Mathematical properties of DNA sequences from coding and noncoding regions

J.A. García[a,b,*] and M.V. José[b]
[a]*Laboratory of Theoretic Biology, Research Department, La Salle University, Benjamín Franklin 47, Col. Condesa, México, D.F. 06140, México,*
[b]*Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Apartado Postal 70228, Ciudad Universitaria, 04510 México D.F. México*

Recibido el 23 de septiembre de 2004; aceptado el 28 de enero de 2005

Several nonlinear techniques have been applied to analyze DNA sequences. As a result, some mathematical properties that distinguish both coding and noncoding regions have emerged. We review and apply some of these techniques selecting some examples and comparing our results with previously published data. We also discuss the main controversies that have been raised in terms of the different taken approaches, particularly the presence or absence of long-range correlations in coding regions.

Keywords: Long-range correlations; DNA mathematical analysis; fractal dynamics.

Se han aplicado diversas técnicas no lineales para analizar secuencias de ADN. Como consecuencia, han surgido ciertas propiedades matemáticas que permiten distinguir a las secuencias codificantes de las no codificantes. En este artículo se revisan y aplican algunas de estas técnicas, seleccionando algunos ejemplos y comparando nuestros resultados con datos previamente publicados. Asimismo, se discuten las principales controversias que han surgido en términos de las diferentes estrategias metodológicas consideradas, en particular, se discute la presencia o ausencia de correlaciones a largo alcance en regiones codificantes.

Descriptores: Correlaciones a largo alcance; análisis matemáticos; DNA; dinámica fractal.

PACS: 87.10.+e; 05.40.+j

1. Introduction

DNA is the molecule in which life organisms store information for their biological processes. In a DNA strand, it is possible to find sequences which can be transcribed to complementary RNAs, such as tRNAs, rRNAs and mRNAs. These sequences, also known as genes, are the coding regions in DNA. Between genes (intergenic regions), we can find regulatory sequences for transcription control. In the case of eukaryotic cells, besides, the vast majority of genes are not continuous: not expressed sequences, known as introns, lie between expression-coding sequences, known as exons. Thus, both intergenic regions and introns are noncoding DNA.

As a replicating information unit, DNA has fascinated not just biologist, but also, other scientists, like physicists, chemists, mathematicians and astrobiologists. The former, have made a lot of contributions to DNA understanding and, recently, they have applied several mathematical techniques for analyzing coding and noncoding regions.

In this paper, we apply, compare, and review some mathematical methods, which have been commonly used in order to reveal signature properties between coding and noncoding DNA sequences. As there are a lot of controversies among some of the results, we include numerical experiments and discuss the different interpretations among them.

2. Biochemistry of DNA

DNA is a double anti-parallel helix builded by concatenating nucleotide blocks. Each nucleotide has a nitrogenous base, a deoxyribose and a phosphate group. The bases are inside the molecule, while the phosphates are in contact with the hydrophilic medium.

DNA has four nitrogenous bases: adenine (A), thymine (T), cytosine (C) and guanine (G). There is complementarity between both DNA strands, as an A on one strand, always binds with a T on the other, and a C always binds with a G. The binding between the bases is through hydrogen bonds: two between A and T and three betwee C and G. Thus, following some physicochemical properties, DNA bases have been classified using three different dichotomies:

(a) Purines (R), A and G; or Pyrimidines (Y), T and C;

(b) Weak (W), A and T; or Strong (S), C and G; and

(c) Amines (M), A and C; or Ketones (K), T and G.

As the DNA backbone is constant (*i.e.* a chain of deoxyriboses bound by phosphodiester bonds), its biological properties reside in the sequence of bases along one strand (as the other is complementary). In this sense, DNA can be seen as a four letter alphabet, or could be transformed to a distance series. In the current paper, we review the main mathematical techniques used to study DNA sequences, considering the latter case.

3. DNA mapping

In order to apply signal processing techniques to DNA analysis, a DNA sequence must be transformed to a distance series. There have been several approaches to accomplish this. Herein, we present the main three.

3.1. Binary representation

The easier approach is to transform a DNA sequence to a binary sequence using one of the three conventions mentioned before (*e.g.* all weak bases -A and T- are changed by 0, and all strong bases -C and G- are changed by 1). The obtained series could now be subject to further mathematical analysis [22].

3.2. DNA random walk

This technique could be seen as a particular case of a binary representation. Consider a conventional one-dimensional random walk model, in which a theoretical walker crosses a DNA strand [37]. The walker starts at position $n = 0$ and gives one step up $[u(n) = +1]$ with each pyrimidine, and one step down $[u(n) = -1]$ with each purine. To graphically represent the walking, one usually plots the cumulative walk $y(n)$, against the position n as shown on Fig. 1a for the first 50,000 nitrogenous bases of the coding genome of *Borrelia burgderfori*.

The mathematical techniques reviewed here have been used in order to differentiate between coding and noncoding regions. In the current paper, we apply these techniques to: a whole coding bacterial genome, obtained by concatenating all the coding genes of *Borrelia burgderfori* in its original order and orientation [31]; the human beta globin chromosomal region (HUMHBB), a mainly non-coding sequence; and two control sequences: a shuffled version of the original coding genome of *Borrelia burgderfori*, and a synthetic DNA of one million bases, obtained by randomly sampling with replacement the four bases [9]. These control sequences do not represent intergenic regions; they are just representations of pure stochastic processes, in order to be able to distinguish between sequences with information (for protein synthesis), and corresponding random sequences. Thus, on Fig. 1 the four DNA walk displacements are shown. Note that both

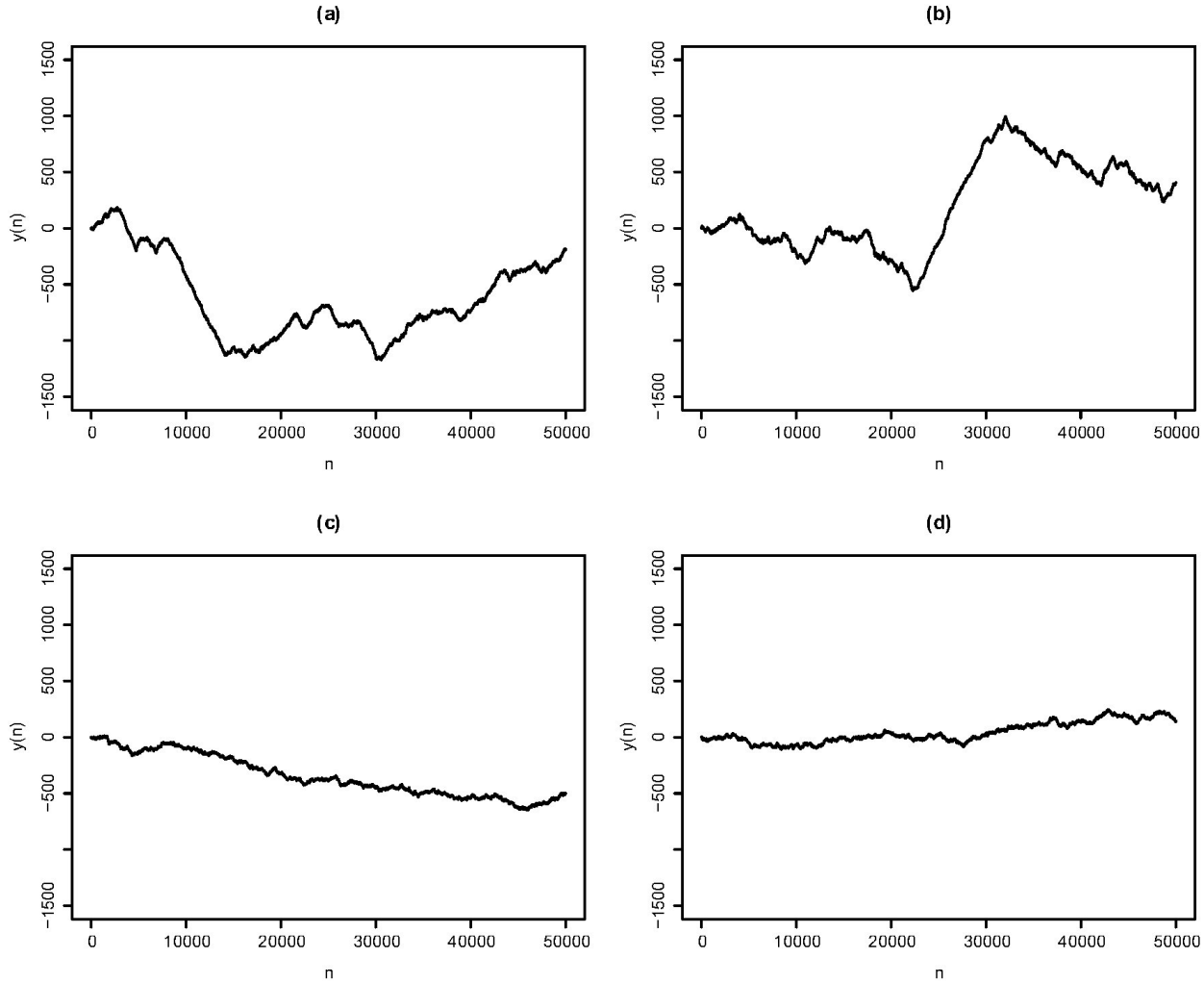

FIGURE 1. DNA walk displacement $y(n)$ against nucleotide distance n for the first 50000 nitrogenous bases in: (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB (human beta globin chromosomal region); (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

the coding and non-coding sequences presents jagged contours with local regions rich rich in either purines or pyrimidines (see Fig. 1a and b), while the control sequences present less variable displacements (see Fig. 1c and d). This kind of pattern on DNA could be related to biological structure.

3.3. Actual distance series

An alternative to binary representations is the generation of actual distance series obtained by calculating the number of characters between identical $n-$tuples (n = mono, du, tri, etc.) along the whole DNA sequence [31]. For example, to generate the distance series for the ATG triplet, the actual position of this sequence is first identified using the three different reading frames, and then the number of bases which lie between consecutive ATGs is computed. By using this approach, less information is lost in comparison with the binary representations, due to an oversimplification process in the latter.

4. DNA mathematical properties

4.1. Periodicities

Shepherd found purine-pyrimidine rhythms on viral genomes using actual distance series from binary DNA representations [33]. He used the RY convention to transform the original sequence, and then looked for the actual distance series between different combinations. As an example, the results for the triplet YRY in the coding genome of *Borrelia burgdorferi* are shown on Fig. 2a.

As shown on Fig. 2a maxima occurs regularly every three bases. This rhythm was preserved for all the studied combinations with an exception of an irregularity in $n = 13$ for $Y.R$ counts [33]. In contrast, there is no pattern found in both the noncoding sequence (see Fig. 2b), and the control sequences (see Figs. 2c and d).

It is worth mentioning that by looking at such periodicities, Shepherd found the sequence RNY as the most prevalent of all other sequences, thus he hypothesized that not only this sequence was an ancestor of the actual universal

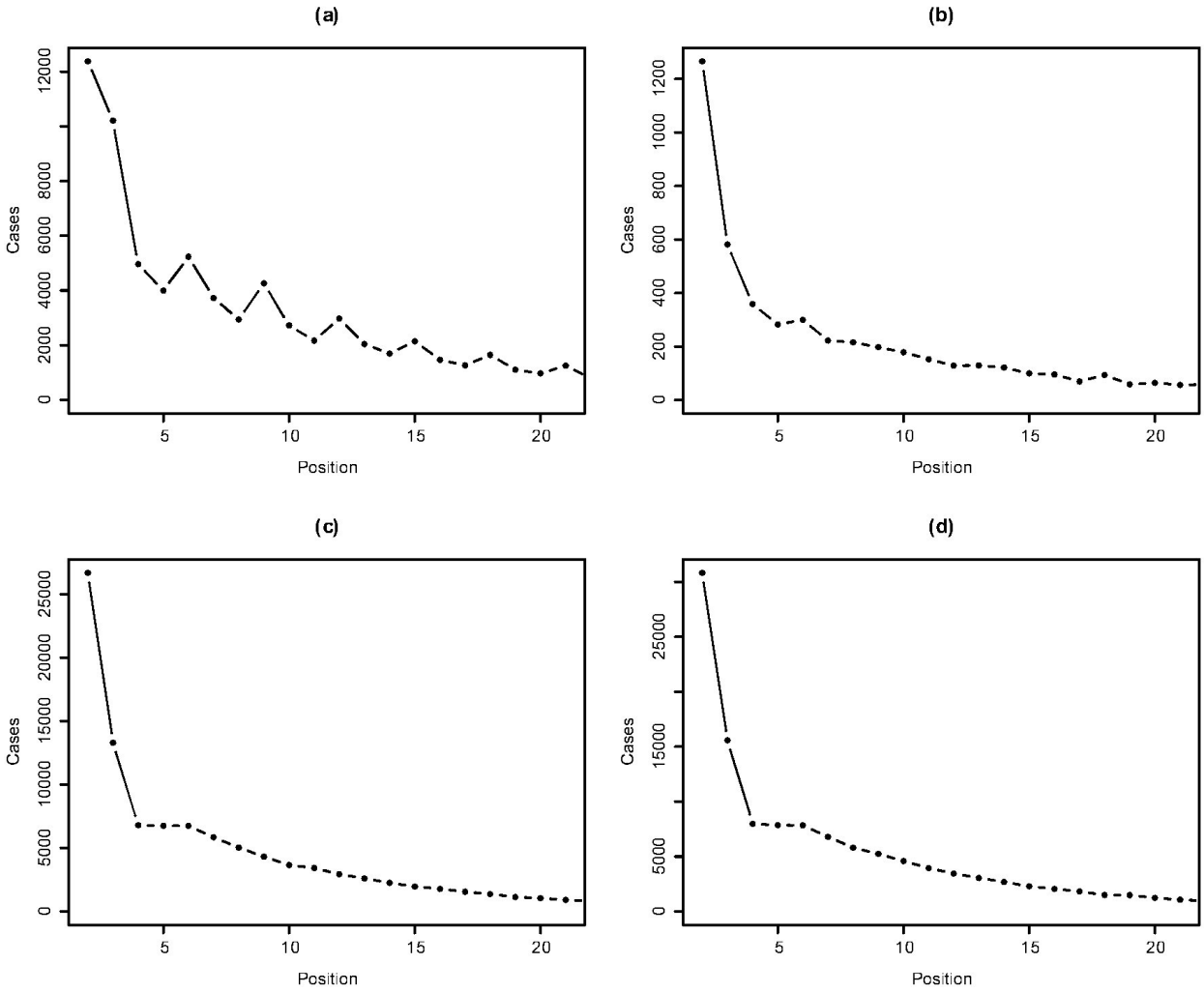

FIGURE 2. Distribution of distances. The number of cases of triplet YRY for the first 21 distances are plotted for (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

code [7, 20], but also that vestiges of this pattern are still detectable on current genomes [34]. Although this hypothesis was challenged by Wong & Cedergren [40], and also by Jukes [18], several cases have been found with RNY prevalence in actual genomes [18].

Arquès and Michel used a similar approach to look for periodicities in coding and noncoding regions [3]. They studied sequences from virus, prokaryotes, and eukaryotes, and looked for the $i-$motiff $m_i = YRY(N)_iYRY$, with i in the range [0,99], *i.e.* two triplets YRY separated by any i bases. They found that the motif $YRY(N)_6YRY$ had preferential occurence on the vast majority of the studied sequences [3]. Also two kinds of periodicities were found, the previously mentioned periodicity three, $P3$ (in both coding regions and also in noncoding regions from virus and mithocondria); and a periodicity two, $P2$ was identified in eukaryotic introns. The last periodicity was attributed to regulatory functions [2].

Although using a mutual information function to distance series (joint probabilities of finding the symbol A_i and k characters downstream the symbol A_j) Herzel and Große concluded that the nonuniform codon usage in protein coding sequences is responsible for the period-three oscillations [12], this periodicity has also been found in whole chromosomal bacterial genomes, with no relation with protein translation process [16]. Thus, we believe that the periodicity three is an intrinsic property of coding sequences, independently of the codon usage.

4.2. Autocorrelation function (ACF)

ACF allows us to prove the null hypothesis over individual data independence in a time series. Let $x(a)$ be a time series, and $x(a-\tau)$ the same series with a τ position delay. The general ACF computes the correlation between $x(a)$ and $x(a-\tau)$ using the following equation:

$$ACF = \langle x(a)x(a-\tau)\rangle - \langle x(a)\rangle\langle x(a-\tau)\rangle \qquad (1)$$

In equivalence with the correlation coefficient of Pearson, a value of $ACF = 1$ is indicative of a complete positive autocorrelation between the series; an $ACF = -1$ is indicative of a complete negative autocorrelation between the series; and an $ACF = 0$ is indicative of independence between the series, *i.e.* it is related with Gaussian white noise.

ACF has been used by Arquès and Michel to study both the $YRY(N)_6YRY$ preference in different kinds of genomes with random mutations [4], and to identify subsets of triplets having a preferential occurrence frame [5]. Following a similar approach, we obtain actual distance series for the ATG triplet, calculating the number of cases for each distance, and then computing the ACF. The results are shown on Fig. 3.

As shown on Fig. 3, while a coding sequence presents an oscillatory decaying pattern with a clear-cut rhythmical alternation of points, which are at distances of multiples of three (Fig. 3a), a noncoding sequence has no apparent periodicity (Fig. 3b), with a pattern similar with stochastic processes (see Fig. 3c and d). The dynamics observed in the coding DNA sequence, is typical of an scale-invariant power-law behavior.

4.3. Nearest neighbor nucleotide patterns

Several physicochemical properties of DNA depend on the interactions between consecutive bases, thus, the identification of patterns from nearest neighbor bases could help in the characterization of nucleotide sequences [22].

Although, the group of Kornberg was the first to measure nearest neighbor frequencies on DNA [17], it was not until recently, when some patterns were identified from the analysis of whole genomes [24, 25].

Nussinov counted the number of different dinucleotides, and found two kinds of patterns: (a) unequal frequencies of appearance of some asymmetric pairs, and (b) preferences of certain nucleotides with specific nearest neighbors over equivalent dinucleotides [24]. In the first case, she found that asymmetries AT $>$ TA; CT $>$ TC; TG $>$ GT; and, GC $>$ CG occur in all the examined genomes, including both prokarytes and eukaryotes. On Table I, the counts differences for these duplets are shown.

As shown on Table I, the highest differences in the counts were detected on the coding genome (original) of *Borrelia burgdorferi*. There is one order of magnitude in the difference between a coding and a noncoding sequence (HUMHBB), and three orders of magnitude in the difference between a coding sequence and its corresponding shuffled version. Furthermore, in the case of the pure random control (synthetic genome), a switch in the relative counts was detected.

TABLE I. Differences in nearest neighbor counts.

Duplet	Original genome	HUMHBB	Shuffled genome	Synthetic genome
AT – TA	13557	589	62	-72
CT – TC	2016	616	34	208
TG – GT	15572	1205	94	135
GC – CG	17577	1648	24	-393
average	12180.5	1014.5	53.5	-30.5

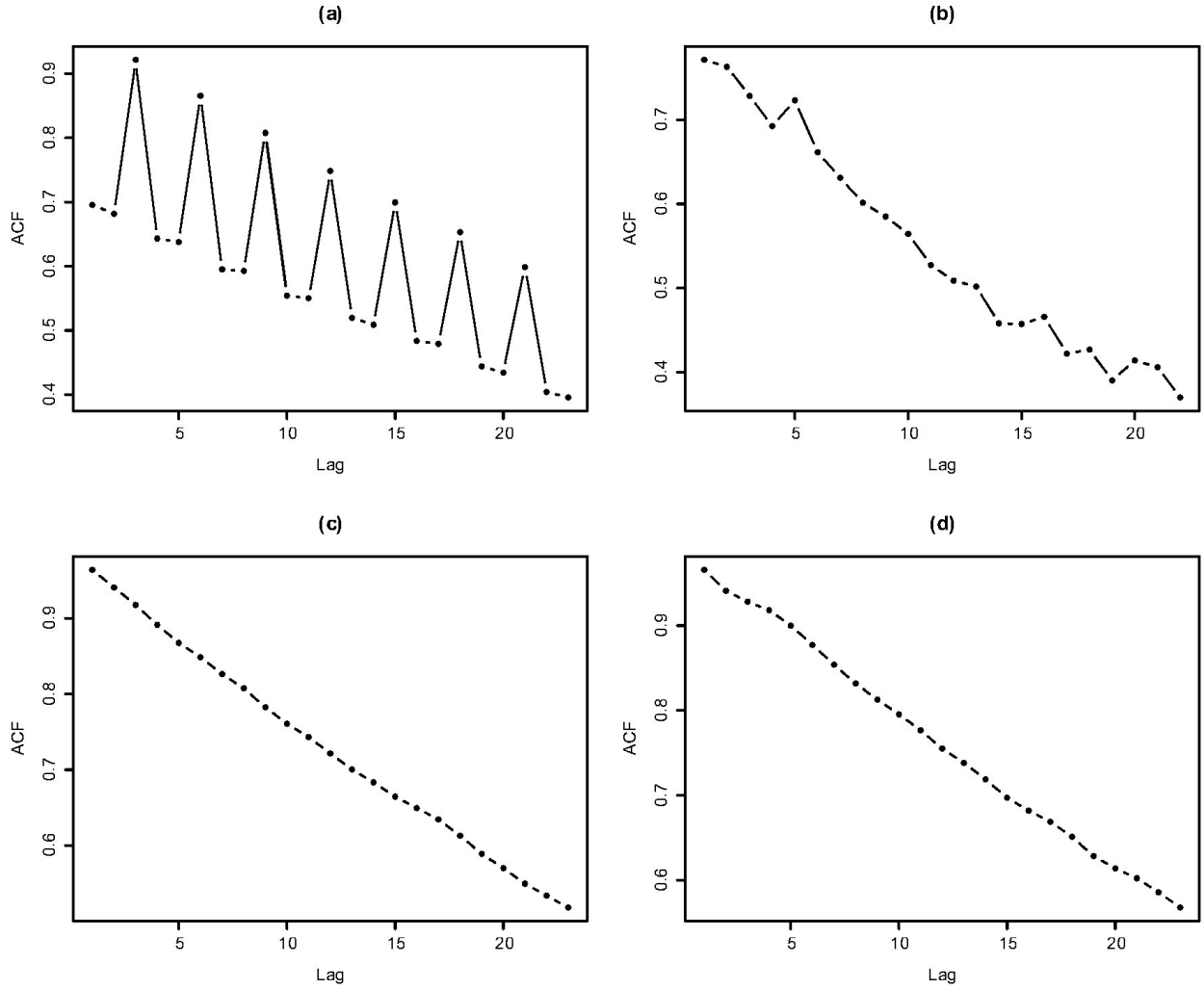

FIGURE 3. Autocorrelation function analysis (ACF). The ACF is computed from the distribution of ATG (number of cases vs. position) for (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

4.4. Long-range correlations

A power-law behavior of the form $y = f(x) = Ax^{\alpha}$, where α is the scaling exponent, and A is the normalization constant, is related with processes exhibiting self-similar properties (fractal dynamics), such as time series with long-range correlations [37]. As DNA sequences can be transformed to distance series, it is feasible to characterize long-range correlations in both coding and noncoding regions.

Using the one-dimensional random walk model, discussed before, Peng, *et al.* applied different, but related techniques to study long-range correlations in DNA [27, 28, 35, 36]. Their first approach was the use of the root-mean square fluctuation, $F(l)$ about the average of the displacement, defined as:

$$F(n) = \sqrt{\overline{[\Delta y(n) - \overline{\Delta y(n)}]^2}} \qquad (2)$$

where $\Delta y(n) = y(n_0 + n) - y(n_0)$, and the bars indicate the arithmetic mean over all positions n in the gene. There are two possible scenarios:

(a) for both pure random process, and for local correlations, $F(n) \sim n^{1/2}$; and

(b) for correlations with no characteristic length (long-range correlations); their fluctuations are described by the power law, $F(n) \sim n^{\alpha}$, with $\alpha \neq 1/2$.

Using this method (known as "min-max"), Peng, *et al.* found long-range correlations in noncoding regions, in contrast with coding regions where $\alpha \approx 0.5$ [27].

The above results have been widely discussed, as some authors argue that there is no difference between coding and noncoding regions [39]. One of the main critics, was the finding that there is heterogeneity in the random walk series obtained from DNA, thus, it has been claimed that DNA presents "patchiness" [19]. Patchiness can be clearly detected on Figs. 1a and b, in which the "walker" moves far away from the origin (compared with Fig. 1d which resembles a pure random process). Due to DNA patchiness, methods like ACF and the root-mean square of fluctuations are not valid,

as they depend on averages, which in turn, change over the DNA sequence.

In order to avoid the effect of DNA patchiness, Peng, *et al.*, improved their method by detrending the fluctuations over different windows or boxes [28]. This technique was called detrended fluctuation analysis (DFA).

In DFA, firstly a sequence of length N is divided into N/l nonoverlapping boxes, each containing l nucleotides, and then the local trend in each box, is calculated. Afterwards, the detrended walk, $y_l(n)$ is obtained with the difference between the original random walk $y(n)$, and the local trend. Next, the variances about the detrended walks are computed; and finally, the averages of these variances over all the boxes $(F_d^2(l))$ are calculated [28].

The reported results of Peng, *et al.*, were essentially the same as before [28], *i.e.* long-range correlations were detected on noncoding regions. In the case of the analyzed coding sequence, a crossover in the slope was detected, with an $\alpha = 0.51$ for the first part of the curve (in equivalence with pure random sequences).

There have been other approaches in order to eliminate local patchiness in DNA. Arneodo, *et al.*, introduced the use of the wavelet transform modulus maxima (WTMM) to study long-range correlations in DNA sequences [1]. In the WTMM the scaling properties of a time series is investigated in terms of their wavelet coefficients. By applying the WTMM to a DNA random walk series, from both coding and noncoding sequences, they also found long-range correlations in noncoding sequences and uncorrelated steps indistinguishable from the Brownian motion in coding sequences [1].

In contrast with the above results, other authors have found long-range correlations in coding sequences [6, 30]. In particular, instead of starting with the random walk series, Voss [39], and Sousa Vieira [38] calculated the power spectrum into equal-symbol correlation series, whereas Mohanty and Narayana Rao applied factorial moments to series representing the excess or deficit of purines over pyrimidines [23]. In these cases, long-range correlations were identified in large coding sequences.

In order to illustrate the long-range correlations in DNA sequences, here, we applied DFA to both, a time series obtained from the one-dimension random walk method, as well as a time series obtained from the actual distance series of triplet ATG. The results for the latter, are shown on Fig. 4.

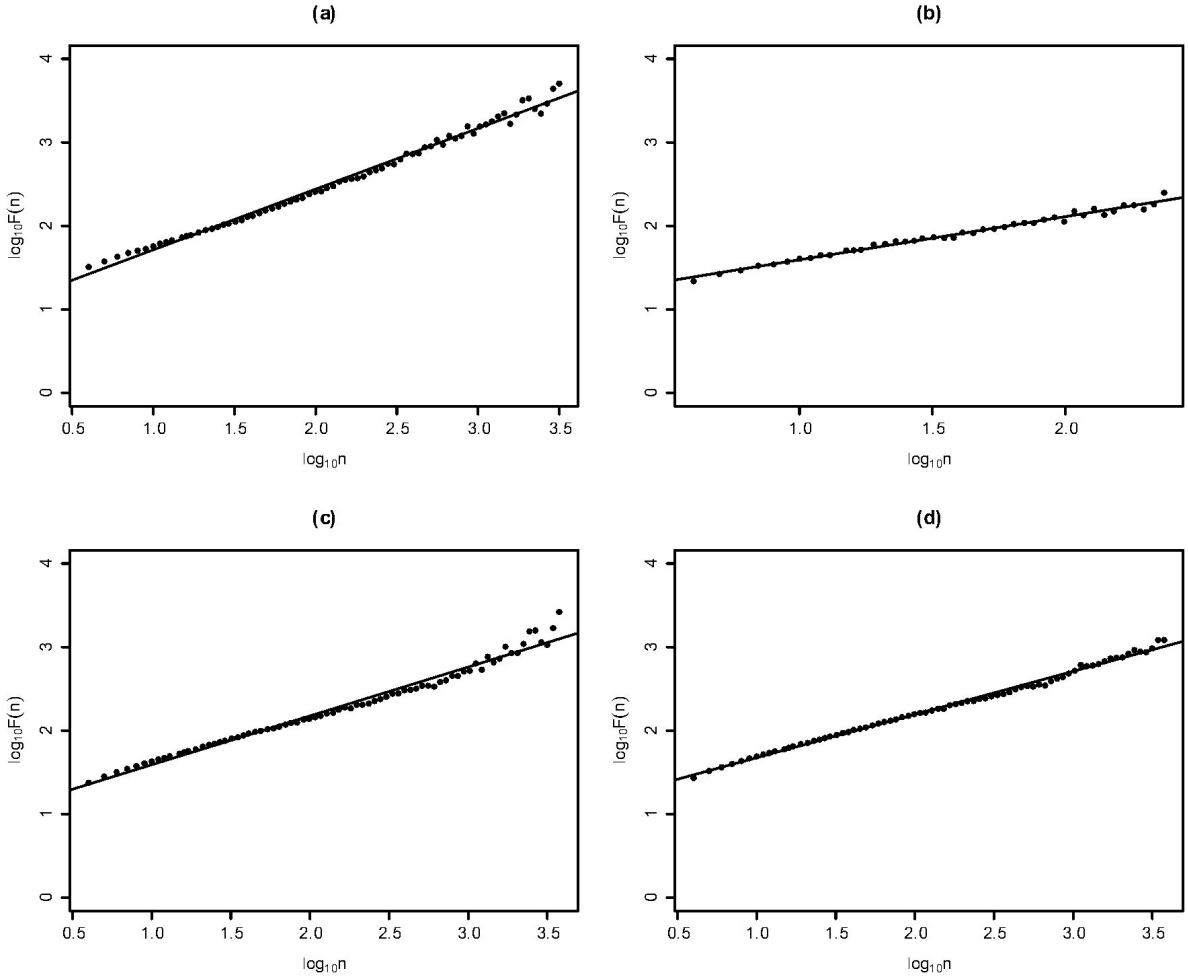

FIGURE 4. Detrended fluctuation analysis (DFA) for actual distance series of triplet ATG in (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome. The corresponding scaling exponents are shown on Table II.

TABLE II. Calculated scaling exponents α, for the studied cases.

Genome	Random walk	ATG distance
	α	α
Original (coding)	0.62*	0.73
HUMHBB (noncoding)	0.67	0.52
Shuffled	0.50	0.59
Synthetic	0.50	0.51

*Overall α, see text for explanation.

As shown on Fig. 4a, there was not a crossover in the slope of the curve. On Table II, we present the comparison of the obtained scaling exponents α, for both time series.

In accordance with Peng, *et al.*, [28], we detected crossovers on the coding sequence. In our case, two crossovers were identified (not shown), thus three different scaling exponents could be obtained: $\alpha_1 = 0.68$, with $n = 11$, number of points; $\alpha_2 = 0.52$, with $n = 29$; and, $\alpha_3 = 0.80$, with $n = 36$. On Table II, we present the overall scaling exponent $(n = 76)$.

On the other hand, in contradiction with Peng, *et al.*, [28], we detected long-range correlations in a coding sequence using both kinds of time series as input. In fact, the highest value of α was obtained from the ATG distance series from the coding genome of *Borrelia burgdorferi*. It is worth mentioning, that this sequence does not have any intergenic regions, thus is pure coding. We believe that using actual dis-

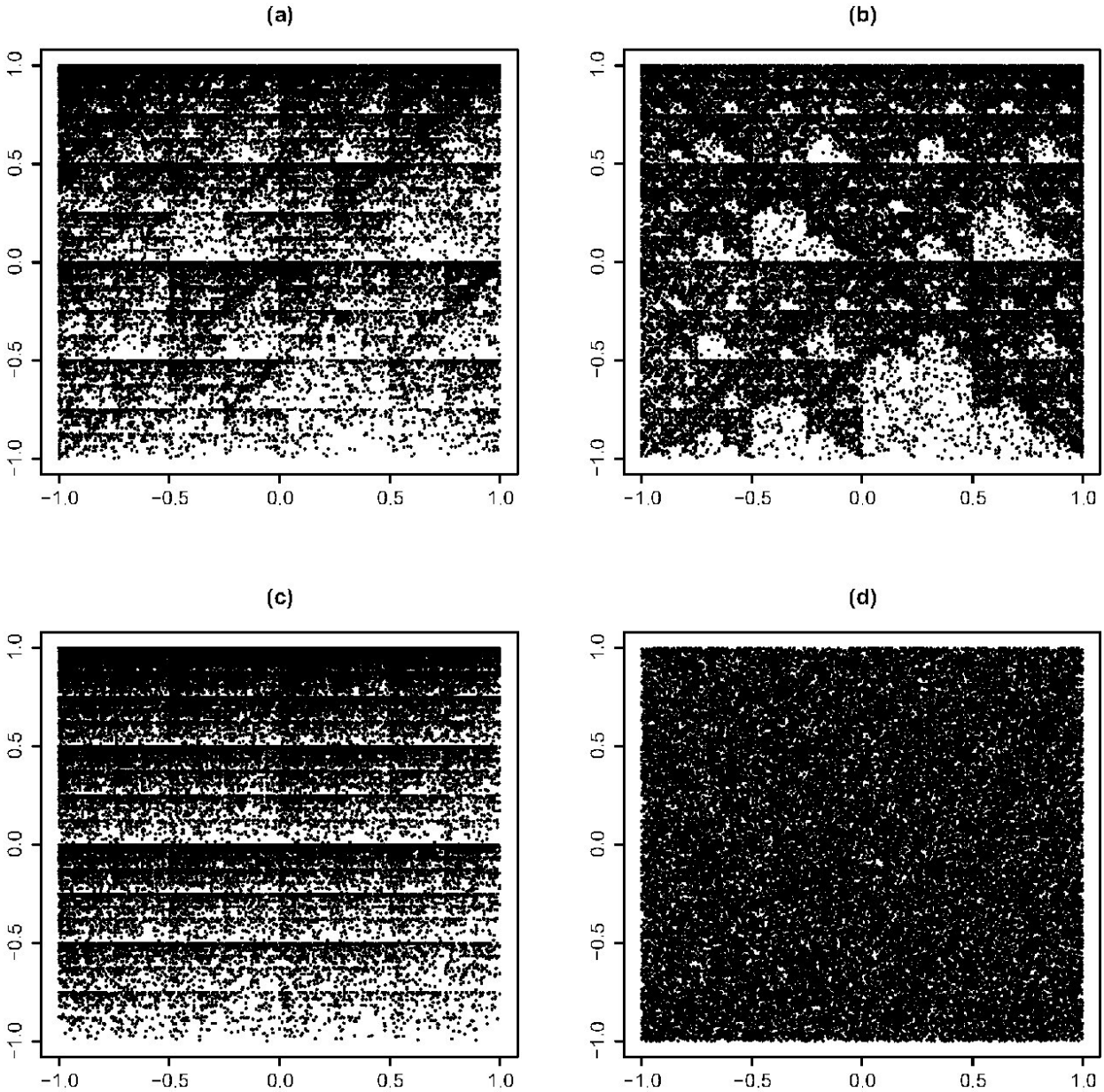

FIGURE 5. Chaos game representation (CGR) applied to DNA sequences. A shifts to the left upper corner; T shifts to the upper right corner; C shifts to the lower left corner; G shifts to the lower right corner. CGR was applied to: (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

tance series of triplets is a better option than using the one-dimension random walk method, as no information is lost (due to binary representations) [31], and is more biological related (due to translation) [9].

Another approach to look for fractal dynamics on DNA sequences has been the use of the chaos game representation (CGR) of gene structure [11, 13, 15, 26]. CGR is a scatter plot derived from a DNA sequence. First a CGR image is divided into squares in which each corner represents one base. Starting from a random point (*e.g.* (0,0)) the next point is plotted in the mid point from the straight line which connects the current point with one of the corners, determined by the DNA sequence. On Fig. 5, we illustrate CGRs from the DNA sequences discussed in the current paper. As expected, there is no pattern neither on the synthetic genome (Fig. 5d) nor on the shuffled genome (Fig. 5c). A fractal dynamics was observed on both the coding (Fig. 5a) and noncoding sequences (Fig. 5b), although it was more clear on the latter. CGRs were applied for the first 50,000 bases in the corresponding sequences.

4.5. Information content

Information can be measured in terms of the number of binary digits, *i.e.* by the logarithm of the number of possible messages [29]. This measure of information is called the Shannon entropy [32]:

$$H_n = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{3}$$

The term entropy is due to its relation with certain formulations of statistical mechanics where p_i is the probability of a system being in the cell i of its phase space [32].

In the case of two possible outcomes, with probabilities p and $q = 1 - p$, the Shannon entropy reaches its maximum value when $p = q$. This result can be generalized for any n, number of probabilities, thus, H_n is a maximum when all the p_i are equal, which is the most uncertain situation.

Several authors have used the Shannon entropy to analyze DNA information content [8, 14, 21]. Here, in order to illustrate the information content quantification in different sequences, we computed the Shannon entropy for the frequencies of all possible triplets (64), in the previously mentioned genomes. The results are shown on Table III.

Note that the entropy value from the synthetic genome had the expected value for a pure random process ($H_n \approx 6$). On the other hand, although the coding genome of *Borrelia burgdorferi* had the minimum entropy (*i.e.* more information content), its value was very closed to its shuffled version; this was unexpected, as the shuffling was made by nucleotides and not by triplets.

TABLE III. Shannon entropy for triplets frequencies.

Genome	H_n
Original (coding)	5.60
HUMHBB (noncoding)	5.82
Shuffled	5.63
Synthetic	5.99

Another alternative to calculate entropies from nonlinear time series, is the maximum entropy method (MEM), which is based upon the power spectrum of autocorrelation coefficients [10]. The MEM has been recently used to study information content in series of amino acids obtained from translating whole bacterial chromosomes [9]. In this case, the information content was proportionally related with the maximum entropy.

5. Concluding remarks

It is important to mathematically distinguish coding DNA sequences from non-coding ones, because through these kind of tools it is possible to identify quickly potential genes in the genome data bases, saving valuable time for a better experimental design. Furthermore, mathematical characterization of DNA sequences could help in the understanding of structural relationships among different genes along the chromosomes.

Although there has been controversies among the presence of long-range correlations in coding DNA, we have shown that the use of actual distance series between triplets is a better approach than the random walker DNA representation, as less information is lost, and a better characterization is made. When the analyses are carried out based upon the actual distance series, the presence of long-range correlations in coding sequences is clear, and its in accordance with the CGR of the same sequence.

The presence of periodical rhythms in the ACF, long-range correlation, and more information content in coding DNA sequences suggests that, although spontaneous mutations and horizontal genetic transfer occurs at random, there should be some kind of structural rules which favor the natural selection of sequences in which these properties are maintained.

Acknowledgments

We are very grearful to Julio Collado and Imelda López for their valuable feed back and for reading the manuscript. M.V.J. was financially supported by PAPIIT-IN205702, UNAM, México.

∗. Corresponding address: Laboratory of Theoretic Biology, Research Department, La Salle University, Benjamin Franklin 47, Col. Condesa, México, D.F. 06140, México Fax:(52-55)5515-7631. e-mail:jgarcia@ci.ulsa.mx

1. A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, *Phys. Rev. Lett.* **74** (1995) 3293.
2. D.G. Arquès and C.J. Michel, *Nucleic Acids Res.* **15** (1987) 7581.
3. D.G. Arquès and C.J. Michel, *J. theor. Biol.* **143** (1990) 307.
4. D.G. Arquès and C.J. Michel, *Math. Biosci.* **123** (1994) 103.
5. D.G. Arquès and C.J. Michel, *Biosystems* **44** (1997) 107.
6. C.A. Chatzidimitriou-Dreismann and D. Larhammar, *Nature* **361** (1993) 212.
7. M. Eigen and P. Schuster, *The hypercycle. A principle of natural self-organization* (Springer-Verlag, Berlin, Alemania, 1979)
8. L. Frappat, C. Minichini, A. Sciarrino, and P. Sorba, *Phys. Rev. E* **68** (2004) 061910.
9. J.A. García, S. Alvarez, A. Flores, T. Govezensky, J.R. Bobadilla, and M.V. José *Physica A* **342** (2004) 288.
10. M. Ghil *et al.*, *Rev. Geophys.* **40** (2002) 1.
11. N. Goldman, *Nucleic Acids Res.* **21** (1993) 2487.
12. H. Herzel and I. Große, *Phys. Rev. E* **55** (1997) 800.
13. K.A. Hill, N.J. Schisler, and S.M. Singh, *J. Mol. Evol.* **35** (1992) 261.
14. D. Holste, I. Große, and H. Herzel, *Phys. Rev. E* **64** (2001) 1.
15. H.J. Jeffrey, *Nucleic Acids Res.* **18** (1990) 2163.
16. M.V. José, J.A. García, J.R. Bobadilla, and T. Govezensky. International Conference on Biological Physics, Gothenburg (2004).
17. J. Josse, A.D. Kaiser, and A. Kornberg, *J. Biol. Chem.* **236** (1961) 864.
18. T.H. Jukes, *J. Mol. Evol.* **42** (1996) 377.
19. S. Karlin and V. Brendel, *Science* **259** (1993) 677.
20. J. Konecny, M. Schöniger, and L. Hofacker, *J. theor. Biol.* **173** (1995) 263.
21. R.N. Mantegna *et al.*, *Phys. Rev. E* **52** (1995) 2939.
22. P. Miramontes *et al.*, *J. Mol. Evol.* **40** (1995) 698.
23. A.K. Mohanty and A.V.S.S. Narayana Rao, *Phys. Rev. Lett.* **84** (2000) 1832.
24. R. Nussinov, *Nucleic Acids Res.* **8** (1980) 4545.
25. R. Nussinov, *J. Biol. Chem.* **256** (1981) 8458.
26. J.L. Oliver, P. Bernaola-Galván, J. Guerrero-García, and R. Román-Roldán, *J. theor. Biol.* **160** (1993) 457.
27. C.-K. Peng *et al.*, *Nature* **356** (1992) 168.
28. C.-K. Peng *et al.*, *Phys. Rev. E* **49** (1994) 1685.
29. J.R. Pierce, *An introduction to information theory. Symbols, signals and noise* (Dover Publications, Inc., New York, USA, 1980)
30. V.V. Prabhu and J.M. Claverie, *Nature* **359** (1992) 782.
31. J. Sánchez and M.V. José, *Biochem. Biophys. Res. Comm.* **299** (2002) 126.
32. C.E. Shannon, *Bell Syst. Tech. J.* **27** (1948) 379.
33. J.C.W. Shepherd, J. Mol. Evol. **17** (1981) 94.
34. J.C.W. Shepherd, *Proc. Natl. Acad. Sci. USA* **78** (1981) 1596.
35. H.E. Stanley *et al.*, *Physica A* **191** (1992) 1.
36. H.E. Stanley *et al.*, *Physica A* **205** (1994) 214.
37. H.E. Stanley *et al.*, In: *Fractal geometry in biological systems. An analytical approach*, eds. P.M. Iannacone and M. Khokha, (CRC Press, New York, USA, 1996) p. 15.
38. M.S. Vieira, *Phys. Rev. E* **60** (1999) 5932.
39. R.F. Voss, *Phys. Rev. Lett.* **68** (1992) 3805.
40. J.T.F. Wong and R. Cedergren, *Eur. J. Biochem.* **159** (1986) 175.