THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 237, No. 6, June 1962
Printed in U.S.A.

Enzymatic Synthesis of Deoxyribonucleic Acid*

XI. FURTHER STUDIES ON NEAREST NEIGHBOR BASE SEQUENCES IN DEOXYRIBONUCLEIC ACIDS

M. N. SWARTZ† T. A. TRAUTNER,‡ AND ARTHUR KORNBERG

From the Department of Biochemistry, Stanford University School of Medicine, Palo Alto, California

(Received for publication, January 26, 1962)

The examination of nearest neighbor base sequences in deoxyribonucleic acid (DNA) with the technique described by Josse, Kaiser, and Kornberg (1) established (*a*) that DNA from a given source directs the synthesis of a product in which the four bases occur next to one another in the 16 possible arrangements, not at random but in a pattern of frequencies unique for that DNA; and (*b*) that enzymatically synthesized DNA, and by inference the primer DNA, shows complementary pairing of adenine to thymine and of guanine to cytosine between two strands of opposite polarity, as proposed in the Watson and Crick model.

This paper describes experiments with the same technique undertaken to explore the following questions. (*a*) Does the replication of a single stranded DNA (with noncomplementary base composition) proceed by base pairing as in double stranded DNA? (*b*) Are the nearest neighbor sequence patterns in the DNA's from different tissues and tumors of a given species the same? (*c*) Do sequence patterns in the DNA's of various biological forms reveal any relationships among or within groups of organisms?

EXPERIMENTAL PROCEDURE

Materials

Substrates and Enzymes—Labeled and unlabeled deoxynucleoside triphosphates, micrococcal DNase, and calf spleen diesterase were prepared as described previously (2–5). The DNA-synthesizing enzymes, *Escherichia coli* polymerase and the T2 phage-induced polymerase, used in most experiments were prepared from Fraction VII, refractionated on diethylaminoethyl cellulose (DEAE-cellulose) (2) or on phosphocellulose resin (6); the specific activities of these enzymes were approximately 1,500. *E. coli* DNA endonuclease (7) (carboxymethyl cellulose fraction, 2,500 units per ml) and *E. coli* DNA phosphodiesterase (8) (DEAE fraction, 20,000 units per ml) were kindly supplied by Dr. I. R. Lehman. Crystalline pancreatic DNase was purchased from the Worthington Biochemical Corporation.

DNA Preparations—The DNA's used in these experiments generally had ϵ_M values (based on deoxypentose) ranging between 6.3 and 7.5; their protein contents were in most cases less than 6% as determined by the method of Lowry *et al.* (9). Unless otherwise noted, DNA's were isolated by homogenizing tissues or cells in a Waring Blendor in 0.15 M NaCl + 0.01 M sodium citrate and centrifuging according to the procedure of Kay, Simmons, and Dounce (10). The crude tissue fractions thus obtained were again stirred in the Waring Blendor and then subjected either to repeated shaking with chloroform-octanol (11) and precipitation with 95% ethanol, or to treatment with sodium lauryl sulfate according to the method of Kay, Simmons, and Dounce (10); in some cases both procedures were used. When necessary, purified DNA's were treated with pancreatic RNase to remove RNA.

DNA's of bacteriophages T1 (T1 phage was kindly provided by H. Modersohn) and T5 were prepared by phenol treatment of virus stocks initially purified by differential centrifugation and freed from bacterial DNA by treatment with pancreatic DNase. *Paracentrotus lividus* (sea urchin) DNA was obtained by extraction with 3 M NaCl of a sperm homogenate, kindly supplied by Dr. R. Hinegardner. Release of DNA from bull sperm, kindly provided by Dr. S. W. Mead, required treatment with 1 N NaOH for 5 hours at 37°; after neutralization, the DNA was precipitated with 95% ethanol and further purified.

We are indebted to Dr. R. L. Sinsheimer for DNA of bacteriophage ΦX 174 (ΦX); to Dr. N. Sueoka for *Tetrahymena pyriformis* DNA, prepared essentially by the Marmur method (12), and for *Cancer borealis* (crab) testis DNA; to Dr. R. Sager for *Chlamydomonas* DNA prepared by a sodium lauryl sulfate procedure; and to Dr. K. Burton for *Echinus esculenta* (sea urchin) DNA.

Methods

Nearest Neighbor Frequency Analysis—This method (1) in volves enzymatic replication of a given DNA primer and employs as substrates, deoxyribonucleoside triphosphates in which the sugar-esterified 5′-phosphate is labeled with P^{32}. Subsequent cleavage between the phosphate and carbon 5′ of the synthesized polynucleotide chains yields P^{32}-labeled 3′-mononucleotides; P^{32} introduced into the DNA by the substrate nucleotide consequently labels the adjacent nucleotide, its "nearest neighbor." By determining the P^{32} content of each of the four 3′-mononucleotides isolated from the digested product of a reaction with a given labeled substrate, one can calculate the frequency with which this nucleotide is linked to each of the four nucleotides found in the DNA.

The methods of enzymatic synthesis of DNA, digestion to 3′-mononucleotides, separation of 3′-mononucleotides, and calculation of nearest neighbor frequencies were similar to those reported

* This investigation was supported by research grants from the National Institutes of Health, United States Public Health Service.

† Present address, Massachusetts General Hospital, Boston, Massachusetts.

‡ Fellow of Deutsche Forschungsgemeinschaft; present address, Institut fur Genetik, Universitat Koln, Koln-Lindenthal, Germany.

earlier (1). As before, the extent of DNA synthesis generally represented a 20% increment over the amount of primer added.

Determination of Experimental Error—An estimate of error in determination of nearest neighbor frequencies was obtained by duplicate analyses of mouse thymus, mouse lymphoma, starfish testis, and salmon sperm DNA samples. Standard deviations, expressed as percentage of the mean ("coefficients of variation"), were calculated for each of the 16 dinucleotide sequences in each of the four duplicate runs; the average of the coefficients of variation obtained for each sequence was found to vary between 2.33 and 10.0%. The over-all average coefficient of variation was 5.8%.

Deviations in Total Isolated Ap and Tp Nucleotides—The total amount of Tp (TpA + TpT + TpG + TpC) has consistently exceeded the total Ap (ApA + ApT + ApG + ApC) by 2 to 20%. This deviation was not observed in our earlier studies (1), which, although largely confined to bacterial DNA's, also included calf thymus DNA. Since the enzymatic digestions and chromatographic procedures were complete and quantitative, it was possible that the character of the polymerase preparation used or the state of the DNA primer might be responsible for this deviation. These factors were explored by (*a*) the use of different polymerase preparations and (*b*) pretreatment of the DNA primer with heating or nuclease cleavage.

Mouse ascites tumor cell DNA was analyzed with three different *E. coli* polymerase preparations, and heated salmon sperm DNA, with the distinctive polymerases from normal and T2 phage-infected cells. The coefficients of variation in these analyses did not exceed the experimental error (see above).

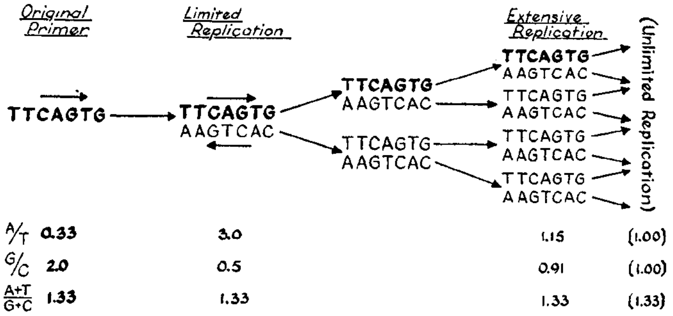

FIG. 1. Scheme for replication of single stranded DNA. An arbitrarily selected sequence of bases in a hypothetical single stranded DNA primer is designated in *bold print*. The base ratios for limited and extensive replication refer to the values for the newly synthesized DNA molecules, designated by *standard print*.

TABLE I

Composition of products after limited and extensive replication of ΦX DNA

Base	Composition determined by chemical analysis*	Composition determined by nearest neighbor analysis				
		20% synthesis		600% synthesis		
		Predicted from chemical analysis	Observed	Predicted† from chemical analysis	Predicted† from 20% synthesis	Observed
A	0.246	0.328	0.310	0.287	0.276	0.271
T	0.328	0.246	0.242	0.287	0.276	0.293
G	0.242	0.185	0.202	0.214	0.224	0.213
C	0.185	0.242	0.246	0.214	0.224	0.224

* See (13).

† Based on unlimited replication (see Fig. 1).

A comparison of heated with native calf thymus DNA showed no significant change in sequence frequencies in the earlier study (1) or in the present one. With heated salmon sperm DNA compared with native, the coefficients of variation for three of the sequences (GpA, TpC, and CpT) differed by more than 10%. Pretreatment of calf thymus DNA with *E. coli* endonuclease, a variable contaminant of polymerase preparations, had a profound effect upon the sequence frequencies; after endonuclease treatment until 53% of the nucleotides were released, coefficients of variations between treated and untreated DNA were greater than 10% for six of the sequences. However, prior action by pancreatic DNase, *E. coli* phosphodiesterase, or a polymerase preparation on calf thymus DNA did not seem to alter the sequence frequencies significantly.

It must be concluded that, although distortions of sequence frequencies can be introduced by alteration of the primer, the basis for the systematic deviations between the Tp and Ap analyses encountered in these studies is not yet clear. These deviations, however, do not alter the principal conclusions which may be drawn from the experiments to be reported.

RESULTS

Replication of Single Stranded Primer—The DNA of phage ΦX has been demonstrated by Sinsheimer (13) to be single stranded and to be further distinguished by the absence of equivalence between A and T and between G and C. The DNA is a primer for polymerase and can lead to 10-fold or greater net synthesis of a product which has the characteristics of double stranded DNA (14). Two nearest neighbor analyses were carried out, one under conditions of *limited replication* (20% increase over the amount of primer), and the other with *extensive replication* (600% increase).

In *limited replication*, DNA synthesis will be directed mainly by original primer molecules. In order to minimize participation of newly synthesized, double stranded molecules as primers, synthesis was restricted to a 20% increase over the initial primer added. If such replication follows the base composition of the primer, the nucleotide composition of the newly synthesized product, identified by its P^{32} label, should be the complement of the base composition of the primer; *i.e.* the A content of the product should be equal to the T content of the primer, and the T content of the product to the A content of the primer (Fig. 1). The A:T ratio of the product should therefore be the reciprocal of the A:T ratio of the primer. Similarly, the G:C ratio of the product should be the reciprocal of the G:C ratio of the primer. As a consequence, the ratio, (A + T)/(G + C), of the product should be identical with that of the primer. The results in Table I show that these predictions are fulfilled.

In *extensive replication*, the priming molecules after the initial period are double stranded, and accordingly their replication should yield a product with identical nucleotide composition. Specifically, equivalence of purine to pyrimidine nucleotides in the product (A = T, G = C) should obtain, and the A (or T) value, for example, should be equal to one-half the sum of the A and T values of the primer (or *limited replication* product). Thus, the mole fraction of A in the original strand was 0.246, and that of T, 0.328; consequently, the mole fractions in the complementary strand will be 0.328 for A and 0.246 for T. The over-all mole fractions should be, A = (0.246 + 0.328)/2 and T = (0.328 + 0.246)/2. These predictions are borne out by the results obtained (Table I).

Individually, the nearest neighbor frequencies obtained under both conditions of synthesis support the replication mechanism discussed here. Matching of complementary base pairs (in the P^{32}-labeled product) is observed under conditions of *extensive replication* but not in *limited replication* (Table II). The frequencies of matching nearest neighbor pairs are close to values predicted from the frequencies obtained in *limited replication* by the same reasoning which had been applied to predict over-all base composition in 600% synthesis. Since each of the four sequencies, TpA, ApT, CpG and GpC, is its own match (1), the frequencies of each of these sequences should remain unaltered, whether replication is limited or extensive; the results fit this prediction closely.

Sequence Frequencies in T1 and T5 Bacteriophage DNA—The base composition and the nearest neighbor pattern of the DNA from temperature coliphage λ are similar to those of its host, *E. coli* (1). The DNA's of the virulent T-even phages, T2, T4, and T6, have identical base compositions; their nearest neighbor patterns differ from those of *E. coli* (1). The sequence frequencies in the DNA of T1, a phage intermediate between the typically temperate and virulent phages, and that of T5, a virulent phage similar in many respects to the T-even phages, can be distinguished from one another (Table III) but do not deviate strikingly from values for random association of the nucleotides. The most tenable conclusion to be drawn from these comparisons of sequence patterns is to regard differences as significant but identity as no more revealing than identity of the base compositions.

Sequence Frequencies in DNA of Several Tissues of a Species—The values for DNA's of three different bovine organs and of four mouse tissues, including two tumors, are compared in Tables IV and V. The coefficients of variation for the bovine DNA's and for the mouse DNA's were not significantly different from those obtained for repeated analyses of identical DNA's. Therefore, any differences that may exist among the several bovine DNA's or among any of the mouse DNA samples are within the experimental error of the analysis.

Sequence Frequencies in Crab Testis DNA—Examination of the DNA of crab testes has revealed discrete and distinctive components. Upon density gradient centrifugation of DNA isolated from the testes of five specimens of *Cancer borealis*, Sueoka (15) found two components, the lighter one representing as much as 30% of the total DNA. The buoyant density of the main band corresponded to a G-C content of 42%, which is characteristic of crab DNA. However, the buoyant density of the minor component indicated a low G-C content (20% or less) and was in fact like the density of dAT polymer, an alternating copolymer of A and T synthesized *de novo* by polymerase (16). In order to eliminate the possibility that adventitious materials, such as protein, might be responsible for the low buoyant density of the DNA band, and with the thought that this band might even be a "natural" dAT polymer, Dr. Sueoka asked us to examine nearest neighbor frequency patterns of the crab DNA preparations.

The light component of *C. borealis* primed DNA synthesis at a rate comparable to dAT, but unlike the latter, all four deoxynucleoside triphosphates were required. When dGTP and dCTP were omitted, the rate of synthesis was only 19% of that

TABLE II

Nearest neighbor frequencies of ΦX DNA in limited and extensive replication*

Nearest neighbor sequence	Limited replication (20%): Observed	Extensive replication (600%)	
		Predicted from limited replication†	Observed
ApA, TpT	0.101, 0.069	0.085, 0.085	0.085, 0.099
CpA, TpG	0.096, 0.048	0.072, 0.072	0.070, 0.070
GpA, TpC	0.054, 0.064	0.059, 0.059	0.058, 0.065
CpT, ApG	0.052, 0.069	0.061, 0.061	0.064, 0.058
GpT, ApC	0.047, 0.068	0.057, 0.057	0.053, 0.053
GpG, CpC	0.040, 0.053	0.046, 0.046	0.041, 0.045
TpA	0.061	0.061	0.059
ApT	0.072	0.072	0.075
CpG	0.045	0.045	0.045
GpC	0.061	0.061	0.061

* Expressed, in this and subsequent tables, as decimal proportions of 1.000.

† Based on unlimited replication (see Fig. 1).

TABLE III

Nearest neighbor frequencies in DNA's of T1 and T5 bacteriophages and E. coli

Nearest neighbor sequence	(0.99)* *E. coli* B_a (1.01)	(0.99) *E. coli* B_b (1.06)	(1.08) Bacteriophage T1 (1.20)	(1.57) Bacteriophage T5 (1.70)
ApA, TpT	0.071, 0.076	0.072, 0.082	0.079, 0.093	0.105, 0.100
CpA, TpG	0.071, 0.071	0.072, 0.070	0.065, 0.069	0.058, 0.057
GpA, TpC	0.055, 0.056	0.056, 0.062	0.062, 0.065	0.054, 0.060
CpT, ApG	0.055, 0.055	0.051, 0.047	0.060, 0.053	0.064, 0.056
GpT, ApC	0.055, 0.054	0.054, 0.055	0.056, 0.054	0.048, 0.052
GpG, CpC	0.056, 0.056	0.051, 0.059	0.046, 0.044	0.035, 0.038
TpA	0.051	0.050	0.057	0.098
ApT	0.068	0.076	0.076	0.103
CpG	0.067	0.068	0.058	0.032
GpC	0.083	0.074	0.063	0.043

* The number in parentheses above each heading is the chemically determined (A + T)/(G + C) ratio for the DNA; the number in parentheses below each heading is the (A + T)/(G + C) ratio determined by nearest neighbor analysis. The two determinations for *E. coli* B involved the use of different DNA primer samples and different polymerase preparations.

TABLE IV

Nearest neighbor frequencies in DNA's of bovine tissues

Nearest neighbor sequence	(1.29)* Thymus (1.28)	(1.24) Liver (1.29)	(1.39) Sperm (1.35)
ApA, TpT	0.080, 0.085	0.079, 0.090	0.084, 0.094
CpA, TpG	0.078, 0.076	0.072, 0.077	0.077, 0.069
GpA, TpC	0.066, 0.072	0.063, 0.071	0.059, 0.069
CpT, ApG	0.074, 0.069	0.079, 0.071	0.073, 0.069
GpT, ApC	0.049, 0.053	0.052, 0.051	0.049, 0.052
GpG, CpC	0.051, 0.060	0.053, 0.057	0.050, 0.060
TpA	0.052	0.055	0.062
ApT	0.074	0.071	0.076
CpG	0.016	0.015	0.014
GpC	0.045	0.046	0.043

* See Table III.

observed with the four triphosphates; when dTTP was also omitted from the incubation mixture, the rate was reduced to less than 0.1%. These results suggested at once that at least a few G and C residues were interspersed in the chains of the light crab DNA.

TABLE V

Nearest neighbor frequencies in DNA's of mouse tissues and tumors

Nearest neighbor sequence	(1.38)* Thymus† (1.43)	(1.38) Liver (1.42)	(1.38) Lymphoma (1.43)	(1.38) Ascites tumor (1.43)
ApA, TpT	0.091, 0.101	0.088, 0.093	0.091, 0.094	0.084, 0.100
CpA, TpG	0.076, 0.083	0.072, 0.078	0.077, 0.079	0.074, 0.074
GpA, TpC	0.060, 0.062	0.061, 0.063	0.059, 0.065	0.059, 0.068
CpT, ApG	0.076, 0.070	0.075, 0.070	0.075, 0.071	0.072, 0.072
GpT, ApC	0.057, 0.053	0.057, 0.054	0.051, 0.053	0.056, 0.050
GpG, CpC	0.051, 0.046	0.051, 0.050	0.052, 0.051	0.048, 0.052
TpA	0.060	0.067	0.063	0.062
ApT	0.072	0.075	0.075	0.078
CpG	0.011	0.009	0.011	0.011
GpC	0.038	0.039	0.037	0.039

* See Table III.

† Each value is the average of two analyses.

TABLE VI

Nearest neighbor frequencies of two DNA components of Cancer borealis (crab) testis

Nearest neighbor sequence	Main component (1.80)*	Light component† (36.6)*
ApA, TpT	0.085, 0.092	0.0127, 0.0126
CpA, TpG	0.067, 0.066	0.0100, 0.0089
GpA, TpC	0.053, 0.054	0.0042, 0.0015
CpT, ApG	0.061, 0.055	0.0004, 0.0018
GpT, ApC	0.057, 0.063	0.0081, 0.0069
GpG, CpC	0.032, 0.038	0.0009, 0.0009
TpA	0.113	0.504
ApT	0.116	0.429
CpG	0.019	0.0007
GpC	0.030	0.0015

* The number in parentheses is the (A + T)/(G + C) ratio determined by nearest neighbor analysis.

† Each value is the average of two analyses.

Nucleotide incorporation into DNA in the first stage of the nearest neighbor analysis was in the ratio, A:T:G:C = 0.84:1.07:0.030:0.028 mμmoles, indicating a G-C content of approximately 3%. A similar value was obtained by determining the actual nearest neighbor frequencies (Table VI).

The most remarkable result of the nearest neighbor analysis is that the light crab DNA appears very similar to the dAT copolymer, with alternating A and T residues comprising 93% of the sequences. However, all 16 possible sequences are observed, and in strikingly nonrandom distribution. The matching of sequences follows the predictions of Watson-Crick base pairing, except in those instances in which the very low frequencies are difficult to measure accurately.

Also included in Table VI are the nearest neighbor frequencies of the main, heavy component of *Cancer borealis* DNA. The ratio, (A + T)/(G + C), was 1.8, compared with a value of 1.6 calculated by Sueoka from the buoyant density. Although the heavy crab DNA showed no gross contamination with the light component, trace amounts might have escaped detection. Since the light component is a better primer, the reaction product of a nearest neighbor analysis of the heavy component would appear to have a relatively higher A + T content (and higher ApT and TpA sequences) than that of the heavy primer.

The possibility might also be considered that the light component is pure dAT and that its contamination by the heavy component is responsible for the presence of G and C residues. This is unlikely, however, since replication of the light component, as mentioned, is markedly reduced when G and C are omitted from the reaction mixture. Furthermore, the nearest neighbor sequences involving G and C are distinctly different in the two DNA components.

Sequence Frequencies in DNA of Various Animal and Plant Species and Bacteria Compared—Analyses of DNA's from several animal and plant species are shown in Tables VII and VIII.

In the previous study (1), the frequency patterns of DNA's from six bacteria, five phages, and calf thymus were analyzed for their fit to frequencies predicted for a random arrangement of bases in DNA molecules. In random ordering of the nucleotides, the frequency of any nearest neighbor pair should be predictable as the product of the frequencies of its constituent mononucleotides (*e.g.* fApT = fTpA = fAp × fTp). Although most of the observed sequence frequencies fell within these predictions, several differed sharply. This is true also in the frequency pat-

TABLE VII

Nearest neighbor frequencies of animal and plant DNA's

Nearest neighbor sequence	(0.56)* *Chlamydomonas* (0.87)	(1.15) Wheat germ (1.21)	(1.65) *Echinus esculenta* (sea urchin) (1.66)	(1.85) *Paracentrotus lividus* (sea urchin) (1.80)	(3.00) *Tetrahymena pyriformis* (3.25)
ApA, TpT	0.060, 0.059	0.072, 0.089	0.100, 0.116	0.110, 0.102	0.153, 0.176
CpA, TpG	0.077, 0.073	0.068, 0.070	0.076, 0.069	0.067, 0.067	0.045, 0.044
GpA, TpC	0.044, 0.046	0.062, 0.071	0.049, 0.064	0.057, 0.059	0.045, 0.048
CpT, ApG	0.057, 0.060	0.067, 0.060	0.064, 0.047	0.059, 0.053	0.052, 0.049
GpT, ApC	0.055, 0.060	0.056, 0.053	0.053, 0.062	0.053, 0.058	0.034, 0.036
GpG, CpC	0.071, 0.074	0.051, 0.059	0.031, 0.046	0.033, 0.038	0.016, 0.017
TpA	0.053	0.058	0.075	0.090	0.127
ApT	0.054	0.075	0.091	0.104	0.133
CpG	0.063	0.039	0.022	0.020	0.007
GpC	0.092	0.050	0.035	0.031	0.020

* See Table III.

TABLE VIII

Nearest neighbor frequencies of animal tissue DNA's

Nearest neighbor sequence	(1.42)* Human spleen (1.47)	Rabbit liver (1.35)	(1.34) Chicken red cell (1.36)	(1.43) Salmon liver (1.33)	Starfish testis (1.45)
ApA, TpT	0.097, 0.097	0.091, 0.096	0.087, 0.097	0.074, 0.083	0.102, 0.104
CpA, TpG	0.074, 0.074	0.073, 0.077	0.078, 0.077	0.077, 0.075	0.072, 0.067
GpA, TpC	0.061, 0.057	0.060, 0.060	0.053, 0.060	0.057, 0.067	0.058, 0.056
CpT, ApG	0.071, 0.070	0.071, 0.069	0.077, 0.068	0.075, 0.069	0.058, 0.058
GpT, ApC	0.049, 0.054	0.051, 0.049	0.050, 0.054	0.062, 0.062	0.064, 0.060
GpG, CpC	0.050, 0.047	0.062, 0.047	0.048, 0.054	0.049, 0.052	0.043, 0.050
TpA	0.067	0.059	0.062	0.068	0.066
ApT	0.081	0.074	0.072	0.073	0.078
CpG	0.010	0.013	0.011	0.017	0.025
GpC	0.043	0.048	0.052	0.040	0.038

* See Table III.

terns presented in this paper; variances from predicted (random) frequencies for the 16 nearest neighbor pairs were found to be considerably greater than experimental error in the cases shown in Table IX. A comparison of variances of isomeric nearest neighbor pairs such as ApT *versus* TpA, for example, reveals sharp differences.

Fig. 2 is an attempt to determine whether variances from random distribution have phylogenetic significance. The *observed* nearest neighbor frequencies are plotted against the values *predicted* from random association, the product of frequencies of constituent nucleotides of a nearest neighbor pair.[1] If there were agreement between observed and predicted values, the points for the sequences whose constitutent nucleotides occur with equal frequency would fall on a straight line through the origin with a slope of 1. As seen in Fig. 2, the sequence, CpG, occurs in animal and plant cells with a frequency which is in-

TABLE IX

Variances from random nearest neighbor frequencies in animal, plant, and bacterial DNA's

Nearest neighbor sequence	12 animal and plant DNA's					6 bacterial DNA's				
	Deviation from random frequency*			Variance × 10^4		Deviation from random frequency*			Variance × 10^4	
	+	0	−	Observed†	Calculated from error‡	+	0	−	Observed†	Calculated from error‡
ApA	11		1	0.74	0.20	4		2	1.35	0.04
TpT	10		2	1.13	0.21	4		2	2.04	0.04
CpC	11	1		0.27	0.14			6	1.87	0.07
GpG	12			0.62	0.16			6	1.83	0.07
CpA	12			2.62	0.01	5	1		0.52	0.05
ApC	1	1	10	0.49	0.09	2		4	1.03	0.05
TpG	11		1	2.12	0.25	6			0.56	0.05
GpT	2		10	0.77	0.09	2		4	1.01	0.05
GpA	3	1	8	1.05	0.01	3		3	0.77	0.05
ApG	8		4	0.97	0.16			6	0.58	0.05
TpC	3	3	6	0.38	0.09	3		3	0.55	0.05
CpT	7	2	3	0.67	0.09			6	0.61	0.05
TpA			12	4.30	0.21			6	3.69	0.04
ApT	1	2	9	0.80	0.19	6			0.13	0.04
CpG			12	6.20	0.32	6			1.88	0.07
GpC	5	1	6	0.59	0.14	6			3.26	0.07

* Key: +, observed frequency > random expectancy; 0, observed frequency = random expectancy (±0.005); −, observed frequency < random expectancy.

† $\frac{\Sigma(f_{\text{obs}} - f_{\text{pred}})^2}{n - 1}$.

‡ $\frac{\Sigma(sf_{\text{pred}})^2}{n - 1}$; n = number of entries; s = coefficient of variation (for animal and plant DNA's, determined as in "Methods"; for bacteria, an average coefficient of variation of 3% was assumed for each nearest neighbor frequency).

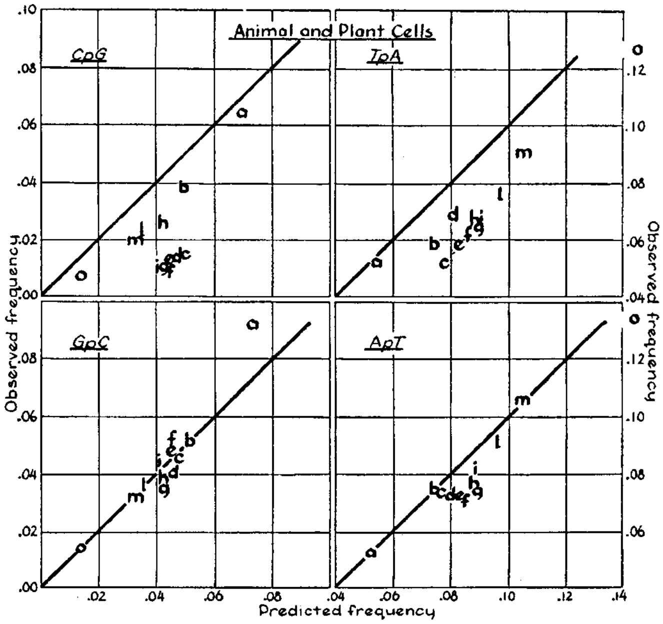

FIG. 2. Sequence frequencies in animal and plant cell DNA's compared with values predicted from random association. The product of frequencies of constituent nucleotides of a nearest neighbor pair is described by the line passing through the origin. The values represented by the *lower case letters* are the observed nearest neighbor frequencies. *a*, *Chlamydomonas*; *b*, wheat germ; *c*, calf thymus; *d*, salmon liver; *e*, rabbit liver; *f*, chicken red cells; *g*, mouse lymphoma; *h*, starfish testis; *i*, human spleen; *l*, *Echinus esculenta*; *m*, *Paracentrotus lividus*; *o*, *Tetrahymena pyriformis*.

[1] This method of plotting the data was suggested by the observation of A. D. Kaiser and R. L. Baldwin that the nearest neighbor frequencies are related to the product of the base frequencies (submitted for publication to the *Journal of Molecular Biology*).

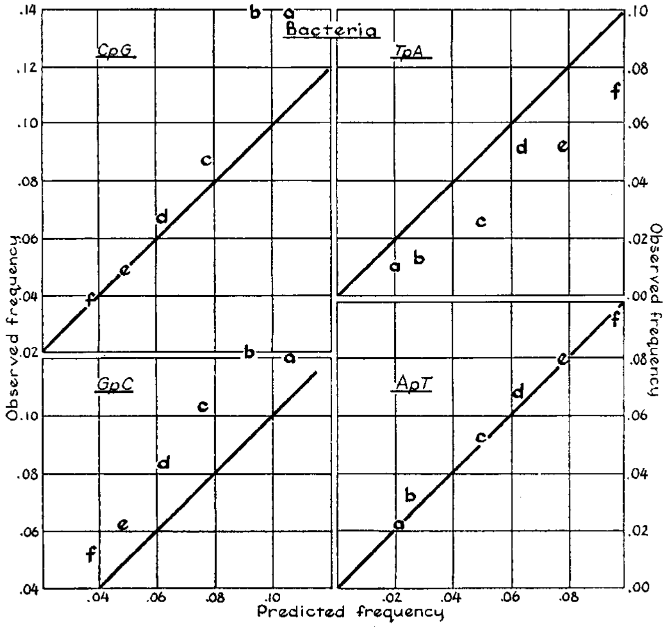

FIG. 3. Sequence frequencies in bacterial DNA's compared with values predicted from random association, as in Fig. 2. *a, Mycrococcus lysodeikticus*; *b, Mycobacterium phlei*; *c, Aerobacter aerogenes*; *d, Escherichia coli*; *e, Bacillus subtilis*; *f, Haemophilus influenzae.*

variably less than random and in some cases only one-third the random value; by contrast, the frequencies of the isomeric sequence, GpC, come very close to the expectation for randomness. In bacteria, the reverse of this pattern appears to be the case (Fig. 3).

In the case of the TpA and ApT sequence frequencies, bacterial DNA's as well as animal and plant cell DNA's show a TpA frequency that is decidedly below the prediction for random association; the ApT frequencies, however, are close to the predicted values (Figs. 2 and 3).

DISCUSSION

Mechanism of Enzymatic Replication of Single Stranded DNA—Replication of the single stranded DNA from phage ΦX was studied by analysis of a product after 20% increase in DNA over the primer added (*limited replication*) and after a 600% increase (*extensive replication*). The base composition of the DNA synthesized under conditions of limited replication was complementary to that of the primer. A comparison of nearest neighbor frequencies after extensive replication with those from limited replication showed that the nearest neighbor sequences also conformed with complementary base pairing with the primer. The fact that a copy of only 20% of the primer chains is representative of the total can be interpreted in two ways: (*a*) only 1 out of 5 molecules primes and is completely replicated or (*b*) replication of an average of $\frac{1}{5}$ of the 5000 nucleotide sequences of *each* molecule is representative of the entire molecule. Density gradient centrifugation experiments indicating that virtually all (>85%) of the ΦX DNA is combined with new DNA when a 20% increase is reached[2] make the first interpretation unlikely and the second preferred.

Earlier viscometric and optical measurements had shown that a double helical molecule is produced from the single stranded primer (14). The chemical determinations presented here indicate that the enzymatically synthesized strand (of the size of the "20% synthesis product") already primes as effectively as the original strand of phage DNA and that both strands of a double helix may prime in the enzymatic replication.

Although these findings are restricted to the product of enzymatic action *in vitro*, they are consistent with Sinsheimer's (17) recent isolation from infected cells of a "replicative form" of ΦX DNA which has many features of a double helix. It would follow from these observations that at least the initial step in replication of ΦX DNA is like that of other DNAs' in the formation of a complementary strand and that some unique process is responsible for including only one type of strand into mature phage.

Nearest neighbor sequence analyses of DNA have now been performed with an RNA polymerase directed specifically by DNA (18–20). The composition and the sequence frequencies of enzymatically synthesized RNA's primed by various DNA's (dAT, dGdC, DNA's of phage ΦX, phage T2, calf thymus) are identical with those obtained with DNA polymerase. Synthesis of this type of RNA seems clearly to involve pairing of uracil to adenine and cytosine to guanine in the manner of DNA synthesis by *E. coli* polymerase.

A bothersome, frequent deviation in the current studies was the absence of matching between certain of the complementary sequences, resulting in a lack of correspondence between the over-all frequencies of Tp and Ap nucleotides. Although the basis for this deviation has not been pinpointed, it is clear that exposure of DNA to one of the *E. coli* nucleases known to be a variable contaminant of the polymerase preparation leads to some distortion in the sequence pattern when this DNA is subsequently used as primer.

Comparison of Sequence Patterns of DNA's Within a Species—Although DNA's isolated from different tissues of the same species have the same base composition, it is conceivable that differences might exist in the arrangement of the bases. A comparison by nearest neighbor analysis of bovine sperm, thymus, and liver revealed no variations beyond the error of the method; the DNA's of two normal tissues and two tumors of the mouse also failed to reveal significant differences.

A surprising development has provided two separable and distinguishable DNA entities within a species. Sueoka identified and isolated two DNA components from crab testis on the basis of distinctive buoyant densities (15). Sequence analysis of the "light" crab DNA uncovered the fact that it contains G and C residues as 3% of the total bases interspersed among A and T residues, which are almost invariably in alternating sequence. This DNA is therefore similar to the simple and well ordered dAT, the copolymer synthesized *de novo* by polymerase. The main crab DNA component, by contrast, has a sequence pattern resembling other animal DNA's of similar base composition. The biological significance of the "light" crab DNA is an intriguing subject and one which immediately suggests studying the distribution of this DNA in other somatic cells and the sperm of this species.

Comparison of Sequence Patterns Among Bacterial, Bacteriophage, Animal, and Plant DNA's—In the previous report, confined largely to the sequence patterns of bacterial and bacteriophage DNA's, it was established in many cases that the nucleotide arrangements did not conform to predictions of ran-

[2] I. R. Lehman, R. L. Sinsheimer, and A. Kornberg, unpublished observations.

dom distributions. In the present work, concerned principally with animal DNA's, it is clear that deviations from random arrangements also obtain. The most striking example is the frequencies of the CpG sequence, which are only approximately one-third of the values calculated from the base composition, whereas those of the isomeric GpC sequence are close to the calculated values. The consistency with which certain nearest neighbor sequences, determined from a wide variety of DNA's, deviate from the random expectancy is a surprising finding if one considers that in each nearest neighbor experiment, on the order of 10^{15} dinucleotides, representative of the dinucleotides of many different primer molecules of a given DNA, are analyzed. This result suggests that determinants other than random arrangement must have governed the development of the sequence pattern of a species.

When all of the sequence frequencies are surveyed, the data appear to distinguish the animal and plant cells from the bacteria and bacteriophages. Other distinctions between and within these groups are suggestive but less striking than the CpG sequence. Were it possible to refine the accuracy of the sequence frequency analyses, these examples might increase and permit more meaningful phylogenetic correlations. The connection between the relative abundance of certain sequences and their phenotypic expression as amino acid sequence remains an ultimate objective.

SUMMARY

1. Replication of the single stranded deoxyribonucleic acid (DNA) from phage ΦX 174 was studied under conditions of limited synthesis (20% increase over the primer) and extensive synthesis (600% increase). The base composition of the products and the nearest neighbor frequencies conform to the model based on pairing of adenine to thymine and of guanine to cytosine between strands of opposite polarity. This experiment further indicates that the enzymatically synthesized DNA strands prime as effectively as the natural strand of phage DNA and that both strands of a double helix may prime in the enzymatic replication.

2. A comparison of the nearest neighbor sequence analyses of the DNA's of several bovine tissues (sperm, thymus, liver) revealed no variations beyond the error of the method; similar results were obtained in a comparison of the DNA's of mouse tissues, including two tumors. However, the two DNA components isolated by Sueoka from crab testis were shown to have distinctive sequence patterns, one remarkably similar to the dAT copolymer synthesized *de novo* by polymerase.

3. A survey of the sequence patterns of a variety of animal and plant DNA's revealed that certain sequence frequencies were strikingly different from predictions based on random arrangement of the bases. The data appear to distinguish animal and plant cells from bacteria by the very low frequency of the cytidyl-(3'-5')-guanosine (CpG) sequence in the former group and the high frequency of the isomeric guanyl-(3'-5')-cytosine (GpC) sequence in the latter group.

REFERENCES

1. Josse, J., Kaiser, A. D., and Kornberg, A., *J. Biol. Chem.*, **236**, 864 (1961).
2. Lehman, I. R., Bessman, M. J., Simms, E. S., and Kornberg, A., *J. Biol. Chem.*, **233**, 163 (1958).
3. Smith, M., and Khorana, H. G., *J. Am. Chem. Soc.*, **80**, 1141 (1958).
4. Cunningham, L., Catlin, B. W., and de Garilhe, M. P., *J. Am. Chem. Soc.*, **78**, 4642 (1956).
5. Hilmoe, R. J., *J. Biol. Chem.*, **235**, 2117 (1960).
6. Aposhian, H. V., and Kornberg, A., *J. Biol. Chem.*, **237**, 519 (1962).
7. Lehman, I. R., Roussos, G. G., and Pratt, E. A., *J. Biol. Chem.*, **237**, 819 (1962).
8. Lehman, I. R., *J. Biol. Chem.*, **235**, 1479 (1960).
9. Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J., *J. Biol. Chem.*, **193**, 265 (1951).
10. Kay, E. R. M., Simmons, N. S., and Dounce, A. L., *J. Am. Chem. Soc.*, **74**, 1724 (1952).
11. Sevag, M. G., Lackman, D. B., and Smolens, J., *J. Biol. Chem.*, **124**, 425 (1938).
12. Marmur, J., *J. Molecular Biol.*, **3**, 208 (1961).
13. Sinsheimer, R. L., *J. Molecular Biol.*, **1**, 43 (1959).
14. Lehman, I. R., *Ann. N. Y. Acad. Sci.*, **81**, 745 (1959).
15. Sueoka, N., *J. Molecular Biol.*, **3**, 31 (1961).
16. Schachman, H. K., Adler, J., Radding, C. M., Lehman, I. R., and Kornberg, A., *J. Biol. Chem.*, **235**, 3242 (1960).
17. Sinsheimer, R. L., *J. Molecular Biol.*, in press.
18. Weiss, S. B., and Nakamoto, T., *Proc. Natl. Acad. Sci. U. S.*, **47**, 1400 (1961).
19. Furth, J. J., Hurwitz, J., and Goldmann, M., *Biochem. and Biophys. Research Communs.*, **4**, 431 (1961).
20. Chamberlin, M., and Berg, P., *Proc. Natl. Acad. Sci. U. S.*, **48**, 81 (1962).