COMMUNICATIONS TO THE EDITOR

Statistical Analysis of DNA Sequences. I

During the last few years a number of exact nucleotide sequences have been established for various DNA molecules, starting with the ϕ-174 sequence determined by Sanger et al.[1] Statistical analysis of the sequences can provide important biological information, especially concerning the evolution of DNA molecules.[2,3] It is also essential for the understanding of the physical properties of DNA, such as the melting curves.[4–6] Prior to Sanger's work, conflicting views were expressed about the overall statistical structure of the sequences. Several authors[7,8] suggested that nucleotide sequences are close to being random, while Wada et al.[9] argued that they are highly nonrandom. In the present communication, we introduce a quantitative measure of the randomness of a sequence and apply it to some recently determined DNA sequences[1,10–15]: ϕ-174, SV-40, MS-2, G-4, FD, hepatitis B-virus, and a segment of human mtDNA. We find that these sequences range from almost random to highly nonrandom.

To simplify the analysis, we have studied the sequences of AT and GC pairs, rather than the complete four-letter sequences. (In this way, we preserve all of the physically important information, since the physical properties of DNA depend predominantly on the sequence of pairs.) We relate number "0" to each GC pair and "1" to each AT pair.

Thus, we obtain a sequence of 0's and 1's having some length L (the number of nucleotide pairs in the corresponding DNA) and some concentration of 1's, p. We treat the sequences as circular, the last nucleotide being followed by the first. The boundary effects introduced by this convention are unimportant for long sequences. Suppose we randomly choose a string of length N (i.e., N subsequent numbers in the sequence starting at some randomly chosen place). In the case of an infinite random sequence ($L = \infty$), the probability of finding r 1's in the string is given by a binomial distribution,

$$P_N(r) = C_N^r p^r (1 - p)^{N-r} \tag{1}$$

where C_N^r is the binomial coefficient. The average and the standard deviations of the variable r are given by

$$\langle r \rangle_N = pN \tag{2a}$$

$$\sigma_{N,\,\text{random}} = [p(1 - p)N]^{1/2} \tag{2b}$$

For $N \gg 1$, the distribution in Eq. (1) approaches the Gaussian one, with the appropriate parameters given by Eq. (2).

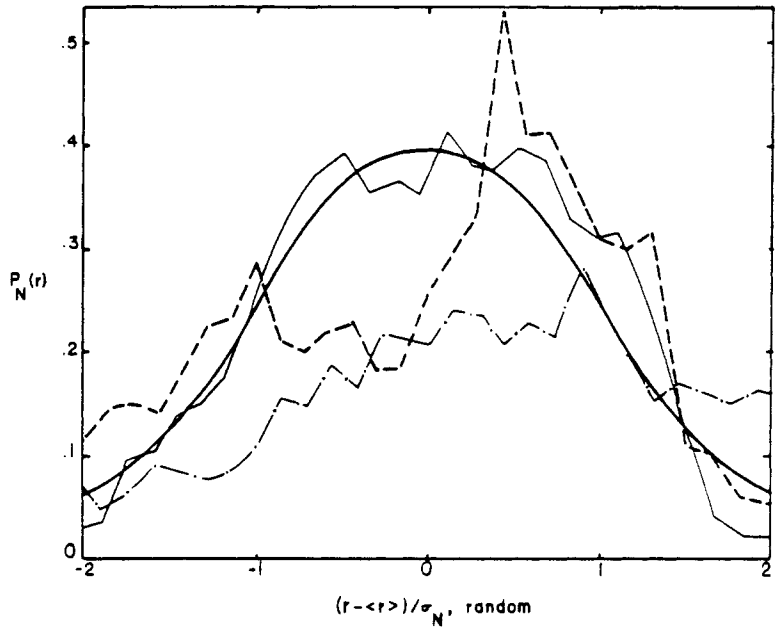
For DNA sequences, $P_N(r)$ can be defined as the frequency of the occurrence of r 1's in strings of length N starting at all possible positions along the molecule. Figure 1 shows $P_N(r)$ with $N = 200$ for the sequences of Refs. 1 and 10–13. All results are appropriately normalized to allow a convenient comparison of the curves with each other and with the Gaussian curve corresponding to a random sequence. We have also calculated $P_N(r)$ for $N = 10, 20 \ldots$, with qualitatively similar results. It should be noted that since our sequences are finite, one cannot expect coincidence between the calculated $P_N(r)$ and the Gaussian curve, but the results already indicate some systematic deviation from randomness.

To allow a quantitative comparison of the degree of randomness of different sequences, we need some integral characteristic of the properties of $P_N(r)$. A suitable quantity is

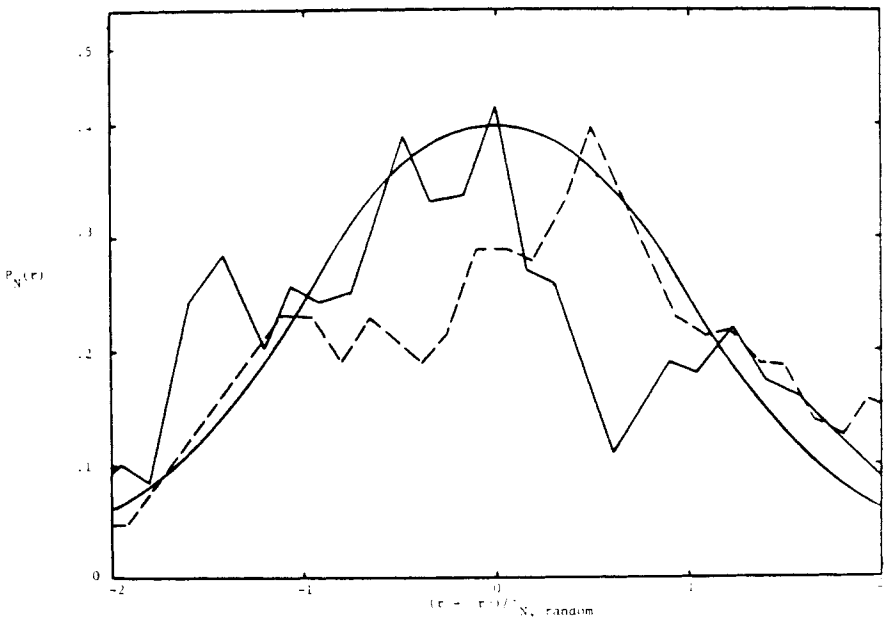
$$\sigma_N = [\langle (r - \langle r \rangle)^2 \rangle]^{1/2} \tag{3}$$

where the averaging is taken over all possible starting positions along the sequence. Figure 2 shows the quantity $\kappa_N = \sigma_N/\sigma_{N,\,\text{random}}$ as a function of N, calculated for all seven DNA sequences for an English text translated into a binary code (each letter and punctuation mark

Biopolymers, Vol. 21, 1687–1690 (1982)
 CCC 0006-3525/82/081687-04$01.40



(a)



(b)

Fig. 1. Calculated $P_N(r)$ as a function of normalized variable $(r - \langle r \rangle)/\sigma_{N,\text{random}}$ for the strings of length $N = 200$ for different sequences: (a) ϕX-174 (solid line), G-4 (dashed line), FD (dashed-dotted line); (b) MS-2 (solid line), SV-40 (dashed line). The smooth curve is the Gaussian distribution.

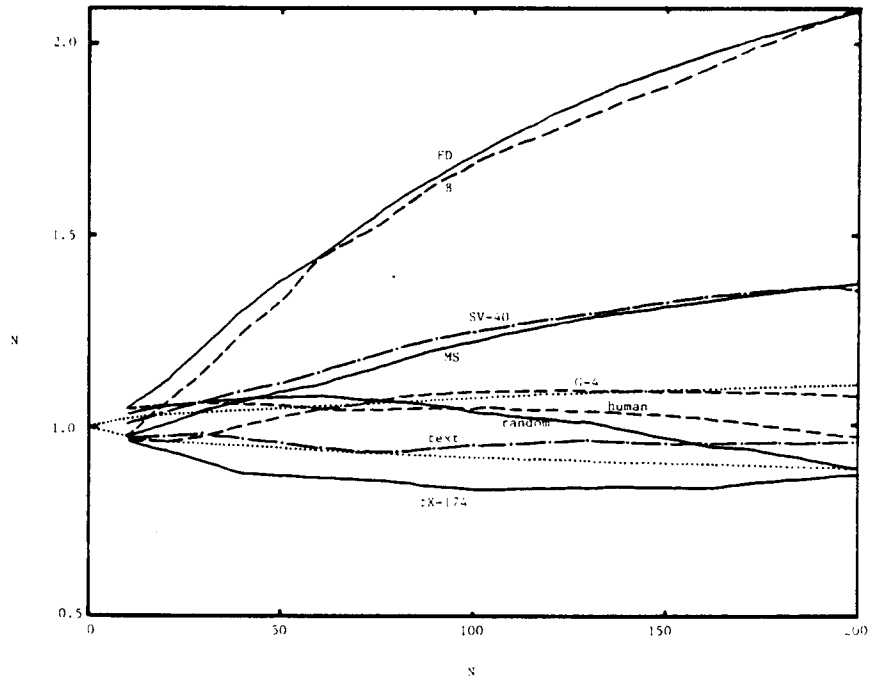


Fig. 2. κ_N as a function of N for various sequences. The dotted lines mark the boundaries of the region $|\kappa_N - 1| < \delta\kappa_N$.

being represented by five binary digits) and for a computer-generated random sequence (with the same p and L as SV-40). As in the case of $P_N(r)$, we cannot expect κ_N to be equal to 1, because of the finite length of the sequences. If an infinite random sequence is cut into pieces of length $L \gg N$, then the value of κ_N calculated for each piece will be slightly different from 1. A sequence of finite length can be called random only with a certain probability. Here, we do not attempt to find the probability distribution $P(\kappa_N)$. The mean fluctuation, $\delta\kappa_N$, is given by $\delta\kappa_N = \frac{1}{2}\delta(\kappa_N^2)$, where

$$\delta(\kappa_N^2) = [\overline{\kappa_N^4} - (\overline{\kappa_N^2})^2]^{1/2} \tag{4}$$

and bars indicate averaging over all possible sequences. A straightforward calculation gives $\delta\kappa_N = (N/3L)^{1/2}$. The boundaries of the domain, $|\kappa_N - 1| < \delta\kappa_N$, are indicated by dotted lines in Fig. 2. ($\delta\kappa_N$ is calculated, using the value of $L = 3575$ for ϕX-174. If we used the length of any other sequence, the difference would be no more than 25%.) We see that the curves for G-4, human mtDNA, English text, and the computer-generated sequence fall within the dotted lines almost everywhere, and thus the corresponding sequences are random, within the accuracy of our method. If we assume that the probability distribution of κ_N is approximately Gaussian, then the probability that a sequence having a certain value of κ_N is random can be estimated as $P_R \sim \pi^{-1/2}|x| \exp(-x^2)$, where $x = (\kappa_N - 1)/\delta\kappa_N$. With $N = 100$, this gives $P_R \sim 0.02$, 8×10^{-4}, 10^{-4}, 2×10^{-32}, and 2×10^{-34} for ϕX-174, MS-2, SV-40, B-virus, and FD, respectively.

Thus, the seven sequences analyzed in this paper range from almost random to highly nonrandom. It should be noted that our method is rather rough, since it takes into account only the second moment of the distribution $P_N(r)$. Our estimates for P_R may change if higher moments are included.

The nucleotide sequence of a DNA molecule can be viewed as a text containing the hereditary information of a living organism. The information content of the text is related to its

regularity. If a long sequence of letters is very regular, it carries very little information per unit length, and we can say that the code is inefficient. On the other hand, an efficiently coded text must look like an almost random sequence of letters. We note also that the degree of randomness of a sequence may be related to the degree of its evolution, i.e., to the number of mutations in the course of the evolution from some primitive initial sequence. It is possible that the sequences are evolving in the direction of greater efficiency, that is of increasing P_R. This question requires further study.

References

1. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison III, C. A., Slocombe, P. M. & Smith, M. (1977) *Nature* **265,** 687–695.

2. Shepherd, J. C. W. (1981) *J. Mol. Evol.* **17,** 94–102.

3. Shepherd, J. C. W. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 1596–1600.

4. Poland, D. & Scheraga, H. A. (1970) *Theory of Helix-Coil Transition in Biopolymers*, Academic Press, New York.

5. Wartell, R. W. & Montroll, E. W. (1972) *Adv. Chem. Phys.* **22,** 129–203.

6. Azbel, M. Ya. (1979) *Phys. Rev.* **A20,** 1671–1684.

7. Lazurkin, Yu. S., Frank-Kamenetskii, M.D. & Trifonov, E. N. (1970) *Biopolymers* **9,** 1253–1306.

8. Frank-Kamenetskii, M.D. (1977) *Nature* **269,** 729–730.

9. Wada, A., Tachibana, H., Goton, O. & Takanami, M. (1976) *Nature* **263,** 439–440.

10. Reddy, V. B., Thimmappaya, B., Dhar, R., Subramarian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. & Weissman, S. M. (1978) *Science* **200,** 494–502.

11. Fiers, W., Coutreras, R., Duerinck, F., Haegeman, G., Iserentaut, D., Merregaert, I., Min Jou, W., Molemans, F., Raeymaekers, A., Van de Berghe, A., Volckaert, G. & Ysebaert, M. (1976) *Nature* **260,** 500–507.

12. Godson, G. H., Barrell, B. G., Staden, R. & Fiddest, J. C. (1978) *Nature* **276,** 236–247.

13. Beck, E., Sommer, R., Auerswald, E. A., Kuzz, C., Zink, B., Osterburg, G., Schaller, H., Sugimoto, K., Sugisaki, H., Okamoto, T. & Takanami, M. (1978) *Nucleic Acids Res.* **5,** 4495–4503.

14. Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P. & Charnay, P. (1979) *Nature* **281,** 646–650.

15. Eperon, I. C., Anderson, S. & Nierlich, D. P. (1980) *Nature* **286,** 460–467.

M. Y. AZBEL*

Department of Physics, University of Pennsylvania
Philadelphia, Pennsylvania 19104

Y. KANTOR

Physics Department, Tel-Aviv University
Tel-Aviv, Israel

L. VERKH

Department of Biophysical Sciences
State University of New York
Buffalo, New York 14214

A. VILENKIN

Physics Department, Tufts University
Medford, Massachusetts 02155

Received October 9, 1981
Accepted March 29, 1982

* On sabbatical from the Department of Physics, Tel-Aviv University, Israel.