Nonlinear modeling technique for the analysis of DNA chains

J. Barral P.,[1,2] A. Hasmy,[1] J. Jiménez,[3] and A. Marcano[3]
[1]*Laboratorio de Física Estadística de Sistemas Desordenados, Centro de Física, IVIC, Apartado Postal 21827, Caracas 1020A, Venezuela*
[2]*Departamento de Biología Celular y Molecular, Facultad de Ciencias, Universidade da Coruña, Campus a Zapateira, S/N 15071 A Coruña, Spain*
[3]*Laboratorio de Fenómenos no Lineales, Escuela de Física, Facultad de Ciencias, Universidad Central de Venezuela, Apartado Postal 52120, Caracas 1050 A, Venezuela*
(Received 6 July 1999)

Using a simple computational procedure, we examine DNA chains from different species in order to prove their nonlinear deterministic structures. This procedure applies a nonlinear modeling technique based upon quantitative comparison of the neighborhoods from similar DNA subsegments of size *d*. Our results reveal that noncoding regions exhibit a deterministic signature at sizes larger than a characteristic dimension d_c. Applications to evolutionary categories and recognition of different DNA regions are discussed.

PACS number(s): 87.10.+e, 05.45.−a, 47.20.Ky, 87.14.Gg

As the human and mouse genome projects are in a phase of systematic sequencing, computational tools based on concepts used in many science fields have recently played a prominent role. To cite a few examples, we could mention gene identification [1,2], assignment of tentative functions to particular sequences [3–6], and elucidation of their structure [6–17]. The increasing interest in DNA chains is due to their fundamental importance in living organisms, since all information of the species evolution is contained in these macromolecules. A relevant contribution to the study is due to statistical methods, usually employed in physics, to determine the nature of a series of events, namely Markovian approximations [18], correlation functions, and Fourier transform [6,8,9], etc. However, these methods do not give specific information of how different regions are characterized, and also fail to distinguish one given species from another. For instance, Markovian approximations describe a genome in terms of *k*-tuple overlapping series of nucleotides (where *k* is the Markovian order) and might ignore some correlations. Fourier transforms only detect periodicity and possible correlations, but the information associated with these correlations lacks relevant details about the composition of DNA chains. On the other hand, scientists in the field are trying combinations of different methods for the recognition of coding and noncoding DNA regions (based on techniques such as those mentioned above) in order to improve the accuracy for prediction of different packages, which actually reach approximately 90% of accuracy [19,20]. For these reasons, alternative tools able to give different ideas and estimators concerning the structure of DNA chains represent an important contribution in the field.

DNA consists of two complementary chains, each of them being a linear polynucleotide consisting of four nitrogenated bases: adenine (A), cytosine (C), guanine (G), and thymine (T). A is paired with T and G with C in the complementary chain. A gene is a sequence of DNA that is essential for a specific function. Traditionally, a gene was defined as a segment of DNA that codes for a polypeptide chain or specifies a functional RNA molecule. Recent molecular studies, however, have altered our perception of genes, making a somewhat vague definition necessary. According to Li [21], a gene is a sequence of genomic DNA or RNA that performs a specific function. Performing the function may not require the gene to be translated or even transcribed. At present, three types of genes are recognized: (i) protein-coding genes, which are transcribed into RNA and subsequently translated into proteins, (ii) RNA-specifying genes, which are only transcribed, and (iii) regulatory genes. According to a narrow definition, the third category includes only untranscribed sequences. The protein-coding genes in bacterias differ from those in eukaryotes in several aspects. The eukaryotic protein-coding gene consists of coding and noncoding parts. The noncoding parts are distinguished according to their location in: *flanking* sequences (upstream or downstream of the protein coding region), and *introns* (which are ignored during the processing of the mRNA molecule) [21]. All sequences that remain in the mature RNA following splicing are referred to as *exons*. Protein genes in bacterias do not contain introns, and may be arranged consecutively to form a unit of gene expression (operon). A large proportion of eukaryotic DNA is apparently nonfunctional, a large part of this DNA is accounted for by noncoding parts, the intergenic regions and introns. The genome of eukaryotes is known to contain various types of repetitive DNA. Repetitive DNA is any piece of nucleotide sequence which is repeated several to many times in the genome. The function of repetitive DNA is unknown and some classes of this DNA seem to be nonfunctional (junk DNA) [22].

In this paper, we report results concerning the nonlinear deterministic structure of nuclear DNA chains obtained by applying a nonlinear modeling (NM) technique. We found that, while coding (exonic) regions behave as uncorrelated random chains, noncoding regions exhibit deterministic signatures.

The NM method applied here to DNA chains of different species has been previously used successfully to distinguish between chaos and noise in time series [23–26]. This is possible because it explores, quantitatively, similarity along the sequence at subchain vicinities of equal subchains of size *d* (the embedding dimension).

Departing from an arbitrary data series $x_1x_2x_3 \ldots x_N$, the NM technique works by organizing the series in d-dimensional delay-register vectors:

$$X_1 \equiv (x_1, x_2, \ldots, x_d),$$
$$X_2 \equiv (x_2, x_3 \ldots, x_{d+1}),$$
$$\vdots \qquad (1)$$
$$X_{N-d+1} \equiv (x_{N-d+1}, \ldots, x_N),$$

which correspond to a catalog of all possible segments of d consecutive data values. Next, for each vector $X_p = (x_p, x_{p+1}, \ldots, x_{p+d-1})$; $(1 \leq p \leq N-d)$, one searches for its nearest neighbor $X_{H(p)} = (x_{H(p)}, x_{H(p)+1}, \ldots, x_{H(p)+d-1})$ and then compares how close the data values x_{p+d} and $x_{H(p)+d}$ are following these two vectors.

For a DNA sequence, for instance ACCATTGACT . . . , each data value x_i will consist of one of four symbols A, C, G, or T. In order to compare the closeness of a pair of data points x_i and x_j we use a Hamming-like metric:

$$h(x_i, x_j) = \begin{cases} 0 & \text{if} \quad x_i = x_j \\ 1 & \text{if} \quad x_i \neq x_j . \end{cases} \qquad (2)$$

Therefore, $h(A,C) = h(A,G) = h(A,T) = h(C,G) = h(C,T) = h(G,T) = 1$ and $h(A,A) = h(C,C) = h(G,G) = h(T,T) = 0$. As a natural extension, the closeness of a pair of vectors X_i and X_j will be measured by

$$H(X_i, X_j) = \sum_{k=0}^{k=d-1} h(x_{i+k}, x_{j+k}), \qquad (3)$$

and the nearest neighbor $X_{H(p)}$ of a given vector X_p is randomly selected among the solutions of $H(X_p, X_j)$ such that it is a minimum for $j \neq p$.

Once the nearest neighbor $X_{H(p)}$ has been determined, we compute the local error: $\epsilon_p \equiv h(x_{p+d}, x_{H(p)+d})$ and from this, the overall mean error in the chain:

$$\langle E \rangle = \frac{1}{N-d} \sum_{p=1}^{N-d} \epsilon_p = \frac{1}{N-d} (\epsilon_1 + \epsilon_2 + \cdots + \epsilon_{N-d})$$
$$= \frac{1}{N-d} [h(x_{1+d}, x_{H(1)+d}) + h(x_{2+d}, x_{H(2)+d}) + \cdots + h(x_N, x_{H(N)})], \qquad (4)$$

where, as was already mentioned, the subscript $H(1)$ corresponds to the vector $X_{H(1)} = (x_{H(1)}, x_{H(1)+1}, x_{H(1)+2}, \ldots, x_{H(1)+d-1})$ which is the nearest neighbor of $X_1 = (x_1, x_2, \ldots, x_d)$ in the metric defined by Eqs. (2)–(3); $H(2)$ corresponds to the vector $X_{H(2)} = (x_{H(2)}, x_{H(2)+1}, x_{H(2)+2}, \ldots, x_{H(2)+d-1})$ which is the nearest neighbor of $X_2 = (x_2, x_3, \ldots, x_{d+1})$ in the metric defined by Eqs. (2) and (3), etc.

For uncorrelated random chains, there is no relation between any value x_{p+d} and the vector X_p, and in that case the error in Eq. (4) can be approximated by $p(A)[1-p(A)] + p(C)[1-p(C)] + p(G)[1-p(G)] + p(T)[1-p(T)]$, $p(\cdot)$, being the probability of occurrences for the symbol $(\cdot)$. Consequently, for such series, the error in Eq. (4) will not depend on the embedding dimension d. In particular, for a uniform process $(p(A) = p(C) = p(G) = p(T) = 0.25)$, $\langle E \rangle = 0.75$.

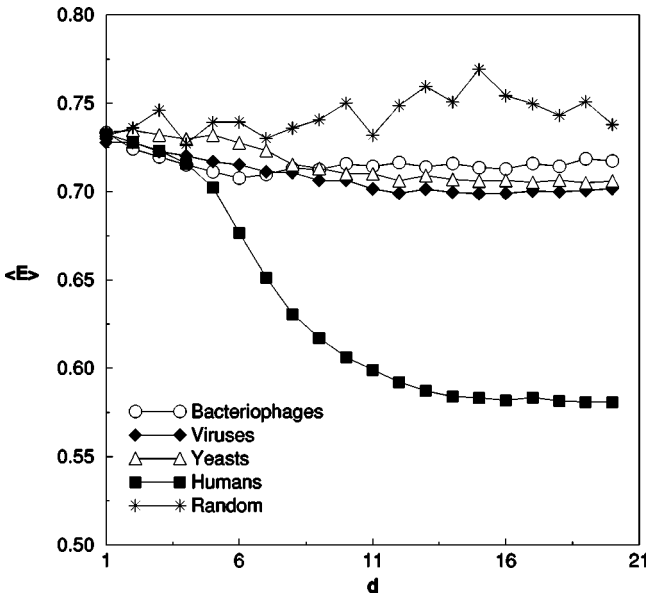

FIG. 1. Average error $\langle E \rangle$ versus the embedding dimension d for bacteriophages, viruses, yeasts, and humans. For comparison we show the case of a random sequence.

Several DNA sequences corresponding to a classification of the genome complexity as bacteriophages, eukaryotic viruses, yeasts, and humans have been analyzed by means of the NM method. For these species, the average error $\langle E \rangle$ has been computed as a function of the embedding dimension d. These results are reported in Fig. 1. Note that while in the numerically simulated random sequence, $\langle E \rangle$ does not depend upon d and is kept around 0.75, the curves corresponding to DNA chains decrease as d is increased from 1 to a characteristic embedding dimension d_c, where the average error $\langle E \rangle$ reaches a minimum value $\langle E_{min} \rangle$. For $d > d_c$ we did not observe significant variation of $\langle E \rangle$ up to $d = 50$, which corresponds to the largest d value we considered. All evolutionary categories were averaged over three different sequences whose GenBank accession numbers (Gan) are: U24159, Z47794, and J02495 for bacteriophages; Z86009, L22858, and J01917 for viruses; X59720, D50617, and Z47047 for yeasts, and U47924, U07000, and M26434 for humans.

This behavior can be interpreted as the existence of an underlying deterministic rule in the structure of nucleotide sequences for a given length scale, which could be hard to detect with other methods. In fact, even for some DNA chains which exhibit a broad-band power spectra similar to that of white noise (as one of the human sequences considered here), our approach shows a dependency of $\langle E \rangle$ on d, implying a deterministic structure, rather than the apparent random one. An interesting result that turns out by inspection of the $\langle E(d) \rangle$ curves is that the smallest value for $\langle E \rangle$ corresponds to the human case which is the more complex genome.

Coding and noncoding regions have been analyzed sepa-

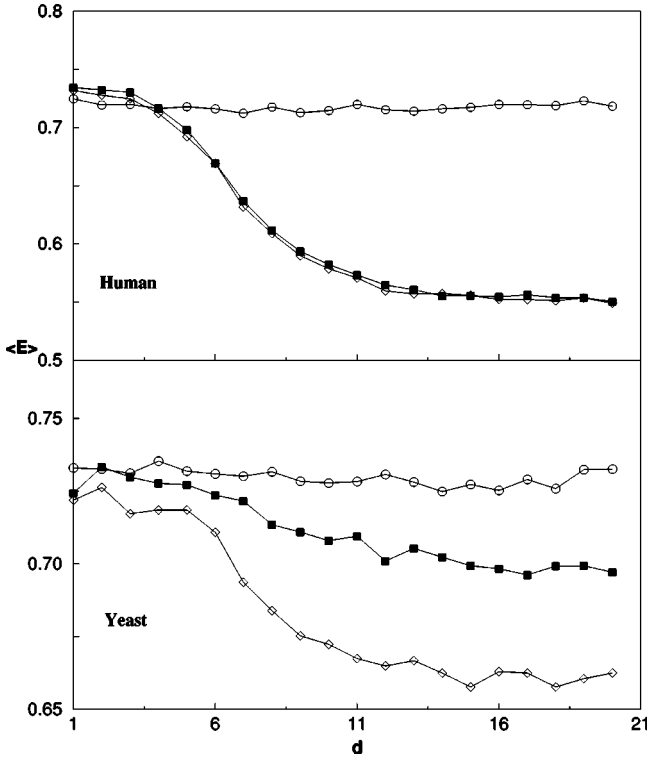

FIG. 2. Average error $\langle E \rangle$ versus the embedding dimension d for coding (circles), noncoding (diamonds), and complete (squares) DNA sequences corresponding to a human (Gan=U47924), and a yeast DNA chain (Gan=D50617).

rately in order to determine the origin of the deterministic behavior shown in Fig. 1. Figure 2 shows the comparative study of DNA sequences from humans (a) and yeast (b). Black symbols denote the overall sequence analyzed. In both cases, the curves corresponding to coding regions (open circles) behave as random chains (i.e., $\langle E \rangle \approx 0.75$ for all d). This happens in spite of typical information contained in these regions which are manifested as a three periodicity [we recall that this periodicity is associated to the nucleotide triplets (codons) that specify one of 20 aminoacides in coding regions [3,27,28]]. Furthermore, noncoding curves (open diamonds) exhibit $\langle E \rangle$ values lower than those from corresponding complete DNA chains. Note that $\langle E_{min} \rangle$ is still smaller in the human case. The difference between the complete genome curve and the noncoding one in yeasts, is due to the amount of coding regions in this species, approximately 70%, in contrast to the human case whose DNA chains contain approximately 10% of exons [29]. Note that from the above results one can conclude that genes containing only coding regions (as bacterias) will behave as random chains.

Although curves reported in Figs. 1 and 2 consider chains with $N=32{,}768$, similar findings are observed for smaller chains. In order to investigate how these ideas could be used to recognize regions in DNA chains, we have assigned to each position of the sequence, the $\langle E \rangle$ value corresponding to a subsequent defined window (i.e., corresponding to the following l nucleotides, l being the size of the window) and we stored the value of $\langle E \rangle$ for $d=16$ where, as observed in Fig. 1, all curves saturate to their corresponding $\langle E_{min} \rangle$. In Fig. 3 we report $\langle E(d=16) \rangle$ for a complete human [Fig. 3(a)] and yeast [Fig. 3(b)] DNA sequence. In these figures, shadowed zones correspond to coding DNA regions as determined by GRAIL software [30]. We compared these results with those obtained using window sizes of 500 and 2000, and we found that the difference was negligible. On the other hand, in Fig. 3(a) one can observe that the minima of these curves are located in noncoding regions (clear re-

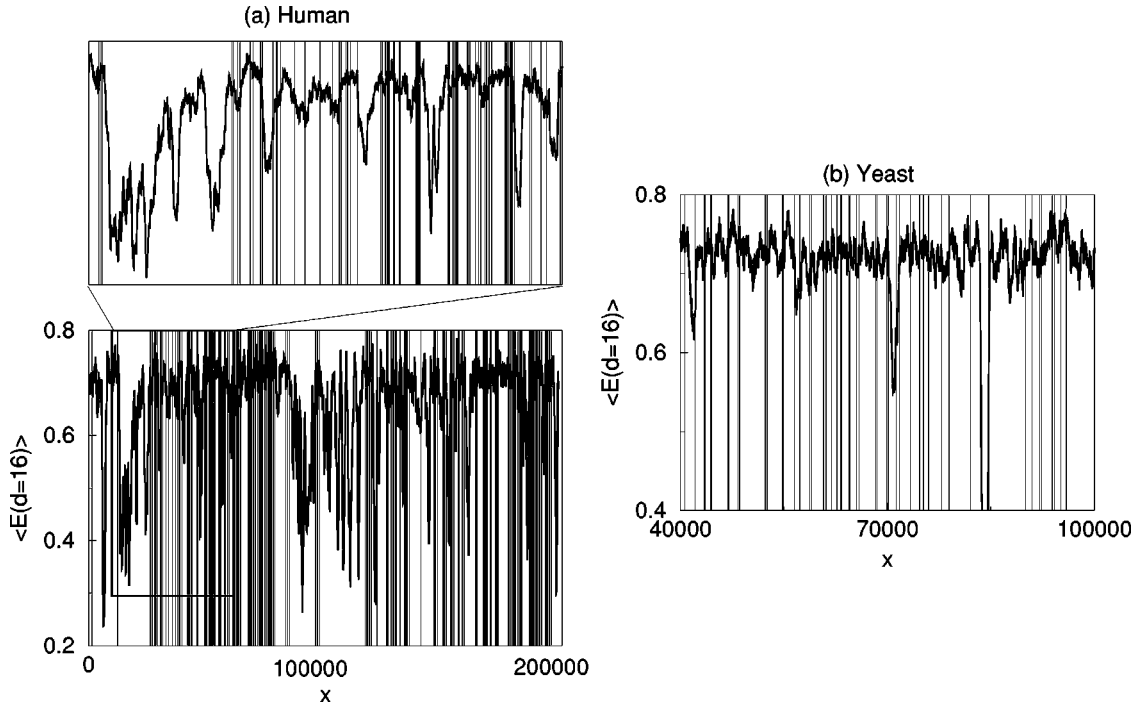

FIG. 3. Average error for $d=16$, $\langle E(d=16) \rangle$, versus their position x inside the chain (solid curve) for a (a) human (a magnified zone is illustrated in the top of this figure), and (b) yeast. Shadowed zones denote coding regions as determined by GRAIL software.

gions), and the wider regions have the deeper minima. In fact, a histogram of $\langle E(d=16)\rangle$ values for different intervals (corresponding to a variation of the noncoding zone width) corroborates this (data not shown).

In Fig. 3(b) it is more difficult to distinguish typical minima, since noncoding regions are much smaller than coding (70%) in yeasts. This is mainly due to the considered window size, since small noncoding parts overlap with exons during the statistical analysis, i.e., the curves are governed by the latter regions. From this, it is clear that further work is needed to find a better criterion to define an optimal window size. We think that combination of this method with other criteria used in the recognition of DNA regions will provide the key to better approaches.

In some DNA chains, from yeasts and humans, and with calculations of cross correlations and Fourier transforms, we tried without success to obtain information on the deterministic structure of these systems. Only the sensitivity of the NM method was able to detect this. We emphasize that this determinism does not concern the typical repetitive DNA regions [22], but concerns an averaged redundancy of approximate repeats of DNA chains. In fact, the majority of $\langle E\rangle$ values smaller than 0.75 in Fig. 3 corresponded to windows (of size = 1000) where DNA chains do not show repeated regions as reported in GenBank. The origin of long range correlations in DNA chains has been a subject of controversies [6,13–17]. From our calculations, one could suppose that these correlations may be due to a deterministic structure of noncoding regions. Information of the DNA structure obtained by the NM technique differs from that provided by standard methods used previously to detect correlations; for instance, while determinism is not detected by the standard methods, the NM technique applied here fails to detect typical periodicity at short ranges in exons. Therefore, it seems that complementarity between different methods is the best way to characterize DNA structures. In particular, we have shown that the NM method provides a powerful tool for obtaining information concerning how nucleotides in noncoding regions are organized. Another interesting point that we would like emphasize is that the characteristic embedding dimension d_c and $\langle E_{min}\rangle$ give different results for each evolutionary category (see Fig. 1). The latter might be helpful to distinguish between evolutionary categories, but this point needs to be explored further.

We thank E. Medina D. and H. Naveira F. for valuable comments about the manuscript. One of us (J.B.P.) would like to acknowledge financial support from Universidade da Coruña and Consellería de Educación e Ordenación Universitaria de la Xunta de Galicia.

[1] J. W. Fickett, Trends Genet. **12**, 316 (1996).
[2] J.-M. Calvarie, Hum. Mol. Genet. **6**, 1735 (1997).
[3] J. C. W. Shepherd, Proc. Natl. Acad. Sci. USA **78**, 1596 (1981).
[4] F. S. Collins, Proc. Natl. Acad. Sci. USA **92**, 10 801 (1995).
[5] D. A. Benson *et al.*, Nucleic Acids Res. **24**, 1 (1996).
[6] W. Li and K. Kaneko, Nature (London) **360**, 635 (1992).
[7] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[8] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[9] V. R. Chechetkin and A. Y. Turygin, J. Theor. Biol. **178**, 205 (1996).
[10] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, Phys. Rev. A **45**, 8902 (1992).
[11] M. de Sousa Vieira and H. J. Herrmann, Europhys. Lett. **33**, 409 (1996).
[12] N. V. Dokholyan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, Phys. Rev. Lett. **79**, 5182 (1997).
[13] S. Nee, Nature (London) **357**, 450 (1992).
[14] J. Maddox, Nature (London) **358**, 103 (1992).
[15] P. J. Munson, R. C. Taylor, and G. S. Michaels, Nature (London) **360**, 636 (1992).
[16] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, F. Sciortino, and H. E. Stanley, Phys. Rev. Lett. **71**, 1776 (1993).
[17] R. F. Voss, Phys. Rev. Lett. **71**, 1777 (1993).
[18] A. Krogh, in *Computational Methods in Molecular Biology*, edited by S. Salzberg, D. Searls, and S. Kasif (Elsevier Science B.V., Amsterdam, 1998).
[19] M. Burset and R. Guigó, Genomics **34**, 353 (1996); C. Burge and S. Karlin, J. Mol. Biol. **268**, 78 (1997); M. Q. Zhang, Proc. Natl. Acad. Sci. USA **94**, 565 (1997).
[20] Different package performances are also reviewed by W. Li at http://linkage.rockefeller.edu/wli/gene/
[21] W. H. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997); W. H. Li and D. Graur, *Fundamentals of Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1991).
[22] M. Nei, *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987), pp. 120–145.
[23] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).
[24] G. Sugihara and R. May, Nature (London) **344**, 734 (1990).
[25] D. M. Rubin, Chaos **2**, 525 (1992).
[26] P. García, J. Jiménez, A. Marcano, and F. Moleiro, Phys. Rev. Lett. **76**, 1449 (1996).
[27] J. W. Fickett, Nucleic Acids Res. **10**, 5303 (1982).
[28] W. Lee and L. Luo, Phys. Rev. E **56**, 848 (1997).
[29] J. Barral P. and A. Hasmy (unpublished).
[30] E. C. Uberbacher and R. J. Mural, Proc. Natl. Acad. Sci. USA **88**, 11 261 (1991).