenomenon applies at least to the 3 residues to the 5'-side of CpG. Results for sequences ended with -C-G and also for sequences ended with -T-G are shown in Table 38. These give distributions of bases on the 5'-side of TpG quite close to those expected from the total base composition of the DNA.

These conclusions are not dependent on assumptions about the relations of sequences in MVM DNA to those in a random chain; the model discussed above merely provides a convenient framework for analysis of the results.

The available data were examined further to see whether any deductions could be made on the nature of the asymmetry to the

5'-side of CpG. Fig.28 shows plots for all-T and all-C runs to the 5'-sides of CpG and TpG sequences. The results from these plots are also shown in Table 38. First, the results for runs next to TpG are quite close to those expected from the total base composition of the DNA. The results for all-T runs next to CpG are also as expected from the overall base composition, while all-C runs adjacent to CpG are only slightly less frequent than predicted. The all-T and all-C runs next to CpG together give a (c+t) value of 0.42, as against that 0.37 predicted from the $(Py)_n$-C-G data. This value of 0.42 is close to the DNA base composition prediction of 0.43-0.46 from nearest-neighbour analyses.

These discrepancies can be accounted for in three ways. First, the data for all-T and all-C runs may be less precise, as low total frequencies per DNA molecule were involved. Second, the pyrimidine runs of low frequency implied by the analysis of Fig.27 may be mixed C and T species. This possibility cannot be tested with the available data. The last possibility is that the value of 0.37 for the pyrimidine frequency in the region to the 5'-side of CpG may be erroneous. This is unlikely since data from two independent sources agree quantitatively.

It therefore appears that the bias in base composition is real, and several points follow. First, the bias might be peculiar to sequences ended with CpG, or it might, for instance, also be found for CpA. This point could be tested with data for pyrimidine runs next to A, but the available data, derived from $^{14}C/^{32}P$ ratios, is

not adequate for this purpose.

Next, it is of interest to ascertain whether similar trends exist in calf thymus DNA. Here the relevant data are incomplete. The only data for sequences of the type $(Py)_n$-C-G and $(Py)_n$-T-G are for n equalling 1 and 2. The ratio of Pu-Py-C-G to Pu-C-G gives a pyrimidine frequency value of 0.42, while the corresponding value for TpG is 0.49, against expected values of 0.50 in each case. Since each of these values is based on two points only they must be treated with caution. However, they are consistent with the view that sequences to the 5'-side of TpG have the base composition expected from values for the total DNA, while those to the 5'-side of CpG are purine rich.

Other data are also consistent with this view of the CpG surroundings. Plots of all-C runs next to CpG and all-T runs next to TpG give frequencies of 0.16 for C next to CpG and 0.26 for T next to TpG (Fig. 29). Finally, u.v. data for all-C and for all-T runs give frequencies for C and T of 0.23 and 0.28 respectively, close to the values expected from the total base composition.

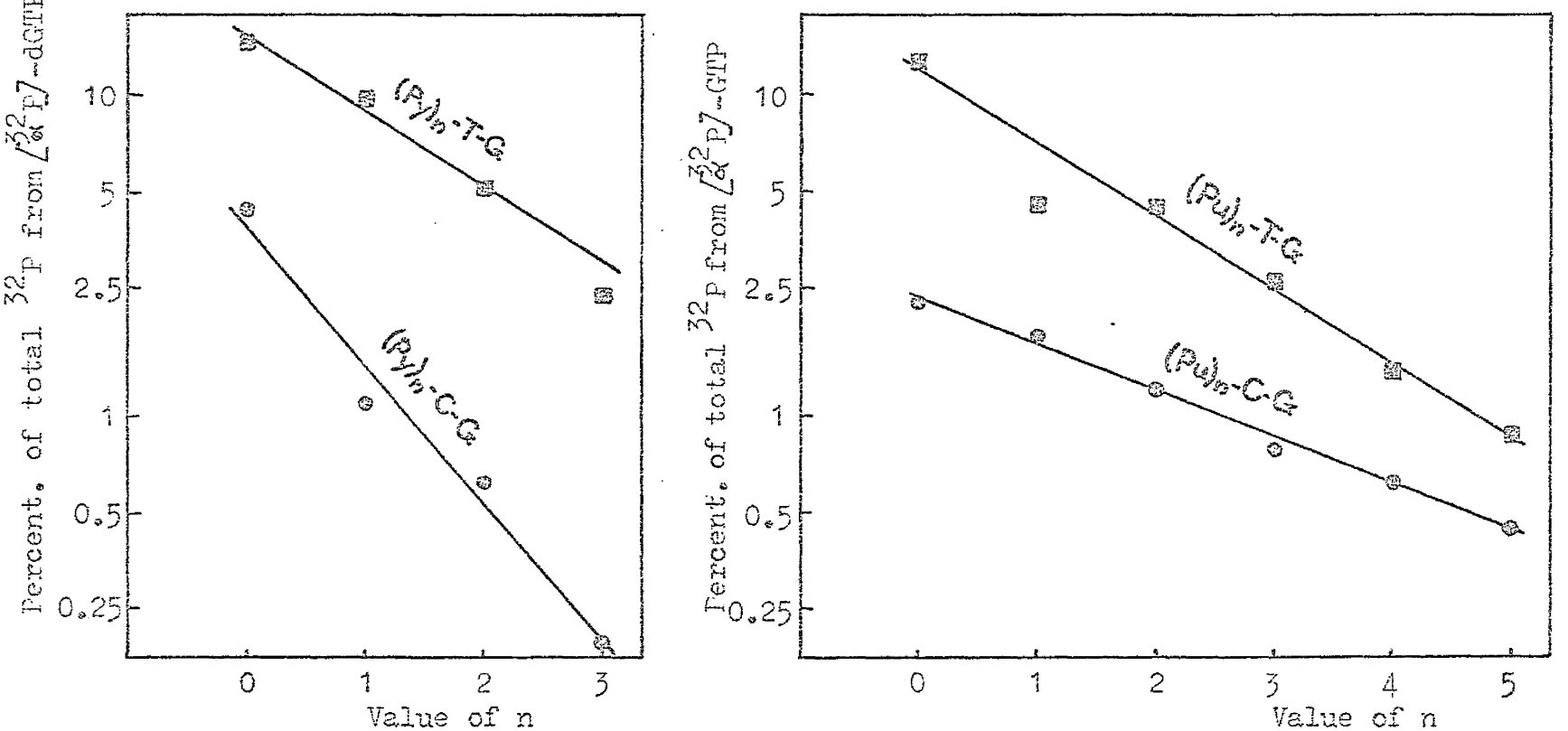
These data are not conclusive, but are consistent with a base frequency distribution to the 5'-side of CpG, in calf thymus DNA, similar to that found in MVM DNA. In double-stranded DNA, such a sequence bias can be either symmetric (that is, the bias can be found on both strands adjacent to a pair of CpG dinucleotides), or asymmetric (i.e. found on one strand only). In the first case, the phenomenon would certainly be detected by exhaustive experiments similar to those

TABLE 38. DATA FROM LOGARITHMIC PLOTS

Series	$\log^{-1}$(slope)
$(Py)_n$--C--G	0.37
$(Py)_n$--T--G	0.51
$(Pu)_n$--C--G	0.67
$(Pu)_n$--T--G	0.60
$(C)_n$--C--G	0.17
$(T)_n$--T--G	0.30
$(T)_n$--C--G	0.25
$(C)_n$--T--G	0.23

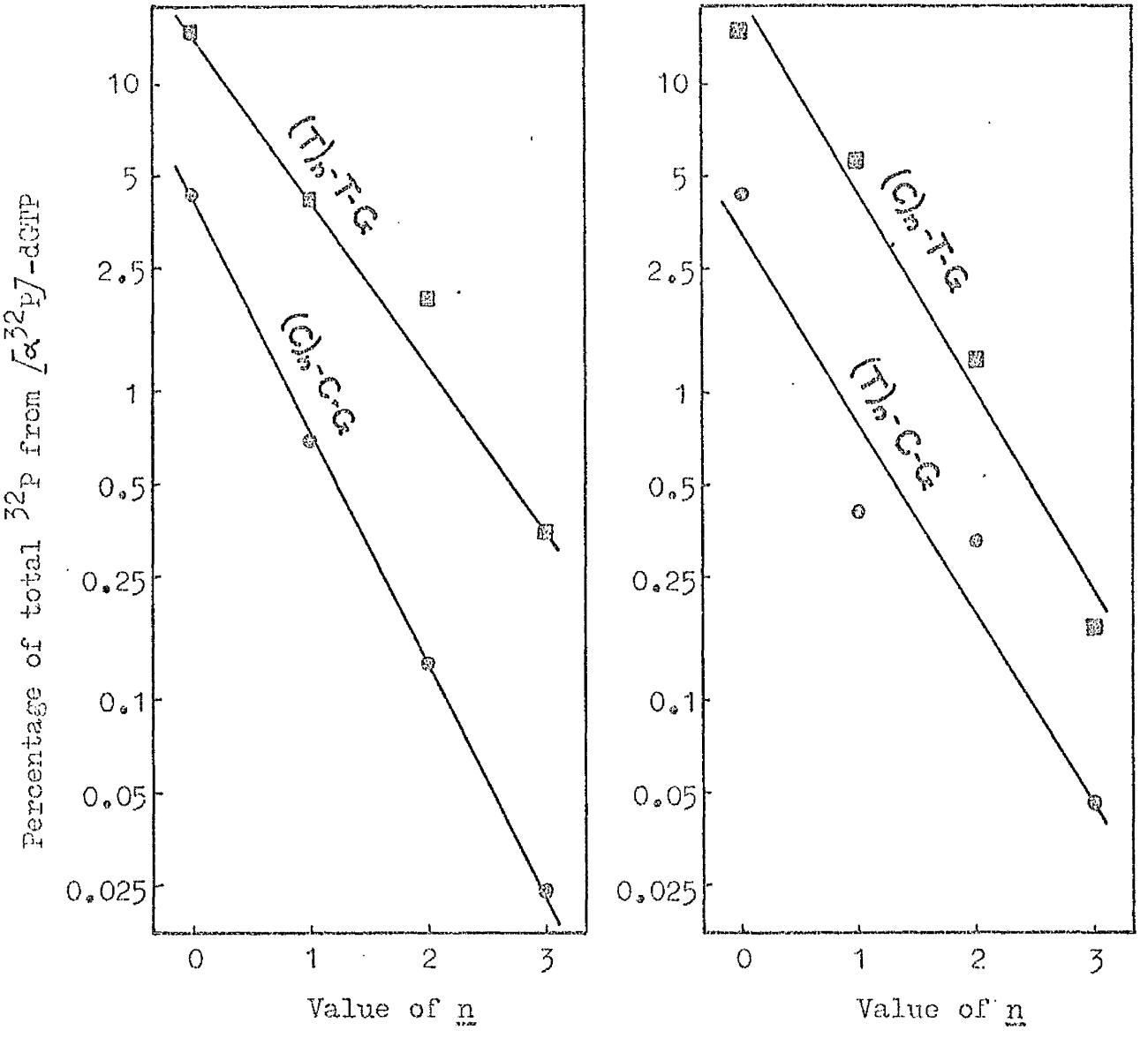
These data refer to MVM DNA (-) strand, and were obtained from the plots in Figures 27 and 28. The best straight line through each set of points was found by an unweighted least-squares procedure. The values of $\log^{-1}$(slope) give a measure of the frequency in each series of the base species in parentheses.

FIGURE 27. FREQUENCIES OF OCCURRENCE IN MVM DNA OF THE SERIES $(Py)_n$-C-G, $(Py)_n$-T-G, $(Pu)_n$-C-G AND $(Pu)_n$-T-G



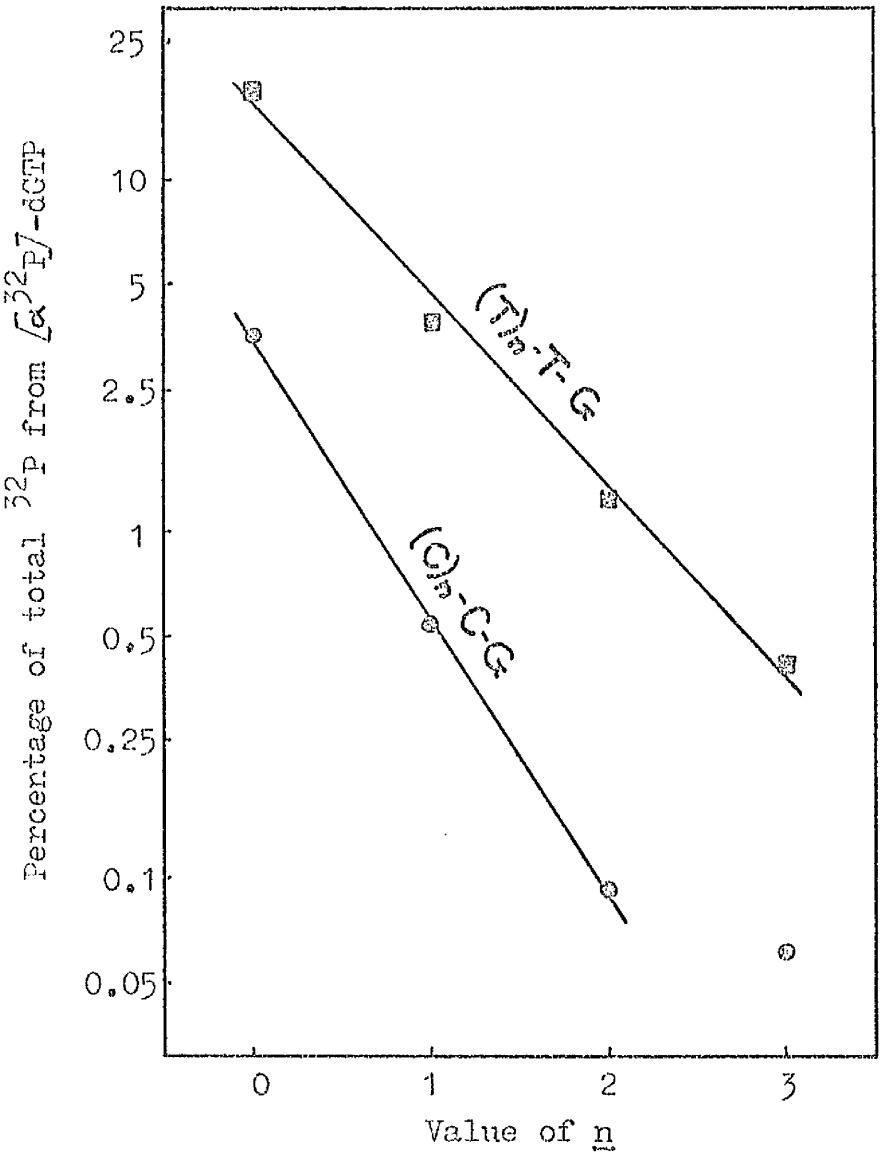
The frequency of occurrence of each species, expressed as a percentage of the total ^{32}P incorporated from $[\alpha^{32}P]$-dGTP or $[\alpha^{32}P]$-GTP, is plotted on a logarithmic scale against a measure of the length of the species. These data represent the (-) strand of MVM DNA.

FIGURE 28. FREQUENCIES OF OCCURRENCE IN HVR DNA OF THE SERIES $(C)_n$-C-G, $(T)_n$-T-G, $(T)_n$-C-G AND $(C)_n$-T-G



See Fig. 27 for details.

FIGURE 29. FREQUENCIES OF OCCURRENCE IN CALF THYMUS DNA OF THE SERIES $(C)_n$-C-G AND $(T)_n$-T-G



This Figure was constructed in the same way as Figs 27 and 28, and represents calf thymus DNA.

already used. In the second case, any such trend on one strand might not be visible using a double-stranded DNA.

As mentioned earlier, Westphal (1970) has described a method for fractionating the strands of SV40 DNA. Sequences in this system could be investigated for each strand either by using the separated strands as templates for DNA polymerase, or by utilising directly the RNA produced by asymmetric transcription of the SV40 DNA with E.coli RNA polymerase (Westphal, 1970). With such a small, defined, double-stranded DNA, there is an additional possibility for the asymmetric model described above:- the asymmetry at different CpG sites might be present always on one strand of the DNA, or it could be found on both strands.

In conclusion, sequences to the 5'-side of CpG in MVM (-) strand DNA are purine rich. The nature of these low frequency pyrimidine sequences has not been defined. This phenomenon extends to 3 or 4 residues to the 5'-side of CpG. It is not known whether this bias is peculiar to CpG or is found for all CpPu sequences. These results imply that in MVM (+) strand the CpG sequences lie to the 5'-side of pyrimidine rich regions. While the data for calf thymus DNA are less extensive, the results available suggest that a similar bias exists in this case also.

4.4.3 Pyrimidine Runs in Calf Thymus DNA

The data obtained from the determination of pyrimidine runs ended with G in calf thymus DNA can be used to extend other

observations on the nature of pyrimidine sequences in calf thymus DNA.

In the pyrimidine run experiments, the u.v. backgrounds gave estimates of the proportions of calf thymus DNA pyrimidines found in various isostichs. These data represented the total base content of each isostich species, and to obtain the "frequency of occurrence" units used elsewhere in this discussion, it is necessary to divide the base content of each isostich fraction by the isostich number, and express each value obtained in this way as a fraction of the total for all isostichs i.e. the weight average frequencies are converted into number average frequencies.

These data are then plotted as log F_n against n as before (Fig.30). As was previously noted by Spencer & Chargaff (1963b), the frequencies of the first three isostichs are lower than expected for a random sequence chain, but isostichs more than four units long are more frequent than expected. This feature is peculiar to vertebrate DNA. The analysis can be extended. The first three isostichs give a line parallel to the random expectation line, and so do isostichs VI to VIII. Isostichs IV and V form the transition. In other words, the relations between the frequencies of the first three isostichs are those predicted by random combination considerations. This is true also for isostichs VI, VII and VIII. These data indicate that there is in the DNA a greater number of some pyrimidine sequences four and five units than would be expected on a random basis. This excess depresses the relative frequency of

the first three isostichs and elevates the frequencies of longer isostichs which contain the high frequency sequences as part of their total length.

The data obtained by labelling calf thymus DNA with [α ^{32}P]-dGTP allow a similar plot to be made for runs ended with G (Fig.31). This shows the same features as the plot for all runs and, by subtraction, it can be shown that a similar situation exists for runs ended with A. This analysis of calf thymus DNA does not, therefore, reveal the source of the asymmetry. As is also shown in Fig.31, this pattern is not shown by MVM DNA, which gives a wider scatter of frequencies. This scatter is perhaps to be expected for such a small DNA.

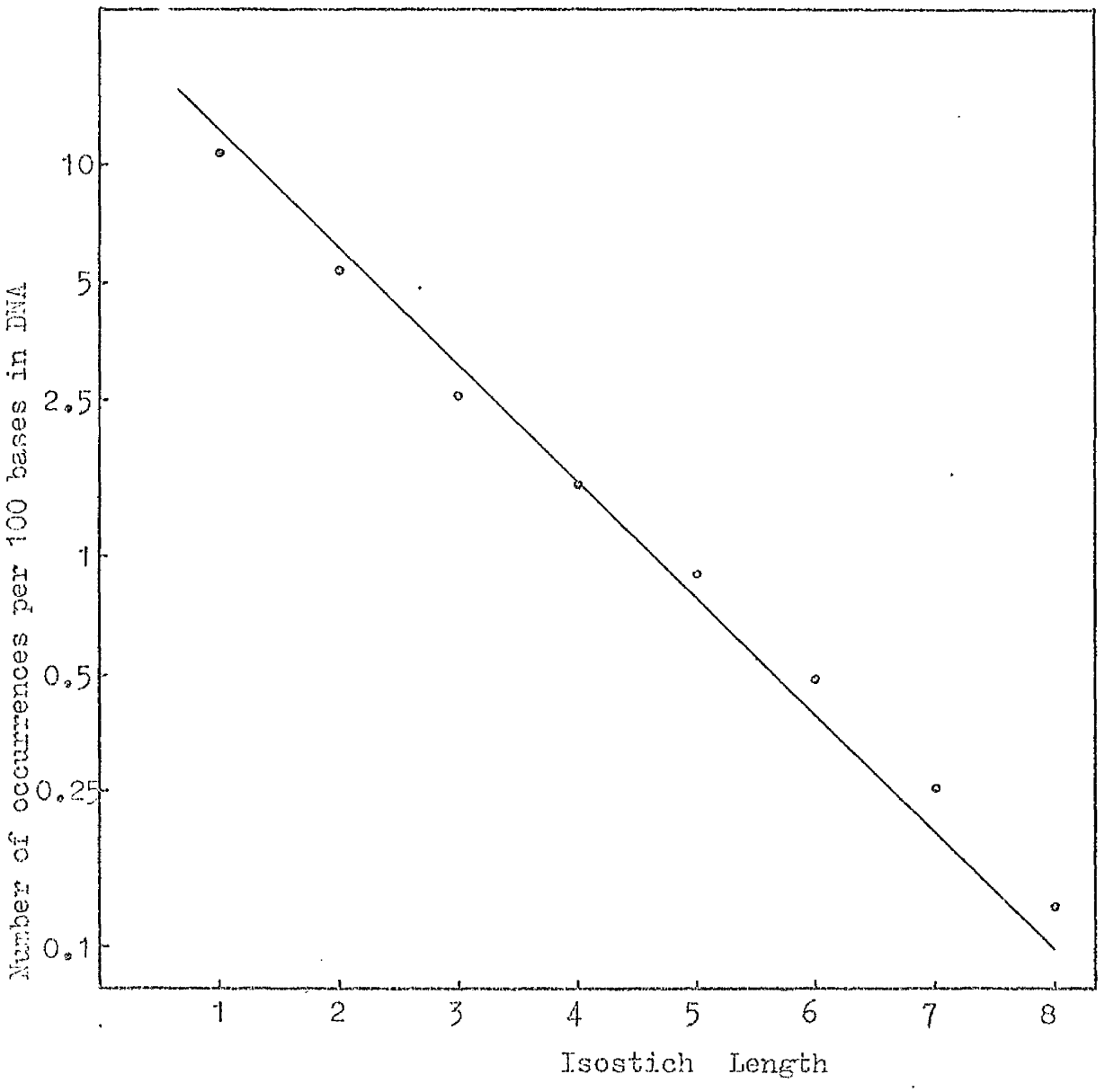
The separation of isostichs into base composition fractions should give some indication of the nature of these base sequences. In Table 39 are shown, for isostichs I to IV, the relative proportions of the fractions in each isostich. The first column shows the relative amounts of the various species expected in a random sequence double-stranded DNA of 42% (G+C) content. The second column shows the relative amounts of u.v. fractions, averaged from the data given in Table 18, and the third column, also from Table 18, shows the relative frequencies of species ended with G. The u.v. data show that the excess sequences in the longer isostichs are C rich sequences. This applies also to the [^{32}P]-G results. In particular, the frequency of Pu-C-C-C-C-G approaches that expected on a random basis.

TABLE 39. PYRIMIDINE ISOSTICHS IN CALF THYMUS DNA

Species	Random	All Runs	Runs next to G
C	42	41	16
T	58	59	84
C_2	18	18	5
CT	49	53	58
T_2	34	30	37
C_3	7	13	2
C_2T	33	33	30
CT_2	42	33	44
T_3	20	22	24
C_4	3	7	2
C_3T	17	21	17
C_2T_2	36	32	34
CT_3	33	27	34
T_4	12	12	14

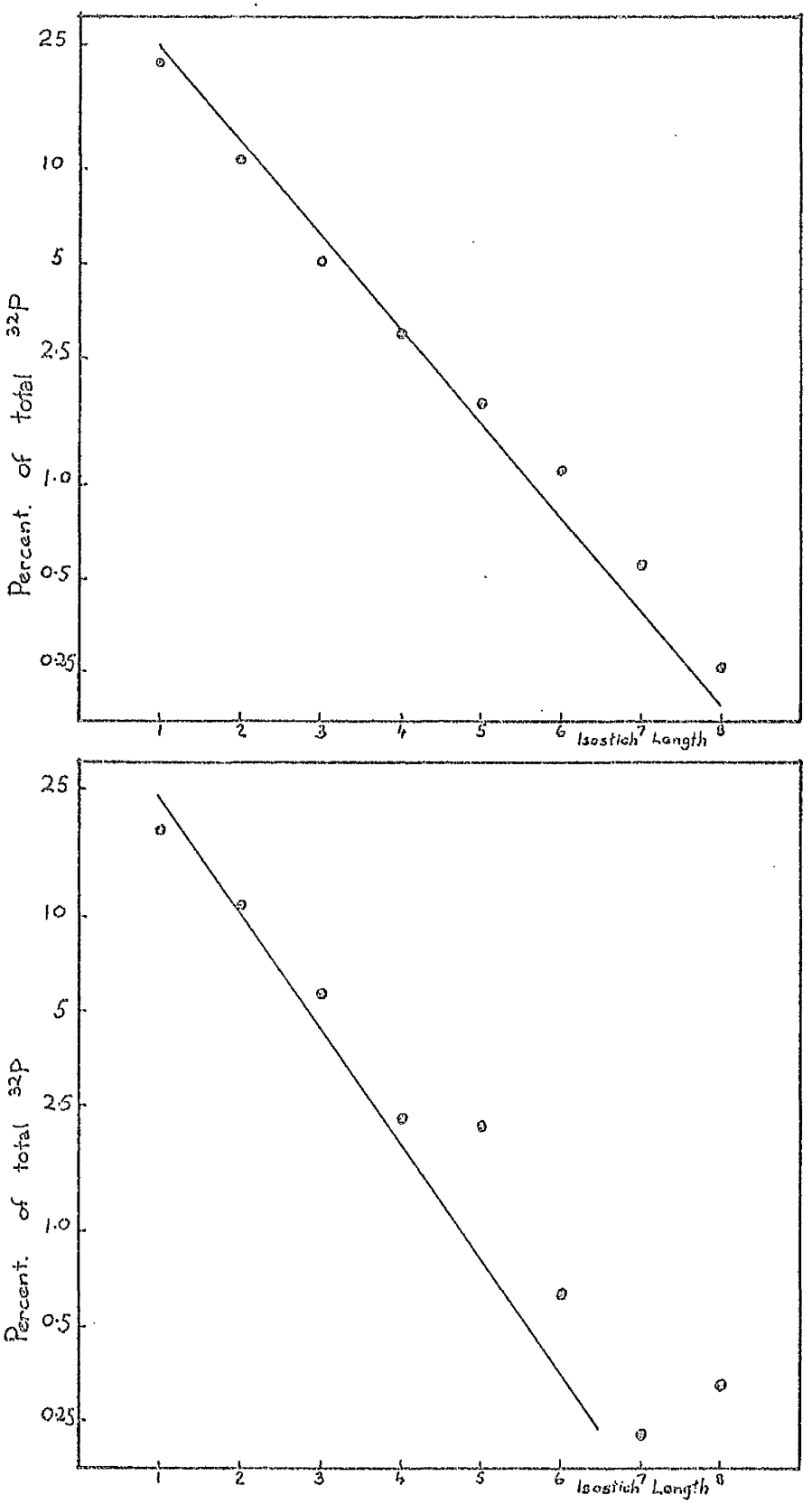
The first column shows the proportions of base composition fractions of pyrimidine isostichs expected in a random sequence, double-stranded DNA of 42% (G+C); frequencies are as percentages of the total material in each isostich. The second and third columns , in the same way, show data for total pyrimidine isostichs, and data for isostichs with G as 3'-neighbour, in calf thymus DNA. These data are based on Table 18.

FIGURE 30. FREQUENCIES OF OCCURRENCE OF PYRIMIDINE ISOSTICHS IN CALF THYMUS DNA



The frequency of each pyrimidine isostich in calf thymus DNA, as number of copies per 100 bases in the DNA, is plotted on a logarithmic scale against the isostich number. The filled circles represent the experimental values, and the straight line represents the random frequency expectation for isostichs from a double-stranded DNA.

FIGURE 31. FREQUENCIES OF OCCURRENCE OF PYRIMIDINE ISOSTICHS [illegible]



This figure is constructed in the same way as Fig. 30. The frequency of occurrence of each fraction is expressed as a percentage of the ^{32}P incorporated into DNA from α-^{32}P-dGTP. The lines represent random expectation; for this purpose MVM (−) strand DNA is assigned an (A+G) content of 57.6%, from DNA polymerase nearest-neighbour analysis.

The excess of observed over expected frequencies of isostichs longer than 5 units in calf thymus DNA is thus due to the high frequency of C rich runs. This phenomenon does not apply to the trends discussed in the last section, since those were concerned with shorter sequences. An interesting, though isolated, observation is that the frequency of Pu-C-C-C-C-G approaches random expectation. It is possible to use these data to make a model for the distribution of CpG in calf thymus DNA as follows: this is speculative in as much as it agrees with the observed facts but is not proven by them.

In calf thymus DNA the low frequency of the first two isostichs can be accounted for by supposing that the only sequences restricted are those with C from CpG as 3'-terminal. It has been observed by Burton & Petersen (1960) and by Doskocil & Sorm (1962) that in acid digests of calf thymus DNA the species CpT is more common than TpC: this observation also can be explained quantitatively as a consequence of low CpG sequences. The low level of CpG and the excess of longer sequences rich in C may, therefore, be complementary aspects of the same phenomenon.

4.4.4 Methylation and Function

Doskocil & Sorm (1962) estimated that, in calf thymus DNA, in the sequences of form Pu-C-Pu 11% of the total C was methylated, and was next to G, and in Pu-C-C-Pu 9% of the 3'-C was methylated, and

was next to G. It can be calculated from the present results that in calf thymus DNA, 16% of Pu-C-Pu is found as Pu-C-G, and 12% of Pu-C-C-Pu as Pu-C-C-G. These estimates agree quite well with those of Doskocil & Sorm (1962) assuming that all MeC occurs in the sequence MeC-G. However, the higher values for CpG sequences may indicate that not all C in CpG is methylated, as was previously supposed. A rigorous test of this possibility would require determinations of the two species made on the same experimental material and by similar methods.

The simplest hypothesis to explain the methylation pattern is that CpG comprises the whole methylation site for the DNA methylases of mammals. Since the data on the surroundings of CpG show that CpG sequences are not found in only one or a few longer sequences, this hypothesis seems reasonable.

The data do not give any clue to the function of the low CpG levels: to say that it serves as an unique methylation site only removes the problem of function one step. They do, however, indicate that, if the CpG sequences present have a positive function, then this function is independent of neighbouring sequences, unless in a way depending on a bias of base composition rather than an unique base sequence. The data also do not give any information on the role of the CpG sequences in protein-specifying DNA. Conceivably this could eventually be resolved by sequence studies on mammalian mRNAs.

SUMMARY

1. Current concepts of genome structure and function in bacteria, viruses and eucaryotic cells are discussed. The technique of nearest-neighbour analysis and its applications are described.

2. Nearest-neighbour patterns and base compositions of three parvovirus DNAs indicate that these DNAs are single stranded. Many features of these nearest-neighbour patterns are similar to those of vertebrate DNA and papovavirus DNA.

3. The nearest-neighbour patterns of DNAs from eight human adenoviruses are very similar. Apart from total base composition changes, no differences were found between DNAs of the non-oncogenic group (Ad 2, 4 and 27), the weakly-oncogenic group (Ad 7, 11 and 21) and the highly-oncogenic group (Ad 12 and 18).

4. Nearest-neighbour patterns for the DNAs of Eubacteria are similar; a distinct pattern is found for DNA from the photosynthetic bacterium Rhodospirillum rubrum.

5. The nearest-neighbour patterns of DNA from Aspergillus nidulans and from Drosophila melanogaster are close to random. DNA from Rana catesbeiana has a highly non-random pattern with a low frequency of the sequence CpG. This pattern is characteristic

of DNAs from all classes of vertebrates examined.

6. The nearest-neighbour pattern of mouse main band DNA is close to that of the total DNA, but mouse satellite DNA shows quite a different pattern. Other features of the satellite DNA are correlated with the nearest-neighbour analysis.

7. Theoretical models are presented for the nearest-neighbour pattern of Escherichia coli DNA, and it is shown that these models account adequately for the observed pattern. Models for vertebrate DNA show that the low CpG frequency exerts a strong influence on the overall pattern.

8. Experiments are described investigating the sequence environment of CpG dinucleotides in vertebrate and virus DNAs. All four bases are found as immediate neighbours of CpG in parvovirus DNA and calf thymus DNA. In parvovirus DNA the sequences to the 3'-side of CpG are pyrimidine rich, and this may also apply to calf thymus DNA.

9. Some features of the set of pyrimidine runs from calf thymus DNA are discussed.

10. The significance of the findings with CpG and possible functions of the low CpG phenomenon are discussed.

REFERENCES

ADAMS, J.M., Jeppesen, P.G.N., Sanger, F. & Barrell, B.G. (1969). Nature, Lond. 223, 1009.

APOSHIAN, H.V. & Kornberg, A. (1962). J. biol. Chem. 237, 519.

ARIMA, T., Uchida, T. & Egami, F. (1968a). Biochem. J. 106, 601.

ARIMA, T., Uchida, T. & Egami, F. (1968b). Biochem. J. 106, 609.

AUGUST, J.T., Shapiro, L. & Eoyang, L. (1965). J. molec. Biol. 11, 257.

BAUERLE, R.H. & Margolin, P. (1966). Proc. natn. Acad. Sci. U.S.A. 56, 111.

BECKWITH, J.R., Singer, E.R. & Epstein, W. (1966). Cold Spring Harbor Symp. quant. Biol. 31, 393.

BELLETT, A.J.D. (1967). J. molec. Biol. 27, 107.

BILLETER, M.A., Dahlberg, J.E., Goodman, H.M., Hindley, J. & Weissman, C. (1969). Nature, Lond. 224, 1083.

BRITTEN, R.J. & Davidson, E.H. (1969). Science, N.Y. 165, 349.

BRITTEN, R.J. & Kohne, D.E. (1968). Science, N.Y. 161, 529.

BURGESS, R.R. (1969). J. biol. Chem. 244, 6160.

BURGESS, R.R., Travers, A.A., Dunn, J.J. & Bautz, E.K.F. (1969). Nature, Lond. 221, 43.

BURTON, K. & Petersen, G.B. (1960). Biochem. J. 75, 17.

CAIRNS, J. (1963). Cold Spring Harbor Symp. quant. Biol. 28, 43.

CALLAN, H. (1967). J. Cell. Sci. 2, 1.

CERNY, R., Mushynski, W.E. & Spencer, J.H. (1968). Biochim. Biophys. Acta 169, 439.

CHAMBERLIN, M. & Berg, P. (1964). J. molec. Biol. 8, 297.

CHEONG, L., Fogh, J. & Barclay, R.K. (1965). Fedn Proc. Fedn Am. Socs exp. Biol. 24, 596.

CLAUSEN, T. (1968). Analyt. Biochem. 22, 70.

CRAWFORD, L.V. (1964). Virology 22, 140.

CRAWFORD, L.V. (1966). Virology 29, 605.

CRAWFORD, L.V., Follett, E.A.C., Burdon, M.G. & McGeoch, D.J. (1969). J. gen. Virol. 4, 37.

CUNNINGHAM, L., Catlin, B.W. & de Garilhe, M.P. (1956). J. Am. chem. Soc. 78, 4642.

DARNELL, J.E. (1968). Bacteriol. Rev. 32, 262.

DE CROMBRUGGHE, B., Perlman, R.L., Varmus, H.E. & Pastan, I. (1969). J. biol. Chem. 244, 5828.

DE WACHTER, R. (1968). J. Chromatog. 36, 109.

DE WACHTER, R. & Fiers, W. (1969). Nature, Lond. 221, 233.

DOERFLER, W. & Kleinschmidt, A.K. (1970). J. molec. Biol. 50, 579.

DOSKOCIL, J. & Sorm, F. (1962). Biochim. biophys. Acta 55, 953.

DOUGHERTY, E.C. (1957). J. Protozool. 4 (suppl.), 14.

DOVE, W.F. (1968). A. Rev. Genet. 2, 305.

DUESBERG, P.H. & Robinson, W.S. (1967). J. molec. Biol. 25, 383.

DuPRAW, E.J. (1968). Cell and Molecular Biology, New York: Academic Press.

ECHOLS, H., Garen, A., Garen, S. & Torriani, A. (1961). J. molec. Biol. 3, 425.

EDELMAN, G.M. & Gall, W.E. (1969). A. Rev. Biochem. 38, 415.

EGAMI, F., Takahashi, K. & Uchida, T. (1964). In Progress in Nucleic Acid Research and Molecular Biology, vol.3, p.59. Ed. by Davidson, J.N. & Cohn, W.E. New York: Academic Press.

EPSTEIN, W. & Beckwith, J.R. (1968). A. Rev. Biochem. 37, 411.

FINAMORE, F.J., & Volkin, E. (1958). Exp. Cell Res. 15, 405.

FITCH, W.M. (1964). Proc. natn. Acad. Sci. U.S.A. 52, 298.

FLAMM, W.G., McCallum, M. & Walker, P.M.B. (1967). Proc. natn. Acad. Sci. U.S.A. 57, 1729.

FLAMM, W.G., Walker, P.M.B. & McCallum, M. (1969). J. molec. Biol. 40, 423.

FOX, C.F., Robinson, W.S., Haselkorn, R. & Weiss, S.B. (1964). J. biol. Chem. 239, 186.

GAREN, A. & Otsuju, N. (1964). J. molec. Biol. 8, 841.

GEIDUSCHEK, E.P. & Haselkorn, R. (1969). A. Rev. Biochem. 38, 647.

GILBERT, W. & Müller-Hill, B. (1966). Proc. natn. Acad. Sci. U.S.A. 56, 1891.

GREEN, M., Pina, M., Kimes, R., Wensink, P.C., MacHattie, L.A. & Thomas, C.A., Jr. (1967). Proc. natn. Acad. Sci. U.S.A. 57, 1302.

GUSSIN, G.N. (1966). J. molec. Biol. 21, 435.

HALL, J.B. & Sinsheimer, R.L. (1963). J. molec. Biol. 6, 115.

HASTINGS, J.R.B. & Kirby, K.S. (1966). Biochem. J. 100, 532.

HAY, J. & Subak-Sharpe, H. (1968). J. gen. Virol. 2, 469.

HAYASHI, M., Hayashi, M.N. & Spiegelman, S. (1963). Proc. natn. Acad. Sci. U.S.A. 50, 664.

HAYES, W. (1968). The Genetics of Bacteria and their Viruses, 2nd ed. Oxford : Blackwell Scientific Publications.

HELLEINER, W.E. & Cohen, H.K. (1970). In preparation.

HENNIG, W. & Walker, P.M.B. (1970). Nature, Lond. 225, 915.

HENSHAW, E.C. (1968). J. molec. Biol. 36, 401.

HILL, L.R. (1966). J. gen. Microbiol. 44, 419.

HILMOE, R.J. (1960). J. biol. Chem. 235, 2117.

HOFFMANN-BERLING, H., Marvin, D.A. & Durwald, H. (1963). Z. Naturforsch. 18b, 876.

HUBERMAN, J.A. & Riggs, A.D. (1968). J. molec. Biol. 32, 327.

JACKSON, J.F., Kornberg, R.D., Berg, P., Rajbhandary, U.L., Stuart, A., A., Khorana, H.G. & Kornberg, A. (1965). Biochim. biophys. Acta 108, 243.

JACOB, F. & Monod, J. (1961). Cold Spring Harbor Symp. quant. Biol. 26, 193.

JACOBSON, K.B. (1964). J. Chromatog. 14, 542.

JEPPESEN, P.G.N., Nichols, J.L., Sanger, F. & Barrell, B.G. (1970). Cold Spring Harbor Symp. quant. Biol. 35, in press.

JONES, A.S., Tittensor, J.R. & Walker, R.T. (1966). Nature, Lond. 209, 296.

JOSSE, J., Kaiser, A.D. & Kornberg, A. (1961). J. biol. Chem. 236, 864.

JOSSE, J. & Swartz, M. (1963). In Methods in Enzymology, vol. 6, p.739. Ed. by Colowick, S.P. & Kaplan, N.O. New York: Academic Press.

KAISER, A.D. & Baldwin, R.L. (1962). J. molec. Biol. 4, 418.

KAY, E.R.M., Simmons, N.S. & Dounce, A.L. (1952). J. Am. chem. Soc. 74, 1724.

KAYE, A.M. & Winocour, E. (1967). J. molec. Biol. 24, 475.

KILHAM, L. & Olivier, L.J. (1959). Virology, 7, 428.

KILHAM, L. (1961). Proc. Soc. exp. Biol. Med. 106, 825.

KIMES, R. & Green, M. (1970). J. molec. Biol. 50, 203.

KING, J.L. & Jukes, T.H. (1969). Science, N.Y. 164, 788.

KIT, S. (1961). J. molec. Biol. 3, 711.

KROON, A.M. (1969). In Handbook of Molecular Cytology, p.943. Ed. by Lima-De-Faria, A., Amsterdam: North Holland Publishing Co.

KURNICK, I. & Herskowitz, A. (1958). J. Cell. Comp. Physiol. 39, 281.

LACY, S. & Green, M. (1964). Proc. natn. Acad. Sci. U.S.A. 52, 1053.

LACY, S. & Green, M. (1965). Science, N.Y. 150, 1296.

LACY, S. & Green, M. (1967). J. gen. Virol. 1, 413.

LASKOWSKI, M., Sr. (1967). In Methods in Enzymology, vol.12, part A, p.281. Ed. by Grossman, L. & Moldave, K. New York: Academic Press.

LWOFF, A. & Tournier, P. (1966). A. Rev. Microbiol. 20, 45.

McCARTHY, B.J. (1965). In Progress in Nucleic Acid Research and Molecular Biology, vol.4, p.129. Ed. by Davidson, J.N. & Cohn, W.E. New York: Academic Press.

MAAS, W.K. & Clark, A.J. (1964). J. molec. Biol. 8, 365.

MAGASANIK, B. (1961). Cold Spring Harbor Symp. quant. Biol. 26, 249.

MAITRA, U., Cohen, S.N. & Hurwitz, J. (1966). Cold Spring Harbor Symp. quant. Biol. 31, 113.

MARKHAM, R. & Smith, J.D. (1952). Biochem. J. 52, 552.

MARMUR, J. (1961). J. molec. Biol. 3, 208.

MARSHALL, R.E., Caskey, C.T. & Nirenberg, M. (1967). Science, N.Y. 155, 820.

MARTIN, M.A. & Hoyer, B.H. (1967). J. molec. Biol. 27, 113.

MAY, P., Niveleau, A., Berger, G. & Brailovsky, C. (1967). J. molec. Biol. 27, 603.

MESELSON, M., Stahl, F.W. & Vinograd, J. (1957). Proc. natn. Acad. Sci. U.S.A. 43, 581.

MIKULSKI, A.J., Sulkowski, E., Stasiuk, L. & Laskowski, M., Sr. (1969). J. biol. Chem. 244, 6559.

MORRISON, J.M., Keir, H.M., Subak-Sharpe, H. & Crawford, L.V. (1967). J. gen. Virol. 1, 101.

MORSE, D.E. & Yanofsky, C. (1968). J. molec. Biol. 38, 447.

NICHOLS, J.L. (1970). Nature, Lond. 225, 147.

OHSAKA, A., Mukai, J-L., & Laskowski, M., Sr. (1964). J. biol. Chem. 239, 3498.

PASTAN, I. & Perlman, R.L. (1968). Proc. natn. Acad. Sci. U.S.A. 61, 1336.

PEREIRA, H.G., Huebner, R.J., Ginsberg, H.S. & Van der Veer, J. (1963). Virology, 20, 613.

PERLMAN, R.L., de Crombrugghe, B. & Pastan, I. (1969). Nature, Lond. 223, 810.

PERUTZ, M.F. & Lehmann, H. (1968). Nature, Lond. 219, 902.

PINA, M. & Green, M. (1965). Proc. natn. Acad. Sci. U.S.A. 54, 547.

RECHLER, M.M. & Martin, R.G. (1970). Nature, Lond. 226, 908.

RICHARDS, G.M. & Laskowski, M., Sr. (1969). Biochemistry, Easton 8, 4858.

RICHARDSON, C.C., Schildkraut, C.L., Aposhian, H.V. & Kornberg, A. (1964). J. biol. Chem. 239, 222.

ROBINSON, D.M. & Hetrick, F.M. (1969). J. gen. Virol, 4, 269.

ROBISON, G.A., Butcher, R.W. & Sutherland, E.W. (1968). A. Rev. Biochem. 38, 149.

SALIVAR, W.O., Tzagoloff, H. & Pratt, D. (1964). Virology 24, 359.

SALOMON, R., Kaye, A.M. & Herzberg, M. (1969). J. molec. Biol. 43, 581.

SANGER, F., Brownlee, G.G. & Barrell, B.G. (1965). J. molec. Biol. 13, 373.

SHAPIRO, H.S. (1968). In Handbook of Biochemistry, p. H52. Ed. by Sober, H.A. Cleveland: The Chemical Rubber Co.

SHAPIRO, H.S. & Chargaff, E. (1957). Biochim. biophys. Acta 26, 608.

SHAPIRO, H.S. & Chargaff, E. (1963). Biochim. biophys. Acta 76, 1.

SHAPIRO, L. & August, J.T. (1965). J. molec. Biol. 11, 272.

SINSHEIMER, R.L. (1959). J. molec. Biol. 1, 43.

SINSHEIMER, R.L. & Lawrence, M. (1964). J. molec. Biol. 8, 289.

SKALKA, A., Fowler, A.V. & Hurwitz, J. (1966). J. biol. Chem. 241, 588.

SMITH, B.J. (1970). J. molec. Biol. 47, 101.

SOUTHERN, E.M. (1970). Nature, Lond. 277, 794.

SPENCER, J.H., Cape, R.E., Marks, A. & Mushynski, W.E. (1968). Can. J. Biochem. 46, 627.

SPENCER, J.H., Cape, R.E., Marks, A. & Mushynski, W.E. (1969). Can. J. Biochem. 47, 329.

SPENCER, J.H. & Chargaff, E. (1963a). Biochim. biophys. Acta 68, 9.

SPENCER, J.H. & Chargaff, E. (1963b). Biochim. biophys. Acta 68, 18.

STEITZ, J.A. (1969). Nature, Lond. 224, 957.

SUBAK-SHARPE, H., Burk, R.R., Crawford, L.V., Morrison, J.M., Hay, J. & Keir, H.M. (1966). Cold Spring Harbor Symp. quant. Biol. 31, 737.

SUBAK-SHARPE, H. & Hay, J. (1965). J. molec. Biol. 12, 924.

SUBAK-SHARPE, H., Shepherd, W.M. & Hay, J. (1966). Cold Spring Harbor Symp. quant. Biol. 31, 583.

SUEOKA, N. (1961). Proc. natn. Acad. Sci. U.S.A. 47, 1141.

SWARTZ, M.N., Trautner, T.A. & Kornberg, A. (1962). J. biol. Chem. 237, 1961.

SZYBALSKI, W. (1970). In RNA Polymerase and Transcription, p.209. Ed. by Silvestri, L. Amsterdam : North Holland Publishing Co.

THIMANN, K.V. (1963). The Life of Bacteria, 2nd ed. New York: The Macmillan Co.

THOMAS, C.A., Jr. & MacHattie, L.A. (1967). A. Rev. Biochem. 36, 485.

TOMLINSON, R.V. & Tener, G.M. (1963). Biochemistry, Easton 2, 697.

TOOLAN, H.W. (1960). Science, N.Y. 131, 1446.

TRAVERS, A.A. (1970). Nature, Lond. 225, 1009.

USATEGUI-GOMEZ, M., Toolan, H.W., Ledinko, N., Al-Lami, F. & Hopkins, M.S. (1969). Virology 39, 617.

UZIEL, M. & Cohn, W.E. (1965). Biochim. biophys. Acta 103, 539.

VAN DER EB, A.J., Van Kesteren, L.W. & Van Bruggen, E.F.J. (1969). Biochim. biophys. Acta N30, 530.

WALKER, P.M.B. & McLaren, A. (1966). Nature, Lond. 211, 486.

WARING, M. & Britten, R.J. (1966). Science, N.Y. 154, 791.

WATSON, J.D. & Crick, F.H.C. (1953). Nature, Lond. 171, 737.

WEIGERT, M.G., Galluci, E., Lanka, E. & Garen, A. (1966). Cold Spring Harbor Symp. quant. Biol. 31, 145.

WEISS, S.B. & Nakamoto, T. (1961). Proc. natn. Acad. Sci. U.S.A. 47, 1400.

WESTPHAL, H. (1970). J. molec. Biol. 50, 407.

WOESE, C.R. (1967). The Genetic Code. New York: Harper & Row.

WOESE, C.R. (1970). In Organisation and Control in Prokaryotic and Eucaryotic Cells, p.39. Ed. by Charles, H.P. & Knight, B.C.J.G. Cambridge: Cambridge University Press.

WYATT, G.R. (1951). Biochem. J. 48, 584.

SOME BASE SEQUENCE CHARACTERISTICS OF DEOXYRIBONUCLEIC ACIDS

Duncan J. McGeoch

Nearest-Neighbour Analyses : Some aspects of base sequences in DNAs from different sources were examined by nearest-neighbour analysis.

The nearest-neighbour patterns and base compositions of three parvovirus DNAs indicate that these DNAs are single-stranded. Many features of these patterns are similar to those for vertebrate DNAs and papovavirus DNAs.

The DNAs from eight human adenoviruses give closely similar nearest-neighbour patterns. Apart from total base composition changes, no large differences were detected between DNAs of the non-oncogenic group, the weakly oncogenic group and the higly oncogenic group.

Nearest-neighbour patterns of DNAs from different Eubacteria are similar; DNA from the photosynthetic bacterium Rhodospirillum rubrum gives a distinct pattern.

The nearest-neighbour patterns of DNA from Aspergillus nidulans and from Drosophila melanogaster are close to random. DNA from Rana catesbeiana gives a highly non-random pattern with a low frequency of the sequence CpG. This pattern is characteristic of DNAs from all classes of vertebrates examined.

The nearest-neighbour pattern of mouse main band DNA is close to that of the total DNA, but mouse satellite DNA gives a different pattern. Other features of the satellite DNA can be correlated with the nearest-neighbour results.

Theoretical models have been constructed for the nearest- neighbour

pattern of Escherichia coli DNA, which account adequately for the observed pattern. Models for vertebrate DNA show that the low CpG frequency exerts a strong influence on the overall pattern.

The CpG Shortage in Vertebrate and Virus DNAs : In vertebrate DNA and in the DNAs of some small viruses of animals the dinucleotide CpG occurs with exceptionally low frequency. This phenomenon has been investigated by examining the longer sequences in which CpG does occur. The general approach adopted was to make, in vitro, radioactive RNA or DNA using animal or virus DNA as template for suitable polymerases. Different aspects of the sequence structure of the template DNA were then examined by specific degradation and fractionation of the labelled copy.

It has been found that, in single-stranded DNA from the parvovirus MVM and in double-stranded DNA from calf thymus, all four bases are found as immediate neighbours of CpG sequences. In MVM DNA the sequences to the 3'-side of CpG are pyrimidine rich; this probably applies also to calf thymus DNA. Therefore, if all the CpG sequences present have a positive function, this function is not dependent on neighbouring sequences, unless in a manner utilising a bias of base composition rather than an unique sequence.

Calf thymus DNA contains a non-random distribution of pyrimidine runs; an analysis of this phenomenon indicates a possible relation with the CpG shortage.

Department of Biochemistry,
University of Glasgow,
December 1970.