Eur. Phys. J. B **1**, 259–263 (1998)

THE EUROPEAN PHYSICAL JOURNAL B

Nucleotide composition effects on the long-range correlations in human genes

A. Arnéodo[1,a], Y. d'Aubenton-Carafa[2], B. Audit[1], E. Bacry[3], J.F. Muzy[1], and C. Thermes[2]

[1] Centre de Recherche Paul Pascal, avenue Schweitzer, 33600 Pessac Cedex, France
[2] Centre de Génétique Moléculaire du CNRS[b], Allée de la Terrasse, 91198 Gif-sur-Yvette, France
[3] Centre de Mathématiques Appliquées, École Polytechnique, 91128 Palaiseau, France

Received: 18 August 1997 / Accepted: 29 October 1997

Abstract. We use the wavelet transform to investigate the fractal scaling properties of coding and noncoding human DNA sequences. We find that the strength of the long-range correlations observed in the introns increases with the guanine-cytosine (GC) content, while coding sequences show no such correlations at any GC content. However, we demonstrate that long-range correlations can be detected when the coding sequences are undersampled by retaining the third base of each codon only. This strongly suggests that the observed correlations are not likely to be due to insertion-deletion mechanisms. We comment about the origin of these correlations in terms of putative dynamical processes that could produce the isochore structure of the human genome.

PACS. 87.10.+e General, theoretical, and mathematical biophysics (including logic of biosystems, quantum biology, and relevant aspects of thermodynamics, information theory, cybernetics, and bionics) – 05.40.+j Fluctuation phenomena, random processes, and Brownian motion – 72.70.+m Noise processes and phenomena

The immense progress made recently in molecular biology has revealed that genomes are of extraordinary complexity [1]. The sequencing of DNA has shed some light on one of the main characteristic features of the genome organization namely its local and global compositional heterogeneity. The possible relevance of scale invariance and fractal concepts to the structural complexity of genomic sequences is the subject of considerable increasing interest [2]. During the past few years, there has been intense discussion about the existence, the nature and the origin of long-range correlations in DNA sequences. Different techniques including mutual information functions [3], autocorrelation functions [4], power spectra [5], "DNA walk" representation [2,6], Zipf analysis [7] were used for statistical analysis of DNA sequences. But despite the effort spent, there is still some continuing debate on rather struggling questions. In that respect, it is important to corroborate the fact that the reported correlations are not just an artefact of the nonuniformity in composition of genes [4,8]. Furthermore, it is still an open question whether the long-range correlation properties are different for protein-coding (exonic) and noncoding (intronic, intergenic) regions of a nucleotide sequence [2–7,9].

One of the main obstacles to long-range correlation analysis in DNA sequences is the mosaic structure of these sequences which are well known to be formed of "patches" ("strand bias") of different underlying composition [10]. These patches appear as trends in the DNA walk landscapes and are likely to introduce some breaking of the scale invariance [8]. Recently, the wavelet transform (WT) has been proposed as a very powerful technique for fractal analysis of DNA sequences [11]. By considering analyzing wavelets that make the "WT microscope" blind to low-frequency trends, one can reveal and quantify the scaling properties of DNA walks. In a preliminary work [11], by applying the so-called wavelet transform modulus maxima (WTMM) method to the analysis of various genomic sequences mainly selected in the human genome, we have found that the fluctuations in the patchy landscapes of both coding and noncoding DNA walks are homogeneous with Gaussian statistics. The main consequence of this result is the justification of using a single exponent, namely the Hurst or roughness exponent H, to characterize the fractal underlying hierarchical organization of DNA sequences. Moreover, the WTMM method [11] has provided strong indications that the fluctuations in noncoding regions behave like fractional Brownian motions ($H > 1/2$), whereas those of coding regions cannot be distinguished from uncorrelated Brownian walks ($H = 1/2$). The aim of this Letter is to push further this analysis and to show

[a] e-mail: arneodo@crpp.u-bordeaux.fr
[b] Laboratoire associé à l'Université Pierre et Marie Curie

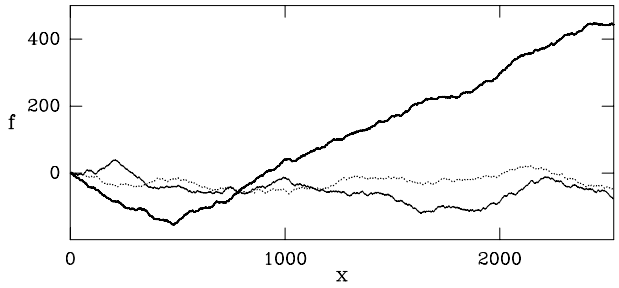

Fig. 1. DNA walk representation of one intron ($L = 2530$) of the human MHC class II HLA-DC-BETA gene: (—) Purine (AG)- Pyrimidine (CT) distinction, (▬▬) Weak (AT)- Strong (GC) coupling distinction, (·····) Amino (AC)- keto (GT) distinction.

that the situation is not that dichotomous since long-range correlations are also present in exons and CDSs but somehow hidden in their inner codon structure. Moreover, we will provide evidence that these correlations, as well as those found in introns, are related to the GC content of the considered sequence.

We concentrate our study on the statistical analysis of 121 DNA sequences selected in the human genome, with the requirement that their overall length L be greater than 2000 nucleotides, so that the range of scales available to fractal scaling be large enough to make the analysis meaningful with respect to finite-size effects. We took the sequences from the EMBL data bank and processed separately 47 coding (individual exons, CDSs) and 74 noncoding (individual introns) regions. To graphically portray these sequences, we follow the so-called "DNA walk" analysis [6] which requires first to convert the four-letter (A, C, G, T) DNA text into a binary sequence $\chi(i)$, by replacing, for example, each purine (A, G) by $+1$ and each pyrimidine (C, T) by -1. The graph of the DNA walk defined by the cumulative variable $f(x) = \sum_{i=1}^{x} \chi(i)$ is plotted in Figure 1 for one intron ($L = 2530$) of the human MHC class II HLA-DC-beta gene. As illustrated on this figure, in order to test the robustness of our results with respect to the chosen binary coding, we will systematically repeat our analysis for the two complementary pair-base identifications [9], namely strong (C, G) *versus* weak (A, T) bonding distinction and the hybrid rule which consists in identifying (A, C) and (G, T). Note that the low frequency trends that appear in the corresponding noisy patchy landscapes are clearly code dependent. To investigate the fractal scaling properties of a DNA sequence, we apply the WTMM method [12] which is a natural generalization of the classical box-counting technique, the wavelets playing the role of "generalized oscillating boxes". This amounts first to wavelet transform the graph $f(x)$:

$$T_\psi[f](x,a) = \frac{1}{a}\int_{-\infty}^{+\infty} f(y)\psi\left(\frac{y-x}{a}\right)\mathrm{d}y, \qquad (1)$$

where x is the space parameter, a (>0) the scale parameter and ψ the analyzing wavelet. In order to break free from the intrinsic patchiness of DNA sequences, we will use as analyzing wavelet, the Mexican hat ($\psi^{(2)}(x) = \dfrac{\mathrm{d}^2}{\mathrm{d}x^2}\mathrm{e}^{-\frac{1}{2}x^2}$) that has two vanishing moments [11]. Since the observed trends are mainly linear, higher order wavelets yield similar results. A partition of the space-scale half-plane is provided by the WT skeleton $\mathcal{S}(a)$ defined, at each scale a, by the set of all the points x_i that correspond to local maxima of $|T_\psi(x,a)|$ considered as a function of x. A statistical characterization of the fluctuations in roughness of a multifractal landscape, can be achieved by investigating the scaling behavior of some partition functions [11,12]:

$$Z(q,a) = \sum_{x_i \in \mathcal{S}(a)} |T_\psi(x_i,a)|^q \sim a^{\tau(q)}, \qquad (2)$$

where $q \in \mathbb{R}$. Then by Legendre transforming $\tau(q)$ one gets the singularity spectrum $D(h) = \min_q(qh - \tau(q))$, defined as the fractal dimension of the set of points x where locally the roughness exponent is h. As originally revealed in reference [11], the fluctuations in DNA landscapes are homogeneous as characterized by a linear spectrum $\tau(q) = qH - 1$, where H is the Hurst exponent. A test of this monofractality consists in checking that $H(q) = \lim_{a\to 0^+} \log_2[(aZ(q,a))^{\frac{1}{q}}]/\log_2 a$ actually does not depend on q. As pointed out in reference [12], there is an alternative way to proceed which requires first the computation of some Boltzmann weight: $\tilde{T}_\psi(q,x_i,a) = |T_\psi(x_i,a)|^q / Z(q,a)$, from which one can calculate:

$$h(q,a) = \sum_{x_i \in \mathcal{S}(a)} \tilde{T}_\psi(q,x_i,a)\log_2(|T_\psi(x_i,a)|). \qquad (3)$$

Then, one can estimate the generalized roughness exponents $h(q) = \partial\tau/\partial q = \lim_{a\to 0^+} h(q,a)/\log_2 a$, and check whether the fluctuations are homogeneous ($h(q) = H, \forall q$) or multifractal ($h(q)$ dependent on q). Since the estimates of the roughness exponent H from respectively $H(q)$ and $h(q)$ yield similar results, we will concentrate, in the following, on the computation of $h(q)$.

We report in Figure 2, the results of the application of the WTMM method when averaging respectively over our statistical samples of 74 introns and 42 CDSs. To systematically examine the effect of variations in GC content on the scaling properties, some care is needed in the averaging procedure in order to remove the bias induced by the nonuniform distribution of GC content in our statistical sample [13]. The results obtained for the mean generalized roughness exponents $\bar{h}(q)$ confirm the observations reported in reference [11]: $\bar{h}(q)$ does not display any significant q-dependence for $-2 \leq q \leq 2$, corroborating the fact that the fluctuations in coding as well as noncoding DNA walks are homogeneous. When using the purine-pyrimidine coding, on a range of scale extending over about a decade (from a size of 15 to about 150 nucleotides), $\bar{h}(q,a)$ computed as averages over both CDSs and introns display a quite convincing scaling behavior: $\bar{h}(q,a) \sim H\log_2 a$ with $H = H_{\mathrm{C}} = 0.50 \pm 0.01$ and $H = H_{\mathrm{NC}} = 0.58 \pm 0.02$ respectively. This distinction is enlightened if we plot, as in Figure 2a for $q = 0$,

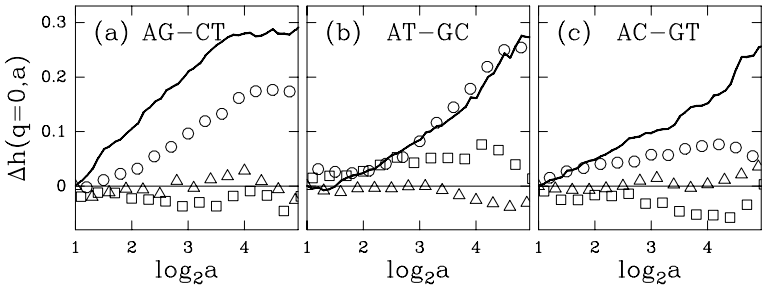

Fig. 2. $\Delta h(q,a) = \bar{h}(q,a) - \bar{h}_{\rm C}(q,a)$ *versus* $\log_2 a$ (Eq. (3)) for $q = 0$, when averaged over 42 coding and 74 non coding human DNA sequences: introns (——), CDSs (—), coding subsequences relative to position 1 (△), 2 (□) and 3 (○) of the bases within the codons. The analyzing wavelet is the Mexican hat $\psi^{(2)}$.

$\Delta h(q,a) = \bar{h}(q,a) - \bar{h}_{\rm C}(q,a)$, which is shown to increase linearly *versus* $\log_2 a$ with a slope $\Delta H_{\rm NC} = H_{\rm NC} - H_{\rm C} \simeq 0.08$. We thus confirm the presence of long-range correlations in non coding sequences. Nevertheless, because of the "period three" codon structure of coding DNA, it is natural to investigate separately the 3 subsequences relative to the position (1, 2 or 3) of the bases within their codons [14]. We have build up these subsequences from the 35 largest CDSs and we have repeated the WTMM analysis. As shown in Figure 2a, the data for $\Delta h(q,a)$ *versus* $\log_2 a$ display a rather flat behavior for both the subsequences relative to positions 1 and 2 which indicates that the measured roughness exponents $H_{\rm C1} = 0.51 \pm 0.01$ and $H_{\rm C2} = 0.50 \pm 0.02$ are undistinguishable from $H_{\rm C}$ and therefore from 1/2. Surprisingly, the data for the subsequence relative to position 3 exhibit an unambiguous linear increase with a slope $\Delta H_{\rm C3} \simeq 0.07$ which reflects the fact that $H_{\rm C3} = 0.57 \pm 0.02$, *i.e.*, a value which is very close to the exponent estimated for introns. This observation suggests that this third coding subsequence is likely to possess the same degree of long-range correlations as non coding sequences. As illustrated in Figure 2b, similar results are obtained when considering the weak-strong bonding coding: we get $H_{\rm NC} = 0.62 \pm 0.04$ and $H_{\rm C} = 0.53 \pm 0.03$. Note that the agreement between the $\Delta h(q=0,a)$ data for respectively the third coding subsequence and the introns is again rather good. The data for both the first and the second coding subsequences do not display a similar dependence *versus* $\log_2 a$. One actually gets $H_{\rm C1} = 0.52 \pm 0.02$ and $H_{\rm C2} = 0.54 \pm 0.02$, *i.e.*, values that are very close to $H_{\rm C}$, while $H_{\rm C3} = 0.63 \pm 0.04 \simeq H_{\rm NC}$. In Figure 2c are reported the results obtained when considering the hybrid coding (A, C) *versus* (G, T). In that case, one gets $H_{\rm NC} = 0.58 \pm 0.03$ and $H_{\rm C} = 0.51 \pm 0.01$, *i.e.*, values that are quite consistent with previous estimates. While the effect seems to be still present, the hybrid coding is undoubtedly the coding for which the long-range correlations do not emerge very clearly in the third coding subsequence.

In Figures 3 and 4 are reported the results of a similar statistical analysis with the purine-pyrimidine coding, when classifying the DNA sequences into categories that correspond to a given GC content. The idea of looking for a link between the correlation properties and the GC content of the sequences results from the remark that the WTMM method indeed fails to distinguish a few introns from actual exons [11]. For example, two introns of the human factor XIIIb subunit gene ($L = 9952$ and $L = 2874$)

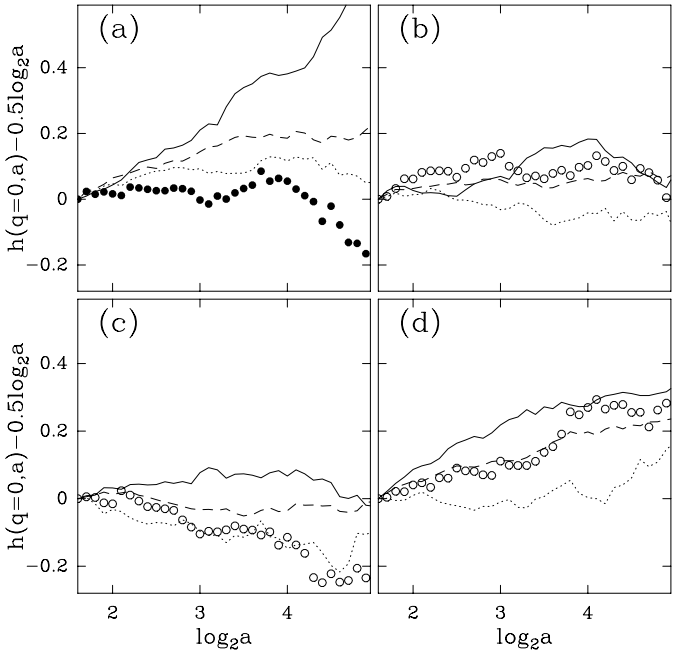

Fig. 3. $\Delta h(q,a) = h(q,a) - 0.5 \log_2 a$ *versus* $\log_2 a$ (Eq. (3)) for $q = 0$, as computed for a given GC content of the DNA sequences and using the Purine-Pyrimidine distinction. (a) Introns: (·····) GC% = 31.6 ± 0.4, $L = \sum_i L_i = 15210$ nucleotides; (- - -) GC% = 48.6 ± 1.0, $L = 17137$; (—) GC% = 63.3 ± 0.9, $L = 10449$; (•) intron (nb 8) of the human factor XIIIb subunit gene with GC% = 31.2, $L = 2874$. Coding subsequences relative to position 1 (b), 2 (c) and 3 (d) of the bases within the codons: (·····) GC% = 38.1 ± 2.9, $L = 4759$; (- - -) GC% = 50.8 ± 2.8, $L = 28521$; (—) GC% = 62.5 ± 2.1, $L = 16558$; (○) exon (nb 26) of the human apoB-100 gene with GC% = 41.0, $L = 7571$. Same analyzing wavelet as in Figure 2.

have respectively the Hurst exponents $H = 0.49 \pm 0.02$ and $H = 0.50 \pm 0.02$. Similarly, one intron ($L = 10986$) of the human retinoblastoma susceptibility gene has an exponent $H = 0.51 \pm 0.02$ which is again very close to 1/2. These introns actually correspond to DNA sequences with a low GC content (from 31% to 36%). As shown in Figure 3a, when comparing $h(q=0,a)$ with $0.5 \log_2 a$, *i.e.*, the scaling behavior expected for uncorrelated sequences, one notices some significant tendency of the curves to become steeper when continuously increasing the GC content. The corresponding values of the roughness exponent

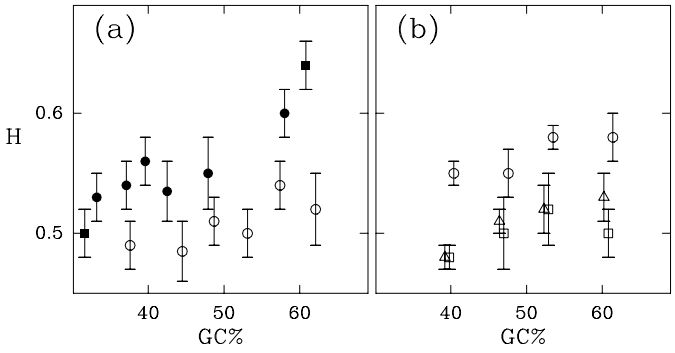

Fig. 4. WTMM estimate of the roughness exponent H *versus* the GC content of the DNA sequence. (a) Introns: (•) $L \simeq 50000$, (■) $L \simeq 15000$; CDSs: (○) $L \simeq 50000$. (b) Coding subsequences relative to position 1 (△), 2 (□) and 3 (○) of the bases within the codons: $L \simeq 20000$.

H are reported in Figure 4a. H clearly increases from values close to 1/2 at low GC content ($\simeq 30\%$) up to values significantly larger than 0.6 at high GC content ($> 60\%$). In Figure 4a are also shown the estimates of the roughness exponent for the coding sequences. Whether the CDS be poor or rich in GC, it does not seem to possess strong long-range correlations as indicated by an exponent H close to 1/2. Figure 4b is devoted to the results of similar analysis of the three coding subsequences relative to the position (1, 2 or 3) of the bases within the codons. For the first and second subsequences, one gets results quite consistent with the estimates obtained for the overall CDS sequences: whatever the GC content, the exponent H does not significantly depart from the value $H = 1/2$. Note that the data do not exclude a slow increase of H_{C1} and H_{C2}. For the third subsequence, H is found to increase up to values close to 0.60 at high GC content, which brings the clue that this subsequence exhibits [GC]-dependent long-range correlations very much like those observed in Figure 4a for introns. Unfortunately, for these subsequences, the overall statistics is rather poor which makes a quantitative comparison between the third coding subsequences and the introns uncertain. Nevertheless, as shown in Figures 3b, 3c and 3d, only the data for the third coding subsequence exhibit a significant increase in the steepness of the curves when going from low to high GC content. The curves remain much flattened for the first and second subsequences, even though some small, but maybe meaningful, increase is perceptible for both of them. In order to investigate the possibility that these observations might result from the exon concatenation in the CDSs, we have analyzed individual human exons (for statistical reason only the largest ones). These exons exhibit the same features than the CDSs, as it is examplified in Figure 3 by the apoB-100 largest exon.

The results reported in this work clearly show that the GC content is likely to be relevant to the long-range correlation properties observed in both intronic and exonic DNA sequences. The evolution of DNA sequences in terms of GC content has attracted a lot of interest during the past few years [10,13–16]. Several mechanisms can be proposed to account for the observed long-range correlations in the GC rich intronic sequences.

(i) Besides punctual mutations, genomic sequences are subject to a number of insertion-deletion events of DNA fragments of widely variable sizes. These events are much less frequent in exonic regions due to the strong constraints imposed by their coding properties. The insertion-deletion mechanisms could be responsible for the observed long-range correlations [3]. However, insertions-deletions occur in intronic sequences presenting a low GC content, which were just shown to present no long-range correlations. Furthermore, the correlations observed between the third bases of the codons, but not between adjacent nucleotides, are unlikely to result from (rare) insertion-deletion events which generally involve several adjacent nucleotides in order to maintain the coding phase.

(ii) The human genome is well known to be compartimentalized into wide specific domains with uniform GC content, called isochores [10]; appreciable scatter of the average GC content is actually observed when comparing different domains. Another hypothesis is to consider that the processes operating to create the GC rich isochores lead to the appearance of long-range correlations. Thanks to the functional constraints acting on the coding sequences embedded in these GC rich regions, these processes should be less active on the exons, with a concomittant lack of long-range correlations as compared to the surrounding introns. Since these constraints are less stringent on the third base of the codons, this would explain the long-range correlations observed between these nucleotides in high GC containing exons. In human genes, the frequencies of the third base of the codons are highly correlated with the neighbouring intronic GC content [17]. This property favors the hypothesis that the exonic correlations are produced by the same mechanisms which lead to intronic correlations.

The choice of human genes in this study was mainly dictated by the large sample of available coding (CDS, exons) and noncoding (introns) sequences and also by their widespread GC content variability [13,15]. From the similarity of the compositional patterns of the human genome with those of the genomes of mammals and warmblooded vertebrates, it is likely that the observations reported here also extend to these other genomes. The exploration of genomes of various organisms including unicellular eukaryotes and prokaryotes is currently under progress.

This research was supported by the GIP GREG (project "Motifs dans les Séquences") and by the Ministère de l'Éducation Nationale, de l'Enseignement Supérieur, de la Recherche et de l'Insertion Professionnelle ACC-SV (project "Génétique et Environnement").

References

1. W.H. Li, D. Graur, *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA, 1991).
2. H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, S.M. Ossadnik, C.K. Peng, M. Simons, Fractals **1**, 283 (1993).
3. W. Li, Int. J. Bif. & Chaos **2**, 137 (1992).
4. M.Ya. Azbel, Phys. Rev. Lett. **75**, 168 (1995); H. Herzel, I. Groβe, Physica A **216**, 518 (1995).
5. R.F. Voss, Phys. Rev. Lett. **68**, 3805 (1992); Fractals **2**, 1 (1994).
6. C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature **356**, 168 (1992).
7. R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E **52**, 2939 (1995); S. Havlin, S.V. Buldyrev, A.L. Goldberger, R.N. Mantegna, C.K. Peng, M. Simons, H.E. Stanley, Fractals **3**, 269 (1995).
8. S. Nee, Nature **357**, 450 (1992); C.A. Chazidimitriou-Dreismann, L. Larhammar, Nature **361**, 212 (1993); S. Karlin, V. Brendel, Science **259**, 677 (1993); B. Borstnik, D. Pumpernik, D. Lukman, Europhys. Lett. **23**, 389 (1993).
9. S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E **51**, 5084 (1995).
10. G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, F. Cuny, M. Meunier-Rotival, F. Rodier, Science **228**, 953 (1985); G.A. Churchill, Bull. Math. Biol. **51**, 79 (1989); J.W. Fickett, D.C. Torney, D.R. Wolf, Genomics **13**, 1056 (1992).
11. A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995); A. Arneodo, Y. d'Aubenton-Carafa, E. Bacry, P.V. Graves, J.F. Muzy, C. Thermes, Physica D **96**, 291 (1996).
12. J.F. Muzy, E. Bacry, A. Arneodo, Int. J. of Bif. & Chaos **4**, 245 (1994); A. Arneodo, E. Bacry, J.F. Muzy, Physica A **213**, 232 (1995).
13. G. Bernardi, Annu. Rev. Genet. **23**, 637 (1989).
14. P. Allegrini, M. Barbi, P. Grigolini, B.J. West, Phys. Rev. E **52**, 5281 (1995).
15. B. Aïssani, G. D'Onofrio, D. Mouchiroud, K. Gardiner, C. Gautier, G. Bernardi, J. Mol. Evol. **32**, 493 (1991); G. D'Onofrio, D. Mouchiroud, B. Aïssani, C. Gautier, G. Bernardi, J. Mol. Evol. **32**, 504 (1991); G. D'Onofrio, G. Bernardi, Gene **110**, 81 (1992).
16. P. Liò, S. Ruffo, M. Buiatti, J. Theor. Biol. **171**, 215 (1994).
17. P. Sharp, G. Matassi, Curr. Opin. Gent. Dev. **4**, 851 (1994).