CBPF-NF-031/95

A Model for Nucleotide Sequences

André M.C. de Souza[1,2] *and Celia Anteneodo*[1]

[1]Centro Brasileiro de Pesquisas Físicas - CBPF
Rua Dr. Xavier Sigaud, 150
22290-180 - Rio de Janeiro-RJ, Brasil

[2] 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. We show that long-range correlations may be favored by the occurrence of intrastrand interactions which give a non-linear characteristic to the sequence.

Key-words: Long-range correlations, DNA-walk.

Introduction

Since the evolutionary history is registered in the genetic material of modern organisms, some of that history could be reconstructed from the analysis of the nucleotide sequences. Therefore, the study of the characteristics of existing genomes may enlighten our understanding about the processes by which they have evolved. In order to study the stochastic properties of nucleotide sequences, it has been recently proposed the method of "DNA walks" (Peng *et al.*, 1992), basically consisting in the association of a random-walk to a given sequence. This method allows to study the fluctuations of the 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, being $\alpha = 0.5$ for uncorrelated sequences.

Many sequences of genes and cDNA have already been mapped onto unidimensional "DNA-walks" (Peng *et al.*, 1992; Peng *et al.*, 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, a controversy has been generated about the characterization of the sequences exhibiting such long-range correlations. 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 while noncoding sequences present more complex fluctuations, some authors (Nee, 1992; Prabhu and Claverle, 1992; Chatzidimitriou-Dreismann and Larhammar, 1993) found no consistent differences in the α exponent for coding and noncoding sequences and also 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, having recently been shown that the correlation properties allow the identification of coding regions in DNA (Ossadnik *et al.*, 1994).

Anyway, 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 (Tavaré and Giddings, 1989) who has estimated the order of DNA sequences treated 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, 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 was 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, in particular nucleic acids (Ponnuswamy and Gromiha, 1994). These forces between residues within the polymer itself, give place, 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 and in the tentative of finding 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, in discrete Markov chains possessing a finite state space, in which short-range nucleotide interactions are introduced, being relevant the interactions between first neighours and those between more distant neighbours.

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 which base component is a pyrimidine (either cytosine or tymine) occurs at position i, whereas if $S_i = -1$, a purine (either adenine or guanine) occurs at that position. We construct the linear chain assuming that S_{i+1} will be equal to S_j with probability p and different of S_j with probability $(1-p)$, where $j = i$ with probability q and $j \neq i$ with probability $(1-q)$, for $i > j$. By convention, we assume that $S_\circ = 1$.

The probability distribution of the variable S_i is given by:

$$S_i(x) \;=\; p_i\,\delta(x-1) \;+\; (1-p_i)\,\delta(x+1), \tag{1}$$

where p_i is the probability of S_i being equal to $+1$. It is easy to find that it obeys the following recursive relation:

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

with $p_\circ = 1$. The variable $q \in$ [0,1] weights the correlation between first neighbours. Since 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 $j \ge l_c$, 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.

The variable which represents the excess of pyrimidines over purines in a sub-chain of length l is:

$$Y(l) \equiv \sum_{i=1}^{l} S_i. \tag{3}$$

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

$$\langle Y(l)\rangle = 2\sum_{i=1}^{l} p_i \;-l. \tag{4}$$

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

$$F^2(l) \equiv \overline{\left[\Delta Y(l) - \overline{\Delta Y(l)}\right]^2}, \tag{5}$$

where $\Delta Y(l) \;\equiv\; Y(l+l_\circ) - Y(l_\circ)$ and the bar indicates average computed for all the positions $l_\circ$ in the sequence $(1 \le l_\circ \le L-l)$.

In the particular case $q = 1$, the exact solution of the recurrence equation (2) is:

$$p_i \;=\frac{(2p-1)^i \;+\; 1}{2}. \tag{6}$$

From Eq. 5 and using Eq. 6, we calculated $\langle F^2(l)\rangle$ which is equivalent to an average over a large number of statistically independent realizations of the model sequence. We find the following asymptotic behaviour $(1 \ll l \ll L)$:

$$\sqrt{\langle F^2(l)\rangle} \sim \begin{cases} 0 & \text{if } p=1 \\ (\frac{p\,l}{1-p})^{1/2} & \text{if } 0<p<1 \\ 0 & \text{if } p=0,\, l \text{ even} \\ 1 & \text{if } p=0,\, l \text{ odd} \end{cases} \tag{7}$$

From the asymptotic behaviour, we conclude that there is no long-range correlation. The plots α (local slope of $\sqrt{\langle F^2(l)\rangle}$) vs. 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 behaviour of α observed experimentally for highly correlated nucleotide sequences. Since, in this case, the direction of each step depends on the history of the walker, there exits an effect of memory produced in the construction of the chain for $p \neq 1/2$. However, it may be noted, from Eq. 6, that, for $0 < p < 1$, the process corresponds to a stationary process (Dougherty, 1990) in which $\lim_{i\to\infty} p_i = 1/2$, 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 Figs. 1.a and 1.c we show the "dna-walks" of two real sequences: human β-cardiac myosin heavy chain gene and human antithrombin III gene, respectively. In Figs. 1.b and 1.d we show typical walks obtained for different values of the model parameters (p, q and μ). 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 well as the mappings 1.c and 1.d. The plots of the local slope of the fluctuation function vs. 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, while 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. 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 short-range correlations between first neighbours is not sufficient to give place to long-range correlations, as shown by the analysis of the case $q = 1$. On the other hand, when short-range interactions between more distant neighbors are introduced, long-range correlations may arise.

We tested other alternative rules for generating the artificial sequences. Besides the first neighbour, we have also taken into account either 1) a mean over all the other precedent nucleotides or 2) a precedent nucleotide at a fixed distance. For no set of the parameters of these two alternative models we were able to obtain $\alpha(l)$ with a behaviour similar to that of highly correlated actual sequences. Thus, we conclude that a broad distribution of the distance to the second interacting neighbour is required for the uprising of long range correlations.

From the analysis of the case $q = 1$ we also observe that, as in actual sequences, the exponent α is not a constant over all the range of values of l, but it does not mean that there can not be a well-defined exponent, corresponding to the asymptotic behaviour 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. Since real sequences are finite, then we can only say that the observed long-range correlations are of 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, since the presence of biased subdomains also give rise to exponents greater than 0.5 (Nee, 1992). The behaviour of real sequences could result from the combination of some “patching” mechanism and a process such as described in this work. There is also a possibility that long-range correlations observed in real sequences arise purely from short-range 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 which determine the chain properties and, particularly, its stability (Aida and Nagata, 1986; Ponnuswamy and Gromiha, 1994). Besides, mechanisms, such as recombination, involved in the formation of new genetic material, are associated to the formation of loops which favor the interaction between distant neighbours. The set up of these intrastrand links at some stage of the evolution of a nucleotide sequence, either creation of a new sequence or growing up 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 (Grosberg *et al.*, 1993), there seem 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 which 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 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 which originate them.

Captions for figures

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 $l_c = 4$; (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 $l_c = 2$ (d).

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.c; (b) corresponds to sequences 1.b and 1.d. Full lines correspond to real sequences and dotted lines to artificial ones.

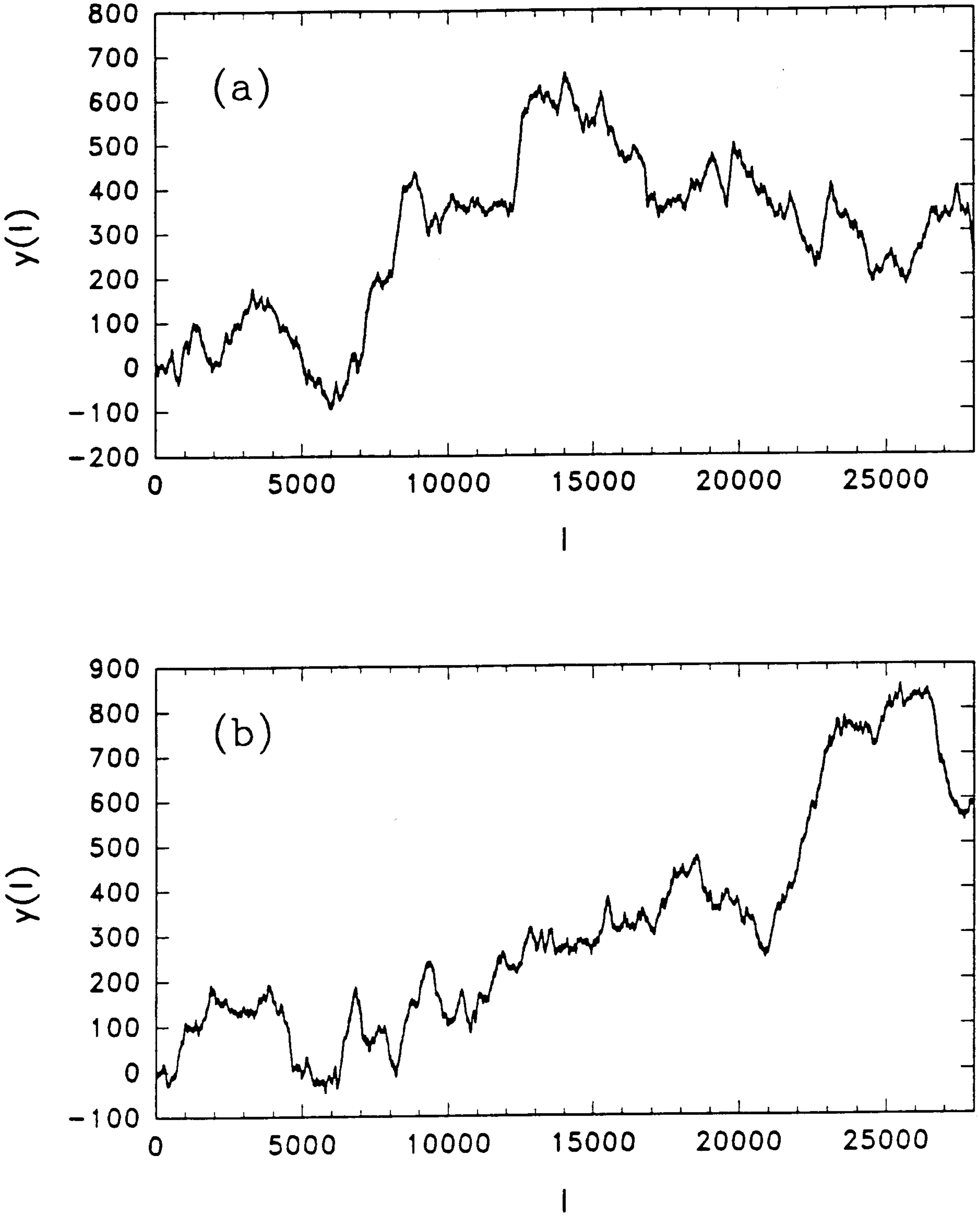

(a)
y(l)
800
700
600
500
400
300
200
100
0
−100
−200
0
5000
10000
15000
20000
25000
l
(b)
y(l)
900
800
700
600
500
400
300
200
100
0
−100
0
5000
10000
15000
20000
25000
l

Figure 1

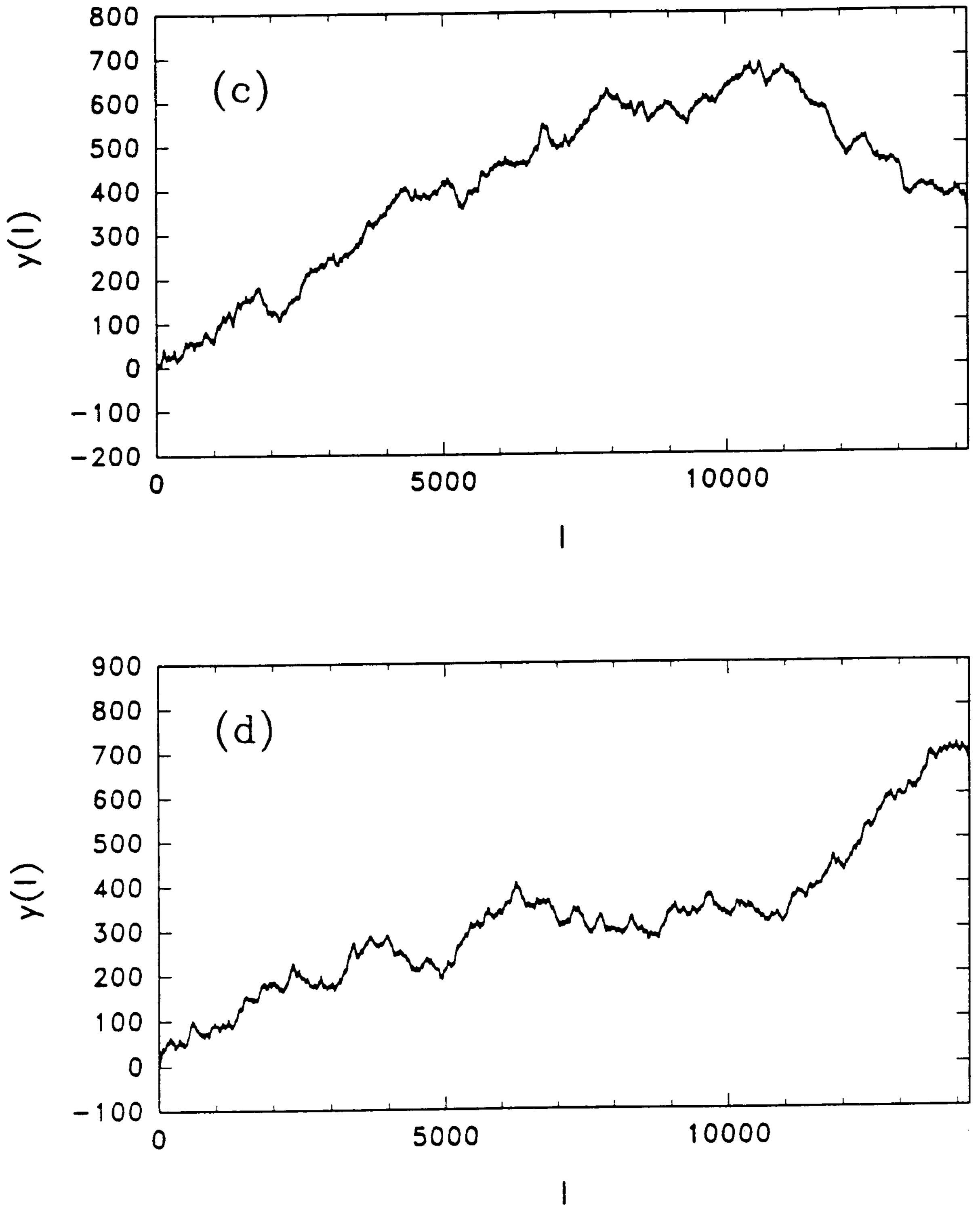
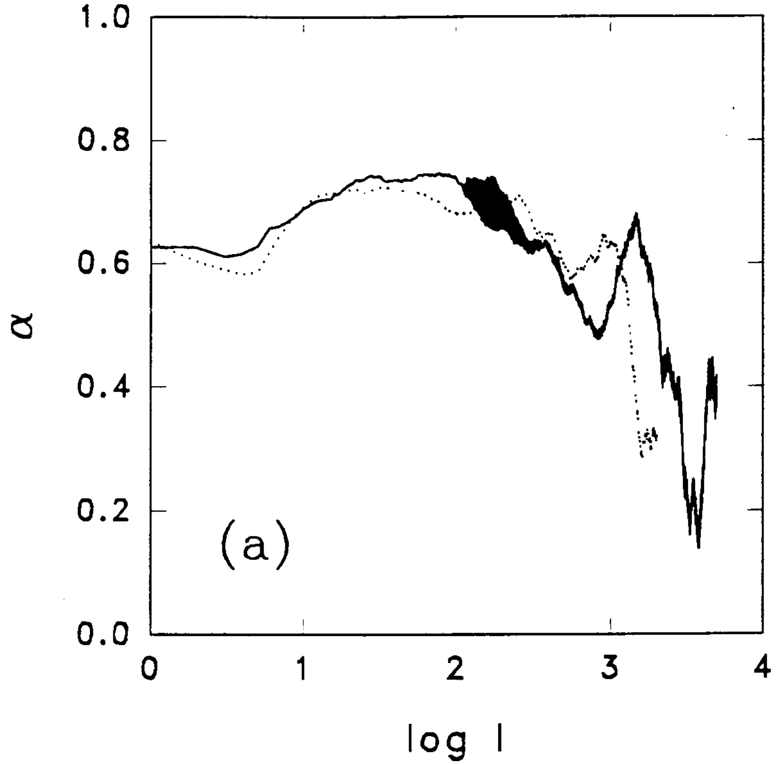
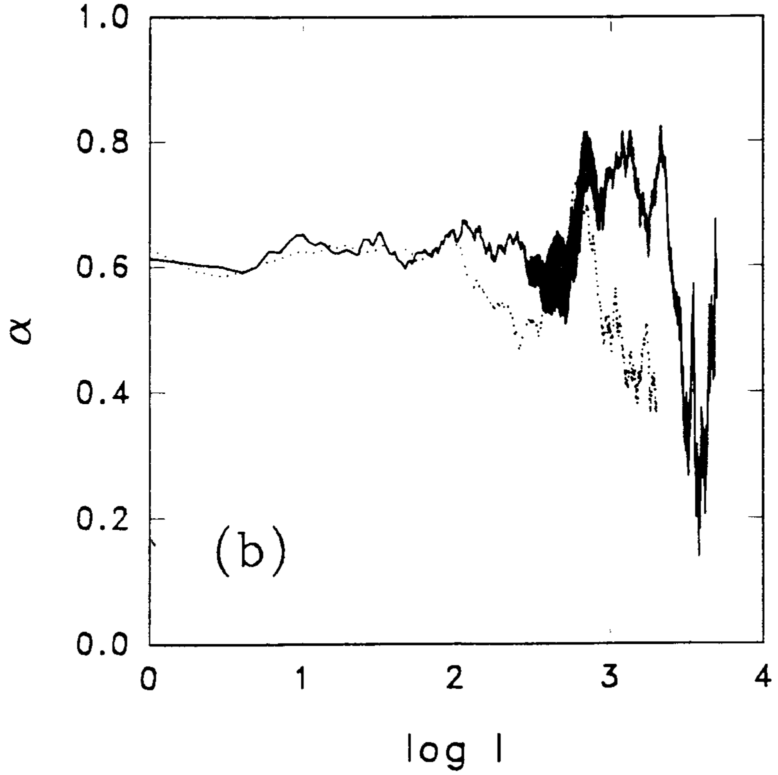

(c)
y(l)
800
700
600
500
400
300
200
100
0
−100
−200
0
5000
10000
l
(d)
y(l)
900
800
700
600
500
400
300
200
100
0
−100
0
5000
10000
l

Figure 1


1.0
0.8
0.6
0.4
0.2
0.0
α
(a)
0
1
2
3
4
log l


1.0
0.8
0.6
0.4
0.2
0.0
α
(b)
0
1
2
3
4
log l

References

Aida, M. & Nagata, C. (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.

Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M. & Stanley, H. E. (1993). Generalized Lévy-walk model for DNA nucleotide sequences. *Physical Review E* **47**, 4514-4523.

Chatzidimitriou-Dreismann, C. A. & Larhammar, D. (1993). Long-range correlations in DNA. *Nature* **361**, 212.

Dougherty, E. R. (1990). Topics in applied probability. *Probability and Statistics for Engineering, Computing and Physical Sciences*, pp. 265-308. Prentice-Hall Int. Ed., New Jersey.

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

Karlin, S. & Brendel, V. (1993). Patchiness and correlations in DNA sequences. *Science* **259**, 677-680.

Li, W. & Kaneko, K. (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., Buldyrev, S. V., Goldberger, A. L., Havlin, Mantegna, R. N., Peng, C.-K., Simons, M. & Stanley, H. E. (1994). Correlation approach to identify coding regions in DNA sequences. *Biophys. J.* **67**, 64-70.

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Simons, M. & Stanley, H. E. (1993). Finite-size effects on long-range correlations: Implications for analyzing DNA sequences. *Physical Review E* **47**, 3730-3733.

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

Ponnuswamy, P. K. & Gromiha, M. M. (1994). On the conformational stability of oligonucleotide duplexesan tRNA molecules. *J. Theor. Biol.* **169**, 419-432.

Prabhu, V. V. & Claverle, J.-M. (1992). Correlations in intronless DNA. *Nature* **359**, 782.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. (1982). Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* **162**, 729-773.

Tavaré, S. & Giddings, B.W. (1989). Some statistical aspects ofthe primaty structure of nucleotide sequences. In: *Mathematical Methods for DNA Sequences* (Waterman M.S. ed.) pp. 117-132. CRC Press, Boca Raton.

Uberbacher, E. C. & Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. U.S.A.* **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.