ELSEVIER

Physica A 300 (2001) 271–284

PHYSICA A

www.elsevier.com/locate/physa

Multifractal analysis of DNA sequences using a novel chaos-game representation

J.M. Gutiérrez[a], M.A. Rodríguez[b,*], G. Abramson[c]

[a] *Departamento de Matemática Aplicada, Universidad de Cantabria, 39005 Santander, Spain*

[b] *Instituto de Física de Cantabria, CSIC-Universidad de Cantabria, 39005 Santander, Spain*

[c] *Consejo Nacional de Investigaciones Científicas y Técnicas, Centro Atómico Bariloche,*
8400 S. C. de Bariloche, Argentina

Received 1 February 2001

Abstract

We present a generalization of the standard chaos-game representation method introduced by Jeffrey. To this aim, a DNA symbolic sequence is mapped onto a singular measure on the attractor of a particular IFS model, which is a perfect statistical representation of the sequence. A multifractal analysis of the resulting measure is introduced and an interpretation of singularities in terms of mutual information and redundancy (statistical dependence) among subsequence symbols within the DNA sequence is provided. The multifractal spectrum is also shown to be more sensitive for detecting dependence structures within the DNA sequence than the averaged contribution given by redundancy. This method presents several advantages with respect to other representations such as walks or interfaces, which may introduce spurious effects. In contrast with the results obtained by other standard methods, here we note that no general statement can be made on the influence of coding and non-coding content on the correlation length of a given sequence. © 2001 Elsevier Science B.V. All rights reserved.

PACS: 05.45.Df; 87.10.+e; 87.14.Gg

Keywords: Multifractal analysis; DNA sequences; Iterated function system; Chaos game

1. Introduction

The large accumulation of data in the DNA databases has aroused considerable interest in the statistical analysis of DNA sequences in the recent years. Li and collaborators have reviewed the early literature on the topic [1,2], including the pioneering work of

[*] Corresponding author.
E-mail address: rodrigma@ifca.unican.es (M.A. Rodríguez).

PII: S0378-4371(01)00333-8

two of them on the $1/f^{\alpha}$ spectrum of DNA sequences [3]. By mapping the sequence onto a (1D) walk, Peng and others have built a kind of interface, whose statistics were used to probe the range of correlation of the sequences [4,5]. Linguistic features were claimed to have been found in noncoding DNA sequences [6], a point that has provoked controversy [7–10]. Still others have emphasized the fractality hidden in some or other representations of the sequences [11–14].

In this work, we will follow an approach pioneered by Jeffrey [15] whose potential, to our understanding, has not been previously fully acknowledged. The method works by mapping a DNA sequence onto an orbit on the attractor of an iterated function system (IFS), forming a graphical pattern of the DNA sequence. To this aim, Jeffrey considered an IFS model consisting of four transformations t_A, t_G, t_T, t_C with contractive factors $k_i = \frac{1}{2}$, each of them mapping the unit square onto a corner of itself. This IFS model was inspired by the four letter (A, G, T, and C) composition of the DNA sequence, representing the nucleotides adenine, guanine, thymine and cytosine, respectively. In this case, the whole unit square is the attractor of the IFS and it can be filled by feeding the probabilistic algorithm known as the “chaos-game” [16] with a random sequence. However, when using a DNA sequence to feed the chaos-game, a special pattern (a subset of the unit square) is obtained, revealing the structure of the nucleotides sequences by visual inspection. Jeffrey’s study limited itself to a few general (graphical) features of the attractor thus generated. These were proven to originate simply in the distribution of mono-, di-, and trinucleotide probabilities in the sequence [17]. However, in spite of the limitations of this representation, it has found several interesting applications, such as representing amino-acid sequences, determining protein structures, and characterizing the evolution of species [18–20].

Jeffrey’s method only deals with the graphical pattern generated on the IFS attractor by the chaos-game orbit. However, the measure generated on the attractor by this orbit provides much more information of the DNA sequence (see [21,22] and references therein). This measure is a perfect statistical representation of the sequence and presents several advantages with respect to other representations such as walks or interfaces, since it probes directly into the distribution of subsequences, independent of any mapping into a walk, interface, etc. Longer sequences provide more minute details and allow the analysis of the attractor up to longer resolution scales. Some attempts for generalizing Jeffrey’s representation by taking advantage of this singular measure have been recently proposed. For instance, Gutiérrez et al. [21] introduce a multifractal analysis of measures corresponding to chaos-game representations of DNA sequences and describe its application for the analysis of long DNA sequence correlation. On the other hand, Tino [23] generalizes Jeffrey’s representation to deal with n-symbol alphabets and uses Rényi entropy to establish a correspondence between the statistical properties of a symbolic sequence and some information related properties of the corresponding measure.

In this paper, we present a new multifractal method which can be easily interpreted in terms of statistical dependence using the concept of redundancy from information theory. To this aim, a novel chaos-game representation is introduced by assigning scaling factors to the transformations according to the frequencies of the associated symbols within the sequence under study. Therefore, a different IFS representation is associated with each symbolic sequence, including statistical information about the symbols in the representation. To our knowledge, other chaos-game representations previously reported in the literature consider a fixed scaling factor for all the transformations and, therefore, the following analysis is not possible in those cases. We show how our representation provides a connection between the multifractal spectrum of the resulting measures and the mutual information, or redundancy, of nucleotides separated by a prescribed distance within the DNA sequence, thus characterizing the statistical dependence structure of subsequences within the symbolic sequence. Moreover, the spectrum of singularities displays the contribution to the averaged redundancy of different combinations of symbols forming precise subsequences within the DNA sequence, thus allowing a more detailed analysis of the statistical structure of these sequences. As illustrated by several examples, the resulting multifractal analysis allows us to study some important properties of DNA, such as the existence of long-range correlations in coding and non-coding DNA sequences previously analyzed with other methods.

In Section 2, the chaos-game representation model is presented, and several advantages over other representations are discussed. In Section 3, we briefly describe the multifractal analysis used in this paper and analyze the connection with the concepts of mutual information and redundancy. Finally, some results of the methodology, when applied to some DNA coding and non-coding sequences, are presented in Section 4, establishing special emphasis on those results which have been obscured by other standard procedures.

2. An alternative IFS representation for symbolic sequences

The chaos-game representation introduced by Jeffrey uses IFS models to represent symbolic sequences. An IFS is a finite collection, $t_1,\ldots,t_m$, of linear contractive maps with contraction factors $k_1,\ldots,k_m$ on $\mathbb{R}^n$ [24,25]. The contractivity property determines the existence of a subset $A\subset\mathbb{R}^n$, called the attractor of the IFS, which exhibits self-similar structure, as it is the union of m affine copies $t_i(A)$, $i=1,\ldots,m$ of itself, as given by $A=\bigcup_{i=1}^m t_i(A)$. This property provides a convenient framework for symbolic analysis of sequences from an alphabet $\{a_1,\ldots,a_m\}$, since any subsequence $a_{i_1}a_{i_2}\ldots a_{i_L}$ of length L corresponds to a succession of applications of the maps which determine a region of the attractor $t_{i_L}(\ldots t_{i_2}(t_{i_1}(A)))$ of size $k_{i_1}k_{i_2}\ldots k_{i_L}$ called an order-L iterator of the attractor (where each of the letters a_i has been associated with the corresponding transformation t_i). Therefore, the attractor can be partitioned at different resolution scales by the m^L order-L iterators.

Jeffrey's 2D representation has the advantage of visual appeal but, as we shall show later, it hinders an analytical treatment of the obtained measure. For this reason, for a finite alphabet consisting of m symbols we consider an alternative 1D representation consisting of m non-overlapping similitudes of factor $k_i=1/m$, which maps the unit

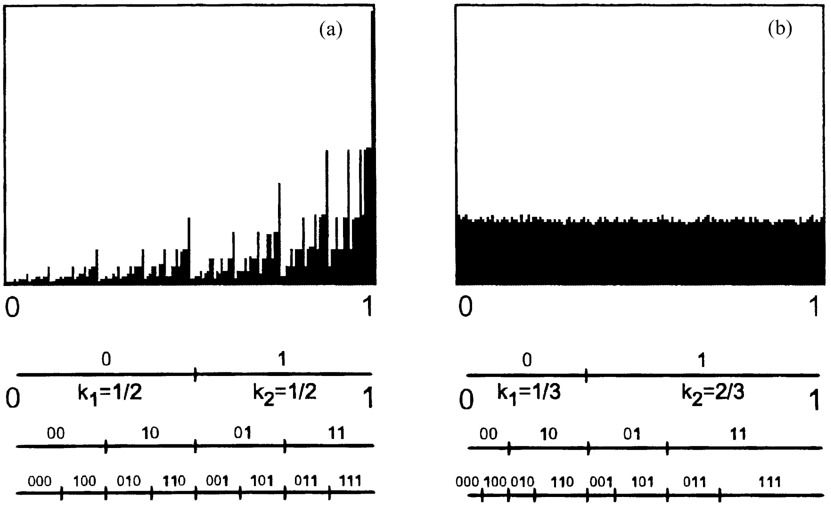

Fig. 1. Standard $k_1 = k_2 = \frac{1}{2}$ (a) and detrended $k_1 = \frac{1}{3}$, $k_2 = \frac{2}{3}$ (b) IFS representation models (below) and multifractal measures (above) corresponding to a binary sequence with $\frac{1}{3}$ zeroes content and $\frac{2}{3}$ ones content.

interval onto itself [23]

$$t_i(x) = \frac{1}{m}x + \frac{i-1}{m}, \quad i = 1,\ldots,m\,, \tag{1}$$

with contractivity factors $k_i = 1/m$. Therefore, the image of a sequence $a_{i_1}a_{i_2}\ldots a_{i_L}$ lies in one of the m^L order-L iterators of size $1/m^L$ in which the unit segment is divided. In other words, the analysis of the attractor up to resolution $1/m^L$ is equivalent to the analysis of the subsequences of length L contained in the sequence. Fig. 1(a) illustrates the case $m = 2$ with different iterators associated with sequences up to $L = 3$, at three resolution levels; the multifractal measure, generated by a 500 000 random binary symbolic sequence with $\frac{1}{3}$ zeroes content and $\frac{2}{3}$ ones content is also shown.

The main advantage of this 1D representation is that we can easily implement a detrended measure replacing the equal contraction factors $k_i = 1/m$ by $k_i = p_i$, where p_i is the normalized frequency of the ith symbol on the sequence under study. This process will cancel the bias of the representation due to different symbols content in the sequence under study and, then, a random uncorrelated sequence will generate a uniform measure on the attractor of the detrended IFS. Therefore, we introduce the following IFS representation for a given symbolic sequence:

$$t_i(x) = p_i x + l_i, \quad \text{where } l_i = \sum_{j=0}^{i-1} p_j; \quad i = 1,\ldots,m, \text{ with } p_0 = 0\,. \tag{2}$$

For instance, Fig. 1(b) shows the detrended IFS representation for the symbolic sequence analyzed in Fig. 1(a) and the corresponding uniform multifractal measure; in this case, the scaling factors $k_1 = \frac{1}{3}$ and $k_2 = \frac{2}{3}$ are considered for the transformations.

Thus, we can easily detect non-random structure within the symbolic sequence through the deviation of the actual measure from a uniform one. A quantitative measure of this deviation can be obtained from a multifractal analysis of the measure. Since, the 1D detrended IFS is still formed by similitudes, a simple analytical treatment is possible for obtaining the multifractal spectrum [26,27] (note that this is not always possible when dealing with 2D representations, like those used in Ref. [21]).

3. Multifractal analysis

Multifractals are self-similar measures that can be regarded as densities on some domain. In most of the cases, these measures cannot be characterized by a unique scaling exponent (such as a fractal dimension), but an entire spectrum of local scaling exponents, or singularities, is needed. We treat the measure $\mu(x)$ defined by a DNA sequence on the support A of the IFS model (1) or (2) as a multifractal. Multifractal formalism analyzes fractal properties of those subsets $E(\alpha)$, where the measure has a given local scaling exponent, or singularity, i.e., subsets formed by points x, where $\mu(B(x,r)) \approx r^{\alpha}$ for small r (where $B(x,r)$ is the r-ball centered at x). In the case of measures generated by the chaos-game algorithm on the support of an IFS model (IFS measures), the singularities can be expressed in terms of the order-L iterators of the support.

$$\alpha \approx \frac{\log \mu(I_n(x))}{\log(r_n)}, \tag{3}$$

where $I_n(x)$ is the set of order-L iterators of size r_n containing x.

The main objective of the multifractal analysis of a measure μ defined on a fractal support is characterizing the dimensions, $f(\alpha)$, and the structure associated with the sets $E(\alpha)$ as a function of the singularities α. This information is gathered in the multifractal spectrum $\alpha - f(\alpha)$.

In general, the calculation of the multifractal spectrum is not an easy task [28–30]; however, in the special case of IFS measures, there exists a simple and efficient computational procedure for obtaining the spectrum of singularities [26,27]. This method relies on the use of a coarse grained representation of the attractor given by the m^L order-L iterators of the attractor associated with sequences of length L; we start from the generating function

$$\kappa_n(q) = \sum_{i=1}^{m^L} \mu_i^q, \tag{4}$$

defined for each value $q \in \mathbb{R}$. For a multifractal measure, this function scales as

$$\kappa_n(q) \approx (k^{-L})^{\tau(q)} \Rightarrow \tau(q) \approx -\frac{\log(\sum_i \mu_i^q)}{L\log(k)}, \tag{5}$$

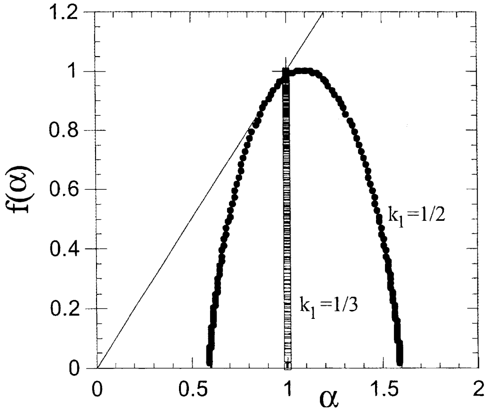

Fig. 2. Multifractal spectra for the two measures shown in Fig. 1; values for parameter q are taken in the range $(-30, 30)$.

in a representative range $L \in (l_1, l_2)$, which determines a scaling regime characterized by the exponent $\tau(q)$ (where k is the contractivity factor of the IFS model). This function provides a simple way to introduce the singularities by considering the following parametric form

$$\alpha(q) = \frac{\partial \tau(q)}{\partial q} \approx -\frac{\log(\sum_i \mu_i^q)}{L(\sum_i \mu_i^q)\log(k)} \,. \tag{6}$$

In this case, the f-value associated with $\alpha(q)$ is given by $f(\alpha(q)) = q\,\alpha(q) - \tau(q)$. This gives a parametrization $(\alpha(q), f(\alpha(q)))$ of the multifractal spectrum that can be computationally obtained in a simple way.

For instance, Fig. 2 shows the multifractal spectrum obtained using the above method for the two measures shown in Fig. 1. The uniform measure generated by the case $k_1 = \frac{1}{3}$ is, as expected, delta-shaped.

Therefore, the multifractal formalism provides appropriate techniques for analyzing DNA IFS-based measures, by obtaining the spectrum of singularities with a simple computational procedure. In this case, the probability of appearance of a certain subsequence is characterized, in the multifractal analysis, by the singularities α_i associated to each point x_i (whose address is a given subsequence). So, by analyzing the attractor up to a certain length scale 2^L, we can probe the distribution of subsequences of length L, where k is the contractivity factor of the IFS model (for instance, for the IFS representation (1) with $m = 2$, $k = \frac{1}{2}$).

It is important to remark here that, in practice, it is not always the case that an observed measure has a unique scaling regime, but several different scaling regimes could be found associated with changes produced in the system at different critical scales. As we shall see later, this is the case of DNA sequences where correlations in the sequence induce changes in the scaling, leading to different multifractal regimes at different lengths, while symbol correlations persist. In fact, this lack of scaling can be used to detect dependence structures within the sequence. Furthermore, we can quantify the decay of correlations in the DNA sequence by fixing a scale $k = 1/2^L$, that corresponds to subsequences of length L, and by computing directly the spectrum $f_L(\alpha)$ at that scale using (6). The way in which the spectrum varies provides a quantification of the non-random structures existing within the DNA sequences. Note that, as opposite to other dimension-like and entropy-like measures, inspired by the information-theoretical measures that are widely used in linguistics [31,32], the analysis of local singularities does not focus on a single exponent, but considers all the scaling behaviors within the sequence. Moreover, some of these measures, such as the entropy, can be obtained from a single point in the spectrum.

To compute the variation of the spectrum more efficiently we introduce a skipping parameter s by reading the sequence for every sk nucleotides. From a sequence $a_1a_2\ldots a_n$ of length n, we extract $s-1$ sequences of length n/s $a_1a_s\ldots, a_2a_{s+1}\ldots, a_{s-1}a_{2s-1}\ldots$. These are used to feed the chaos-game algorithm. We have observed that, as the skipping length s is increased, the spectrum becomes narrower, mainly from the contributions of positive q's. Since the growth of s produces a loss of correlations, it is clear that the narrowing spectra approach the delta-shaped spectrum of a uniform measure, associated with an uncorrelated random sequence (remember that we are using the detrended representation introduced in the previous section). This narrowing can be observed in plots of $\Delta\alpha = \alpha(q_c) - \alpha(-q_c)$ (see below), for some value of q_c (normally the extreme values). This function decays as s grows until it is almost indistinguishable from that corresponding to a random reference sequence.

Moreover, the resulting multifractal spectrum can be interpreted in terms of statistical dependence by using the concepts of mutual information and redundancy [33]. The mutual information can be considered a non-linear analog of the correlation between two random variables X_1 and X_2 with a joint probability function $p(x_1,x_2)$ and marginal probabilities $p(x_1)$ and $p(x_2)$. The mutual information of the variables is defined as

$$I(X_1,X_2) = \sum_{x_1}\sum_{x_2} p(x_1,x_2)\log\frac{p(x_1,x_2)}{p(x_1)p(x_2)} . \tag{7}$$

$I(X_1,X_2)$ is symmetric, non-negative and equal to zero if and only if X_1 and X_2 are independent. Therefore, this definition provides a convenient framework for analyzing statistical independence (which is more general than simple correlation) in symbolic DNA sequences. The generalization of this concept to more than two variables leads to the following definition of redundancy, which has similar properties:

$$Re(X_1,\ldots,X_n) = \sum_{x_1}\cdots\sum_{x_n} p(x_1,\ldots,x_2)\log\frac{p(x_1,\ldots,x_2)}{p(x_1)\ldots p(x_n)} . \tag{8}$$

Now, note that from (3) we have

$$\alpha_{i_1,\ldots,i_L} = \frac{\log\mu(I)}{\log(r)} = \frac{\log p(a_{i_1},\ldots,a_{i_L})}{\log(k_{i_1}k_{i_2}\ldots k_{i_L})} = \frac{\log p(a_{i_1},\ldots,a_{i_L})}{\log(p(a_{i_1})\ldots p(a_{i_L}))} . \tag{9}$$

for an arbitrary iterator $I(x)=t_{i_L}(\ldots t_{i_2}(t_{i_1}(A)))$ of size $k_{i_1}k_{i_2}\ldots k_{i_L}$, where $\mu(I)=p(a_{i_1},\ldots,a_{i_L})$ denotes the probability of finding a subsequence $a_{i_1}\ldots a_{i_L}$ within the sequence, and $p(a_i)$ denotes the probability of finding a symbol a_i in the sequence. Then,

$$\log p(a_{i_1},\ldots,a_{i_L})=\alpha_{i_1,\ldots,i_L}\log(p(a_{i_1})\ldots p(a_{i_L}))\Rightarrow \\ p(a_{i_1})\ldots p(a_{i_L})=p(a_{i_1},\ldots,a_{i_L})^{1/\alpha_{i_1,\ldots,i_L}}\,. \tag{10}$$

Substituting (10) into (8) leads to

$$Re=\sum_{i_1}\cdots\sum_{i_L}\left(1-\frac{1}{\alpha_{i_1,\ldots,i_L}}\right)p(a_{i_1},\ldots,a_{i_L})\log p(a_{i_1},\ldots,a_{i_L})\,. \tag{11}$$

Therefore, values of $\alpha_{i_1,\ldots,i_L}=1$ correspond to subsequences $a_{i_1}\ldots a_{i_L}$ which do not contribute to the increase of the redundancy, $p(a_{i_1},\ldots,a_{i_L})=p(a_{i_1})\ldots p(a_{i_L})$, whereas values lower or higher than 1 indicate a positive or negative contribution to the redundancy of the variables, respectively. This result shows that Eq. (11) can be interpreted in connection with the multifractal spectrum of the measure (which displays the distribution of singularities on the DNA measure) for providing a general picture about the independence relationships among sequences of symbols within the DNA sequence. Moreover, if we select only the set of subsequences corresponding to a certain interval of singularities, then the new redundancy of the new restricted "language" can still be obtained using (11).

Note that by using sequences with skipped symbols the above analysis can also provide information about the existence of short and long-range dependencies within coding and non-coding DNA sequences; this problem has been extensively analyzed in the literature leading to some controversy [7–10].

4. Results and discussion

We are interested in analyzing the measure generated by DNA sequences on the IFS representation (2); for the sake of comparison with other methods, in this paper, we have considered both a four-letter A, C, G, T alphabet ($m=4$) and a two letter pyrimidines and purines alphabet ($m=2$), obtaining similar results; therefore, for the sake of simplicity, we shall use the pyrimidines and purines representation (two letter alphabet) throughout the paper.

In order to assert the generality of the results, we have applied this analysis to a large number of representative genomic sequences across the phylogenic spectrum, with high and low coding content. No general results seem to hold for the long-range correlation character of coding and non-coding sequences, in contrast with other results reported in the literature that found long-range correlation in non-coding sequences and short-range correlation in coding sequences. In the light of our results, we think that

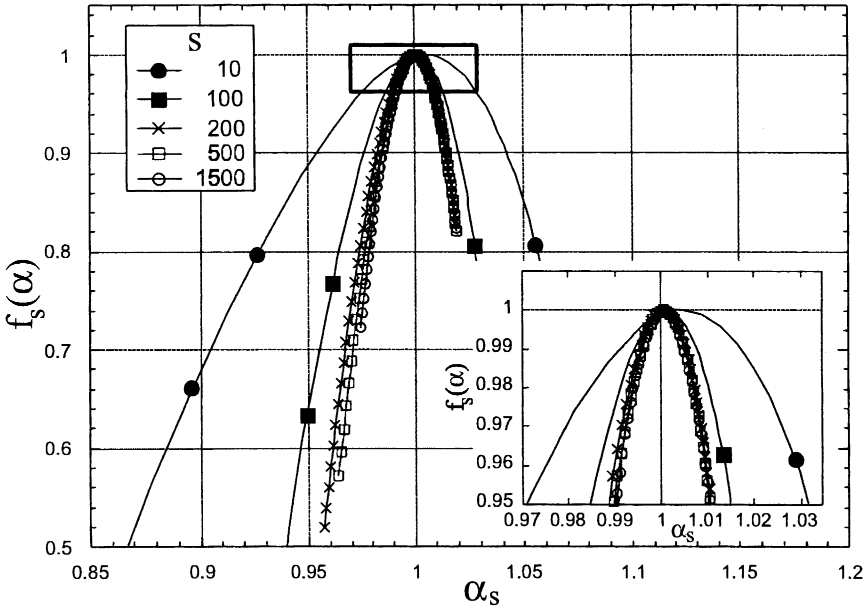

Fig. 3. $f_s(\alpha)$ spectra for the human β-globin (low coding content) for several values of the skipping length s. The sequence of spectra collapses for values of s larger than 200. The inset shows a magnification of the area contained in the square, corresponding to the q-values closer to zero. The spectra are obtained with $q \in (-30,30)$.

this characteristic feature appears to be rather related to the positions in the phylogenic tree of each particular chain.

With the aim of illustrating the lack of general results we present two examples of previously analyzed DNA sequences with high and low coding contents, respectively. We found long and short-range correlations, respectively, in contradiction with other general results reported in the literature. We first consider the intron-containing human β-globin sequence (GenBank name HUMHBB, 73326 base pairs) which has low coding content and has been previously analyzed using a random walk in Ref. [4], reporting a mean fluctuation exponent 0.71 from a scaling region of four orders of magnitude; this suggests the presence of a long-range correlation. Fig. 3 shows a sequence of $f_s(\alpha)$ spectra for some values of the skipping length s, ranging from $s = 10$ to 1500. The resulting sequences collapse for values of s larger than 200 and become indistinguishable from those corresponding to a random reference sequence (obtained by randomly rearranging the symbols in the sequence).

This fact is also shown in Fig. 4, where $\Delta\alpha$ is plotted against the skipping length s for the case $q_c = 30$ (the extreme value), leading to a value $\Delta\alpha$ equal to the width of the spectrum of singularities. The log–log inset shows two different regions. First, $\Delta\alpha$ decays following a power law up to $s = 200$. Afterwards, a plateau is reached, where its value becomes indistinguishable from that corresponding to a random reference

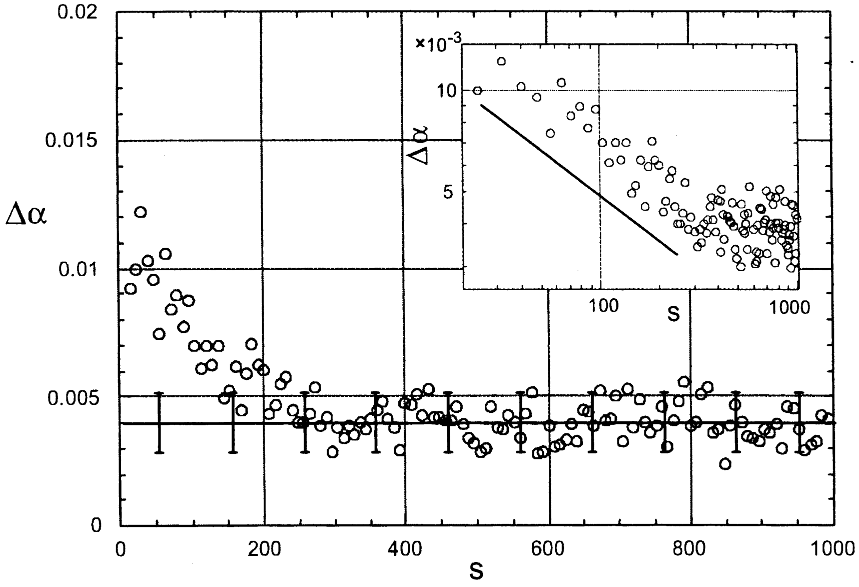

Fig. 4. $\Delta\alpha$ versus s, from the spectra of Fig. 3; the error bars correspond to 20 realizations of an uncorrelated reference sequence obtained by randomly rearranging the nucleotides in the original chain. The inset shows a double logarithmic plot of the same figure, with a power law decay up to $s \sim 200$ and a roughly constant regime beyond.

sequence. We conclude that it makes no sense to ascribe a long-range correlation character to this sequence beyond a few hundred nucleotides.

The lack of scaling in the distribution of subsequences can also be observed by using the same skipping criteria with other standard representations. In Fig. 5, we plot the scaling exponent of the mean square fluctuation of a DNA walk for different skipping lengths. This figure looks very similar to the one obtained with the multifractal method. Namely, after a few hundred nucleotides, the exponent reaches a plateau of value 0.5, corresponding to an uncorrelated random walk. A similar result has been recently obtained computing the fractal dimension associated with DNA random walks [14].

On the other hand, we can see that $\Delta\alpha$ corresponding to the sequences with high coding content decays by oscillating with a characteristic period three, that arises from the codon structure. In Fig. 6, we illustrate this fact by considering a complete genome containing mostly coding regions (complete genome of *E. coli*, approx. 4 Mbp). This oscillation has been previously observed in other contexts [14,34]. The oscillation has the effect such that the decaying function presents two branches, one above the other. The upper branch corresponds to the values of s that are a multiple of 3, and keep the skipping sequence roughly in the same "frame" (in the genetic sense). So, the enhancement of the correlation in this branch can be associated with an effect due to the coding part of the sequence (as previously reported in [35,36]). In agreement

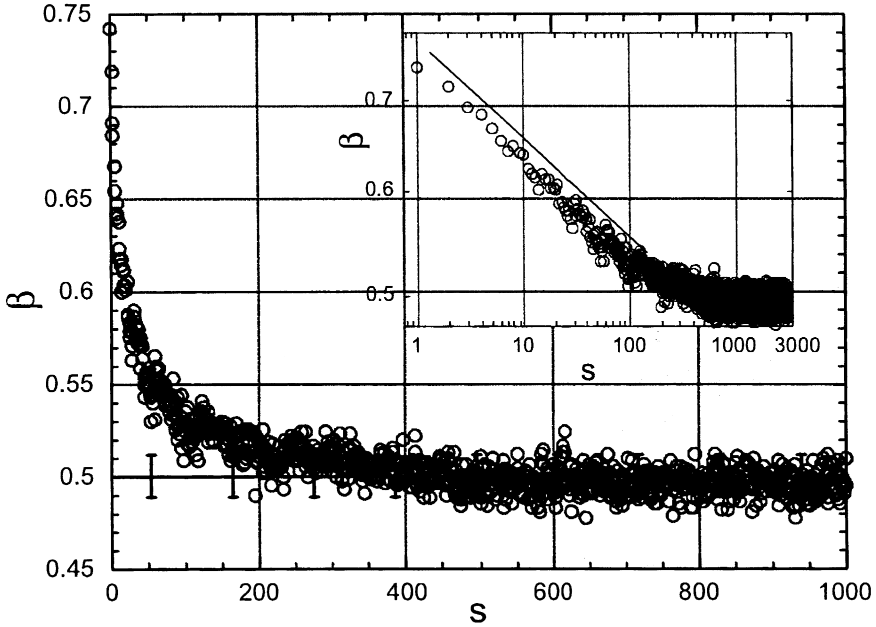

Fig. 5. Scaling exponent β of the mean square fluctuation function $F(l)$ ($F(l) \sim l^{\beta}$) as a function of the skipping length s, for a DNA walk corresponding to the same sequence as Figs. 3 and 4 (see [4] for a description of the DNA walk); the error bars correspond to 20 realizations of an uncorrelated reference sequence obtained by randomly rearranging the nucleotides in the original chain. The inset gives a double logarithmic plot, showing how an uncorrelated regime is attained at skipping length $s \sim 400$.

with this interpretation we can observe in the log–log representation (right inset in Fig. 6) an initial plateau which extends to amount hundred bases (the order of proteins' length). Beyond this plateau, correlations decay slightly faster than in the lower branch, where coding effects are not enhanced by skipping. In both branches, a long-range loss of memory can be observed. These results are in contrast with those obtained from a standard detrended DNA walk analysis which gives the exponent 0.51 from a scaling region up to $L = 1000$ [5], suggesting that no long-range correlation is associated with this sequence with a high coding content. In this case, the method proposed in this paper allows us to analyze in more detail the decay of correlations, discovering that the power law holds up to a kilo-base where it is indistinguishable from a reference random sequence. Fig. 7 compares the standard redundancy and the singularities width of the sequence, as a function of the skipping parameter s. From this figure, we can see how the standard "averaged" redundancy quickly decays to zero, while the spectrum width $\Delta\alpha$ still detects a dependent combination of symbols within the sequence, associated with extreme singularity values, as explained above.

There are two points to be stressed here, under the light of these results. First, the fact that there is not a single well-defined spectrum at all scales shows that a simple scaling in the distribution of subsequences within a DNA sequence does not exist. To

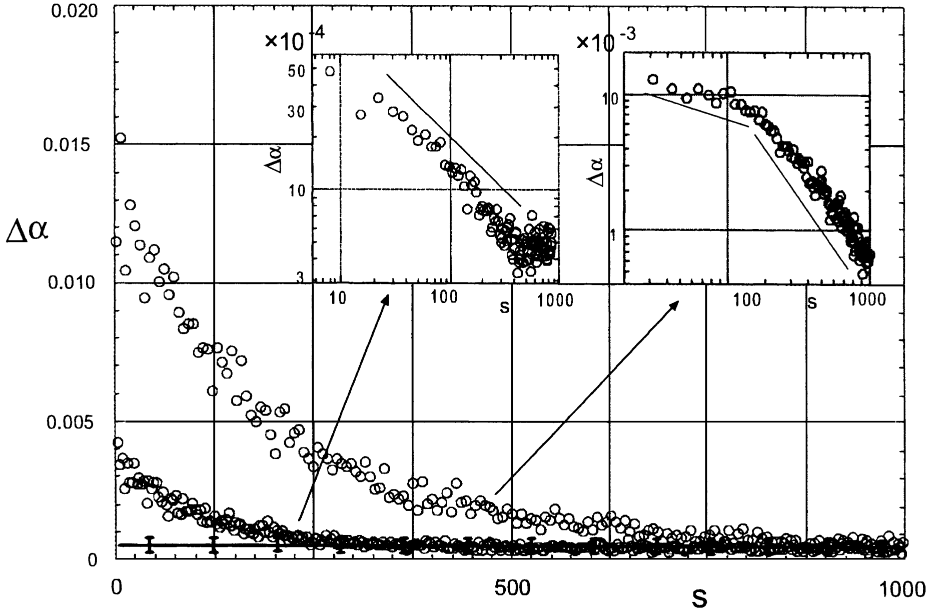

Fig. 6. $\Delta\alpha$ versus s for the complete *E. coli* genome (high coding content); the error bars correspond to 20 realizations of an uncorrelated reference sequence obtained by randomly rearranging the nucleotides in the original chain. The two insets show the power-law decay of the branches.

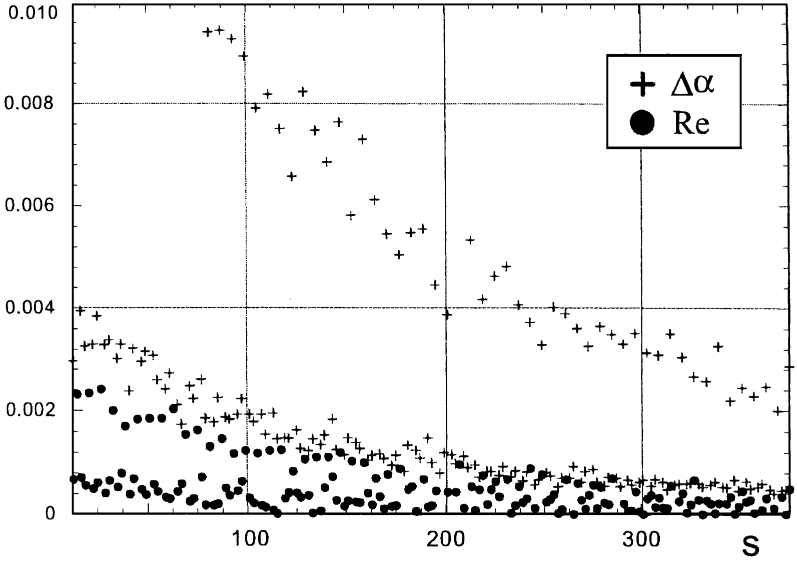

Fig. 7. $\Delta\alpha$ and redundancy (Re) versus s for the complete *E. coli* genome.

put it more dramatically, we can say that any dimension of the multifractal, say D_1, decays with the scale at which we probe. If D_1, that is an exponent and should be of the same characterization of the scaling, decays, then all we can say is that the only meaningful exponent is that of the random sequence. Second, we observe that the attractor—the distribution of subsequences—is nevertheless not trivial, since it presents the same characteristic feature just described.

We need to emphasize that these results have been obtained with a method that directly probes into the distribution of subsequences of the DNA. This raises the question of how to interpret previous results by other authors, where long range correlations have been reported, in particular in the non-coding sequences. We are not sure as to what the answer could be, but we do envisage several possibilities. When a sequence is mapped onto a random walk or interface, some kind of uncontrolled effect is likely to be added to the simple repetition of sequences. For instance, in the mapping to interfaces, sequences of different singularities can produce segments with equal width, and contrarily, segments with different widths can be constructed from sequences with similar singularities. No connection between singularities of sequences and widths of interfaces seems to exist. Moreover, random walks or interfaces are processes that result from the integration of a numerical series directly mapped from the coding sequence. This integration could originate from an inertia of the method to extrapolate a valid exponent at some short scale, to other scales where it is no longer valid. This effect has been well described in the analysis of superrough interfaces [37]. On the other hand, this effect could be maintained for long scales since, in these processes, longer sequences need better statistics, contrarily to the IFS mapping, where longer sequences produce more minute details.

Acknowledgements

This work was partially supported by DGES Research Grant No. PB96-0378-C02-02. We acknowledge partial financial support from the ANPCyT (Argentina) through its project PICT 03-000000-00988.

References

[1] W. Li, T.G. Marr, K. Kaneko, Physica D 75 (1994) 392.
[2] W. Li, Comput. Chem. 21 (4) (1997) 257.
[3] W. Li, K. Kaneko, Europhys. Lett. 17 (1992) 655.
[4] C.-K. Peng et al., Nature 356 (1992) 168.
[5] C.-K. Peng et al., Phys. Rev. E 49 (1994) 1685.
[6] R.N. Mantegna et al., Phys. Rev. Lett. 73 (1994) 3169.
[7] N.E. Israeloff, M. Kagalenko, K. Chan, Phys. Rev. Lett. 76 (1996) 1976.
[8] S. Bonhoeffer et al., Phys. Rev. Lett. 76 (1996) 1977.
[9] R. Voss, Phys. Rev. Lett. 76 (1996) 1978.
[10] R.N. Mantegna et al., Phys. Rev. Lett. 76 (1996) 1979.
[11] C.L. Berthelsen, J.A. Glazier, M.H. Skolnick, Phys. Rev. A 45 (1992) 8902.

[12] A.S. Borovik, A.Yu. Grosberg, M.D. Frank-Kamenetskii, J. Biomol. Struct. Dyn. 12 (1994) 655.

[13] G. Abramson, P.A. Alemany, H.A. Cerdeira, Phys. Rev. E 58 (1998) 914.

[14] G. Abramson, H.A. Cerdeira, C. Bruschi, Biosystems 49 (1999) 63.

[15] H. Joel Jeffrey, Nucl. Acid. Res. 18 (1990) 2163.

[16] H.-O. Peitgen, H. Jürgens, D. Saupe, Chaos and Fractals, New Frontiers of Science, Springer, New York, 1992.

[17] N. Goldman, Nucl. Acid. Res. 21 (1993) 2487.

[18] K.P. Pleissner, L. Wernisch, H. Oswald, E. Fleck, Electrophoresis 18 (15) (1997) 2709.

[19] A. Fiser, G.E. Tusnady, I. Simon, J. Mol. Graphics 12 (4) (1994) 302.

[20] K. Hill, S.M. Singh, Genome 40 (3) (1997) 342.

[21] J.M. Gutiérrez et al., in: M. Barbi, S. Chillemi (Eds.), Chaos and Noise in Biology and Medicine, World Scientific Publishing, Singapore, Vol. 315, 1998.

[22] H. Bailin, Physica A 282 (2000) 225.

[23] P. Tino, IEEE Trans. Systems Man Cybern. 29 (1999) 386.

[24] M.F. Barnsley, Fractals Everywhere, 2nd Edition, Academic Press, New York, 1990.

[25] K. Falconer, Fractal Geometry, Wiley, New York, 1993.

[26] J.M. Gutiérrez, A. lglesias, M.A. Rodríguez, Fractals 5 (1996) 17.

[27] J.M. Gutiérrez, M.A. Rodríguez, Chaos Solitons Fractals 11 (5) (2000) 675.

[28] T.C. Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia, B.I. Shraiman, Phys. Rev. A 33 (1986) 1141.

[29] H.G.E. Hentschel, I. Procaccia, Physica D 8 (1983) 435.

[30] L. Olsen, Advan. Math. 116 (1995) 82.

[31] W. Ebeling, G. Nicolis, Chaos Solitons Fractals 2 (1992) 635.

[32] F. Ledrappier, L.S. Young, Young. Ann. Math. 122 (1985) 540.

[33] C. Granger, J.L. Lin, J. Time Series Analysis 15 (1994) 371.

[34] P. Allegrini et al., Phys. Rev. E 52 (1995) 5281.

[35] J.W. Fickett, Nucleic Acids Research 10 (1982) 5303.

[36] H. Herzel, I. Grosse, Physica A 216 (1995) 518.

[37] J.M. López, M.A. Rodríguez, R. Cuerno, Physica A 246 (1997) 329.