NON-RANDOMNESS IN BASE SEQUENCES OF DNAs*

J. Moacanin

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Robert Simha

Department of Chemistry
University of Southern California
Los Angeles, California

Received March 24, 1966

Experimental analyses of sequences in a series of natural and synthetic DNAs have led investigators to the conclusion that the arrangement of the four bases in the chain departs significantly from a random distribution (Jones, et al., 1957; Burton and Petersen, 1960; Shapiro and Chargaff, 1957 and 1960; Josse, et al., 1961; Swartz, et al., 1962). A statistical analysis by Simha and Zimmerman (1961 and 1962) showed that sequence data for a copolymer of adenine and uracil as well as for calf thymus DNA must be interpreted on the basis of higher than nearest neighbor effects. That is, the frequency of occurrence of a specific base arrangement cannot be predicted from the composition alone, but has to be expressed in terms of conditional probabilities which describe departures from randomness.

We consider in more detail Petersen's (1963) recent results on calf thymus DNA, which are more extensive than those (Burton and Petersen, 1960) analyzed previously (Simha and Zimmerman, 1962). Some results on other animal, plant and bacterial DNAs are discussed in support of the generality of the conclusions.

*Supported by the National Aeronautics and Space Administration, Contract NAS7-100 with the Jet Propulsion Laboratory, California Institute of Technology, and by the Office of Naval Research, Contract Nonr 228 (21), NR 108-405 with the University of Southern California.

Discussion

For our purpose we shall use the previously introduced hierarchy of correlation coefficients ρ_{JK}, Θ_{JKM}, Θ_{IJKM}, - - -, (Simha and Zimmerman, 1962) which express, respectively, the deviations from randomness, nearest neighbor effects, penultimacy, etc.:

$$\rho_{JK} = \frac{p_{KM}}{f_K} = \frac{N_{KM}}{N_K} \Big/ \frac{N_M}{\sum_i N_i} \tag{1}$$

$$\Theta_{JKM} = \frac{p_{JKM}}{p_{KM}} = \frac{N_{JKM}}{N_{JK}} \Big/ \frac{N_{KM}}{N_K}$$

N_J is the number of J units, N_{JK} of JK doublets, etc. The conditional probability of finding an M preceeded by K is p_{KM}, preceeded by JK is p_{JKM}. Then $\rho_{JK} = 1$ in the purely random case, $\Theta_{JKM} = 1$ when there is a nearest neighbor but not a penultimate effect, and so on. In what follows the experimentally measured frequencies are given in moles of base per 100 g-atom of phosphorous, thus $N_j \equiv J$, $N_{jk} \equiv JK/2$, etc.

The experimental methods of analysis (Burton and Petersen, 1960) yield frequencies of pyrimidine runs C^iT^j (Table 1). These represent sums of frequencies for fixed numbers i and j of C and T nucleotide units, respectively. Thus, for example, C^2T is the sum of frequencies of CCT, CTC, and TCC. Each sequence is initiated and terminated by a purine unit. Hence, the tabulated numbers refer to structures of the general type pu C^iT^j pu summed over all C and T combinations for given i and j, initiated and terminated by either G or A. Data on nearest neighbor frequencies are taken from Josse et al (1961) and shown in Table 2.

The departures of the ρ's from unity and the inequalities AT ≠ TA, TG ≠ GT, etc. (Table 2), make the presence of non-randomness self-evident. Turning our attention to pyrimidine runs (Table 1), we note the differences between measured frequencies and those computed assuming randomness (Simha and Zimmerman, 1962); again, for purine flanked doublets, CT ≠

Table 1. Comparison of measured sequence data on calf thymus DNA and computed estimates assuming nearest-neighbor and random effects

Sequence	Moles of pyrimidine/100 g atoms of DNA P			
	Measured	Computed assuming nearest neighbor effect		Computed assuming random effect
		(From Column 1[a])	(From Table 2[b])	
C	3.92		4.25	5.35
T	6.23		5.88	7.08
C^2	1.99		2.11	2.29
CT	2.96		2.70	3.03
TC	2.23		2.70	3.03
T^2	2.62		3.62	4.01
C^3	0.76	0.76	0.79	0.74
C^2T	2.73	2.77	2.95	2.92
CT^2	2.64	2.90	3.78	3.86
T^3	1.28	0.83	1.67	1.70
C^4	0.31	0.26	0.26	0.21
C^3T	1.39	1.21	1.29	1.11
C^2T^2	1.99	1.91	2.43	2.20
CT^3	1.53	1.17	2.08	1.94
T^4	0.55	0.23	0.68	0.64
C^5		0.08	0.08	0.06
C^4T	0.39	0.47	0.50	0.37
C^3T^2	1.06	1.08	1.23	0.98
C^2T^3	1.27	0.94	1.56	1.30
CT^4	0.82	0.40	1.00	0.86
T^5	0.19	0.06	0.26	0.23

[a] Computed using the first six measured frequencies: C, T, C^2, CT, TC, and T^2. For example, $C^2T = 3\,(\bar{N}_{CCT} + \bar{N}_{CTC} + \bar{N}_{TCC})$, where $\bar{N}_{CCT} = (C^2/2)(CT/2)/(C)$, and similarly for $\bar{N}_{CTC}$ and N_{TCC}

[b] Computed using data from Table 2 following a procedure analogous to computations from Col. 1, except now, $\bar{N}_{CCT} = (N_{AC} + N_{GC}) <p_{CC}> <p_{CT}> (<p_{TA}> + <p_{TG}>)$, and similarly for $\bar{N}_{CTC}$ and $\bar{N}_{TCC}$

Table 2. Nearest-neighbor frequencies N_{JK} correlation coefficient ρ_{JK} and configurational probabilities $< p_{JK} >$ for native calf thymus DNA

Nearest neighbors	N_{JK}, mole %	ρ_{JK}	$< p_{JK} >$
AA	8.9	1.088	0.311
AT	7.3	0.902	0.255
AG	7.2	1.137	0.252
AC	5.2	0.838	0.182
TA	5.3	0.655	0.187
TT	8.7	1.086	0.307
TG	7.6	1.255	0.269
TC	6.7	1.091	0.237
GA	6.4	1.046	0.299
GT	5.6	0.925	0.262
GG	5.0	1.092	0.234
GC	4.4	0.948	0.206
CA	8.0	1.289	0.369
CT	6.7	1.091	0.309
CG	1.6	0.345	0.074
CC	5.4	1.147	0.249

Nucleotide composition: 28.6 mole % A; 28.3 mole % T; 21.4 mole % G; 21.7 mole % C.

TC is to be noted. On the assumption of nearest neighbor effects, but absence of penultimacy, the averaged $<p_{JK}>$ values (Table 2) reduce to p_{JK} and permit C^iT^j values to be estimated (Table 1). Similarly, longer runs were computed using frequencies of the purine flanked singlet and doublet C and T runs. The lack of agreement between the two columns of computed values is apparent. If the assumption of no penultimacy were correct, the two methods of computation should yield the same results. Moreover, neither of the two sets of frequency estimates seem to improve the random model. Thus, these observations offer strong indirect evidence for the presence of penultimate effects. To be sure, experimental uncertainties must be partly responsible for the discrepancies, but a comparison of pyrimidine run frequencies from four separate experiments (Table 3) suggests that the penultimate effects are greater than those caused by experimental errors. Qualitatively similar conclusions are reached from the examination of data on other DNAs.

Table 3

Comparison of Four Sets of Measured Sequence Data on Calf Thymus DNA

Sequence	Moles of Pyrimidine/100 g of DNA P			
	Peterson 1963	Burton 1960	Burton-Petersen 1960	Spencer-Chargaff 1963
C	3.92	3.85	3.8	3.41
T	6.23	6.53	6.2	8.23
C^2	1.99	1.77	1.8	1.38
CT	2.96	3.04	3.1	4.73 (CT + TC)
TC	2.23	2.21	2.0	
T^2	2.62	2.60	2.6	2.48
C^3	0.76	0.77	0.7	0.54
C^2T	2.73	2.90	2.6	2.43
CT^2	2.64	2.73	2.7	2.90
T^3	1.28	1.21	1.3	1.73
C^4	0.31			
C^3T	1.39			1.11
C^2T^2	1.99	1.62	1.9	2.14
CT^3	1.53	1.48	1.6	2.19
T^4	0.55	0.57	0.6	0.87
C^5				
C^4T	0.39			
C^3T^2	1.06			
C^2T^3	1.27			
CT^4	0.82	0.92	0.7	
T^5	0.19	0.20	0.3	

Additional insight can be gained by considering pure T or C runs, because for these the averaging effects caused by the summation over the various isomers are eliminated. Thus, the ratio $\left[T^j/j\right] / \left[T^{j-1}/(j-1)\right]$ yields p_{TT} for $j \geq 2$ and p_{TTT} for $j \geq 3$ in absence or presence of penultimate effects, respectively. Table 4 lists these ratios for a variety of DNAs. Petersen's data on calf thymus show that the ratio is nearly independent of j for $j \geq 3$, giving convincing evidence for penultimacy. These are probably the most reliable results since separate experiments using higher DNA loadings were carried out for the determination of the longer runs such as T^4 or T^5. The reasonable agreement with results from other experiments on calf thymus further supports these conclusions.

Table 4. Comparison of Ratios of T and C Frequencies

DNA	$\frac{T^2/2}{T}$	$\frac{T^3/3}{T^2/2}$	$\frac{T^4/4}{T^3/3}$	$\frac{T^5/5}{T^4/4}$	$\frac{T^6/6}{T^5/5}$	$\frac{T^7/7}{T^6/6}$	$\frac{C^2/2}{C}$	$\frac{C^3/3}{C^2/2}$	$\frac{C^4/4}{C^3/3}$	References
Calf thymus	0.21	0.33	0.33	0.30	0.33	0.36	0.26	0.25	0.36	Petersen 1963
	0.21	0.33	0.35	0.40	-	-	0.24	0.26	-	Burton et al 1960
	0.20	0.31	0.35	0.29	-	-	0.23	0.29	-	Burton 1960
	0.11	0.46	0.37	-	-	-	0.20	0.26	-	Spencer et al 1963
Human Spleen	0.24	0.41	0.44	0.48	-	-	0.23	0.33	-	Shapiro et al 1963
Herring Testis	0.19	0.31	0.28	-	-	-	0.15	0.23	-	Burton 1960
	0.19	0.38	0.26	0.25	-	-	0.17	0.20	0.14	Petersen 1963
Salmon	0.24	0.25	0.35	-	-	-	0.21	0.18	-	Burton 1960
E. esculentus (sea urchin)	0.26	0.33	0.31	0.26	-	-	0.20	0.17	-	Burton 1960
E. coli	0.27	0.23	0.25	0.37	-	-	0.30	0.17	-	Burton 1960
A. faecalis	0.14	0.22	0.17	-	-	-	0.30	0.20	0.17	Burton 1960
P. aeruginosa	0.19	-	-	-	-	-	0.40	0.13	0.17	Burton 1960
Wheat germ	0.21	0.37	0.43	-	-	-	0.21	0.11	-	Spencer et al 1963
Rye germ	0.23	0.31	0.49	-	-	-	0.19	-	-	Spencer et al 1963
Average	0.18	0.32	0.34	0.34						

Inspection of the values for the other DNAs shows that in spite of considerable scatter the same trend is followed, i.e., a lower value for $j = 2$, is followed by higher and more or less constant values for $j \geq 3$. For example, the average values for all the data in the table are 0.18, 0.32, and 0.34 for $j = 2$, 3 and 4, respectively. The data for C runs are less extensive because of the generally low C content and, hence, less conclusive. Recent results (Petersen and Burton, 1964) seem to indicate constancy for the ratios for $2 \leq j \leq 4$ for calf thymus, and decreasing values with increasing j for both herring testis and m. lysodeikticus. The latter results suggest effects higher than penultimacy, but considering the experimental difficulties to even detect the presence of C^4 or higher runs, some caution in the interpretation is in order.

Burton, et al (1963) observed the constancy of the ratios $\left[T^n/n\right] / \left[T^{n-3}/(n-3)\right]$ and assumed these to equal the frequency of randomly distributed non-overlapping T-triplets. We note, however, that with our definitions:

$$\frac{T^n/n}{T^{n-3}/(n-3)} = \frac{T^n/n}{T^{n-1}/(n-1)} \times \frac{T^{n-1}/(n-1)}{T^{n-2}/(n-2)} \times \frac{T^{n-2}/(n-2)}{T^{n-3}/(n-3)} = p^3_{TTT} \qquad (2)$$

Hence, the constancy of left hand side noted by Burton follows simply from the constancy of p_{TTT}.

It should be noted that the comparisons of C^iT^j's are not completely independent, since a given set was calculated using the same nearest neighbor correlation coefficients. Certain pyrimidine frequencies, however, can be examined for consistency in a more direct way. For example, on the assumption of penultimacy, but no higher effects, one can arrive at the following relationship:

$$CT^4/5 = \frac{T^2/2}{T}\left\{\left[\left(\frac{T^3/3}{T^2/2}\right)\Big/\left(\frac{T^2/2}{T}\right) + \lambda\right]\left(CT^3/4\right) - \lambda\frac{T^3/3}{T^2/2}\left(CT^2/3\right)\right\} \qquad (3)$$

where $\lambda \equiv \bar{\theta}_{CTT}\,\bar{\theta}_{TTC}\Big/\left(\bar{\theta}_{CTT} + \bar{\theta}_{TTC} - \bar{\theta}_{TTT}\right)$. Here the barred correlation coefficients are defined as in Eq. (1), where the frequencies N_{JKM} now refer to pyrimidine runs flanked by purines (Simha and Zimmerman, 1962). For the random model the relationship reduces to:

$$CT^4/5 = f_T\,\{2(CT^3/4) - f_T\,(CT^2/3)\} \qquad (4)$$

where f_T is the fraction of T-nucleotide bases. For the nearest neighbor case the equation becomes:

$$CT^4/5 = p_{TT}\,\{2(CT^3/4) - p_{TT}\,(CT^2/3)\} \qquad (5)$$

where $p_{TT} = \langle p_{TT}\rangle$ or $(T^2/2)/T$ when nearest neighbor effects are computed from Table 2 or 1, respectively. Using data on calf thymus (Table 1) and herring testis (Petersen, 1963) to compute the respective CT^4 values, the following results were obtained:

	Calf thymus	Herring testis
Random (Eq. 4)	0.73	0.53
Nearest Neighbor (Eq. 5):		
$(T^2/2)/T$	0.61	0.44
$< p_{TT} >$	0.76	-
Penultimate (Eq. 2 and assume $\lambda = 1$)	0.73	0.60
Observed	0.82	0.59

These results, of course, neither indicate a random base pairing, nor a nearest neighbor effect, since the other respective conditions are not satisfied; e.g., $f_T \neq (T^2/2)/T \neq < p_{TT} >$. To obtain agreement with the observed value for calf thymus we require $\lambda = 2$. From Table 2 one finds that $\bar{\theta}_{TTT} = 1.55$, which leads to the reasonable condition, $\bar{\theta}_{CTT}\ \bar{\theta}_{TTC}/(\bar{\theta}_{CTT} + \bar{\theta}_{TTC} - 1.55) = 2$; note that if $\bar{\theta}_{CTT} = \bar{\theta}_{TTC} = 1$, $\lambda = 2.2$. Similarly, for herring testis agreement with experiment is obtained for $\lambda = 2.84$. Again, one finds that $\bar{\theta}_{TTT} = 1.94$, and,, hence, $\bar{\theta}_{CTT}\ \bar{\theta}_{TTC}/(\bar{\theta}_{CTT} + \bar{\theta}_{TTC} - 1.94) = 2.84$.

In conclusion, the unequivocal evidence for non-randomness along with the failure of nearest neighbor effects alone to account for experimental results, give strong evidence for the existence of at least penultimate effects. Furthermore, without violating any of the conditions for penultimacy and using ad hoc but acceptable numerical values for λ (i.e., a function of $\bar{\theta}_{CTT}$ and $\bar{\theta}_{TTC}$) agreement with experiment can be obtained. Data on triplet frequencies are needed to ascertain the extent of effects higher than penultimacy. We wish to point out, however, that separation of isomers for the triplet pyrimidine runs, C^2T and CT^2 would yield estimates of the $\bar{\theta}_{IJK}$'s. The latter approach should be feasible with the current separation techniques.

Acknowledgement:

We are indebted to Mr. E. F. Cuddihy, Jet Propulsion Laboratory for programming the computations.

References

Jones, A. S., Stacey, M. and Watson, B. E., J. Chem. Soc., 2454 (1957).
Burton, K. and Petersen, G. B., Biochem. J., 75, 17 (1960).
Shapiro, H. S. and Chargaff, E., Biochim. Biophys. Acta, 26, 608 (1957); 39, 62 (1960).
Josse, J., Kaiser, A. D. and Kornberg, A., J. Biol. Chem., 236, 864 (1961).
Swartz, M. N., Trautner, T. A. and Kornberg, A., J. Biol. Chem., 237, 1961 (1962).
Simha, R. and Zimmerman, J. M., J. Polymer Sci., 42, 309 (1960); 51, S39 (1961).
Simha, R. and Zimmerman, J. M., J. Theoret. Biol., 2, 87 (1962).
Burton, K., Biochem. J., 77, 547 (1960).
Petersen, G. B., Biochem. J., 87, 495 (1963).
Spencer, J. H. and Chargaff, E., Biochim. Biophys. Acta, 68, 18 (1963).
Shapiro, H. S. and Chargaff, E., Biochim. Biophys. Acta, 76, 1 (1963).
Petersen, G. B. and Burton, K., Biochem. J., 92, 666 (1964).
Burton, K., Lunt, M. R., Petersen, G. B. and Siebke, J. C., Symposium on Quantitative Biology, 28, 27 (1963).