Biophysical Journal Volume 69 November 1995 1708–1711

A Model for Nucleotide Sequences

A. M. C. de Souza*‡ and C. Anteneodo*

*Centro Brasileiro de Pesquisas Fisicas Rua Xavier Sigaud 150,22290-180, Rio de Janeiro, and ‡Departamento de Física, Universidade Federal de Sergipe, 49100-000, Aracaju-SE, Brazil

ABSTRACT We propose a model for generating "artificial" nucleotide sequences and, by the method of mapping those sequences onto a "DNA-walk," we analyze the presence of correlation between nucleotides. Artificial sequences are constructed considering, basically, interactions between first neighbors and between more distant units. We show that long-range correlations may be favored by the occurrence of intrastrand interactions, which give a nonlinear characteristic to the sequence.

INTRODUCTION

Because the evolutionary history of organisms is registered in their genetic material, some of that history could be reconstructed from the analysis of the nucleotide sequences. Therefore, study of the characteristics of existing genomes may enlighten our understanding of the processes by which they have evolved. The method of "DNA walks" (Peng et al., 1992), basically consisting of the association of a random walk to a given sequence, was recently proposed to study the stochastic properties of nucleotide sequences. This method allows us to study the fluctuations of nucleotide content and to obtain a quantitative measure of the degree of correlation between nucleotides, given by a power exponent α, which characterizes the dependence of the correlation function on the distance along the sequence, which is $\alpha = 0.5$ for uncorrelated sequences.

Many sequences of genes and cDNA have already been mapped onto unidimensional "DNA walks" (Peng et al., 1992, 1993; Buldyrev et al., 1993; Uberbacher and Mural, 1991) and long-range power law correlations were found in several of the analyzed DNA sequences. These long-range correlations have also been detected through alternative approaches (Li and Kaneko, 1992; Voss, 1992). However, the characterization of the sequences exhibiting such long-range correlations has initially generated a controversy. Although coding and noncoding regions of DNA seem to have different statistical characteristics (coding sequences usually consist of a few regions of different strand bias, whereas noncoding sequences present more complex fluctuations), some authors (Nee, 1992; Prabhu and Claverie, 1992; Chatzidimitriou-Dreismann and Larhammar, 1993) found no consistent differences in the α exponent for coding and noncoding sequences and showed that a well-defined fractal exponent does not always exist for a given sequence. On the other hand, other authors (Peng et al., 1992) found long-range correlations in intron-containing genes but not in complementary DNA sequences or intron-less genes, differences that have been confirmed by recent works (Ossadnik et al., 1994; Buldyrev et al., 1995; Arneodo et al., 1995), showing that the degree of correlation may be a good criterion for identifying coding regions.

In any case, whether these correlations have arisen by pure chance or by some nonrandom process remains an open problem (Nee, 1992; Karlin and Brendel, 1993). Considering that there are patterns found more or less frequently than would be expected from random occurrence (Tavaré and Giddings, 1989), it seems that a nonrandom process is involved. Tavaré and Giddings (1989) estimated the order of DNA sequences, treating them as Markov chains, where a chain is of order k if the probability of finding a given nucleotide at a site is determined by the previous k nucleotides. They found that most sequences exhibit orders of dependence higher than zero, which corresponds to the case of independence. From the point of view of molecular mechanisms, base-stacking interactions were shown to constitute a dominant factor in nucleic acid stability and to be highly sequence dependent (Aida and Nagata, 1986). Moreover, non-covalent forces, namely, hydrophobic, hydrogen bonding, van der Waals, and electrostatic, are also responsible for the conformational stability of any chain molecule, particularly nucleic acids (Ponnuswamy and Gromiha, 1994). These forces between residues within the polymer itself give rise, at least locally, to well-defined three-dimensional structures found not only in RNA, but also in single strands of DNA (Sanger et al., 1982). Taking into account these features, in the attempt to find an explanation for the observed statistical properties of nucleotide sequences, we develop in the present work a simple model for generating artificial sequences. The model consists, basically, of discrete Markov chains with a finite state space in which short-range nucleotide interactions are introduced, where the relevant interactions are those between close neighbors and those between more distant ones.

Received for publication 28 December 1994 and in final form 7 August 1995.

Address reprint requests to Dr. Celia Anteneodo, Centro Brasileiro de Pesquisas Físicas, Rua Dr. Xavier Sigaud 150, Rio de Janeiro 22290-180 RJ, Brazil. Tel.: 055-021-541-0337, X193; Fax: 055-021-541-2047; E-mail: celia@cat.cbpf.br.

0006-3495/95/11/1708/00 $2.00

MODEL AND RESULTS

We assume a finite linear chain in which each site i ($i = 0, 1, \ldots, L$) is occupied by a binary random variable $\{S_i\}$. If $S_i = +1$, a monomer whose base component is a pyrimidine (either cytosine or thymine) occurs at position i, whereas if $S_i = -1$, a purine (either adenine or guanine) occurs at that position. Each element S_i in the chain is chosen considering a previous element S_k, where $k = i - 1$ with probability q and $k = i - j$, for $1 < j \leq i$, with probability $1 - q$. That is, q measures the correlation between first neighbors. Because intrastrand interactions may take place through the formation of loops and the probability of finding a loop of length j in a very long linear polymer is $P(j) \propto j^{-\mu}$ for $j \geq l_c$, with μ a positive real number and l_c an integer which represents a lower cut-off distance (Buldyrev et al., 1993), then j is chosen according to this distribution of probabilities. Once the interacting site k is chosen, we take $S_i = S_k$ with probability p (hence, $S_i = -S_k$ with probability $1 - p$). We construct the linear chain by following this procedure and assuming, by convention, that $S_o = 1$.

For analytical calculations, it is convenient to consider the probability p_i of S_i being equal to 1. Considering that $p_i = p_k\, p + (1 - p_k)(1 - p)$, where k is either $i - 1$ or $i - j$ as defined above, it is easy to find that p_i obeys the following recursive relation:

$$p_i = (2p - 1)[qp_{i-1} + (1 - q)p_{i-j}] + 1 - p \qquad 1 \leq i \leq L, \tag{1}$$

with $p_o = 1$.

The variable that represents the excess of pyrimidines over purines in a subchain of length l is

$$Y(l) = \sum_{i=1}^{l} S_i. \tag{2}$$

Its mean value $\langle Y(l)\rangle$, equivalent to the mean net displacement after l steps in a random walk, is

$$\langle Y(l)\rangle = 2\sum_{i=1}^{l} p_i - l. \tag{3}$$

A measure of the correlation of the constructed sequence is provided by the square root of the mean quadratic fluctuation (Peng et al., 1992):

$$F^2(l) \equiv \overline{[\Delta Y(l) - \overline{\Delta Y(l)}]^2}, \tag{4}$$

where $\Delta Y(l) = Y(l + l_o) - Y(l_o)$, and the bar indicates the average computed over all positions l_o in the sequence ($1 \leq l_o \leq L - l$).

In the particular case $q = 1$, the exact solution of the recurrence equation (1) is

$$p_i = \frac{(2p - 1)^i + 1}{2}. \tag{5}$$

From Eq. 4 and using Eq. 5, we calculate $\langle F^2(l)\rangle$, which is equivalent to an average over a large number of statistically independent realizations of the model sequence. We find the following asymptotic behavior ($1 \ll l \ll L$):

$$\sqrt{\langle F^2(l)\rangle} \sim \begin{cases} 0 & \text{if } p = 1 \\ \left(\frac{p}{1-p}\right)^{l/2} & \text{if } 0 < p < 1 \\ 0 & \text{if } p = 0, \quad l \text{ even} \\ 1 & \text{if } p = 0, \quad l \text{ odd} \end{cases} \tag{6}$$

From the asymptotic behavior, we conclude that there is no long-range correlation. Furthermore, the plots of α (local slope of $\sqrt{\langle F^2(l)\rangle}$) versus l start from a value close to p and decrease sigmoidally down to $\alpha = 0.5$. Thus, for $q = 1$ the model does not reproduce the behavior of α observed experimentally for highly correlated nucleotide sequences. Because, in this case, the direction of each step depends on the history of the walker, there exists an effect of memory produced in the construction of the chain for $p \neq \frac{1}{2}$. However, it may be noted, from Eq. 5, that, for $0 < p < 1$, the process corresponds to a stationary process (Dougherty, 1990) in which $\lim_{i\to\infty} p_i = \frac{1}{2}$. So, for long distances the memory effect vanishes.

Let us consider now the more general case ($q \neq 1$) and compare artificially generated sequences with actual ones. In Fig. 1, *a* and *c*, we show the "DNA walks" of two real sequences: human β-cardiac myosin heavy chain gene and human antithrombin III gene, respectively. In Fig. 1, *b* and *d*, we show typical walks obtained for different values of the model parameters (p, q, μ, and l_c). Artificial sequences were generated with the same length as the real sequences to which they are compared. The mappings 1 *a* and 1 *b* show similar fluctuations, as do the mappings 1 *c* and 1 *d*. The plots of the local slope of the fluctuation function versus the logarithm of the distance along the sequence, for the sequences in Fig. 1, are presented in Fig. 2: sequences 1 *a* and 1 *b* are analyzed in Fig. 2 *a*, and sequences 1 *c* and 1 *d* are analyzed in Fig. 2 *b*. There is also a great similarity between the plots obtained from artificial sequences and those obtained from real sequences, in the relevant range (first decades). It may be noted that many real sequences, as those showed here, smoothly decrease for small values of l up to a value close to 5, which, in our model, corresponds approximately to the value of parameter l_c.

DISCUSSION

The presence of interactions between first neighbors is not sufficient to give rise to the observed long-range correlations, as shown by the analysis of the case $q = 1$. On the other hand, when interactions between more distant neighbors are introduced, long-range correlations may arise and the behavior of real sequences can be mimicked.

We tested alternative rules for generating the artificial sequences. Besides the first neighbor, we have also taken into account either 1) a mean over all the other precedent

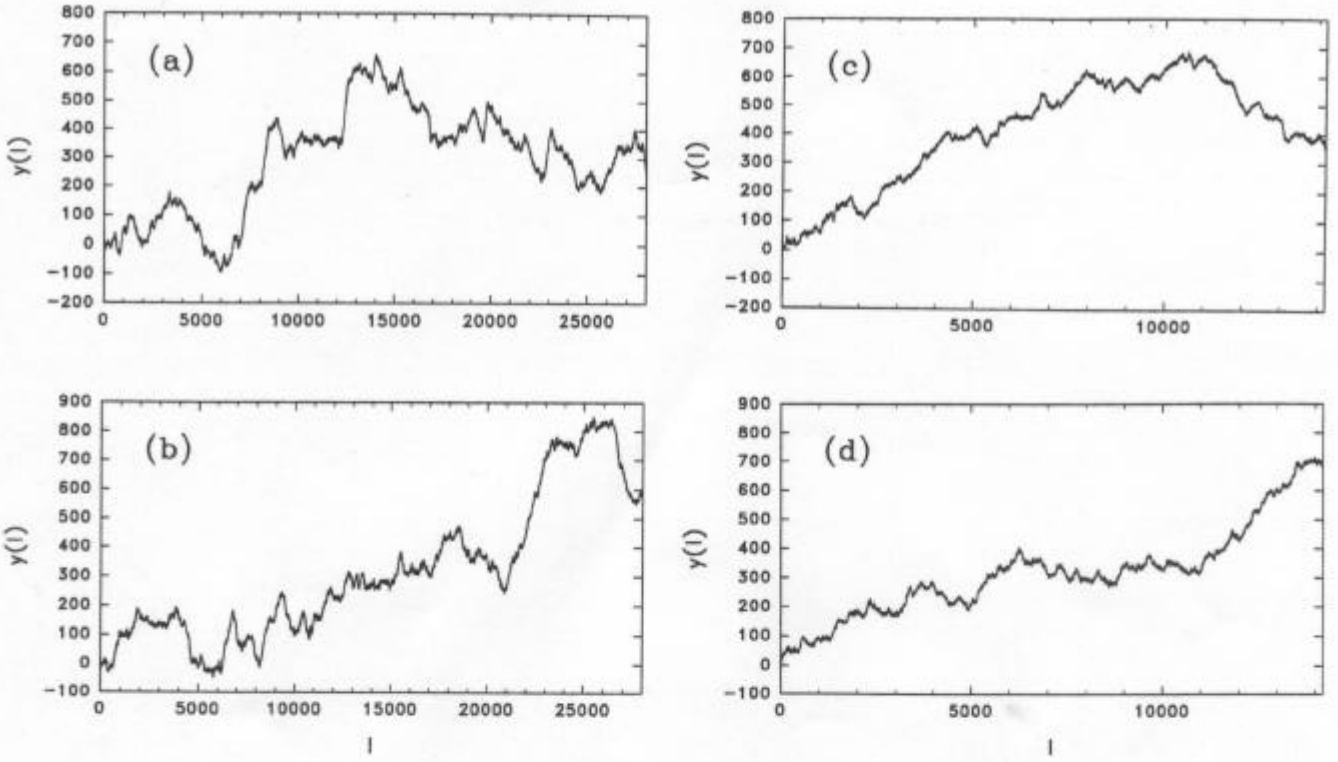

FIGURE 1 DNA-walk displacement $y(l)$ (excess of purines over pyrimidines) vs. nucleotide distance l for: (a) human β-cardiac myosin heavy chain gene (GenBank name: HUMBMYH7); (b) an artificial sequence generated with parameters $p = 0.85$, $q = 0.25$, $\mu = 1.65$, and $l_c = 6$; (c) human antithrombin III gene (GenBank name: HSAT3); and (d) an artificial sequence generated with parameters $p = 0.77$, $q = 0.37$, $\mu = 1.5$, and $l_c = 4$.

nucleotides or 2) a precedent nucleotide at a fixed distance. For no set of the parameters of these two alternative models were we able to obtain $\alpha(l)$ with a behavior similar to that of highly correlated actual sequences. Thus, we conclude that a broad distribution of the distance to the second interacting neighbor is required for the long-range correlations observed in real sequences to arise.

From the analysis of the case $q = 1$ we also notice that, as in actual sequences, the exponent α is not constant over the whole range of values of l, but this does not mean that there can not be a well-defined exponent, corresponding to the asymptotic behavior of F, which may be significantly different from the local values of α. On the other hand, exponents different from 0.5 at finite distances indicate some kind of long-range correlations but do not necessarily mean infinite long-range correlations. Because real sequences are finite, we can only say that the observed long-range correlations are at most on the order of polynucleotide chain length and not infinite.

The mosaic character of DNA, consisting of biased subsequences, could account for apparent long-range correlations. Thus, we should also consider the possibility that correlation arises from the occurrence of statistically different regions, because the presence of biased subdomains also gives rise to exponents greater than 0.5 (Nee, 1992). The behavior of real sequences could result from the combination of some "patching" mechanism and a process such as the one described in this work. But there is also a possibility that long-range correlations observed in real sequences arise purely from interactions between distant neighbors, as shown in the present work. Correlated units may occur in actual sequences by interactions such as hydrogen bonding, hydrophobic, van der Waals, or electrostatic forces, which determine the chain properties and, particularly, its stability (Aida and Nagata, 1986; Ponnuswamy and Gromiha, 1994). Furthermore, mechanisms involved in the production of new genetic material, such as recombination, are associated with the formation of loops that favor the interaction between distant neighbors. The set-up of these intrastrand links at some stage of the evolution of a nucleotide sequence, at either the creation of a new sequence or the enlargement of a preexisting one, may have promoted the observed long-range correlations. On the other hand, it seems that looped structures are more frequent in noncoding regions of DNA, such as intergenic regions, as found for λ bacteriophage DNA (Sanger et al. 1982). Thus, as already pointed out (Grosberg et al., 1993), there seems to be a correlation between spatial arrangement and fractal properties, which would explain why coding/noncoding regions are statistically different, with higher correlations in noncoding regions. Our model is consistent with these considerations. The present work develops a simple model that exhibits the same type of effects as real sequences. In comparison with previous models (Ossadnik et al., 1994; Buldyrev et al., 1993), which also account for some of the features observed in real sequences, our model, because of

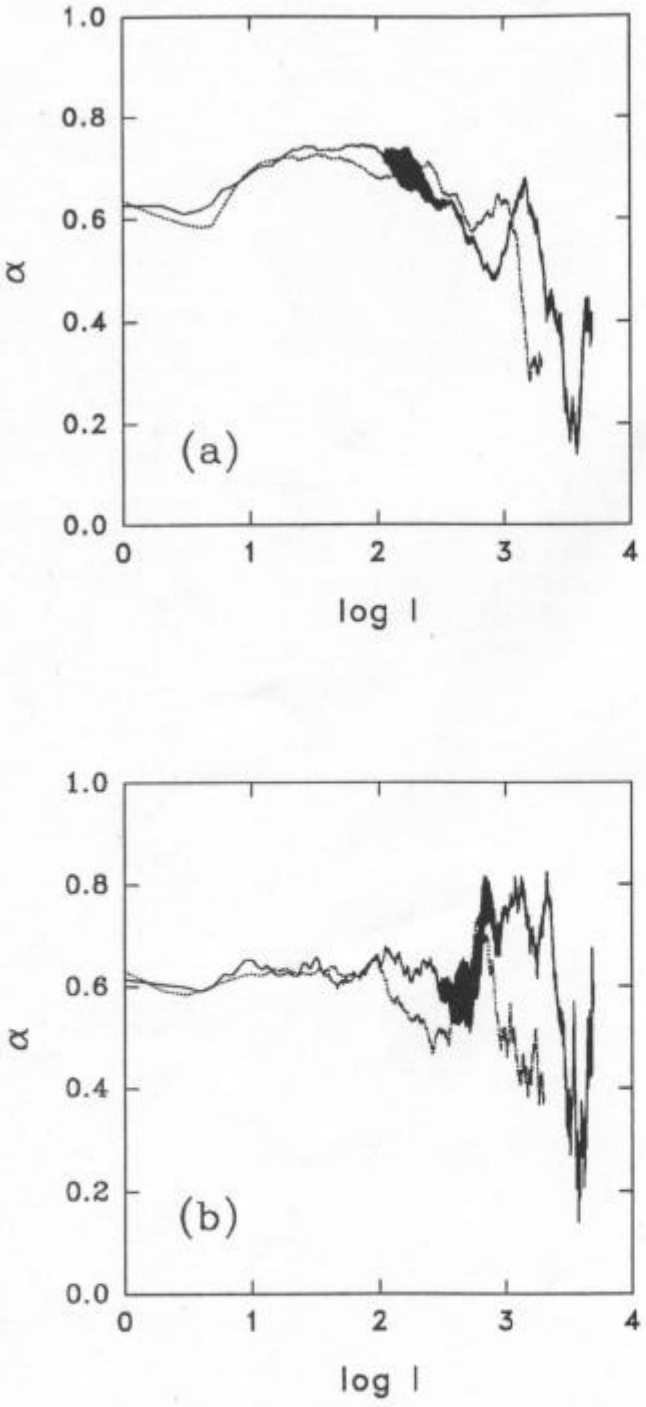

FIGURE 2 Plots of the local slope of the square root mean quadratic fluctuation ($\alpha(l)$) vs. log l for the sequences in Fig. 1. (*a*) corresponds to sequences 1.*a* and 1.*b*. (*b*) corresponds to sequences 1.*c* and 1.*d*. Full lines correspond to real sequences, and dotted lines to artificial ones.

its simplicity, puts into evidence a possible factor (the basic ingredient of the model: interaction between distant units) responsible for the observed correlations, which is not easily evidenced in models with more ingredients. Thus, the exploration of the present model may contribute to a better understanding of the statistical properties exhibited by regions of nucleic acids and of the mechanisms that give rise to them.

We thank the Supercomputing Center of the Universidade Federal do Rio Grande do Sul (CESUP-UFRGS) for the use of the Cray YMP-2E and brazilian agency CNPq for financial support.

REFERENCES

Aida, M., and C. Nagata. 1986. An ab initio molecular orbital study on the stacking interaction between nucleic acid bases: dependence on the sequence and relation to the conformation. *Int. J. Quantum Chem.* 29:1253–1261.

Arneodo, A., E. Bacry, P. V. Graves, and J. F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.* 74:3293–3296.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley. 1993. Generalized Lévy-walk model for DNA nucleotide sequences. *Phys. Rev. E.* 47:4514–4523.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley. 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E.* 51:5084–5091.

Chatzidimitriou-Dreismann, C. A., and D. Larhammar. 1993. Long-range correlations in DNA. *Nature.* 361:212.

Dougherty, E. R. 1990. Topics in applied probability. *In* Probability and Statistics for Engineering, Computing and Physical Sciences. Prentice-Hall, Englewood Cliffs, NJ. 265–308.

Grosberg, A., Y. Rabin, S. Havlin, and A. Neer. 1993. Crumpled globule model of the three-dimensional structure of DNA. *Europhys. Lett.* 23: 373–378.

Karlin, S., and V. Brendel. 1993. Patchiness and correlations in DNA sequences. *Science.* 259:677–680.

Li, W., and K. Kaneko. 1992. Long-range correlation and partial $1/f\alpha$ sepctrum in a noncoding DNA sequence. *Europhys. Lett.* 17:655–660.

Nee, S. 1992. Uncorrelated DNA walks. *Nature.* 357:450.

Ossadnik, S. M., S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons, and H. E. Stanley. 1994. Correlation approach to identify coding regions in DNA sequences. *Biophys. J.* 67:64–70.

Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature.* 356:168–170.

Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley. 1993. Finite-size effects on long-range correlations: implications for analyzing DNA sequences. *Phys. Rev.* E47:3730–3733.

Ponnuswamy, P. K., and M. M. Gromiha. 1994. On the conformational stability of oligonucleotide duplexesan tRNA molecules. *J. Theor. Biol.* 169:419–432.

Prabhu, V. V., and J.-M. Claverle. 1992. Correlations in intronless DNA. *Nature.* 359:782.

Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen. 1982. Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* 162: 729–773.

Tavaré, S., and B. W. Giddings. 1989. Some statistical aspects ofthe primaty structure of nucleotide sequences. *In* Mathematical Methods for DNA Sequences. M. S. Waterman, editor. CRC Press, Boca Raton, FL. 117–132.

Uberbacher, E. C., and R. J. Mural. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA.* 88:11261–11265.

Voss, R. F. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* 68:3805–3808.