

Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis

A. Arneodo,¹ E. Bacry,² P. V. Graves,³ and J. F. Muzy¹

¹Centre de Recherche Paul Pascal, Avenue Schweitzer, 33600 Pessac, France

²Université de Paris VII, Unité Fondamentale de Recherche de Mathématiques, Tour 45-55, 2 Place Jussieu, 75251 Paris Cedex 05, France

³Institut de Biologie et de Génétique Cellulaire, 1 rue C. Saint Saëns, 33077 Bordeaux, France
(Received 22 September 1994)

The fractal scaling properties of DNA sequences are analyzed using the wavelet transform. Because the wavelet transform microscope can be made blind to the “patchiness” of genomic sequences, we demonstrate and quantify the existence of long-range correlations in genes containing introns and noncoding regions. Moreover, the fluctuations in the patchy landscapes of DNA walks are found to be homogeneous with Gaussian statistics.

PACS numbers: 87.10.+e, 05.40.+j, 07.07.-t, 72.70.+m

The possible relevance of scale invariance and fractal concepts to the structural complexity of genomic DNA has been the subject of considerable recent interest. During the past few years, there has been intense discussion about the existence and the nature of long-range correlations within DNA sequences [1–6]. But despite the efforts spent, it is still an open question whether the long-range correlation properties are different for intronless and intron-containing coding regions. On more fundamental ground, there is still continuing debate as to whether the reported long-range correlations really mean a lack of independence at long distances or simply reflect the “patchiness” (bias in nucleotide composition) of DNA sequences [1–6]. One of the main reasons for this controversial situation is that the different techniques (e.g., DNA walk, correlation function, and power spectrum analyses) used, so far, for characterizing the presence of long-range correlations are not well adapted to study patchy sequences [6]. In that respect, there have been some attempts to eliminate local patchiness using *ad hoc* methods such as the so-called “min-max” [1] and “detrended fluctuation analysis” [4(c)] methods. The purpose of this Letter is to advocate the use of a new technique, the *wavelet transform* [7] (WT), which has proven to be well suited for characterizing the scaling properties of fractal objects even in the presence of low-frequency trends [8]. We will proceed to a statistical analysis of DNA sequences by applying the *wavelet transform modulus maxima* (WTMM) method [8(b)] that has been recently proposed as a generalized multifractal formalism for fractal functions.

The WT is a space-scale analysis which consists of expanding signals in terms of *wavelets* that are constructed from a single function, the *analyzing wavelet* ψ , by means of dilations and translations [7]. The WT of a function $s(x)$ is defined as

$$T_\psi(x_0, a) = \frac{1}{a} \int_{-\infty}^{+\infty} s(x) \psi\left(\frac{x - x_0}{a}\right) dx, \quad (1)$$

where x_0 is the space parameter and a (> 0) the scale parameter. The main advantage of using the WT for

analyzing the regularity of a function s is its ability to eliminate polynomial behavior by an appropriate choice of the wavelet ψ . Indeed, if s has, at the point x_0 , a local scaling (Hölder) exponent $h(x_0) \in]n, n + 1[$, in the sense that, around x_0 , $|s(x) - P_n(x)| \sim |x - x_0|^{h(x_0)}$, where $P_n(x)$ is some order- n polynomial, then one can easily prove [8] that $T_\psi(x_0, a) \sim a^{h(x_0)}$, provided the first $n + 1$ moments of ψ are zero [$\int x^m \psi(x) dx = 0$, $0 \leq m \leq n$]. Therefore the WT turns out to be a very powerful tool to detect and characterize singularities, even when they are masked by a smooth behavior. This property will be of fundamental importance to break free from the intrinsic patchiness of DNA sequences without using any *ad hoc* recipe. In this work, we will use the derivatives of the Gaussian function as analyzing wavelets: $\psi^{(N)} = d^N(e^{-x^2/2})/dx^N$ (the first N moments of $\psi^{(N)}$ are vanishing).

The WTMM method [8(b)] is a natural generalization of classical box-counting techniques. It consists of investigating the scaling behavior of some partition functions defined in terms of wavelet coefficients,

$$Z(q, a) = \sum_{l \in \mathcal{L}(a)} \left[\sup_{(x, a') \in l} |T_\psi(x, a')| \right]^q \sim a^{\tau(q)}, \quad (2)$$

where $q \in \mathbb{R}$. The sum is taken over the WT skeleton defined at each scale a by the local maxima of $|T_\psi(x, a)|$ considered as a function of x ; these WTMM are disposed on connected curves called *maxima lines*; the set $\mathcal{L}(a)$ of all maxima lines that exist at scale a indicates how to position the wavelets (“generalized oscillating boxes”) in order to obtain a partition at this scale. In the framework of this wavelet-based multifractal formalism, $\tau(q)$ is the Legendre transform of the *singularity spectrum* $D(h)$ defined as the Hausdorff dimension of the set of points x where the Hölder exponent is h . *Homogeneous fractal functions* (i.e., functions with a unique Hölder exponent h) are characterized by a linear $\tau(q)$ spectrum ($h = \partial\tau/\partial q$). On the contrary, a nonlinear $\tau(q)$ curve is the signature of *nonhomogeneous functions* that displays *multifractal* properties [i.e., $h(x)$ is a fluctuating quantity that depends upon

x]. For some specific values of q , $\tau(q)$ has a well-known meaning. For example, $-\tau(0)$ can be identified to the fractal dimension (capacity) of the set where s is not smooth; $\tau(1)$ is related to the capacity of the graph of the considered function; furthermore, $\tau(2)$ is related to the scaling exponent β of the spectral density: $S(f) = |\hat{s}(f)|^2 \sim f^{-\beta}$, with $\beta = 2 + \tau(2)$. Thus $\tau(2) \neq 0$ indicates the presence of long-range correlations. Previous studies of DNA sequences [1–6] have mainly focused on the estimate of the power-law exponents of the rms fluctuations of the DNA walk or the autocorrelation function, which are simply related to the power spectrum exponent β and thus to $\tau(2)$. For its ability to resolve multifractal scaling via the estimate of the entire $\tau(q)$ spectrum, the WTMM method is a definite step beyond the techniques used so far in the literature. Its reliability has been tested [8(b)] on various experimental and mathematical examples including fractional Brownian motions [9] (FBM). The FBM's $B_H(x)$ are Gaussian stochastic processes of zero mean with stationary increments, which are indexed by a parameter H ($0 < H < 1$) that accounts for the presence ($H \neq \frac{1}{2}$) or the absence ($H = \frac{1}{2}$) of correlations between increments. The FBM's are statistically homogeneous fractals characterized by a single Hölder exponent $h = H$ and thus by a $\tau(q)$ spectrum which is a linear function of slope H : $\tau(q) = qH - 1$.

As a first application of the WTMM method in a biological context, this study is devoted to the statistical analysis of 70 human DNA sequences, extracted from the EMBL data bank and long enough ($L \geq 6000$ nucleotides) to make the fractal analysis meaningful with respect to finite-size effects [4(b)]. We processed separately the entire genes, the coding (individual exons, cDNA's) and the noncoding (individual introns, flanks) regions, provided the length of these subsequences exceed 2000 nucleotides. To graphically portray these sequences, we followed the strategy originally proposed in [1], which consists of transforming them into random walks by defining an incremental variable that associates to the position i the value $\chi(i) = 1$ for purine (A,G) or -1 for pyrimidine (C,T). The graph of the DNA walk defined by the cumulative variable $s(x) = \sum_{i=1}^x \chi(i)$ is plotted in Fig. 1(a) for the human desmoplakin I cDNA. The patchiness of this DNA sequence is patent; one clearly recognizes three regions of different strand bias. Figure 1(b) shows the WT space-scale representation of this DNA signal when using order 1 analyzing wavelet $\psi^{(1)}$. This WT displays a treelike structure from large to small scales, that looks qualitatively similar to the fractal branching observed in the WT representation of Brownian or turbulent signals [8]. In Figs. 1(c) and 1(d) two horizontal cuts $T_{\psi^{(1)}}(x, a)$ are shown at two different scales $a = a_1 = 32$ and $a_2 = 512$ that are represented by the dashed lines in Fig. 1(b). When focusing the WT microscope at small scale $a = a_1$ in Fig. 1(c), since $\psi^{(1)}$ is orthogonal to constants, one filters the local (high frequency) fluctuations

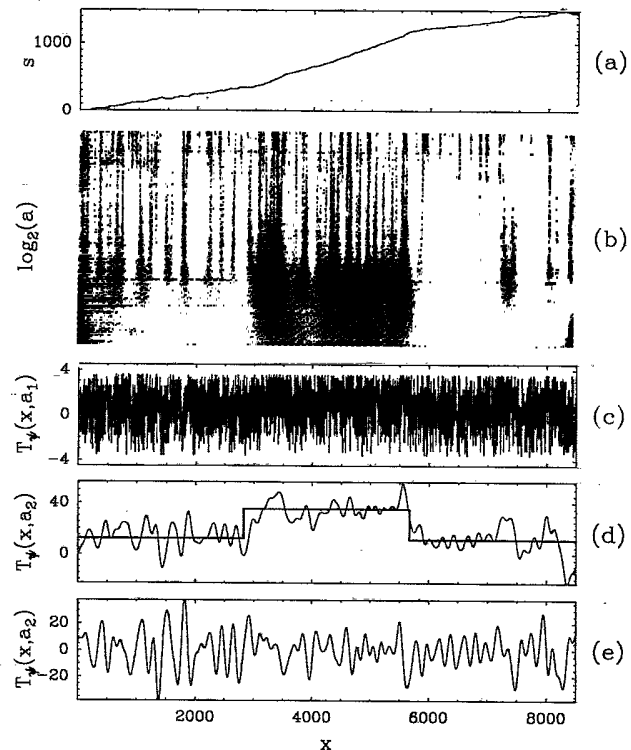


FIG. 1. WT analysis of the human desmoplakin I cDNA ($L = 8499$). (a) DNA walk displacement $s(x)$ (excess of purines over pyrimidines) vs nucleotide distances x . (b) WT of $s(x)$ computed with the analyzing wavelet $\psi^{(1)}$, $T_{\psi^{(1)}}(x, a)$ is coded, independently at each scale a , using 32 grey levels from white [$\min T_{\psi^{(1)}}(x, a)$] to black [$\max T_{\psi^{(1)}}(x, a)$]; small scales are at the top. (c) $T_{\psi^{(1)}}(x, a = a_1)$ vs x for $a_1 = 32$. (d) $T_{\psi^{(1)}}(x, a = a_2)$ vs x for $a_2 = 512$. (e) Same analysis as in (d) but with the analyzing wavelet $\psi^{(2)}$.

of $s(x)$, i.e., the fluctuations over a characteristic length of the order of $a_1 = 32$ nucleotides. When increasing the WT magnification in Fig. 1(d), one realizes that these fluctuations actually occur around three successive linear trends; $\psi^{(1)}$ not being blind to linear behavior, the WT coefficients fluctuate about nonzero constant behavior that correspond to the slopes of those linear trends. Even though this phenomenon is more pronounced when progressively increasing the scale parameter a , it is indeed present at all scales and drastically affects the fractal branching of the WT. In Fig. 1(e) at the same coarse scale $a = a_2$ as in Fig. 1(d), the fluctuations of the WT coefficients are shown as computed with the order-2 wavelet $\psi^{(2)}$. The WT microscope being now orthogonal also to linear behavior, the WT coefficients fluctuate about zero and one does not see the influence of the strand bias anymore. Furthermore, by considering successively $\psi^{(3)}, \psi^{(4)}, \dots$, one can hope to eliminate more complicated nonlinear trends, with the ultimate goal of filtering the fractal underlying structure that might be responsible for the presence of long-range correlations in DNA sequences [10].

In Figs. 2(a)–2(c) typical data are reported coming from the quantitative application of the WTMM method to the DNA walk graph corresponding to an intron-containing sequence ($L = 73\,326$) which has been widely studied in previous works [1,6(b),6(c)]: the human beta-globin intergenomic sequence (gene bank name HHUMHBB). First, let us mention that the patchy structure of this sequence is a little trickier than the one of the cDNA sequence in Fig. 1(a), in so far as it is not so easily amenable to *ad hoc* detrending methods such as the min-max procedure [1]. Figures 2(a) and 2(b) display plots of $\log_2 Z(q, a)$ vs $\log_2 a$ for some values of q ; two sets of data are represented corresponding to computations performed with $\psi^{(1)}$ (●) and $\psi^{(2)}$ (○), respectively. While the scaling behavior expected from Eq. (2) seems to operate over a wide range of scales when using $\psi^{(2)}$, it does not show up so clearly for the data obtained with $\psi^{(1)}$ for which some continuous (nonlinear) crossover phenomenon is observed as the signature of the breaking of the scale invariance by the extent patchiness of the DNA walk [6(b)]. The overall $\tau(q)$ spectrum obtained with $\psi^{(2)}$ using a linear regression fit of the data is compared in Fig. 2(c) to the corresponding spectrum (□) derived for the same sequence but after randomly shuffling the nucleotides. The data for both the true and the shuffled HHUMHBB sequences remarkably fall on straight lines which are the hallmark of homogeneous fractal functions. The corresponding unique Hölder exponents can be estimated by fitting the slope h of these straight lines; the value extracted for the original sequence $h = 0.60 \pm 0.02$ [$\tau(2) = 0.20 \pm 0.02$] is significantly different from the expected value $h = 0.50 \pm 0.02$ [$\tau(2) = 0.00 \pm 0.02$] obtained for the uncorrelated random shuffled sequence (the error bars have been estimated from the fluctuations of h

observed when splitting the original signals in samples of length $L \approx 2000$). We checked the reliability of this measurement by reproducing this WTMM analysis with higher order analyzing wavelets ($\psi^{(3)}, \psi^{(4)}$). Let us point out that the value $h = 0.60$ is clearly lower than previous estimates reported in the literature, e.g., $h = 0.71$ in [1] or $h = 0.67$ in [6(c)]. Note that if we had used the inappropriate analyzing wavelet $\psi^{(1)}$, we would have obtained a similar biased estimate $h = 0.70 \pm 0.03$. Therefore the WTMM method not only corroborates the existence of long-range correlations in this intron-containing sequence [$\tau(2) > 0$], but it also provides a very efficient methodology to correct some systematic overestimates reported in previous works. In Fig. 2(c) the $\tau(q)$ spectrum computed when analyzing (using $\psi^{(2)}$) the coding DNA sequence portrayed in Fig. 1(a) is also shown for comparison. The data (Δ) are almost indistinguishable from those previously obtained from the shuffled sequence; the $\tau(q)$ spectrum is linear with a slope $h = 0.49 \pm 0.02$, as expected for homogeneous fractal fluctuations that do not display long-range correlations.

Actually the linearity of the $\tau(q)$ spectrum turns out to be a general result for all the DNA walks that we considered. In Figs. 2(d)–2(f) the results of a systematic investigation of our statistical sample of 70 human genomic sequences are summarized under the form of histograms of Hölder exponent values. In order to test the robustness of our results with respect to the rule used to map the DNA sequences to DNA walks, we reproduced our original analysis based on purine (A, G) vs pyrimidine (C, T) distinction [Fig. 2(d)], for the two other possible choices of identifying two base pairs [Figs. 2(e) and 2(f)]. Whatever the association rule, the histograms obtained for individual introns are very similar to the histograms obtained for

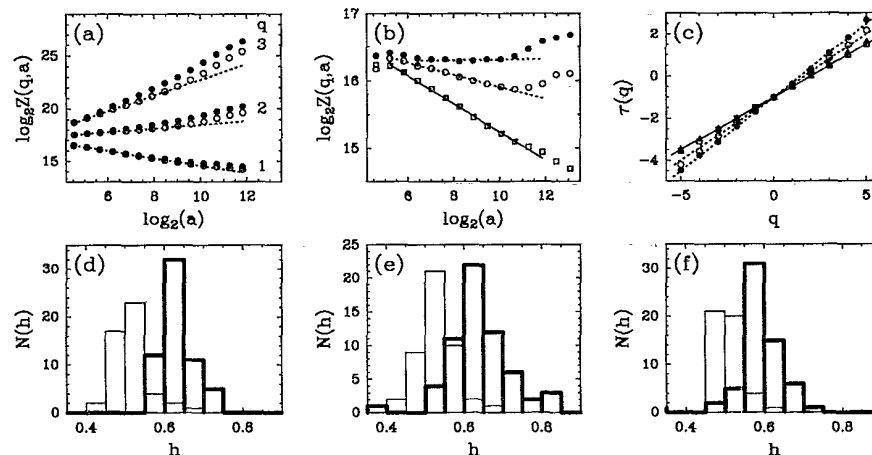


FIG. 2. WTMM analysis of protein coding and noncoding DNA sequences. (a) HHUMHBB: $\log_2 Z(q, a)$ vs $\log_2 a$ for different values of q ; the data correspond to the analyzing wavelets $\psi^{(1)}$ (●) and $\psi^{(2)}$ (○). (b) HHUMHBB: $\log_2 Z(q, a)$ vs $\log_2 a$ for $q = 1.5$; the symbols (●) and (○) have the same meaning as in (a); the symbol (□) corresponds to the data for the shuffled sequence (see text) when using $\psi^{(2)}$; the dashed and solid lines represent the corresponding least-squares fit estimates of $\tau(q)$. (c) $\tau(q)$ vs q for the HHUMHBB sequence when using $\psi^{(1)}$ (●), the HHUMHBB sequence (○), the shuffled HHUMHBB sequence (□), and the human desmoplakin I cDNA (Δ) when using $\psi^{(2)}$. Histograms of Hölder exponent values extracted from the WTMM analysis (with $\psi^{(2)}$) of 70 human DNA sequences, (—) cDNA's and (—) intron-containing sequences, the DNA walk graphs are constructed on the basis of the following identifications: (A, G) vs (C, T) in (d), (C, G) vs (A, T) in (e), and (A, C) vs (G, T) in (f).

the entire genes; they display a rather pronounced peak for $h \approx 0.63$. Despite some slight overlap, the histograms obtained for individual exons and cDNA are systematically shifted towards lower h values and markedly peaked at $h \approx 0.50$. These results are in qualitative agreement with the conclusions of [1,4,5] and suggest that the experimental finding of long-range correlations in the human noncoding sequences is very likely to be a general characteristic feature of nucleotide organization in DNA.

Let us stress that the $\tau(q)$ spectra extracted from both coding and noncoding DNA walks are remarkably well fitted by the theoretical spectrum for FBM's [$\tau(q) = qH - 1$]. Within that prospect, we studied the probability distribution function of wavelet coefficient values $P(T_{\psi^{(2)}}(., a))$, as computed at a fixed scale a in the scaling range. The distributions obtained for both the coding DNA sequence of Fig. 1(a) and the largest intron ($L = 33\,895$) contained in the human retinoblastoma susceptibility gene are shown in Figs. 3(a) and 3(b), respectively. When plotting $\ln P$ vs $T_{\psi}/\sigma(a)$, where $\sigma(a)$ is the rms value at scale a , all the data points computed at different scales fall on the same parabola independently of the nature of the sequence. Thus the fluctuations in the DNA walks are likely to have Gaussian statistics. The presence of long-range correlations in the intron sequence is in fact contained in the scale dependence of the rms $\sigma(a) \sim a^h$, where $h = 0.60 \pm 0.02$ as compared to the uncorrelated random walk value $h = 0.50 \pm 0.02$ obtained for the coding sequence.

To conclude, we have emphasized the WT as a very powerful and reliable tool to characterize the fractal scaling organization of DNA sequences. Our main experimental finding is that the fluctuations in the patchy DNA walks are homogeneous with Gaussian statistics. Those for noncoding and intron-containing DNA sequences display long-range correlations and are well modeled by FBM's with $H > \frac{1}{2}$. In contrast, those for coding sequences cannot be distinguished from classical (uncorrelated steps) Brownian motions. Actually, much more can be learned from the WTMM analysis, especially as far as the nature of the overall superimposed patchy structure of the DNA sequences

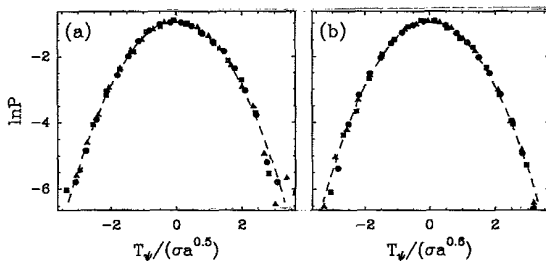


FIG. 3. Probability distribution functions of wavelet coefficient values at fixed scale $a = 32$ (●), 64 (▲), and 128 (■); the analyzing wavelet is $\psi^{(2)}$. $\log_2 P$ is plotted vs $T/\sigma(a)$, where $\sigma(a) = \sigma a^h$ is the rms value. (a) Human desmoplakin I cDNA sequence: $h = 0.5$. (b) Largest intron in the human retinoblastoma susceptibility gene: $h = 0.6$. The dashed lines in (a) and (b) are parabolas characteristic of Gaussian statistics.

is concerned. For instance, very instructive information is contained in the way the scaling range [Figs. 2(a) and 2(b)] depends upon the shape of the analyzing wavelet and the value of q , whether there exists some finite characteristic scale above which either the scaling is broken or some crossover to a different scaling regime is observed. This information is likely to provide decisive tests for the validity of various models proposed to account for the long-range correlations in DNA sequences. On a more fundamental ground, further application of the WTMM analysis to DNA sequences of different evolutionary categories [3] looks promising for future understanding of the role played by introns, repetitive motifs, and noncoding intergenic regions in the nonequilibrium dynamical process [1,2] that produced nucleic acid sequences.

We are very grateful to A. Hénaut and A. Kuhn for helpful discussions. This work was supported by the GIP GREG (project "Motifs dans les Séquences").

- [1] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, *Nature (London)* **356**, 168 (1992).
- [2] W. Li and K. Kaneko, *Europhys. Lett.* **17**, 655 (1992); *Nature (London)* **360**, 635 (1992).
- [3] R. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992); **71**, 1777 (1993); *Fractals* **2**, 1 (1994).
- [4] (a) S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, M. Simons, F. Sciortino, and H.E. Stanley, *Phys. Rev. Lett.* **71**, 1776 (1993); (b) C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, M. Simons, and H.E. Stanley, *Phys. Rev. E* **47**, 3730 (1993); (c) C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Goldberger, *Phys. Rev. E* **49**, 1685 (1994).
- [5] V.V. Prabhu and J.M. Claverie, *Nature (London)* **357**, 782 (1992); P.J. Munson, R.C. Taylor, and G.S. Michaels, *ibid.* **360**, 636 (1992).
- [6] (a) S. Nee, *Nature (London)* **357**, 450 (1992); (b) S. Karlin and V. Brendel, *Science* **259**, 677 (1993); (c) D. Larhammar and C.A. Chatzidimitriou-Dreismann, *Nucleic Acids Res.* **21**, 5167 (1993).
- [7] (a) *Wavelets*, edited by J.M. Combes, A. Grossmann, and P. Tchamitchian (Springer-Verlag, Berlin, 1989); (b) *Wavelets and Applications*, edited by Y. Meyer (Springer-Verlag, Berlin, 1992); (c) *Progress in Wavelet Analysis and Applications*, edited by Y. Meyer and S. Roques (Frontières, Gif-sur-Yvette, 1993).
- [8] (a) A. Arneodo, G. Grasseau, and M. Holschneider, *Phys. Rev. Lett.* **61**, 2281 (1988); in *Wavelets* (Ref. [7(a)]), p. 182; (b) J.F. Muzy, E. Bacry, and A. Arneodo, *Phys. Rev. Lett.* **67**, 3515 (1991); *Int. J. Bifurcation Chaos* **4**, 245 (1994).
- [9] B.B. Mandelbrot and J. Van Ness, *S.I.A.M. Rev.* **10**, 422 (1968).
- [10] Note that when using the analyzing wavelets $\psi^{(N)}$ with $N \geq 2$, the WT does not depend, up to a global multiplicative factor, upon the choice of the values of the DNA walk elementary steps $\chi(i)$.