Long-Range Correlations in Genomic DNA: A Signature of the Nucleosomal Structure

B. Audit,[1,*] C. Thermes,[2] C. Vaillant,[1] Y. d'Aubenton-Carafa,[2] J. F. Muzy,[1] and A. Arneodo[1]

[1]Centre de Recherche Paul Pascal, avenue Schweitzer, 33600 Pessac, France

[2]Centre de Génétique Moléculaire du CNRS, Laboratoire associé à l'Université Pierre et Marie Curie,
Allée de la Terrasse, 91198 Gif-sur-Yvette, France

We use the "wavelet transform microscope" to carry out a comparative statistical analysis of DNA bending profiles and of the corresponding DNA texts. In the three kingdoms, one reveals on both signals a characteristic scale of 100–200 bp that separates two different regimes of power-law correlations (PLC). In the small-scale regime, PLC are observed in eukaryotic, in double-strand DNA viral, and in archaeal genomes, which contrasts with their total absence in the genomes of eubacteria and their viruses. This strongly suggests that small-scale PLC are related to the mechanisms underlying the wrapping of DNA in the nucleosomal structure. We further speculate that the large scale PLC are the signature of the higher-order structure and dynamics of chromatin.

The availability of fully sequenced genomes offers the possibility to study the scale-invariance properties of DNA sequences on a wide range of scales extending from tens to thousands of nucleotides. Actually, scale invariance measurement enables us to evidence particular correlation structures between distant nucleotides or groups of nucleotides. 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 [1,2]. If it is now well admitted that long-range correlations do exist in genomic sequences [3,4], their biological interpretation is still a continuing debate [1,5,6]. Most of the models proposed so far are based on the genome plasticity and are supported by the reported absence of power-law correlations (PLC) in coding DNA sequences [3,4,7]. Recently, from a systematic analysis of human exons, coding sequences (CDS), and introns, we have found that PLC are not only present in noncoding sequences but also in coding regions somehow hidden in their inner codon structure [6,8]. The aim of the present study is to demonstrate that the long-range correlations observed in DNA sequences are more likely the signature of the hierarchical structural organization of chromatin. In contrast to previous interpretations, we thus propose some understanding of these correlations as a necessity for chromosome packaging.

A major problem of fractal analysis applied to DNA sequences is that these display a mosaic structure which is characterized by "patches" resulting from compositional biases with an excess of one type of nucleotide [1,3,7]. When mapping DNA sequences to numerical sequences using the DNA walk representation, these patches appear as trends in the DNA walk landscapes that are likely to break the scale invariance [1–3,7]. In previous works [4], we have emphasized the wavelet transform (WT) as a well suited technique to overcome this difficulty. By considering analyzing wavelets that make the "WT microscope" blind to low frequency trends, any bias in the DNA walk

can be removed and the existence of PLC associated with specific scale invariance properties can be revealed accurately. When exploring sequences selected from the human genome, we have found that the fluctuations in the patchy landscapes of both coding and noncoding DNA walks are monofractal with Gaussian statistics in the small-scale range, which justifies the use of a single exponent H usually called the Hurst or roughness exponent [4]. H values larger than the uncorrelated random walk value $H = 1/2$ correspond to the existence of long-range correlations that we refer to as "persistence." To estimate this exponent, we just have to investigate the behavior across scales (a) of the root mean square (rms) fluctuations of wavelet coefficients: $\sigma(a) \propto a^H$. Wavelet coefficients actually reflect the local variation (over size a) of the concentration of nucleotides. Persistence ($H > 1/2$) therefore means that these concentrations fluctuate more smoothly (over short distances) than for uncorrelated sequences, but in the same time with a larger amplitude (over large distances) around the mean value.

Here we carry out a comparative analysis of the persistence properties for both DNA texts and DNA bending profiles of various eukaryotic, eubacterial, and archaeal genomes [8]. To study the DNA texts, we construct "DNA walks" according to the binary coding method extensively used by Voss [9]; this method decomposes the nucleotide sequence into four sequences corresponding to A, C, T, or G (coding with 1 at the nucleotide position and 0 at other positions). To construct bending profiles that account for the fluctuations of the local DNA curvature, we use the trinucleotide model proposed in Ref. [10] (here called Pnuc) and which was deduced from experimentally determined nucleosome positioning [11]. The first completely sequenced eukaryotic genome *Saccharomyces cerevisiae* (*S.c.*) provides an opportunity to perform a comparative wavelet analysis of the scaling properties displayed by each chromosome. When looking at the global estimate

of $\sigma(a)$ over the DNA walks corresponding to "A" in each of the 16 yeast chromosomes [12] shown in Fig. 1(a), one sees that all present superimposable behavior, with notably the same characteristic scale that separates two dif-

ferent scaling regimes. At small scales, $20 \lesssim a \lesssim 200$ (expressed in nucleotide units), PLC are observed as characterized by $H = 0.59 \pm 0.02$, a mean value which is significantly larger than $1/2$. At large scales, $200 \lesssim a \lesssim 5000$, stronger PLC with $H = 0.82 \pm 0.01$ become dominant with a cutoff around 10 000 bp (a number by no means accurate) above which uncorrelated behavior is observed. A similar wavelet analysis of the bending profiles of the yeast chromosomes [Fig. 1(a)] reveals striking similarities with the curves resulting from the DNA walk analysis, in both the small-scale and the large-scale regimes. These observations are not simply due to a "recoding" of the DNA sequences since when using the DNase table of curvature based on sensitivity of DNA fragments to DNase [13], one notices a significant weakening of the H exponent observed in the large-scale regime ($H \simeq 0.6$). The existence of these two scaling regimes is confirmed in Fig. 2(a), where the probability density functions (pdfs) of wavelet coefficient values of the yeast bending profiles computed at different scales are shown to collapse on a single curve, as predicted by the self-similarity relationship [4] $a^H \rho_a(a^H T) = \rho(T)$, provided one uses the scaling exponent value $H = 0.60$ in the scale range $10 \lesssim a \lesssim 100$ and $H = 0.75$ in the scale range $200 \lesssim a \lesssim 1000$. In the small-scale regime, the pdfs are very well approximated by Gaussian distributions. In the large-scale regime, the pdfs have stretched exponential-like tails. The fact that the self-similarity relationship is satisfied in both regimes corroborates the monofractal nature of the roughness fluctuations of the yeast bending profiles. Similar quantitative results are obtained for the corresponding DNA texts [8]. Let us emphasize that we have also examined a number of eukaryotic DNA sequences from different organisms (human, rodent, avian, plant, and insect) and that we have observed the same characteristic features as those obtained in Fig. 1(a) for *S.c.* [see Fig. 1(c) for the human
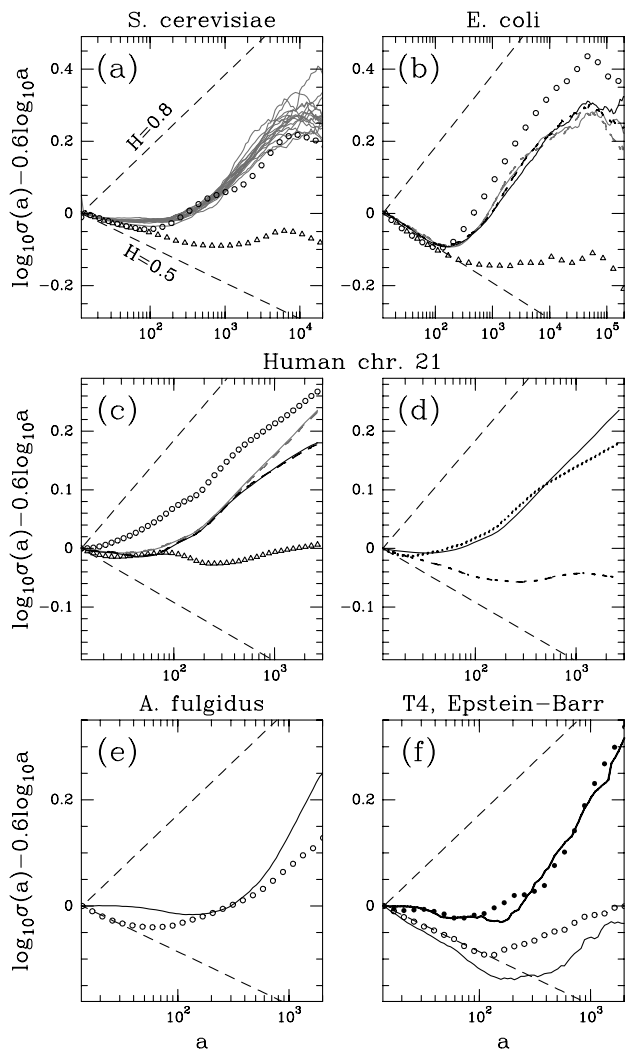

FIG. 1. Global estimate of the rms of WT coefficients: $\log_{10}\sigma(a) - 0.6\log_{10}a$ is plotted versus $\log_{10}a$; the dashed lines corresponding to uncorrelated ($H = 1/2$) and strongly correlated ($H = 0.80$) regimes are drawn to guide the eyes. (Some horizontal line in this logarithmic representation will correspond to $H = 0.6$.) The analyzing wavelet is the "Mexican hat" wavelet [4]. (a) *S. cerevisiae:* "A" DNA walks of the 16 *S. cerevisiae* chromosomes (—) and of the corresponding bending profiles obtained with the Pnuc ($\bigcirc$) and DNase ($\triangle$) coding tables when averaged over the 16 chromosomes. (b) *Escherichia coli:* "A" (grey —), "T" (grey – – –), "G" (black —), and "C" (black – – –) DNA walks and the corresponding Pnuc ($\bigcirc$) and DNase ($\triangle$) bending profiles. (c) *Human chromosome 21:* "A," "C," "G," and "T" DNA walks and corresponding Pnuc and DNase bending profiles; same symbols as in (b). (d) *Human chromosome 21:* comparative analysis of DNA walks for all adenines (—), adenines part of a dinucleotide *AA* ($\cdots$), and isolated adenines not part of a dinucleotide *AA* (– – –). (e) *Archaeglobus fulgidus:* "G" DNA walk (—) and Pnuc bending profile ($\bigcirc$). (f) *Epstein-Barr virus* (black) and *T4 bacteriophage* (grey) "G" DNA walks (solid line) and Pnuc bending profiles (circle).
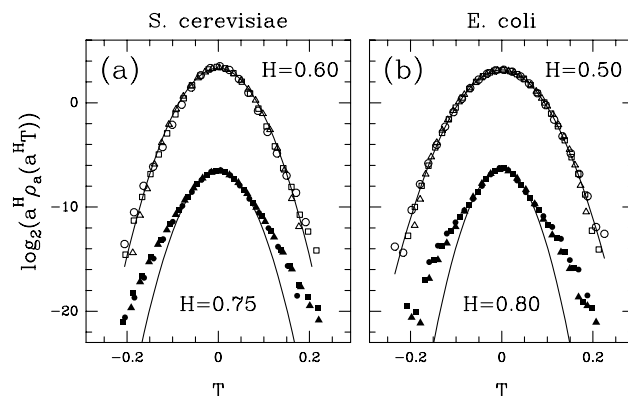

FIG. 2. Probability distribution functions of wavelet coefficient values of Pnuc bending profiles. The analyzing wavelet is the Mexican hat [4]. (a) *Saccharomyces cerevisiae:* $\log_2[a^H \rho_a(a^H T)]$ vs T for the set of scales $a = 12$ ($\triangle$), 24 ($\square$), 48 ($\bigcirc$), 192 ($\blacktriangle$), 384 ($\blacksquare$), and 768 ($\bullet$) in nucleotide units; $H = 0.60$ ($H = 0.75$) in the small (large) scale regime. (b) *Escherichia coli:* same as in (a) but with $H = 0.50$ ($H = 0.80$) in the small (large) scale regime.

chromosome 21]. Note that the crossover between the two PLC regimes is remarkably robust for the four A, C, G, and T DNA walks.

The striking overall similarity of the results obtained with these different eukaryotic genomes prompted us to also examine the scale invariance properties of bacterial genomes. In Fig. 1(b) are reported the results obtained for *Escherichia coli* which are quite typical of what we have observed with other eubacterial genomes (data not shown). Again, there exists a well defined characteristic scale $a^* \simeq 200$ bp that delimits the transition to very strong PLC with $H = 0.80 \pm 0.05$ at large scales. Let us point out that as for *S.c.* [Fig. 1(a)], if one uses the DNase table for human [Fig. 1(c)] and *E. coli* [Fig. 1(b)] sequences, one no longer observes the strong PLC as obtained with the Pnuc table. In Fig. 2(b) are reported the wavelet coefficient pdfs of the *E. coli* Pnuc bending profile that corroborate the existence of a crossover scale between two different monofractal scaling regimes characterized by $H = 0.50 \pm 0.02$ and $H = 0.80 \pm 0.05$, respectively. In order to examine if these properties actually extend homogeneously over the whole genomes, $\sigma(a)$ was calculated over a window of width $l = 2000$, sliding along the bending profiles. The results reported in Fig. 3 for both yeast and *E. coli* confirm the existence of a characteristic scale $a^* \simeq 100-200$ bp which seems to be robust all along the corresponding DNA molecules and this for all investigated genomes in the three kingdoms. Note that analogous results are obtained for the four mononucleotide DNA walks as for the bending profiles [8]. Let us mention that a similar characteristic scale has been previously obtained on the patchiness structure of DNA walks [14].
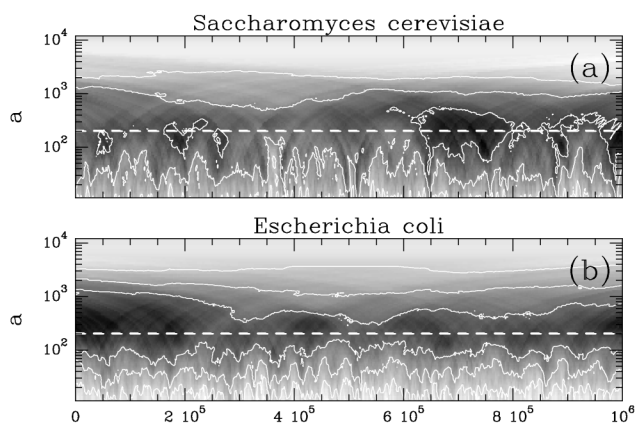

FIG. 3.   Space-scale waveletlike representation (x and a are expressed in nucleotide units) of the local estimate of the rms $\sigma(a, x)$ of the WT coefficients of the A DNA walk. $\sigma(a)$ is computed over a window of width $l = 2000$, sliding along the first 10^6 bp of the *yeast* chromosome IV (a), and of the *Escherichia coli* genome (b). $\log_{10}\sigma(a) - 2/3\log_{10}a$ is coded using 128 grey levels from black (min) to white (max). The horizontal white dashed line marks the scale $a^* = 200$ bp where some minimum is observed consistently along the entire genomes as a separation between two different monofractal scaling regimes (see text).

There exists however an important difference between eukaryotic and eubacterial genomes: no PLC are observed for the latter in the small-scale regime where uncorrelated Brownian motionlike behavior with $H = 1/2$ is observed [Figs. 1(b) and 2(b)]. As discussed in previous works [1,3–5,7], separate analyses of coding and noncoding eukaryotic DNA walks actually show that introns display PLC (with a mean H value of 0.60 ± 0.02) in the small-scale regime, while exons have no such correlations. At this point, it may seem that PLC are inherent to noncoding sequences only, but that is not the case. As shown in Fig. 1(e) for *Archaeoglobus fulgidus*, the wavelet investigation of five archaeal genomes (which are mostly coding) also reveals the presence of small-scale PLC as observed in eukaryotic genomes, although somewhat less pronounced. Note that the strong large-scale PLC are present in all eubacterial, archaeabacterial, and eukaryotic genomes (data not shown) [8,9].

What mechanism or phenomenon might explain the small-scale PLC in eukaryotic genomes? Their total absence in eubacterial genomes raises the possibility that they could be related to certain nucleotide arrangements in the 150 bp long DNA regions which are wrapped around histone proteins to form the eukaryotic nucleosome [15]. Indeed, eubacterial genomic DNA is associated with histonelike proteins (e.g., HU), but no nucleosome-type structure has been detected in these organisms [16]. Along this line, the observation of small-scale PLC in archaeal genomes is consistent with the presence in archaebacteria of structures similar to the eukaryotic nucleosomes [17]. This analysis has also been extended to viral genomes. Small-scale PLC are clearly detected in most eukaryotic viral double-strand DNA genomes as shown for the *Epstein-Barr* virus in Fig. 1(f). This further supports the hypothesis of nucleosome-based PLC since nucleosomes are present on double-strand DNA viruses [18]. Finally, bacteriophage genomes do not present any small-scale PLC [Fig. 1(f) for *T4* bacteriophage and data not shown] as already observed for their eubacterial hosts. This wavelet based fractal analysis of viral and cellular genomes of all three kingdoms sustains the fact that small-scale PLC are a signature of nucleosomal DNA.

To further investigate this PLC nucleosomal diagnostic, we ask whether particular dinucleotides which are known to participate in the positioning and formation of nucleosomes [19] (e.g., AA dinucleotides) would carry PLC specifically associated to eukaryotic genomes. This can be examined if one performs the analysis of different DNA walks generated with (i) all adenines, (ii) only adenines that are part of a dinucleotide AA, and (iii) isolated adenines that are not part of a dinucleotide AA. The analysis of human sequences [Fig. 1(d)] shows that the "isolated A" DNA walk exhibits a clear weakening of the PLC properties at small scale, while the AA DNA walk accounts for a major part of the observed PLC on the A DNA walk, which confirms the nucleosomal signature of small-scale PLC [20].

Several studies have established the presence in genomic sequences of repeated DNA motifs related to bending properties. A 10.2 base periodicity has been observed using either Fourier [21] or correlation function [22] analysis, specifically in eukaryotic genomes where it has been interpreted in relation to nucleosomal structures. However, there is a fundamental difference between this nucleosome diagnostic based on periodicity and our analysis based on scale invariance properties [23] which strongly suggests that the mechanisms underlying the nucleosomal structure of eukaryotic genomes are multiscale phenomena that actually involve the whole set of scales in the 1–200 bp range. In this respect, periodicity and scale invariance should not be considered as opposed to each other but rather complementary. The understanding of small-scale PLC should thus provide further insight into the nucleosome structure and dynamics.

The interpretation of the large scale PLC observed in the DNA bending profiles as well as in the DNA walks for all organisms is an open problem [8]. Actually, all chromosomes are submitted to condensation-decondensation processes (in relation with DNA replication, gene expression, ...) which might result in common dynamical and structural properties [24]. A deep understanding of the large-scale PLC and their interpretation in terms of these constraints remain challenging questions requiring further investigation.

---

*Present address: Computational Genomics Group, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.

[1] H. Stanley *et al.,* Fractals **1**, 283 (1993); W. Li, T. G. Marr, and K. Kaneko, Physica (Amsterdam) **75D**, 392 (1994); R. F. Voss, Fractals **2**, 1 (1994).

[2] L. Larhammar and C. A. Chazidimitriou-Dreismann, Nucleic Acids Res. **21**, 5167 (1993); S. Karlin and V. Brendel, Science **259**, 677 (1993); B. Borstnik, D. Pumpernik, and D. Lukman, Europhys. Lett. **23**, 389 (1993).

[3] S. V. Buldyrev *et al.,* Phys. Rev. E **51**, 5084 (1995).

[4] A. Arneodo *et al.,* Phys. Rev. Lett. **74**, 3293 (1995); A. Arneodo *et al.,* Physica (Amsterdam) **96D**, 291 (1996).

[5] W. Li, Int. J. Bifurcation Chaos **2**, 137 (1992); S. V. Buldyrev *et al.,* Phys. Rev. E **47**, 4514 (1993); Biophys. J. **65**, 2673 (1993); H. P. Herzel *et al.,* Physica (Amsterdam) **249A**, 449 (1998).

[6] A. Arneodo *et al.,* Eur. Phys. J. B **1**, 259 (1998).

[7] C.-K. Peng *et al.,* Nature (London) **356**, 168 (1992).

[8] B. Audit, Ph.D. thesis, Université de Paris VI, 1999.

[9] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[10] D. S. Goodsell and R. E. Dickerson, Nucleic Acids Res. **22**, 5497 (1994).

[11] As pointed out by H. R. Drew and A. A. Travers, J. Mol. Biol. **186**, 773 (1985); J. Yao, P. T. Lowary, and J. Widom, Proc. Natl. Acad. Sci. U.S.A. **87**, 7603 (1990), the bending properties of DNA play an essential role in the compaction processes underlying the higher-order structure of chromatine.

[12] W. Li *et al.,* Genome Res. **8**, 916 (1998).

[13] I. Brukner *et al.,* J. Biomol. Struct. Dyn. **13**, 309 (1995).

[14] G. M. Viswanathan *et al.,* Biophys. J. **72**, 866 (1997).

[15] T. J. Richmond *et al.,* Nature (London) **311**, 532 (1984); J. Widom and A. Klug, Cell **43**, 207 (1985); Y. Saitoh and U. K. Laemmli, Cold Spring Harbor Symp. Quant. Biol. **58**, 755 (1993).

[16] L. D. Murphy and S. B. Zimmerman, J. Struct. Biol. **119**, 336 (1997).

[17] J. N. Reeve, K. Sandman, and C. J. Daniels, Cell **89**, 999 (1997).

[18] M. Coca-Prados, H. Y. Yu, and M. T. Hsu, J. Virol. **44**, 603 (1982).

[19] A. Thaström *et al.,* J. Mol. Biol. **288**, 213 (1999).

[20] Note that this observation is an additional illustration of the fact that recoding does not trivially conserve the correlation law.

[21] J. Widom, J. Mol. Biol. **259**, 579 (1996).

[22] H. Herzel, O. Weiss, and E. N. Trifonov, Bioinformatics **15**, 187 (1999).

[23] Note that persistence ($H > 1/2$) shows up as a power-law behavior of the envelope of the correlation function.

[24] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Europhys. Lett. **23**, 373 (1993).