Fractional Brownian motion in DNA sequences of bacterial chromosomes: a renormalization group approach

M.V. José, T. Govezensky, and J.R. Bobadilla
Theoretical Biology Group, Instituto de Investigaciones Biomédicas,
Universidad Nacional Autónoma de México,
Ciudad Universitaria, 04510, D.F., México,
Tel.:/Fax: 01-52-55-5622-3894,
e-mail: marcojose@biomedicas.unam.mx

Recibido el 29 de junio de 2009; aceptado el 2 de febrero de 2010

A renormalization group (RG) approach shows that the relative dispersion of the distance series of a triplet for each half of most bacterial chromosomes follows an inverse power-law as a function of the window size in a log-log plot. These straight lines indicate that when each half of the bacterial chromosome is analysed a random monofractal is obtained. With this approach, inverse bilateral symmetry of some triplets in the 4 bacterial chromosomes analyzed is also illustrated. Thus, DNA sequences of whole bacterial genomes contain both long-range correlations and random components. In particular the RG approach captures a harmonic modulation of the underlying inverse power-law. The frequency distributions of distances of triplets are also analyzed and they exhibit an oscillatory decaying pattern that displays the well-known 3-base periodicity. It is concluded that the DNA fluctuations of the distance series of triplets are not completely random, like Brownian motion, nor are they the result of processes with short-term correlations. Instead, the inverse power-law reveals that the DNA distance series at any position is influenced by fluctuations that occurred hundreds or thousands of bases apart. This behavior is a consequence of the fractional Brownian nature of the distance series of DNA.

Keywords: Frequency distributions of distances of triplets; bacterial chromosomes; statistical properties of DNA distance series; renormalization group approach; scaling exponents; Hurst exponent.

El enfoque del grupo de renormalización (RG) muestra que la dispersión relativa de una serie de distancias de tripletes para cada mitad de los cromosomas bacterianos sigue una ley de potencia inversa en función del tamaño de la ventana en una grafica log-log. Estas líneas rectas indican que cuando la mitad de cada cromosoma bacteriano es analizado se obtiene un monofractal. Con este método se ilustra que ciertos pares de tripletes exhiben una simetría bilateral inversa en las 4 bacterias estudiadas. Asimismo, las secuencias de ADN de los genomas de bacterias en su conjunto contienen correlaciones de largo alcance y componentes aleatorios. En particular, el enfoque RG captura una modulación armónica de las leyes de potencia inversa. Se analizan también las distribuciones de frecuencias de las distancias de tripletes y se presenta un patrón oscilatorio que muestra la conocida periodicidad de 3. Se concluye que las fluctuaciones de las series de distancia de tripletes del ADN no son al azar, como en el movimiento Browniano, ni son el resultado de correlaciones de procesos a corto plazo. Por el contrario, la forma de las leyes de potencia inversa revela que la serie de distancias en cualquier posición se ve influida por fluctuaciones que tuvieron lugar a cientos o miles de bases de separación. Este comportamiento es consecuencia de la naturaleza fractal browniana de las series de distancias de tripletes en las secuencias de ADN.

Descriptores: Distribución de frecuencias de distancias de tripletes; cromosomas bacterianos; propiedades estadísticas de series de distancias de ADN; renormalización de grupos; exponentes de escalamiento; exponentes de Hurst.

PACS: 87.10.+e; 05.40.+J

1. Introduction

The renormalization group (RG) analysis, introduced in field theory and in critical phase transitions, is a very general mathematical and conceptual tool, which allows one to decompose the problem of finding the macroscopic behavior of a large number of interacting parts into a succession of simpler problems with a decreasing number of interacting parts, whose effective properties vary with the scale of observation [1]. The RG permits one to determine the scaling properties of a system. At the outset, a set of equations that may describe the behavior of the system is assumed. Then the length scale at which the system is being described is changed. By moving away from the system, some of the details are lost. At the new scale, the same set of equations is applied, but possibly with different coefficients. The objective is to relate the set of equations on one scale to the set of equations on the other scale. In this way, the scaling properties of the system can be obtained. The premise of the RG is that exactly at a second order phase transition, the equations describing the system are independent of scale.

The concept of RG is useful for systems that exhibit the properties of scale invariance and self-similarities of the observables at the critical point of the system [2]. The RG approach deals with the concept that a critical point results from the aggregate response of an ensemble of elements. The two main transformations of the RG are decimation and rescaling. When we go from the fine scale to the coarse scale, the process is called decimation. The idea of RG is to decimate the degrees of freedom, while rescaling so as to keep the same scale by calculating, for example, the

relative dispersion, the ratio of the standard deviation to the mean. The procedure can be repeated using groupings of two, three, four and more data points. In this way the fractal dimension that is independent of the degree of coarse-graining can be determined.

The statistical analysis of DNA sequences has been studied for almost 60 years. Several properties have been unveiled using different methods for their analysis. A non-exhaustive selected list of some of the methods and main findings related to the statistical properties of DNA sequences is offered in Table I. The finding that several bacterial chromosomes possess an inverse bilateral symmetry (IBS) was demonstrated by means of the RG approach and to our knowledge this was the first time that this approach was used for analyzing DNA sequences of whole bacterial genomes [3,23]. Furthermore, in a more recent work we have shown that the scaling exponents for a given triplet in several prokaryotes have remained unaltered throughout their evolution [24]. Then, there is a strong evidence of critical scale invariance in the scaling exponents which indicates that not all information of ancestral organisms has been erased at least for the last 3 billion years of evolution [24].

In this work we present the basic ideas of the RG approach and give examples of how this procedure can be applied for analyzing the entire genomes of 4 bacteria.

2. Renormalization group

The purpose of the RG is to translate into mathematical language the concept that the sum is the aggregation of an ensemble of defined sub-sums, each sub-sum defined by the sum of sub-sums and so on. In other words, the RG approach implies that a critical point results from the aggregate response of an ensemble of elements.

Let us assume the renormalization group scaling relation:

$$F(x) = \frac{F(bx)}{a} \tag{1}$$

This relation expresses the property of $F(x)$ being self-affine, *i.e.*, the graph of $F(x)$ on a scale bx has to be scaled down by a factor $1/a$ to obtain the desired function on scale x. Whereas self-similarity refers to the fact that the shapes are identical under magnification, self-affinity expresses the fact that F and x have to be scaled by different amounts for the two views to become identical. Scale invariance means reproducing itself on a different time or space scale. An observable F which depends on a control parameter x is scale invariance under the arbitrary change $x \to bx$ if there is a number $a(b)$ such that Eq. (1) holds.

The solution to (1) is:

$$F(x) = cx^d \quad \text{with} \quad d = \frac{\ln(a)}{\ln(b)} \tag{2}$$

Power-laws are the hallmark of scale invariance as the ratio, $(F(xb))/(F(x)) = b^d$, does not depend on x, *i.e.* the relative value of the observable at two different scales depends simply on the ratio of the two scales.

Sornette [2] has generalized the continuous fractal dimension into what he calls the Discrete Scale Invariance. Considering the solution in (2) we get:

$$\frac{b^d}{a} = 1 = e^{i2\pi n}$$

This leads to:

$$d = \frac{\ln(a)}{\ln(b)} + i\frac{2\pi n}{\ln(b)}, \tag{3}$$

which characterizes the system in terms of complex fractal dimensions. The imaginary part of the fractal dimension translates itself into a log-periodic modulation decorating the leading power law behavior.

3. Aggregating data and relative dispersion (Hurst exponent)

Let us examine how the relative dispersion (RD) changes as a function of the number of adjacent data elements we aggregate. We start by aggregating $n-$adjacent data points, so that the $j-$ element in such an aggregation is given by $Y_j^{(n)} = Y_{nj} + Y_{nj-1} + Y_{nj-2} + ... + Y_{nj-(n-1)}$. Here, Y represents the distance between two identical triplets. In terms of these new data the arithmetic average is defined as the sum over the total number of data points, where the bracket $[\,]$ denotes the closest integer value, and N is the original number of data points, *i.e.*,

$$\overline{Y}^{(n)} = \frac{1}{[N/n]} \sum_{j=1}^{[N/n]} Y_j^{(n)} = n\overline{Y}^{(1)}$$

The variance for a monofractal time series is similarly given by [18]:

$$Var(\overline{Y}^{(n)}) = n^{2H} Var(\overline{Y}^{(1)}), \tag{4}$$

where $H$is the Hurst exponent, and the superscript on the average variable indicates that it was determined using all the original data without aggregation and the superscript on the average variable indicates that it was determined using the aggregation of data points. Thus the relative dispersion (RD) for an aggregated data set is:

$$RD^{(n)} = \frac{\sqrt{Var(\overline{Y}^{(n)})}}{\overline{Y}^{(n)}} = \frac{\sqrt{n^{2H} Var(\overline{Y}^{(1)})}}{n\overline{Y}^{(1)}} = n^{H-1} RD^{(1)} \tag{5}$$

which is exactly an inverse power-law in the aggregation number for the Hurst exponent in the interval $0 \leq H \leq 1$

TABLE I. Analysis of DNA Sequences

Method	Finding	Reference
Frequency analysis	Statistical confirmation of the 1^{st} and 2^{nd} Parity Rules	[4]
Positional Autocorrelation Function	3- and 10-11 base pair periodicities	[5]
Correlation analysis	2^{nd} Parity rule for n-tuples	[6]
Fourier Spectra	Long-range correlations in non-coding DNA sequences	[7]
Detrending Fluctuation Analysis and Fourier Spectra	Long-range correlations in non-coding DNA and lack of correlations in coding sequences	[8,9]
Multifractal spectra and wavelets analysis	Self-similarity and multiple scalings in both coding and non-coding sequences	[9]
Indexes of base composition (C-G)/(C+G)	Asymmetric substitution patterns coincide with Ori and Ter sites of chromosome replication (GC-skews)	[10]
Cumulative skew diagrams	GC-skews	[11,12]
Autocorrelation function	Power-laws in long DNA sequences	[13,14]
Markov models of sequence alignments	Origin and nature of low- and high-order symmetric complementary DNA strands Long range correlations in isochores	[15]
Log-log plots of standard deviation versus fixed window sizes	Long range correlations in isochores	[16]
Generalized Autocorrelation	Long-range correlations of base composition at the 3 codon positions at distances which are multiples of 3 and anticorrelations for distances which are not multiples of 3	[17]
Renormalization group approach	Long-range correlations; inverse bilateral symmetry of whole bacterial chromosomes; critical scale invariance	[3,24]

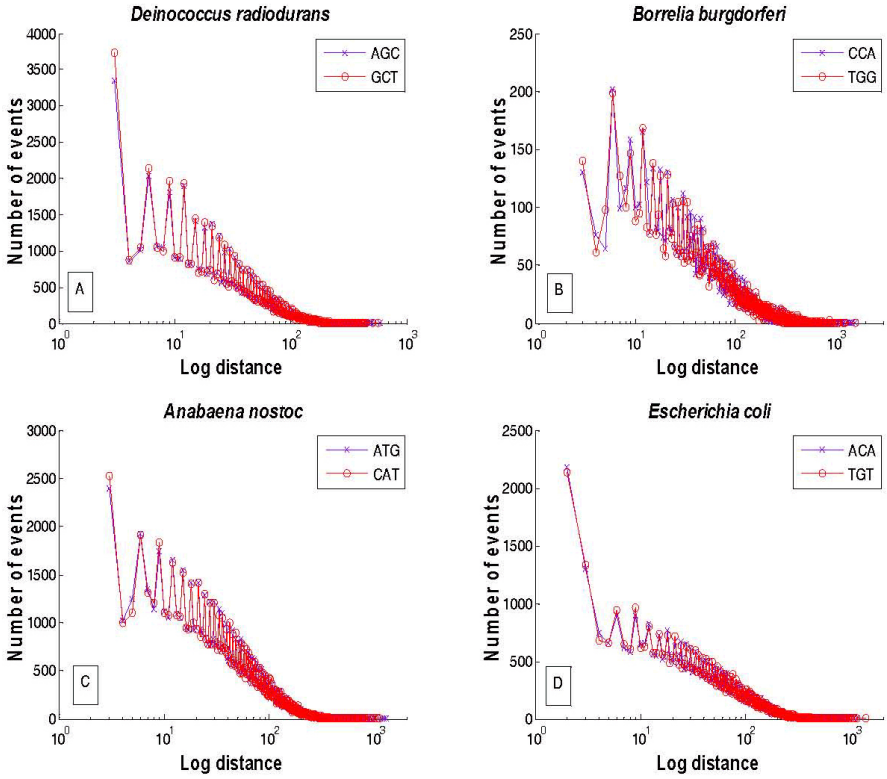

FIGURE 1. Log-linear plot of the frequency distribution of distances of (A) AGC (crosses) and GCT (empty circles) in *D. radiodurans*; (B) CCA (crosses) and TGG (empty circles) in *B burgdorferi*; (C) ATG (crosses) and CAT (empty circles) in *A. nostoc*; (D) ACA (crosses) and TGT (empty circles) in *E. coli*.

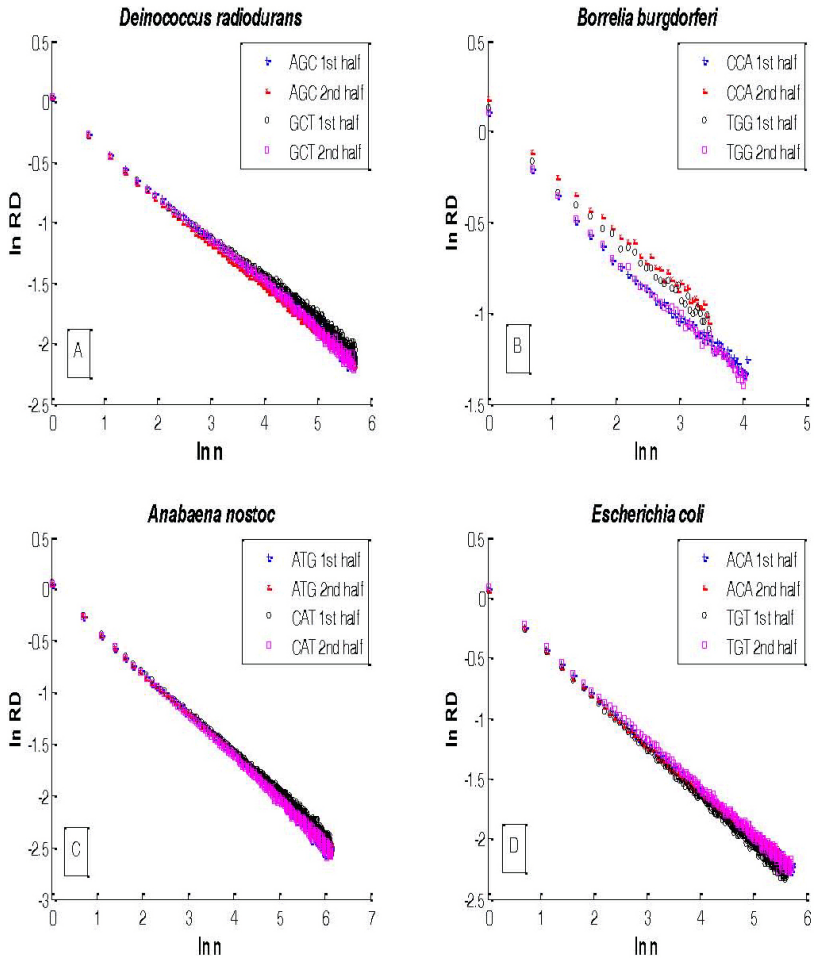

FIGURE 2. Aggregation analysis by chromosomal halves of the triplets (A) AGC and GCT in *D. radiodurans*; (B) CCA and TGG in *B. burgdorferi*; (C) ATG and CAT in and TGT in *E. coli*. All fittings of Eq. (5) have correlation coefficients of $r^2 = [0.97 - 0.99]$.

It is well established [1819] that the exponent in such scaling equations is related to the fractal dimension, D of the underlying distance series by $D = 2 - H$. A simple monofractal time series, therefore, satisfies the inverse powerlaw relation for the RD given by Eq. (4), which can be expressed by the linear regression relation [20]:

$$\ln\, RD^{(n)} = \ln\left(RD^{(1)}\right) + (1 - D)\ln\left(n\right) \qquad (6)$$

4. Methods and results

We illustrate the RG approach with 4 bacteria: *Deinococcus radiodurans*, *Borrelia burgdorferi*, *Anabaena nostoc* and *Escherichia coli*. *D radiodurans* is a bacterium that can live through intense levels of radiation. A human being exposed to 1,000 rads of radiation energy, a dose delivered in the atomic explosions of Hiroshima and Nagasaki, dies within two weeks. At one million rads *Deinococcus* still survives and at 3 million rads a very small number still endure. The extraordinary genomic resilience of this impressive bacterium lies in its ability to repair broke DNA. *B burgdorferi* is an spirochaete that produces Lyme disease, whose symptoms are arthritis-like. *A. nostoc* is a filamentous (metazoan-like) nitrogen-fixing cyanobacterium whose genome is very large (∼6.4 Mb). *E. coli* is the most studied organism in biology and it is an enteric bacteria.

The complete sequences of *D. radiodurans* *B. burgdorferi*, *A. nostoc* and *E. coli*. were retrieved from the NCBI, Genbank resource from the NIH (http://www.ncbi.nlm.nih.gov) with the following corresponding accession numbers: NC_001263, NC_001318, NC_003272; NC_000913.

Instead of using the classical stochastic random walk mapping rules of DNA (*e.g.* the purine–pyrimidine (RY) rule), the distance series of any triplet along chromosomal halves were determined [3,23]. For a given triplet its actual position along the whole chromosome was determined and from this the actual distance series (distance measured in bases) of that particular triplet were obtained. In other words, we can directly visualize how a given triplet is distributed along the whole chromosome. In this work, we divide the bacterial chromosomes into two non-overlapping halves according to the location of the origin of replication usually denoted by "Ori".

In Fig. 1 the frequency distributions of distances at which some triplets (*e.g.* ATG and CAT which are reverse comple-

mentary of each other) are encountered along the bacterial chromosomes of the 4 studied bacteria are presented. Note that the upper envelope of both distributions for each triplet in each bacterium occurs with a period of every 3 bases whereas the lower envelope occurs at distances different from the 3-base periodicity. These distributions are very similar for every pair of codons with its corresponding reverse complementary, and they display an oscillatory decaying pattern.

In Fig. 2 the results of the aggregation analysis of a given triplet and its reverse complementary for both halves of the chromosome of each bacterium are illustrated. The fitting of the data using Eq. (5) produces straight lines whose slopes lie between $-0.5 < 1 - D < 0$ in all bacteria indicating that the corresponding distance series follow an inverse power-law behavior and they are random monofractals. We also remark that log-periodic variation of the data about this power-law behavior can be observed if we amplify the behavior of the relative dispersion for large window sizes, as was clearly shown in a previous work (3). In Table I, the fractal dimensions for each pair of triplets for each bacterium is shown. Note that all scaling exponents lie between 1.51 (*e.g.* AGC in *B. burgdorferi*) and 1.68 (*e.g.* TTA, TAA and TGG in *E. coli*). In regard to the type of symmetry of the bacterial chromosomes, note, for example, that in *D. radiodurans* for the 1st half of the chromosome the slope of the aggregation analysis gives $H = 0.37$ and, for the 2nd half $H = 0.36$ Using the relation, $D = 2 - H$ the corresponding fractal dimension of AGC for the 1st half and for the 2nd half of the chromosome are $D = 1.63$ and $D = 1.64$, respectively. These estimates are in turn obtained for the reverse complementary triplet of AGC which is the triplet GCT, whose fractal dimensions in the 2nd and in the 1st half of the chromosome are $D = 1.65$ and $D = 1.62$, respectively (see Table II). Since the scaling exponent of a given triplet along one half of the chromosome is similar to the scaling exponent of the reverse complement in the other half, then the chromosome of *D. radiodurans* does posses inverse bilateral symmetry. This type of symmetry is also clearly observed for the triplets AGCGCT, ATG-CAT and CCA-TGG in *A. nostoc*, for TTA-TAA, ATG-CAT, and CCA-TGG in *B. burgdorferi*, for AGC-GCT, TTA-TAA, TAG-CTA, and CCA-TGG in *D. radiodurans*, and for ACA-TGT in *E. coli* In any case, for a given triplet the relative dispersion has a dominant inverse power-law with an index given by d and is modulated by a function that varies logarithmically with a fundamental period $\ln(b)$. The magnitudes of the fractal dimension D, reveal that there are long-range correlations in the distance series of a given triplet corresponding to what is called fractional Brownian motion. A particular type of randomness, which seems to maximize the information content, is also displayed by the distance series of triplets.

5. Conclusions

In this work we have used a renormalization group approach to obtain an expression for the aggregated relative dispersion that is the product of an inverse power-law and a modulation function that varies as the logarithm of the aggregation number.

In the literature it often appears that one has only two choices, either a process is a monofractal or it is a multifractal. The latter applied to a distance series would imply that the fractal dimension changes over distances, ultimately leading to a distribution of fractal dimensions [20]. This is not the situation here, however. The aggregated relative dispersion indicates that the process has a preferential scale length, b, in addition to the monofractal behavior determined by the inverse power-law index d [3]. Thus there is the interleaving of two mechanisms, one that is scale free and produces the monofractal, and the other has equal weighting on a logarithmic scale and is sufficiently slow as to not disrupt the much faster fractal behavior [3]. The tying together of the long and the short distance scales is necessary in order to adaptively regulate the complex DNA sequences in a changing environment. The log-periodic modulation of the inverse power-law is a consequence of the correlation function satisfying a renormalization group relation and having a complex fractal dimension [2].

TABLE II.

Triplet	Half	*Anabaena nostoc*	*Borrelia burgdorferi*	*Deinococcus radiodurans*	*Escherichia coli*
AGC	first	1.59	1.51	1.63	1.61
	second	1.56	1.55	1.64	1.59
GCT	first	1.54	1.55	1.65	1.63
	second	1.60	1.58	1.62	1.62
TTA	first	1.63	1.59	1.59	1.68
	second	1.61	1.60	1.61	1.66
TAA	first	1.63	1.62	1.61	1.68
	second	1.57	1.60	1.61	1.65
ACA	first	1.59	1.65	1.58	1.61
	second	1.57	1.60	1.56	1.62
TGT	first	1.64	1.53	1.59	1.60
	second	1.59	1.67	1.61	1.61
ATG	first	1.57	1.68	1.60	1.62
	second	1.58	1.65	1.64	1.60
CAT	first	1.60	1.65	1.59	1.62
	second	1.58	1.68	1.60	1.66
TAG	first	1.58	1.54	1.61	1.62
	second	1.57	1.59	1.59	1.61
CTA	first	1.59	1.55	1.58	1.65
	second	1.56	1.56	1.58	1.61
CCA	first	1.57	1.67	1.59	1.64
	second	1.59	1.67	1.62	1.65
TGG	first	1.57	1.67	1.60	1.66
	second	1.57	1.66	1.61	1.68

Sornette argues that the log-periodicity is a result of what he calls Discrete Scale Invariance, that is, also a consequence of renormalization group properties of the system.

According to the renormalization group theory [1], the nature of bacterial genomes pertains to a class of phenomena, where events at many scales of length make contributions of equal importance. Any scaling analysis of DNA sequences must take into account the entire spectrum of length scales since we are facing a system near its critical point [24]. There seems to be nothing more deterministic than the sum or multiplication of a large number of random variables [22].

The fractional Brownian nature in DNA sequences of bacterial chromosomes as obtained by the RG approach must be clearly considered in sequence analysis and in any further studies on the evolution of prokaryotes and eukaryotes.

Acknowledgments

M.V. José was financially supported by PAPIIT IN205307, UNAM, México and by the Macroproyecto de Tecnologías para la Universidad de la Información y la Computación (MTUIC).

1. K.G. Wilson, *Scientific American* **241** (1979) 158.
2. D. Sornette, *Critical Phenomena in Natural Sciences, Chaos, Fractals, Self-organization and Tools* (Springer, Berlin. 2000).
3. M.V. José, T. Govezensky, and J.R. Bobadilla, *Physica A* **351** (2005) 477.
4. V.V. Prabhu, *Nucleic Acids Res.* **21** (1993) 2797.
5. E.N. Trifonov and J.L. Sussman. *Proc. Natl. Acad. Sci. USA* **77** (1980) 3816
6. V.V. Prabhu. *Nucleic Acids Res.* **12** (1993) 2797.
7. W. Li and K. Kaneko, *Europhys. Lett.* **17** (1992) 655.
8. C.K. Peng *et al.*, N*ature* **356** (1992) 168.
9. S.V. Buldyrevet *et al.*, *Phys. Rev. E* **51** (1995) 5084.
10. A. Arneodo, E. Bacry, P.V. Graves, and J.-F. Muzy, *Phys. Rev. Lett.* **74** (1995) 3293.
11. J.R. Lobry, *Mol. Biol. Evol.* **13** (1996) 660.
12. A. Grigoriev, *Nucleic Acids Res.* **26** (1998) 2286.
13. M.V. de Sousa, *Phys. Rev. E* **60** (1999) 5932.
14. W. Li, G. Stolovitzki, P. Bernáola-Galván, and J.L. Oliver, *Genome Res.* **8** (1998) 916.
15. P.F. Baisnee, S. Hampson, and P. Baldi, *Bioinformatics* **18** (2002) 1021.
16. O. Clay, N. Carels, C. Douady, G. Macaya, and G. Bernardi, *Gene* **276** (2001) 15.
17. P. Bernáola-Galván, P. Carpena, R. Román-Roldán, and J.L. Oliver, *Gene* **300** (2002) 105.
18. J.B. Bassingthwaighte, L. Liebovitch, and B.J. West, *Fractal Physiology* (Oxford University Press, New York. 1994).
19. B.J. West and L. Griffin, *Chaos Solitons Fractals* **10** (1999) 1519.
20. B.J. West, V. Bhargava, and A. Goldberger, *J. Appl. Phys*. **60** (1986) 1089.
21. J. Feder, *Fractals* (Plenum Press, New York. 1988).
22. B.V. Gnedenko and A.N. Kolmogorov, *Distributions for Sum of Independent Random Variables* (Addison Wesley, Reading MA. 1954).
23. J. Sánchez and M.V. José, *Biophys. Biochem. Res. Comm.* **299** (2002) 126.
24. M.V. José, T. Govezensky, J.A. García, and J.R. Bobadilla, *PLoS ONE* **4** (2009) 4340.