Journal of Biological Physics **30:** 33–81, 2004.

Wavelet Analysis of DNA Bending Profiles reveals Structural Constraints on the Evolution of Genomic Sequences

BENJAMIN AUDIT[1,3], CÉDRIC VAILLANT[1], ALAIN ARNÉODO[1,*], YVES D'AUBENTON-CARAFA[2] and CLAUDE THERMES[2]

[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*

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

(**Author for correspondence, e-mail: alain.arneodo@ens-lyon.fr*)

Abstract. Analyses of genomic DNA sequences have shown in previous works that base pairs are correlated at large distances with scale-invariant statistical properties. We show in the present study that these correlations between nucleotides (letters) result in fact from long-range correlations (LRC) between sequence-dependent DNA structural elements (words) involved in the packaging of DNA in chromatin. Using the wavelet transform technique, we perform a comparative analysis of the DNA text and of the corresponding bending profiles generated with curvature tables based on nucleosome positioning data. This exploration through the optics of the so-called 'wavelet transform microscope' reveals a characteristic scale of $100-200$ bp that separates two regimes of different LRC. We focus here on the existence of LRC in the small-scale regime ($\lesssim$ 200 bp). Analysis of genomes in the three kingdoms reveals that this regime is specifically associated to the presence of nucleosomes. Indeed, small scale LRC are observed in eukaryotic genomes and to a less extent in archaeal genomes, in contrast with their absence in eubacterial genomes. Similarly, this regime is observed in eukaryotic but not in bacterial viral DNA genomes. There is one exception for genomes of Poxviruses, the only animal DNA viruses that do not replicate in the cell nucleus and do not present small scale LRC. Furthermore, no small scale LRC are detected in the genomes of all examined RNA viruses, with one exception in the case of retroviruses. Altogether, these results strongly suggest that small-scale LRC are a signature of the nucleosomal structure. Finally, we discuss possible interpretations of these small-scale LRC in terms of the mechanisms that govern the positioning, the stability and the dynamics of the nucleosomes along the DNA chain. This paper is maily devoted to a pedagogical presentation of the theoretical concepts and physical methods which are well suited to perform a statistical analysis of genomic sequences. We review the results obtained with the so-called wavelet-based multifractal analysis when investigating the DNA sequences of various organisms in the three kingdoms. Some of these results have been announced in B. Audit et al. [1, 2].

Key words: chromatin, DNA bending profile, fractals, genomic DNA sequence, long-range correlations, nucleosome, scale-invariance, wavelet transform

1. Introduction

The relation between the primary structure of DNA and its biological function is one of the outstanding problems in modern biology. There are many objective reasons to believe that the functional role of DNA sequences is not only to code for proteins but also to control the spatial structure of DNA in chromatin. In eukaryotic cells, DNA is severely compacted when it folds into chromosomes. The elementary structural unit of chromatin is the nucleosome [3, 4, 5, 6], which consists of a histone protein core enveloped by DNA. At first level of organization, the chromatin fiber is built from a linear array of nucleosomes [7, 8]. In the nucleus, this fiber is further packed into a higher order structure known as the 30 nm fiber [7, 9, 10, 11, 12, 13, 14]. This hierarchical folding of the DNA molecule is likely to imply constraints on the molecule bending and flexibility properties and on the capability of interacting with and of being anchored to protein matrix and scaffold. So far, such constraints have been evidenced to favour the formation and positioning of nucleosomes [15, 16, 17, 18, 19]. These structural properties depend upon the local nucleotides composition and therefore can be seen as statistical features of the DNA primary structure [20, 21, 22, 23, 24, 25, 26]. The actual challenge is thus to find a way to extract these structural informations from an appropriate reading of the DNA text. Since the different orders of packaging in the hierarchical structural organization of DNA are implicated in the accessibility of DNA sequence elements to trans-acting factors that control the processes of transcription and replication [27, 28, 29, 30, 31, 32, 33, 34], there is actually a wealth of structural and dynamical informations to learn in the primary DNA sequence about how DNA works in a living cell.

At first glance, the primary DNA sequences look rather random in the sense that they do not exhibit obvious regular features except some particular patterns like for instance tandem repetitions. Besides the existence of those repeated segments (for reviews see [35, 36]), the major part of the sequences seems hardly distinguishable from uncorrelated or 'Markov like' short-range correlated random sequences. With the specific goal to identify periodic repetitions as well as possible hidden periodicities, Fourier and correlation function techniques have been extensively used to process eukaryotic, eubacterial and archaeal genomes [23, 25, 37, 38, 39, 40]. Several oscillating patterns have been detected mainly when investigating the distributions of di- and tri-nucleotides. A 3-base periodicity is actually observed in both prokaryotic and eukaryotic sequences reflecting the existence of strings of codons in protein-coding regions [20, 23, 38]. Another well studied periodicity is 10-11 bp oscillations that show up in all three kingdoms. Several interpretations have been raised concerning the origin of the observed periodicities since significantly different periodicities have been identified close to the equilibrium helical repeat for free DNA of 10.55 bp. The 10.8-11 bp period identified in coding eukaryotic sequences might be the signature of encoded proteins [25, 41]. Indeed the alternation of hydrophobic and hydrophylic amino acids in α-helices leads to a periodicity of about 3.5 amino acids in protein sequences [42]. There exists another rather well identified periodicity of 10.2-10.4 bp that is unique to the eukaryotic genomes. This periodicity is likely to be the consequence of the wrapping of DNA around the histone octamer [20, 23, 25, 26, 37, 38, 43, 44, 45, 46, 47, 48], since a slight but significant decrease of the helical repeat has been observed experimentally in the nucleosome where the weight average of 8 independent measurements yields a periodicity of 10.39 ± 0.02 bp/turn (see Table II in [49]). This positive supercoiling is mainly seen in the distribution of some dinucleotides such as AA(= TT), GC, and GG(= CC) to some lower extent. Most of these dinucleotides are known to contribute to the intrinsic bending and flexibility properties of the DNA double helix [15, 21, 50, 51, 52]. The periodic positioning of these dinucleotides with a rather definite phase shift between them (e.g., about 5 bases between AA and TT as well as in between AA and GC) contributes in a coherent manner to a global curvature of DNA which is likely to amplify the affinity for the histone octamer and therefore to favour the wrapping of DNA on the histone surface [6, 18]. Larger periodicities are further observed that may also be related to the hierarchical organization of DNA via successive foldings of higher order structured nucleoprotein complexes [24]. Actually, the 200- and 400-base periodicities identified in eukaryotic sequences might correspond to the characteristic sizes of a nucleosome or of a dinucleosome [38]. Other periodicities have been related to the segmented structure of protein-coding sequences or DNA mobile elements [49, 53], or to the periodic distribution of transcription factor sites [39].

In prokaryotes, a periodicity about 10.8-11 bp/turn has been observed that is significantly above the equilibrium helical period [23, 25, 38]. It has been related to unconstrained negative supercoiling which is essential to a number of processes like DNA transcription, replication and condensation [54, 55, 56]. On the contrary to the negative supercoiling revealed in eubacterial genomes, archaeal plasmids were found to be positively coiled [57]. However, the recent discovery and studies of histone proteins in various archaea indicates that the binding of DNA to these histones introduces some toroidal overwinding of DNA to form nucleosome structures [40, 58, 59, 60].

Besides the investigation of hidden periodicities that emerge as statistically significant as compared to the noisy background in the high frequency domain, Fourier transform analysis and correlation function techniques have been also applied to investigate the scale-invariance properties of DNA sequences over a wide range of scales extending from tens to thousand of nucleotides [61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71]. The possible relevance of scale-invariance and fractal concepts to the structural complexity of genomic sequences has been the subject of increasing interest [72, 73, 74, 75, 76]. Scale-invariance measurement enables us to evidence particular long-range correlations (LRC) between distant nucleotides or group of nucleotides that may or may not display hidden periodicities. During the past ten years, there has been intense discussion about the existence, the nature and the origin of LRC in DNA sequences. Besides Fourier and autocorrelation analysis, different techniques including mutual information functions [61, 71, 77, 78], DNA walk representation [64, 75, 79, 80, 81, 82, 83, 84], Zipf analysis [85, 86, 87] and entropies [88, 89, 90, 91] were used for statistical analysis of DNA sequences. A lot of effort has been spent to adress rather struggling questions. In particular, it was of fundamental importance to corroborate the fact that the reported LRC really meant a lack of independence at long distances and were not just an artefact of the compositional heterogeneity of the genome organization [68, 70, 71, 79, 82, 83, 84, 92]. Furthermore, since most of the models proposed in the literature are based on the genome plasticity [61, 75, 93, 94, 95, 96, 97, 98], a rather crucial issue which is still debated is the fact that long-range correlation properties might be different for protein-coding (coding exons) and non-coding (introns, intergene) sequences [61, 62, 63, 64, 65, 66, 67, 69, 70, 71, 75, 80, 81, 83, 85, 86, 92, 99].

Actually, there were many objective reasons for this somehow controversial situation. Most of the investigations of LRC in DNA sequences were performed using different techniques that all had their own advantages and limitations. They all consisted in measuring power-law behavior of some characteristic quantity, e.g., the fractal dimension of the DNA walk, the scaling exponent of the correlation function or the power-law exponent of the power spectrum. Therefore, in practice, they all faced the same difficulties, namely finite size effects due to the finiteness of the sequence [96, 100, 101]. Actually, in these conditions, the definition of the scaling range can be a rather delicate task which may strongly affect the estimate of the scaling exponent. Moreover, some precautions are required when averaging over many sequences in order to improve statistical convergence; in particular some severe criticisms [69, 75, 83] were raised against the biological significance of Voss study [65, 66, 67], since his results correspond to averages over complete genebank database categories which are not of equal taxonomic rank. However, beyond these practical problems, there is also a more fundamental theoretical restriction since the measurement of a unique exponent which characterizes the global scaling properties of a sequence fails to resolve multifractality [102], and thus provides very poor information upon the nature of the underlying LRC (if there are any). Actually, it can be shown that for a homogeneous (monofractal) DNA sequence, the scaling exponents estimated with the techniques previously mentioned, can all be expressed as a function of the so-called Hurst or roughness exponent H of the corresponding DNA walk landscape [75, 102]. $H = 1/2$ corresponds to classical Brownian motion, i.e., uncorrelated random walk. For any other value of H, the steps (increments) are either positively correlated ($H > 1/2$: persistent random walk) or anti-correlated ($H < 1/2$: antipersistent random walk).

One of the main obstacles to LRC analysis in DNA sequences is the genuine mosaic structure of these sequences which are well known to be formed of ‘patches’ (‘strand bias’) of different underlying composition [103, 104, 105, 106]. These patches appear as trends in the reconstructed DNA walk landscape which are likely to introduce some breaking of the scale-invariance [64, 68, 75, 79, 82, 83, 84, 107]. One possibility is that these trends possess some characteristic length scale corresponding to a low frequency component that is not invariant with respect to dilatations. Another possibility is that these trends do not have any characteristic length scale either, but actually display some scale-invariance properties that differ from those of the basic fluctuations; in this case some cross-over should be observed in the scaling regime, the largest scale fluctuations behaving differently from those at small scales. There have been some phenomenological attempts to differentiate local patchiness from long-range correlations using 'ad hoc methods' such as the so-called 'min-max method' [64] and the 'detrended fluctuation analysis' [108, 109]. Alternatively, the wavelet transform (WT) [110, 111, 112, 113, 114] has been proposed as a very powerful technique for fractal analysis of DNA sequences [102, 115] (see also Ref. [116] where the WT is used as a tool for visualizing regular patterns in DNA sequences). As already experienced in various fields, the WT can be seen as a mathematical microscope that is well suited for characterizing the scaling properties of fractal objects and this even in the presence of some polynomial component [113, 117, 118, 119, 120]. By considering analyzing wavelets that make the microscope blind to low frequency trends, one can reveal and quantify the scale invariance properties of DNA walks [102, 115, 121]. In a previous work, by applying the so-called wavelet transform modulus maxima (WTMM) method [119, 120, 122] 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 [102, 115]. The main consequence of this result is the justification of using a single exponent, namely the Hurst or roughness exponent H, to characterize the underlying fractal organization of DNA sequences. From a systematic analysis of human exons, CDS's and introns, we have found that long-range power law correlations are not only present in non-coding sequences but also in coding ones somehow hidden in their inner codon structure [123]. Moreover in both coding and non-coding sequences, the strength of these correlations appears to increase with the GC content of the analyzed sequence [123]. (We refer the reader to the work of Yeramian [124] for an alternative study of coding and non coding regions based on the sequence-specific propensity for the thermal disruption of the double-helix.)

In the present report, we first give an overview of the wavelet based methodology for genomic sequence analysis. Then we use the WT microscope to proceed to a systematic investigation of coding and non-coding (intronic) DNA sequences. Taking advantage of the availability of fully-sequenced genomes, we extend the study of scale-invariance properties of DNA sequences on a much wider range of scales than previously done. We show that there actually exists some crosss-over scale about $100 - 200$ bp that separates two different regimes of scale-invariance properties corresponding to two different regimes of long-range power-law correlations. Here we will mainly focus on the small-scale regime ($\lesssim 200$ bp) where the simultaneous observation of LRC on DNA texts and on the corresponding bending profiles generated with various curvature tables [51, 125], will be shown to provide a quite reliable signature of the existence of a nucleosomal structure [1, 2]. In a forthcoming communication, we will concentrate on the large-scale regime ($\gtrsim$ 200 bp) which looks rather universal in the sense that it is present in all three kingdoms. As an overall message of our wavelet based statistical analysis of genomic DNA, we will discuss the observed LRC regimes as a direct manifestation in the primary DNA sequences of the structural organization of DNA in chromatin. In contrast to previous interpretations mainly based on mechanisms involved in genome dynamics, we will further propose some understanding of these correlations in terms of structural and dynamical constraints for chromosome packaging.

2. Theoretical Concepts and Methodology: Wavelet Analysis of Long-Range Correlations in DNA Sequences

In this section, we use artificial sequences that mimic the distribution of purines and pyrimidines along a DNA sequence, to illustrate the concept of long-range power-law correlations and its relationship to scale invariance properties [64, 75, 102]. We also explain how to quantify LRC from the measurement of the so-called Hurst-exponent H [119, 126]. We discuss the necessity of using wavelet analysis to investigate LRC in real DNA sequences.

2.1. LONG-RANGE CORRELATIONS AND SCALE INVARIANCE PROPERTIES OF SYMBOLIC SEQUENCES

To build synthetic DNA sequences displaying LRC, we use a method which will not be described here and which is based on a two-valued ‘fractional auto-regressive integrated moving average’ (FARIMA) process that has been extensively studied by Audit et *al.* [127]. Let us consider two artificially built sequences. The first one is a purely random sequence with equal probability for each of the nucleotides. The second one is built under the constraint that the purine (or equivalently the pyrimidine) positions be long-range correlated with a Hurst exponent $H = 0.9$. Both these sequences are 262 144 bp long and contain 50% purines. First, we use a bar code representation of the purine along the first 400 bp of the two sequences. Figures 1(a) and 1(b) can readily be distinguished by visual inspection. Stretches of black (resp. white) meaning stretches of purines (resp. pyrimidines) are clearly wider for the correlated sequence (Figure 1(b)) than for the uncorrelated one (Figure 1(a)). Figure 1(b) seems to be more contrasted than Figure 1(a). This qualitative difference is simply the signature of what we will refer to as *persistence* [126, 128, 129]. When the positions of purines are positively correlated, if there is a purine at a position, the probability to have a purine at the following position is enhanced; long-range-correlations mean that this enhancement also depends on the presence of purines at positions on the sequence on arbitrarily large distance.

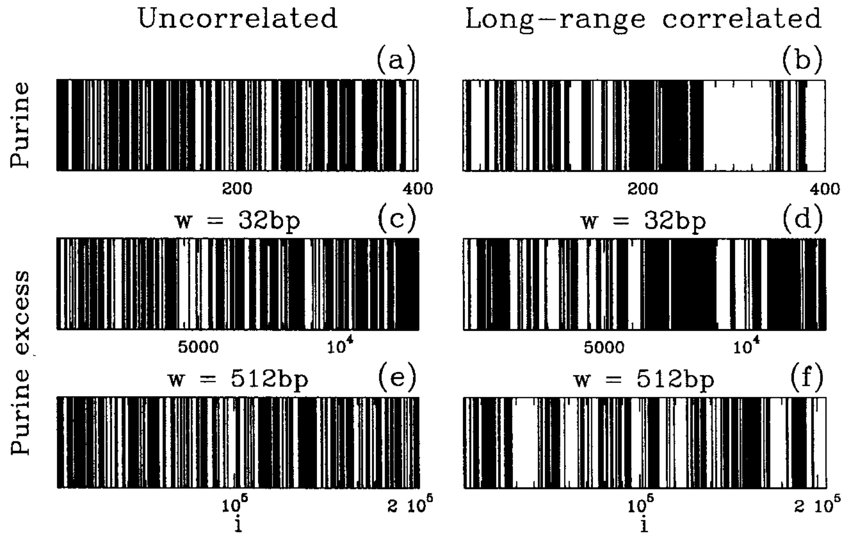

Figure 1. Bar code representation of the purine/pyrimidine content for two artificial DNA sequences. In (a), (c) and (e), we display the analysis of an artificial sequence generated by uncorrelated trials with equal probability for the 4 nucleotides. In (b), (d) and (f), we display the analysis of an artificial sequence obtained under the constraint that the purine (or equivalently the pyrimidine) positions along the sequence be long-range correlated with a Hurst exponent $H = 0.9$ and such that each nucleotide appears with the same probability. In (a) and (b), for each sequence position corresponding to a purine, a black bar of width 1 bp is drawn. In (c) and (d) (resp. (e) and (f)), the sequence has been divided into non overlapping boxes of size 32 bp (resp. 512 bp) and we measure the purine content in each of these boxes. If this concentration is greater than 50%, then a black bar of width 32bp (resp. 512 bp) is drawn at the corresponding position. Notice that for the 6 pictures, the abscissa range is 400 bar widths to ensure that the visual effect obtained is not due to bars of different sizes. Going from (e) to (c) to (a) (resp. (f) to (d) to (b)) is equivalent to zooming in the uncorrelated (resp. long-range correlated) sequence with a black and white 400 pixels resolution camera.

To illustrate this particular structural property induced by these correlations and its relation to scale invariance, let us perform the following visual experiment. Let us decompose the sequence into non overlapping boxes of size w. If the purine content in a box is greater than 50%, then a black bar of width w is drawn at the corresponding position. This experiment amounts to doing a coarse graining on the sequence, replacing each box by a purine or a pyrimidine according to the purine content in that box. The results of this experiment are shown respectively in Figures 1(c) and 1(d) for $\mathrm{w} = 32$ bp and in Figures 1(e) and 1(f) for $\mathrm{w} = 512$ bp. When comparing these results to the previous ones (Figures 1(a) and 1(b)), one can notice an important feature: for a given sequence, the three bar code representations are statistically undistinguishable. This illustrates two important properties of the sequences under analysis:

1. Both the uncorrelated and the correlated sequences are *scale invariant* in the sense that one cannot statistically distinguish the original sequence from those obtained after a coarse graining as we have done here.
2. For the correlated sequence, the fact that it is scale invariant means that the way purines are positioned is the same as the way boxes containing an excess of purines are positioned and this, whatever the box width. We talk about *long-range correlations* (LRC) because the way the boxes are distributed is persistent, that is to say that the correlations between 2 boxes separated by n other boxes are independent of $\mathtt{w}$ and behave as $\propto n^{2H-2}$ (see Eq. (4)). Note that the uncorrelated sequence is also scale invariant because it does not possess correlations at any scale.

At this point, it is important to note that persistence does not mean that there is no, or only little, variation of the purine concentration along the sequence. In other words, persistence does not mean that in the representation of Figure 1, the correlated sequences are all black, or all white (the sequences contain by construction an equal number of black and white boxes). This results from equation (2) which shows that for any given box size $\mathtt{w}$, the standard deviation $\sigma_H(\mathtt{w}) \sim \mathtt{w}^{H-1}$ increases with H. This characteristic feature is illustrated in Figure 2(b) where the amplitude of oscillations of the purine content is larger than for the uncorrelated sequence in Figure 2(a) (note that here $\sigma(1)$ is the same for the two sequences). In other words, persistence of the purine concentration means that it fluctuates 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. Note also that larger H values do not mean larger density values.

It is also important to note that LRC cannot be constructed with a Markov model of finite size memory [130]. Markov chains yield correlation functions that decay exponentially over some characteristic finite size. Hence an artificial sequence built with a Markov model of order m would be undistinguishable from an uncorrelated one as soon as the size of the box $\mathtt{w} \gg m$. Therefore, when one reveals LRC in DNA sequences, one evidences processes that structure objects of size $\mathtt{w}$ along the genome in the same statistical manner at any scale $\mathtt{w}$ of observation.

2.2. QUANTIFICATION OF LONG-RANGE CORRELATIONS IN SYMBOLIC SEQUENCES

In Figure 2, we propose a quantitative analysis of the fluctuations of purine concentration along the same two sequences we have already studied in the previous section. In Figures 2(a), 2(b), 2(c) and 2(d), the purine concentration is plotted for the same box sizes as those used in figure 1(c), 1(d), 1(e) and 1(f) respectively. For the uncorrelated sequence, one clearly notices that the fluctuations around the mean value $1/2$ are much smaller for the largest window width $\mathtt{w}_2 = 512$ bp (Figure 2(c)) than for the smallest window width $\mathtt{w}_1 = 32$ bp (Figure 2(a)). In the case of the long-range correlated sequence, one also notices that the fluctuations of the purine

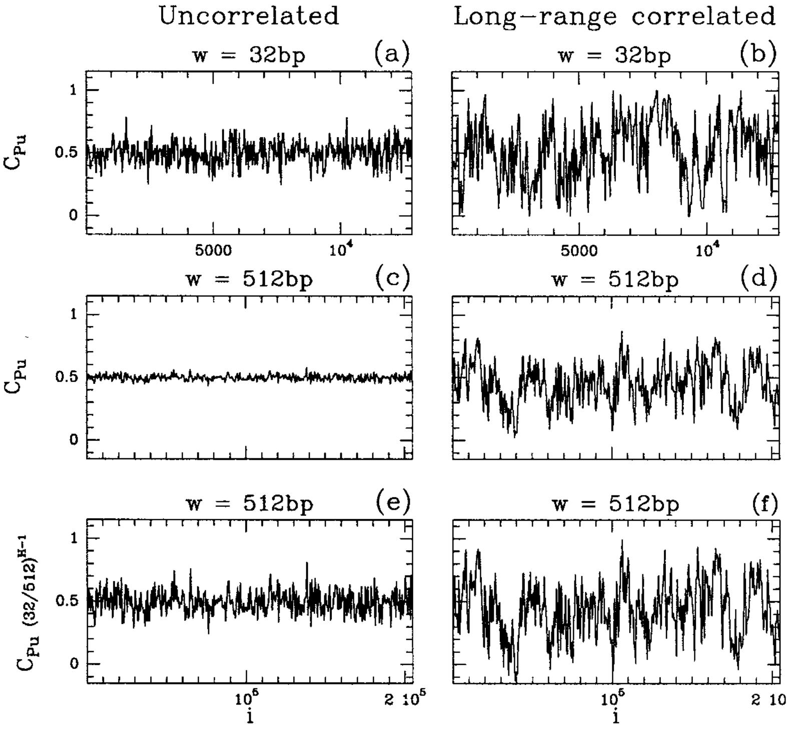

Figure 2. Fluctuations of the purine content C_{Pu} within non overlapping boxes of size w as a function of the box position i for the same two artificial DNA sequences as in Figure 1. In (a) and (b), $\mathrm{w}_1 = 32$ bp; in (c) and (d), $\mathrm{w}_2 = 512$ bp. In (e) and (f), C_{Pu} computed with $\mathrm{w}_2 = 512$ bp, is rescaled by $(\mathrm{w}_1/\mathrm{w}_2)^{H-1}$ according to equation (2) with $H = 0.5$ (e) and $H = 0.9$ (f). For the 6 pictures, the abscissa range is 400 box wide so that each curve is made of the same number of points.

concentration decreases as the window width increases (Figures 2(b) and 2(d)) but to a much smaller extent. We take advantage of this difference to perform a quantitative measure of the scale invariance properties.

For an uncorrelated sequence, the purine concentration measured in a box of width w is simply the arithmetic mean of w independent and identically distributed (i.i.d.) random variables. Its standard deviation $\sigma(\mathrm{w})$ is thus of the form:

$$\sigma(\mathrm{w}) = \frac{\sigma(1)}{\sqrt{\mathrm{w}}}. \tag{1}$$

In the case of a sequence possessing scale invariance properties with a Hurst exponent H, then the standard deviation reads [64, 75, 102, 126, 128, 129]:

$$\sigma_H(\mathrm{w}) = \sigma_H(1)\mathrm{w}^{H-1}. \tag{2}$$

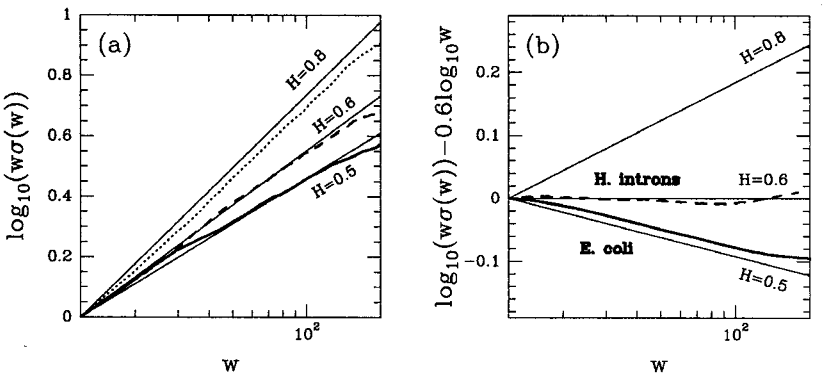

Figure 3. Scale-invariance analysis of purine concentration fluctuations. (a) $\log_{10}(\mathrm{w}\sigma_H(\mathrm{w}))$ vs $\log_{10}\mathrm{w}$ for artificial sequences of length 10 kbp with Hurst exponent $H = 0.5$ (solid line), 0.6 (dashed line) and 0.8 (dotted line) respectively. (b) $\log_{10}(\mathrm{w}\sigma_H(\mathrm{w})) - 0.6\log_{10}\mathrm{w}$ vs $\log_{10}\mathrm{w}$ for the primary DNA sequence of *Escherichia coli* (solid line) and some average over the human intron sequences of length larger than 800 (dashed line) (see Materials and Methods). The solid straight lines corresponding to uncorrelated ($H = 0.5$) and long-range correlated ($H = 0.6$ and $H = 0.8$) sequences are drawn to guide the eyes. Note that for the sake of clarity, the curves in (a) and (b) have been vertically shifted in order to start at the same ordinate for $\mathrm{w} = 10$ bp.

As a visual check of this power-law behavior of the root-mean square (r.m.s.) fluctuations of purine concentration, we have plotted in Figures 2(e) and 2(f) the purine concentration computed for the box size $\mathrm{w}_2 = 512$ bp after some rescaling by $(\mathrm{w}_1/\mathrm{w}_2)^{H-1}$. Once rescaled with the appropriate Hurst exponent value, the purine concentration fluctuations obtained for both the uncorrelated ($H = 0.5$ in Figure2(e)) and the long-range correlated ($H = 0.9$ in Figure 2(f)) sequences have the same overall amplitude and are statistically undistinguishable from the corresponding fluctuations obtained with boxes of smaller size in Figures 2(a) and 2(b) respectively. Note that equation (1) reduces to equation (2) in the particular case $H = 1/2$ characteristic of uncorrelated sequences. The quantitative characterization of scale invariant properties is a straightforward consequence of equation (2). Taking the logarithm of equation (2), one gets:

$$\log_{10}(\mathrm{w}\sigma_H(\mathrm{w})) = H\log_{10}\mathrm{w} + \log_{10}\sigma_H(1). \qquad (3)$$

So, when plotting $\log_{10}(\mathrm{w}\sigma_H(\mathrm{w}))$ as a function of $\log_{10}\mathrm{w}$, the fact that all the data points fall on a linear curve enables us to diagnostic scale invariant properties. Then, from measuring the slope of this straight line, one gets some estimate of the exponent H. In Figure 3(a) is illustrated the estimate of the Hurst exponent of an uncorrelated ($H = 0.5$) and of two long-range correlated ($H = 0.6$ and $H = 0.8$) random sequences. For the three sequences, a straight line of slope H provides a very good fit of the r.m.s. data. In Figure 3(b), we show a more readable presentation that we are going to use all along this manuscript for the analysis of genomic sequences. By plotting $\log_{10}(\mathtt{w}\sigma_H(\mathtt{w})) - 0.6\log_{10}\mathtt{w}$ versus $\log_{10}\mathtt{w}$, we select $H = 0.6$ as the Hurst exponent value of reference for an horizontal linear scaling behavior. This particular value actually corresponds to the LRC observed for the set of sequenced human introns of length $L \geqslant 800$ bp [102, 115]. In Figure 3(b) are also reported for comparison the results obtained when investigating the genome of *Escherichia coli* which are quite typical of what we have observed with other eubacterial genomes. The data points remarkably fall on the straight line corresponding to the Hurst exponent value $H = 0.5$ characteristic of uncorrelated sequences. The straight line corresponding to $H = 0.8$ is drawn as the signature of strongly correlated sequences. The existence of a large-scale regime ($\gtrsim 200$ bp) of strong LRC [1, 2] will be detailed in a forthcoming publication. An important feature is that we actually need to use oscillating boxes, i.e., wavelets, to perform this standard deviation measure on real sequences as the mosaic stucture of genomic DNA sequences may lead to severe bias [102, 115].

Now, if one is interested in the behavior of the correlation function $C(n)$ between nucleotides separated by a distance n, equation (2) implies that

$$C(n) \propto \sigma_H^2(n) \propto n^{2H-2}, \tag{4}$$

i.e. a power-law decrease as a function of the distance n with exponent $2H - 2$ [61, 64, 99, 126]. Hence, the larger $H(< 1)$, the weaker the power-law decrease and the stronger the LRC.

2.3. QUANTIFICATION OF LONG-RANGE CORRELATIONS IN REAL DNA SEQUENCES: ABOUT THE NECESSITY OF USING WAVELET ANALYSIS

2.3.1. *Problems resulting from the Compositional Heterogeneity of Genome Sequences*

As pointed out in the previous section, a way to quantify LRC and scale-invariance properties of symbolic sequences consists in using the so-called ‘variance method’. This method amounts to compute the r.m.s. fluctuations of some nucleotide concentration measured in a box of width $\mathtt{w}$ and to look for some power-law behavior (Eq. 2) from which one extracts an estimate of the Hurst exponent H. But the variance method requires the experimental data to be stationary [126], a prerequisite which is absolutely not satisfied by the DNA sequences. As illustrated in Figures 4(a), 4(c) and 4(e) for the sequence of the bacteriophage λ (see Materials and Methods) when using the ‘Purine’ coding rule, DNA sequences consist of patches of different compositions, namely purine-rich regions which alternate with pyrimidine-rich regions. This genuine mosaic structure of genomic DNA sequences may lead to severe bias in the estimate of H [64, 68, 75, 79, 82, 83, 84]. The fact that the average purine concentration over the first 20000 bp and the last 10000 bp of the bacteriophage λ is 54%, while it is only 46% from 22000 to 37000 bp, has a dramatic consequence in the variance calculation. This rather obvious breaking of stationarity leads to the presence of an additional constant term in $\sigma_H^2(\mathrm{w})$ which induces some departure from scale-invariance:

$$\sigma_H(\mathrm{w}) = \left(A\mathrm{w}^{2H-2} + B\right)^{1/2}, \tag{5}$$

i.e.,

$$\mathrm{w}\sigma_H(\mathrm{w}) = \left(A\mathrm{w}^{2H} + B\mathrm{w}^2\right)^{1/2}. \tag{6}$$

Thus, as shown in Figure 5, when plotting $\mathrm{w}\sigma_H$ $(= \sigma_{WT})$ versus w in a logarithmic representation, to estimate the exponent H from the slope of the experimental data (Eq. 3), one does not observe a well defined straight line but rather a cross-over from a scaling regime with $H \simeq 0.5$ at small scale ($\mathrm{w} \lesssim 35$ bp) where the first term in the right-hand side of equation (5) is dominating, to a trivial regime $H = 1$ at large scale ($\mathrm{w} \gtrsim 300$ bp) as the signature of the non stationarity of the purine concentration signal. Then if one proceeds blindly to a linear regression fit of the data in the range $10 \lesssim \mathrm{w} \lesssim 200$, one gets a value of $H \simeq 0.6$ which is totally misleading. Actually if one rescales, like in Figures 2(e) and 2(f), C_{Pu} obtained for $\mathrm{w}_2 = 512$ bp, by $(\mathrm{w}_1/\mathrm{w}_2)^{H-1}$ with this biased H value, one gets the purine concentration fluctuations shown in Figure 4(e) which do not look statistically the same as those observed at a smaller scale $\mathrm{w}_1 = 32$ bp in Figure 4(a): the non stationarity of the purine concentration is much more pronounced whereas the overall amplitude in each patches is clearly smaller.

2.3.2. *A Solution: The Continuous Wavelet Transform*

To investigate the scaling properties of DNA sequences and the possible existence of LRC, one thus needs a mathematical tool that can master the non stationarity of these genomic data. Actually such a technique does exist and is called the continuous wavelet transform [110, 111, 112, 113, 114, 131]. The WT consists in expanding signals in terms of wavelets which are constructed from a single function, the analyzing wavelet ψ, by mean of translations and dilations. The WT of a distribution μ is defined as [118, 119]:

$$T_\psi[\mu](x, \mathrm{w}) = \frac{1}{\mathrm{w}} \int_{-\infty}^{+\infty} \psi\left(\frac{x-y}{\mathrm{w}}\right) d\mu(y), \tag{7}$$

where x is the space parameter and $\mathrm{w}(> 0)$ the scale parameter. The continuous wavelet transform (WT) can be seen as some generalization of the Fourier transform in the sense that it provides a time-frequency (or space-scale) analysis instead of a mere frequency(or scale) analysis. The analyzing wavelet ψ is generally chosen to be well localized in both space and frequency. Usually, ψ is only required to be of zero mean for the WT to be invertible. In Figures 4(b) and 4(d) are shown the WT of the DNA sequence of the bacteriophage λ as computed at two different scales $\mathrm{w} = 32$ bp and 512 bp, when considering the 'Purine'

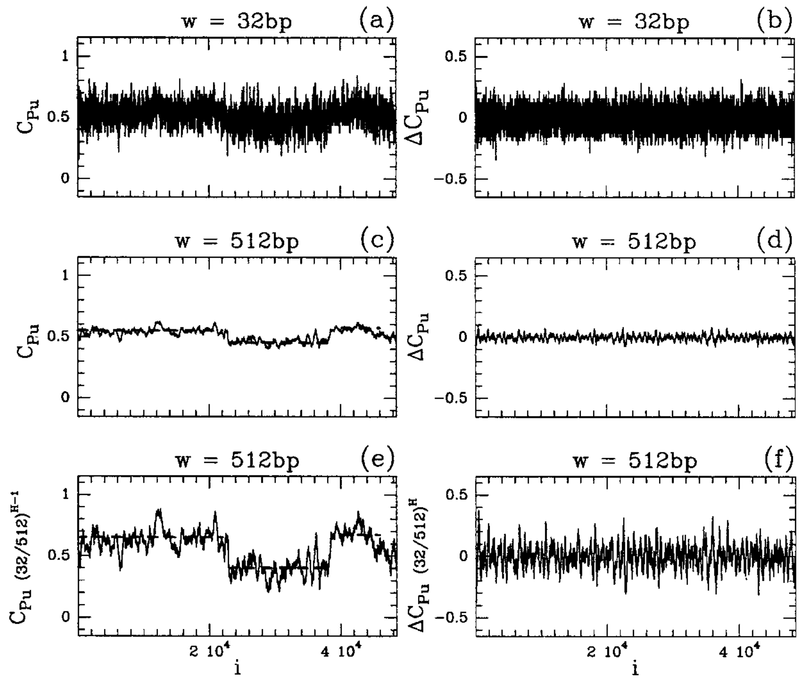

Figure 4. Fluctuations of the purine content within non overlapping boxes (Figure 6(a)) of size w as a function of the box position i for the DNA sequence of the bacteriophage λ ($L = 48502$ bp): (a) C_{Pu} for w $= 32$ bp; (c) C_{Pu} for w $= 512$ bp and (e) C_{Pu} for w $= 512$ bp after being rescaled according to equation (2) with $H = 0.6$. Fluctuations of the corresponding wavelet coefficients as computed with the Haar wavelet (Figure 6(b)): (b) ΔC_{Pu} for w $= 32$ bp; (d) ΔC_{Pu} for w $= 512$ bp and (f) ΔC_{Pu} for w $= 512$ bp after being rescaled according to equation (8) with $H = 0.5$. In (c) and (e) the horizontal dashed segments correspond to the patches of different average purine concentrations.

coding. To perform the WT calculation, we have used the so-called Haar wavelet [110, 114] which is illustrated in Figure 6(b). When applying this piece-wise constant analyzing wavelet to the positive-valued distribution μ of purines along the bacteriophage λ DNA sequence, equation (7) simply amounts to compute the difference of the purine concentrations in two juxtaposed boxes of size w$/2$ at the position x. As shown in Figures 4(b) and 4(d), when doing this difference, one cancels any possible departure of the local mean from the global mean over the entire sequence. Indeed the global mean as well as the local mean of the wavelet coefficients in the various patches of different purine compositions are now all equal to zero. It is in that sense, i.e., *by looking at the variations of concentration instead of the concentration itself*, that the WT restores stationarity: whatever the scale w, the wavelet coefficients sustainly fluctuate about zero all along the entire DNA sequence. Then, as shown in Figure 5, one can proceed to the analysis of the

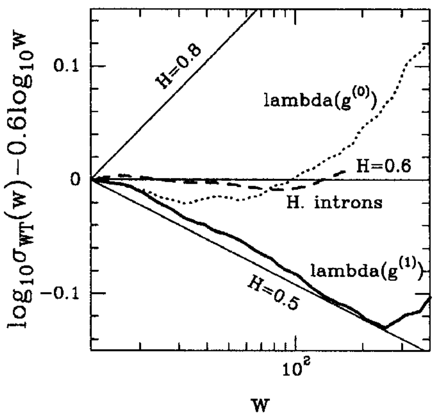

Figure 5. Scale-invariance analysis of the DNA sequence of the bacteriophage λ ($L = 48502$) using the Purine coding: $\log_{10}\sigma_{WT}(\mathrm{w}) - 0.6\log_{10}\mathrm{w}$ vs $\log_{10}\mathrm{w}$. $\sigma_{WT}(\mathrm{w})$ corresponds to the variance of the wavelet coefficients computed at scales w (in bp units) with the zero-order $g^{(0)}$ (smooth box: dotted line) and first-order $g^{(1)}$ (the first derivative of the Gaussian function: solid line) analyzing wavelets illustrated in Figures 6(c) and 6(d) respectively. The results obtained for our set of human introns with $g^{(1)}$ (dashed line) are shown for comparison. Note that when using $g^{(0)}$, $\sigma_{WT}(\mathrm{w})$ actually corresponds to $\mathrm{w}\sigma_H(\mathrm{w})$ (see equations (2) and (8)).

scale-invariance properties by investigating the power-law behavior of the r.m.s. fluctuations of the wavelet coefficients as function of the scale w:

$$\sigma_{WT}(\mathrm{w}) = \sigma_{WT}(1)\mathrm{w}^H. \tag{8}$$

Note that when comparing equation (8) to equation (2), $\sigma_{WT}(\mathrm{w})$ plays the role of $\mathrm{w}\sigma_H(\mathrm{w})$ in equation (3). When using the Haar wavelet $d^{(1)}$ (Figure 6(b)), as well as a smooth version given by the first derivative $g^{(1)}$ of the Gaussian function (Figure 6(d)), as analyzing wavelets, one clearly remedies to the cross-over previously observed when using the inappropriate box function $d^{(0)}$(Figure 6(a)) (or the Gaussian function $g^{(0)}$(Figure 6(c))). As shown in Figure 5, one recovers a well defined scaling behavior from which one can extract the experimental estimate $H = 0.50 \pm 0.02$ from a linear regression fit over the range of scales $10 \lesssim \mathrm{w} \lesssim 200$. As a visual test of the relevance of this measurement, we have plotted in Figure 4(f), the wavelet coefficients computed at scale $\mathrm{w}_2 = 512$ bp, after being rescaled by $(32/512)^H$ with $H = 0.5$ in order to be compared to the wavelet coefficients computed at scale $\mathrm{w}_1 = 32$ bp in Figure 4(b). Both signals in Figures 4(b) and 4(f) have the same overall amplitude and are statistically undistinguishable stationary signals. These observations converge to the conclusion that when using the ‘Purine’ coding, the DNA sequence of the bacteriophage λ does not display any LRC since the estimate of the Hurst exponent cannot be distinguished from the canonical $H = 1/2$ value for uncorrelated sequences. Let us remind that we would have

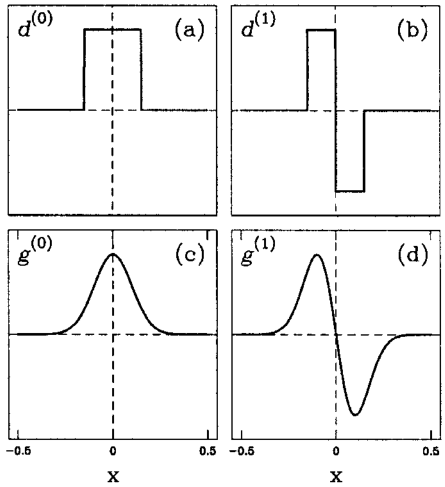

Figure 6. Set of analyzing functions that can be used in equation (7). (a) $d^{(0)}$: the box function; (b) $d^{(1)}$: the Haar wavelet; (c) $g^{(0)}$: the Gaussian function and (d) $g^{(1)}$: the first derivative of the Gaussian function. Note that $g^{(0)}$ and $g^{(1)}$ can be seen as smooth versions of $d^{(0)}$ and $d^{(1)}$ respectively.

reached the opposite conclusion, namely the existence of LRC with $H \simeq 0.6$, when using the unappropriate box-function $d^{(0)}$.

2.3.3. *Mastering Nonlinear Mosaic Structures*

We have just seen that with a first-order analyzing wavelet, one can easily remove piece-wise constant behavior that may be superimposed to the fluctuations of purine concentration. Actually, nothing prevents the heterogeneity in composition of DNA sequences to induce a more complex (possibly nonlinear) mosaic structure that will further perturb the estimate of the scaling exponent H. At this point, let us note that one can use analyzing wavelets of arbitrarily high order. Indeed, the main advantage of using the WT for revealing and characterizing LRC, is its ability to be blind to polynomial behavior, i.e. to low-frequency trends that can mask the existence of scale-invariance properties. Hence, by considering analyzing wavelets $\psi^{(n)}$ that have n_ψ vanishing moments [119, 120]:

$$\int_{-\infty}^{+\infty} x^m \psi^{(n)}(x) dx = 0, \quad \forall m, 0 \leq m < n_\psi, \tag{9}$$

one can make our 'WT mathematical microscope' blind to polynomial behavior up to degree $n_\psi - 1$, with the hope to master more elaborated mosaic structure effects than the rather simple piece-wise constant bias observed in the fluctuations of Purine concentration of the bacteriophage λ DNA sequence in Figures 4(a), 4(c) and 4(e). In our pioneering study of the complexity of DNA sequences using the

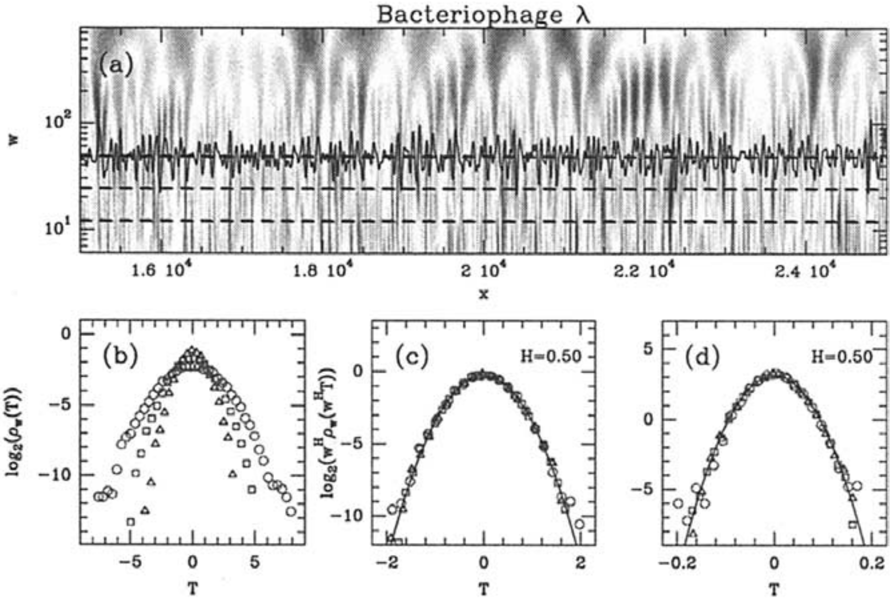

Figure 7. Wavelet transform analysis of the DNA sequence of the bacteriophage λ ($L = 48502$ bp) when using the 'A' coding rule. The analyzing wavelet is the first-derivative of the gaussian function $g^{(1)}$ (Figure 6(d)). (a) Space-scale representation provided by the WT: $T_{g^{(1)}}(x, \mathtt{w})$ is coded, independently at each scale w, using 32 grey levels from white ($min_x T_{g^{(1)}}(x, \mathtt{w})$) to black ($max_x T_{g^{(1)}}(x, \mathtt{w})$). The arborescent structure of the wavelet representation of the bacteriophage λ DNA sequence is typical of fractal signals that display scale invariance properties. (b) Probability density functions $\rho_{\mathtt{w}}(T)$ of wavelet coefficients for the set of scales $\mathtt{w} = 12(\triangle)$, $24(\square)$ and $48(\circ)$ in bp units. (c) $\log_2(\mathtt{w}^H \rho_{\mathtt{w}}(\mathtt{w}^H T))$ vs T for the same data as in (b), when fixing $H = 0.50$. (d) Same representation as in (c) but when considering the 'Pnuc' trinucleotide coding rule. The horizontal dashed lines in (a) correspond to the investigated scales $\mathtt{w} = 12$, 24 and 48 bp from bottom to top; at the scale $\mathtt{w} = 48$ bp, the wavelet coefficients are represented on the top of the corresponding horizontal dashed line.

WT [102, 115] we have mainly worked with the class of analyzing wavelets defined by the successive derivatives of the Gaussian function:

$$g^{(N)}(x) = \frac{d^N e^{-x^2/2}}{dx^N}, \tag{10}$$

for which $n_\psi = N$. In the present work, most of the results reported in the various figures have been obtained with the first-order analyzing wavelet $g^{(1)}(x)$. Of course we have checked that these results are robust when using higher-order analyzing wavelets, in particular when using the so-called Mexican hat $g^{(2)}$ (data not shown).

2.3.4. *Monofractality: An Essential Statistical Property of DNA Sequences*

The WT is a mathematical technique that has been originally introduced for time-frequency analysis of seismic data and acoustic signals [132, 133, 134]. In previous works [102, 115], we have shown that from the space-scale WT representation

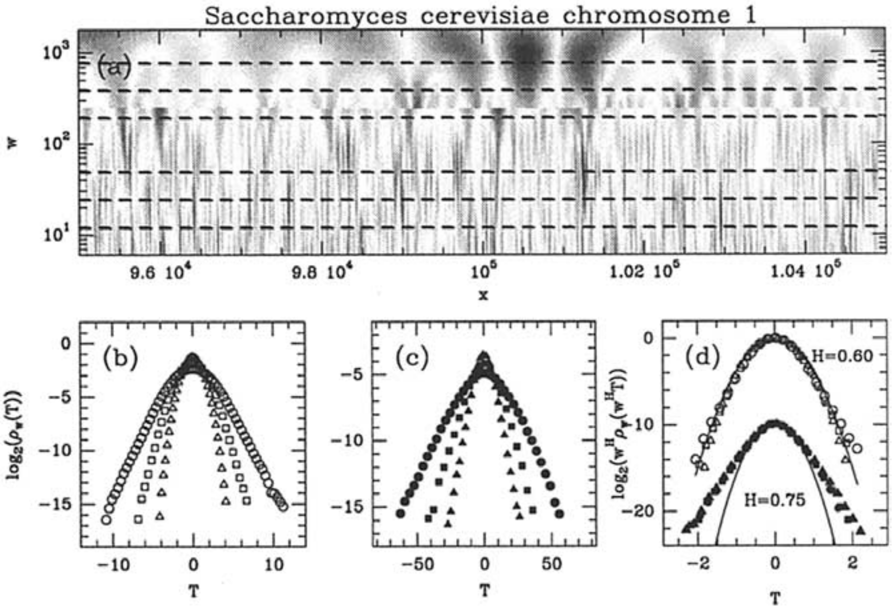

Figure 8. Wavelet transform analysis of the DNA sequence of the chromosome 1 of *Saccharomyces cerevisiae* ($L = 230209$ bp) when using the 'A' coding rule. The analyzing wavelet is the first-derivative of the gaussian function $g^{(1)}$(Figure (6(d))). (a) Space-scale representation obtained when using the same grey level coding of the WT as in Figure 7(a). (b) Probability density functions $\rho_{\mathrm{w}}(T)$ of wavelet coefficients for the set of scales $\mathrm{w} = 12(\triangle)$, $24(\square)$, $48(\circ)$. (c) Same as in (b) for the scales $\mathrm{w} = 192(\blacktriangle)$, $384(\blacksquare)$ and $768(\bullet)$ in bp units. (d) $\log_2(\mathrm{w}^H \rho_{\mathrm{w}}(\mathrm{w}^H T))$ vs T for the three lowest scales shown in (b) when fixing $H = 0.60$ and for the three largest scales shown in (c) when fixing $H = 0.75$. The horizontal dashed lines in (a) correspond to the investigated scales $\mathrm{w} = 11$, 24, 48, 192, 384 and 768 bp from bottom to top.

of 'DNA walks', one can bring the experimental proof of the monofractal nature of DNA walk landscapes. In Figure 7(a) is shown a 32-grey level coding of the space-scale WT representation of the bacteriophage λ sequence when using the 'A' mononucleotide coding (see Materials and Methods) and $g^{(1)}$ (Figure 6(d)) as analyzing wavelet. A way to investigate the evolution of the statistics across scale consists in computing the probability density function (pdf) $\rho_{\mathrm{w}}(T)$ of wavelet coefficient values at different scales. As shown in Figure 7(b), when increasing w from 12 to 24 up to 48, the corresponding histograms become wider and wider with a r.m.s. which behaves as predicted by equation (8) with a power-law exponent $H = 0.50 \pm 0.02$. Moreover, when rescaling the wavelet coefficients by the r.m.s. value at the corresponding scales, all the data computed at different scales collapse on a single curve as shown in Figure 7(c). This is the demonstration that the WT pdfs satisfy the self-similarity relationship [102, 115]:

$$\mathrm{w}^H \rho_{\mathrm{w}}(\mathrm{w}^H T) = \rho_1(T), \tag{11}$$

the hallmark of monofractality. The way the moments of ρ_{w} behave as a function of w requires the knowledge of a single scaling exponent only, namely the Hurst exponent H. This contrasts with multifractal signals for which a continuum of scaling exponents is required to account for the evolution across scales of the shape of the WT pdf [113, 119, 120]. A second important message brought by Figure 7(c) is the fact that in a semi-logarithmic representation, all the data fall on a same curve which is well approximated by a parabola as predicted for Gaussian statistics. Thus, as explored through the optics of the WT microscope, the basic fluctuations in the spatial distribution of A nucleotides of the bacteriophage λ are likely to be Gaussian. At this point, let us emphasize that these two experimental observations, namely monofractality with $H = 0.5$ (no LRC) and Gaussian statistics, are quite characteristic features of the bacteriophage λ genome sequence that are recovered with all the mononucleotide, dinucleotide and trinucleotide coding rules used in this work. As an illustration, the results obtained with the 'Pnuc' trinucleotide coding associated to bending properties of nucleosomal DNA (see Materials and Methods), are reported in Figure 7(d).

The progress made in sequencing programs allowed us to investigate completely sequenced genomes [1, 2]. In Figure 8, we report the results of the WT analysis of the Chromosome 1 ($L = 230209$ bp) of *Saccharomyces cerevisiae (S.c.)* when using $g^{(1)}$ (Figure 6(d)) as analyzing wavelet. These results obtained with the 'A' coding rule are again quite representative of the whole set of data obtained when using different mono-, di- and tri- nucleotide coding rules. When investigating the evolution across the scales of the pdf of wavelet coefficients in Figures 8(b) and 8(c), one reveals the existence of a characteristic scale $\mathrm{w}_c \simeq 200$ bp that separates two different scaling regimes which both satisfy the self-similarity relationship (11), provided one uses the scaling exponent value $H = 0.57$ in the scale range $10 \lesssim \mathrm{w} \lesssim 100$ and $H = 0.75$ in the scale range $200 \lesssim \mathrm{w} \lesssim 1000$ (Figure 8(d)). In the small-scale regime, the pdfs are very well approximated by Gaussian distributions. In the large-scale regime, the pdfs of the wavelet coefficients of the yeast chromosome 1 have fat stretched exponential like tails (Figure 8(d)). The fact that the self-similarity relationship is satisfied in the small- as well as in the large-scale regimes corroborates the monofractal nature of the yeast DNA sequences in these two regimes.

3. Results

We report in this section the results of a wavelet based statistical analysis of the scale-invariance properties of genomic sequences that belong to the three kingdoms, namely eukaryotic, eubacterial and archaeal genomes as well as sequences of DNA and RNA viruses [1, 2]. To set the general framework of our study, we will start investigating the 16 chromosomes of the *Saccharomyces cerevisiae* (*S.c.*) which will allow us to perform a scaling analysis on a wide range of scales extending from tens to thousands of nucleotides. After exhibiting the overall general features that are more or less common to all genomes, we will specify the particular issues we want to address in this study.

3.1. SACCHAROMYCES CEREVISIAE

The first completely sequenced eukaryotic genome *Saccharomyces cerevisiae* provides an opportunity to perform a comparative wavelet analysis of the scale-invariance and possible LRC properties displayed by each chromosome. When looking at the global estimate of the r.m.s. of WT coefficients $\sigma_{WT}(\mathtt{w})$ obtained for each of the 16 yeast chromosomes, when using the 'A' mononucleotide coding rule, one sees in Figure 9(a) that all present superimposable behavior, with notably the same characteristic scale $\mathtt{w}_c = 200$ bp, that separates two different scaling regimes. Note that the 16 yeast chromosomes also display other compositional features [105]. At small scales, $10 \lesssim \mathtt{w} \lesssim 200$ (expressed in bp units), LRC are observed as characterized by $H = 0.57 \pm 0.03$, a mean Hurst exponent value which is significantly larger that the theoritical prediction $H = 1/2$ for uncorrelated sequences. At large scales, $200 \lesssim \mathtt{w} \lesssim 5000$, stronger LRC with $H = 0.82 \pm 0.02$ become dominant with a cutoff around 10000 bp (a number by no means accurate) above which uncorrelated behavior is observed. In Figure 9(b) are reported the results of some test of the robustness of the above observations when using different mononucleotide coding rules. The first remarkable feature is that the data for the 'A' and 'T' codings are quite undistinguishable as well as the data for the 'G' and 'C' codings. This will justify that, in the following, we will systematically present results corresponding to the average over the 'A' and 'T' codings on the one hand and over the 'G' and 'C' codings on the other hand. While each of these mononucleotide codings display a characteristic scale $\mathtt{w}_c \simeq 200$ bp that separates two scaling regimes as observed in Figure 9(a) for the 'A' coding, there is however some difference between the 'G' (+'C') coding and the 'A' (+'T') coding. This difference arises mainly in the small-scale regime ($10 \lesssim \mathtt{w} \lesssim 200$) where the estimate of the Hurst exponent turns out to be definitely smaller $H = 0.53 \pm 0.03$ for the 'G' (+'C') coding than the value $H = 0.57 \pm 0.03$ obtained with the 'A' (+'T') coding. The existence of a characteristic scale $\mathtt{w}_c \simeq 200$ bp that separates two different monofractal scaling regimes corroborates the results reported in Figure 8. The probability density functions (pdfs) of wavelet coefficient values of the DNA sequence of the yeast chromosome 1 ($L = 230209$ bp), computed at different scales using the 'A' coding rule (Figure 8(b) and 8(c)), are shown to collapse on a single curve when rescaling the wavelet coefficients by $\mathtt{w}^H$, provided one uses the scaling exponent value $H = 0.60$ in the scale range $10 \lesssim \mathtt{w} \lesssim 100$ and $H = 0.75$ in the scale range $200 \lesssim \mathtt{w} \lesssim 1000$ (Figure 8(d)). In the small-scale regime, the pdfs are very well approximated by Gaussian distributions whereas these pdfs exhibit fat stretched exponential like tails in the large-scale regime. A similar change in the nature of the statistics of wavelet coefficients is observed with all four mononucleotide codings as well as with di- and tri-nucleotide codings.

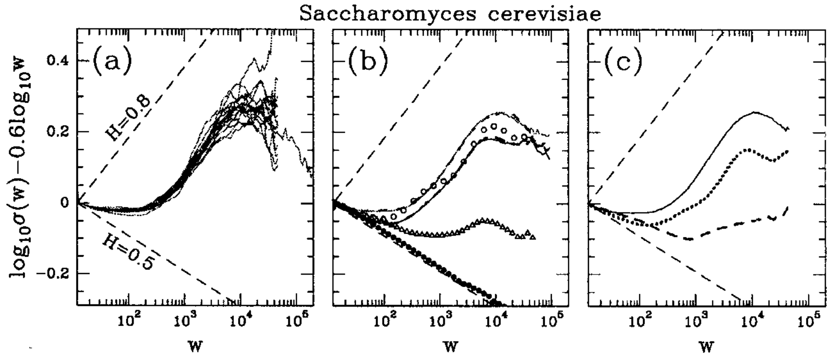

Figure 9. Global estimate of the r.m.s. of WT coefficients of the 16 chromosomes of *S. cerevisiae* : $\log_{10}\sigma_{WT}(\mathrm{w}) - 0.6\log_{10}\mathrm{w}$ is plotted versus $\log_{10}\mathrm{w}$. The analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). (a) Comparative analysis of the 16 chromosomes when using the 'A' mononucleotide coding. (b) Comparative analysis of the 'A' (grey solid line), 'T' (grey dashed line), 'G' (black solid line) and 'C' (black dashed line) mononucleotide codings with the 'Pnuc' (circles) and 'DNase' (triangles) trinucleotide codings; the black dots correspond to a randomly shuffled Pnuc table (see text). (c) Comparative analysis of the 'A' (+'T') mononucleotide coding (solid line) with the 'AA' (='TT') dinucleotide coding (dotted line) and the 'A_{iso}' (='T_{iso}') mononucleotide coding (dashed line). In (b) and (c), the results correspond to averaging over the 16 chromosomes. For coding rules see Materials and Methods.

As shown in Figure 9(b), a comparative wavelet analysis of the yeast DNA sequences using the 'Pnuc' coding rule [51] reveals striking similarities with the curves resulting from the mononucleotide coding rules, and this both in the small-scale ($H = 0.54 \pm 0.01$) and in the large-scale ($H = 0.75 \pm 0.02$) regimes [1, 2]. To ensure that these observations are not simply due to a 'recoding' of the DNA sequences, but rather to the proper values of roll angles used to determine the bending profile of the axis of the double helix, we have randomly changed the Pnuc table that maps trinucleotides to roll-angle values. The new table is obtained using a Gaussian distribution of same mean, variance and symmetries as the original table. As shown in Figure 9(b), this results in the vanishing of the observed LRC; now $H = 0.50 \pm 0.01$ at scales $\mathrm{w} \lesssim 1000$ bp, which strongly suggests that these LRC are likely to reflect persistent structural scale-invariance properties. An additional evidence that the 'Pnuc' trinucleotide coding is not a trivial recoding of the DNA sequences is brought by the data obtained when using the DNase table of curvature [125]. As shown in Figure 9(b), one notices a significantly weakening of the LRC exponent observed in the large-scale regime($H \simeq 0.6$ with the DNase coding instead of $H \simeq 0.8$ with the 'Pnuc' coding), a result which is also true but less flagrant in the small-scale regime. We will see in the following, that some significant weakening is also observed in the small-scale regime of DNA sequences of other eukaryotic sequences when using the 'DNase' coding rule. This demonstrates that on the contrary to what the intuition could tell us, the agreement between the LRC properties observed with the ‘Pnuc’ coding and the mononucleotide codings is not a trivial observation [1, 2].

To strengthen our interpretation of the observed LRC in terms of structural constraints, we have performed, in parallel, the wavelet analysis of DNA sequences using some dinucleotide codings which are known to contribute to the intrinsic bending and flexibility properties of the DNA double helix. As an illustration, we report in Figure 9(c) the results obtained with the ‘AA’ (= ‘TT’) coding when averaging over the 16 yeast chromosomes. When comparing to the results obtained with the ‘A_{iso}’ (= ‘T_{iso}’) coding rule (i.e., A (T) that are not part of a dinucleotide AA (TT)), one observes a clear weakening of the LRC properties with the ‘A_{iso}’ (= ‘T_{iso}’) coding, while the ‘AA’ (= ‘TT’) coding accounts for a major part of the LRC observed with the ‘A’ and ‘Pnuc’ codings. Let us point out that this observation will be the cornerstone of our analysis of LRC in the small-scale regime of genomic sequences, in relation with the existence of nucleosomes.

We have extended this wavelet based analysis of the scaling properties of DNA sequences to various fully sequenced eukaryotic, eubacterial and archaebacterial genomes and also to viral DNA and RNA sequences [1, 2]. A general observation is the existence of a characteristic scale $\mathrm{w}_c \simeq 100 - 200$ bp that separates two different monofractal scaling regimes whatever the coding rule used to digitize the DNA sequences. In the large-scale regime ($200 \lesssim \mathrm{w} \lesssim 1000$), when using the ‘Pnuc’ coding rule as well as the four elementary mononucleotide coding rules, strong LRC ($H \simeq 0.8 \pm 0.1$) are systematically observed in most DNA sequences whatever the organism, the kingdom and the coding or non-coding nature of the sequence under study. To what extent this long-range correlated large-scale regime is universal, is still the subject of current research. Here, we will mainly concentrate our study on the small-scale regime ($10 \lesssim \mathrm{w} \lesssim 100{-}200$ bp), with the specific goal to demonstrate that the LRC observed with the ‘Pnuc’ coding rule provide a rather original signature of the presence of nucleosomes. As control of this ‘nucleosomal hypothesis’, we will systematically investigate a number of eubacterial genomes for possible LRC.

For a sake of simplicity, we will systematically report the results of our wavelet-based LRC analysis using the representation illustrated in Figure 3(b). The data corresponding to different mono-, di- and tri-nucleotide coding rules will be compared on the range of scales $10 \lesssim \mathrm{w} \lesssim 400$ (in bp units). When the curve corresponding to some coding (e.g., ‘GG’ coding) will be missing or cut at very small scales, this will mean that the density of the mono-, di- or tri-nucleotide under consideration (e.g., GG) is too small for the investigation of the LRC properties over this range of scales to make any sense. In order to minimize this possible lack of statistics in the small-scale regime, we will mainly consider the ‘A_{iso}’ (=‘T_{iso}’), ‘G_{iso}’ (=‘C_{iso}’), ‘AA’ (=‘TT’) and ‘GG’ (=‘CC’), instead of the corresponding individual mono- and di-nucleotide codings. We will not carry out a systematic investigation of the statistics of wavelet coefficients in the small-scale regime as

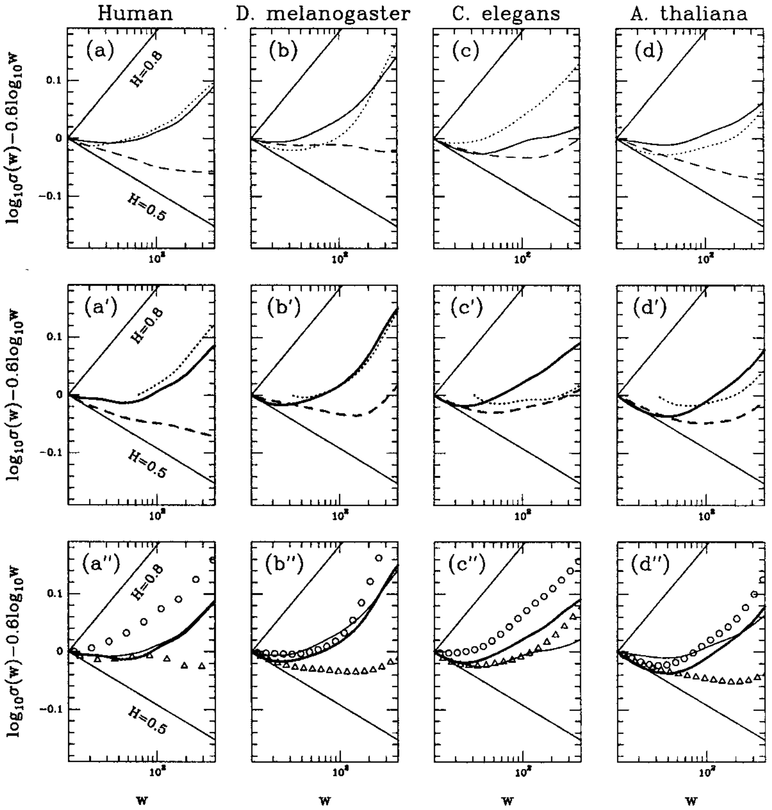

Figure 10. Global estimate of the r.m.s. of WT coefficients of the human chromosome 21 (a-a″), *Drosophila melanogaster* (b-b″), *Caenorhabditis elegans* (c-c″) and *Arabidopsis thaliana* (d-d″) genomes: $\log_{10}\sigma_{WT}(\mathrm{w}) - 0.6\log_{10}\mathrm{w}$ is plotted versus $\log_{10}\mathrm{w}$. The analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). (a-d) Comparative analysis of the ‘A’ (+‘T’) mononucleotide coding (solid line) with the ‘AA’ (=‘TT’) dinucleotide coding (dotted line) and the ‘A_{iso}’ (=‘T_{iso}’) mononucleotide coding (dashed line). (a’-d’) Comparative analysis of the ‘G’ (+‘C’) mononucleotide coding (black solid line) with the ‘GG’ (=‘CC’) dinucleotide coding (black dotted line) and the ‘G_{iso}’ (=‘C_{iso}’) mononucleotide coding (black dashed line). (a″-d″) Comparative analysis of the ‘Pnuc’ coding (circles), the ‘DNase’ coding (triangles) with the ‘A’ (+‘T’) mononucleotide coding (solid line) and the ‘G’ (+‘C’) mononucleotide coding (black solid line). For the coding rules, see Materials and Methods.

illustrated in Figures 7 and 8 for the bacteriophage λ and *S. cerevisiae* respectively (Preliminary results suggest that some departure from Gaussian pdfs may sometimes be observed with some coding rules mainly for eukaryotic genomes).

3.2. EUKARYOTIC GENOMES

In Figure 10 and Table I are reported the results of a wavelet transform analysis of various eukaryotic genomes in the small-scale monofractal scaling regime. In Figure 10 are illustrated the data for the r.m.s. of WT coefficients of the human chromosome 21 (Figure 10(a-a$''$)), *Drosophila melanogaster* (Figure 10(b-b$''$)), *Caenorhabditis elegans* (Figure 10(c-c$''$)) and the *Arabidopsis thaliana* (Figure 10(d-d$''$)) genomic sequences using various mono-, di- and tri-nucleotide codings. The considered analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). In Table I are reported the estimates of the Hurst exponent H when performing a linear regression fit of the data over the range $10 \leq \mathtt{w} \leq 100$. As a first general observation, there exist significant LRC in every examined eukaryotic DNA sequence when using the 'Pnuc' coding rule (Figure 10(a$''$-d$''$)). For example, one gets the following estimate of the Hurst exponent for the eukaryotic sequences illustrated in Figure 10: $H = 0.67 \pm 0.04$ (Human chromosome 21), 0.62 ± 0.03 (*Drosophila melanogaster*), 0.66 ± 0.06 (*Caenorhabditis elegans*) and 0.60 ± 0.07 (*Arabidopsis thaliana*), i.e. values which are all significantly larger than the theoretical prediction $H = 1/2$ for uncorrelated sequences. Some LRC are also observed when using the 'DNase' coding rule in Figure 10(a$''$-d$''$), but they are systematically weaker than those identified with the 'Pnuc' coding rule. For the sake of comparison, one gets the following estimates $H = 0.59 \pm 0.04$ (Human chromosome 21), 0.56 ± 0.03 (*Drosophila melanogaster*), $H = 0.59 \pm 0.05$ (*Caenorhabditis elegans*) and 0.55 ± 0.02 (*Arabidopsis thaliana*). Let us point out that, in contrast to the above observation, the estimates obtained for the yeast genome are quite comparable: $H = 0.54 \pm 0.01$ with the 'Pnuc' coding and $H = 0.54 \pm 0.02$ with the 'DNase' coding.

Another rather general observation is the fact that LRC are also observed with each of the four mononucleotide codings (Figures 10(a$''$-d$''$)). However, quite systematically, the data obtained with the 'Pnuc' coding yield a larger (when they are not in good agreement) estimate of the strength H of the LRC than the values obtained with the mononucleotide codings. This is particularly true in Figure 10(c$''$) for the *Caenorhabditis elegans* genome, where the mononucleotide codings provide results which are in remarkable agreement with the data obtained with the 'DNase' trinucleotide coding but which are definitely smaller than the corresponding 'Pnuc' data. The remarkable feature in this case is that the 'AA' (='TT') dinucleotide coding in Figure 10(c) reproduces quite well the 'Pnuc' data. Indeed the estimate $H = 0.63 \pm 0.05$ is the largest value obtained with this dinucleotide coding overall our set of eukaryotic genome sequences. Let us emphasize that the 'A_{iso}' (='T_{iso}') coding in Figures 10(a-d) strongly deviates from the 'AA' (='TT') coding and fails to account for the strength of the LRC exhibited with the 'Pnuc' coding. It is also rather clear in Figures 10(a'-d'), that the 'G_{iso}' (='C_{iso}') coding does not participate to a large extent to the LRC revealed by the 'Pnuc' coding. Note that when the density of 'GG' (= 'CC') dinucleotides allows us to investigate

LRC, these LRC are quantitatively similar to the ones observed with the 'Pnuc' coding (Figures 10(a'-d')).

As a final observation, let us point out that, as reported in Table I, the main features recognized in the results illustrated in Figure 10 are quite characteristic of other eukaryotic genomes such as warm and cold blooded vertebrates as well as invertebrates and plants. A systematic investigation of the WT coefficient pdfs such as done in Figure 8 for the yeast chromosome 1, confirms a definite change of shape of these pdfs for a characteristic scale $\mathtt{w}_c$ which may be closer to 100 bp for certain organisms than to 200 bp as observed for the yeast chromosomes.

3.3. EUBACTERIAL GENOMES

In Figure 11 and Table I are reported the results of a wavelet transform analysis of the scale-invariance properties of complete eubacterial genomes that belong to the following groups: *Proteobacteria*, *Gram-Positive*, *Spirochaetes*, *Cyanobacteria*, *Thermotogales* and *Chlamydiae*. In Figure 11 are illustrated the data for the r.m.s. of WT coefficients of some selected complete eubacterial genomes that are representative of the results obtained for other genomic sequences in these various groups. The visualized range of scales is the same as for the eukaryotic sequences in Figure 10, as well as the analyzing wavelet $g^{(1)}$. In Table I are reported the estimates of the Hurst exponent H when performing a linear regression fit of the data over the same range of scale as before, i.e., $10 \leq \mathtt{w} \leq 100$. In eubacterial genomes, the characteristic scale $\mathtt{w}_c$ that separates the small-scale and the large-scale monofractal regimes is better defined and slightly greater than what we have observed for common eukaryotic genomes, i.e., $\mathtt{w}_c$ is more likely about 200 bp (see also Figure 3). Note that this scale is about the size of the persistence length of the DNA heteropolymer while the characteristic scale observed for eukaryotic genomes is more compatible to the $100 - 150$ bp long DNA regions which are wrapped around histone proteins to form the eukaryotic nucleosomes [4, 6].

The main observation when examining the data in Figure 11, is that whatever the coding rule used to digitize the eubacterial DNA sequences, one does not observe any evidence of a possible existence of LRC. As one can check quantitatively in Table I, the estimates of the Hurst exponent H all fall in the range $0.48 \lesssim H \lesssim 0.52$ and therefore cannot be distinguished from the canonical value $H = 1/2$ for uncorrelated sequences. Indeed, all the curves in Figure 11 are quite parallel, if not almost superimposed, to the theoritical straight-line corresponding to $H = 1/2$. The results reported in Figure 11(a″-d″), provide a remarkable demonstration that whatever the coding tables used to modelling DNA local bending and flexibility properties, one does not get any footprint of the possible existence of LRC. Both 'Pnuc' and 'DNase' codings yield similar quantitative estimates of $H = 0.50 \pm 0.02$ than those obtained with the four mononucleotide codings. This contrasts with what we have observed for eukaryotic sequences that do contain nucleosomes and that systematically exhibit LRC in the small-scale regime.

Table I. Values of the Hurst exponent H in the small-scale regime. This exponent is estimated using our wavelet based method as described in Theoretical Concepts and Methodology. The numbers in each line correspond to a linear regression fit of $\log_{10} \sigma_{WT}(\mathtt{w})$ versus $\log_{10} \mathtt{w}$, in the $10(20) - 100$ bp range, for the indicated sequence or set of sequences. The error bars are estimated from the fluctuations of the local slope of the data in this range of scales. Each column indicates the coding rule that is used (Materials and Methods). n.a., non attributed, when the statistical sample is not large enough to allow reliable measurements

	PNuc	DNase	AA $(= TT)$	A_{iso} $(= T_{iso})$	A $(+T)$	GG $(= CC)$	G_{iso} $(= C_{iso})$	G $(+C)$
Homo sapiens	0.67 ±0.04	0.59 ±0.03	0.64 ±0.02	0.55 ±0.01	0.61 ±0.03	n.a.	0.54 ±0.02	0.59 ±0.03
Homo sapiens *Introns*	0.68 ±0.02	0.60 ±0.02	0.64 ±0.01	0.54 ±0.02	0.60 ±0.02	n.a.	0.54 ±0.01	0.60 ±0.01
Homo sapiens *Exons*	0.55 ±0.01	0.49 ±0.02	n.a.	0.50 ±0.01	0.53 ±0.01	0.54 ±0.01	0.51 ±0.01	0.54 ±0.01
Exons *GC%*$<$ 50	0.53 ±0.04	0.50 ±0.02	0.50 ±0.05	0.50 ±0.01	0.53 ±0.02	n.a.	0.50 ±0.02	n.a.
Exons *GC%*$<$ 60	0.55 ±0.02	0.50 ±0.01	n.a.	0.52 ±0.02	0.53 ±0.01	0.56 ±0.03	0.53 ±0.03	0.58 ±0.03
Danio rerio	0.61 ±0.05	0.58 ±0.03	0.58 ±0.05	0.58 ±0.03	0.60 ±0.03	n.a.	0.56 ±0.03	0.57 ±0.06
Drosophila *melanogaster*	0.62 ±0.03	0.56 ±0.03	0.60 ±0.05	0.59 ±0.02	0.63 ±0.05	n.a.	0.56 ±0.01	0.61 ±0.06
Caenorhabditis *elegans*	0.66 ±0.06	0.59 ±0.05	0.63 ±0.05	0.56 ±0.02	0.59 ±0.05	n.a.	0.57 ±0.04	0.62 ±0.07
Arabidopsis *thaliana*	0.60 ±0.07	0.55 ±0.02	0.58 ±0.05	0.55 ±0.01	0.60 ±0.03	n.a.	0.54 ±0.02	0.58 ±0.07
Saccharomyces *cerevisiae*	0.54 ±0.01	0.54 ±0.02	0.54 ±0.01	0.55 ±0.01	0.57 ±0.03	n.a.	0.52 ±0.01	0.53 ±0.03
Herpesviridae	0.57 ±0.01	0.52 ±0.01	n.a.	0.53 ±0.01	0.53 ±0.01	0.57 ±0.02	0.53 ±0.01	0.59 ±0.01
Adenoviridae	0.57 ±0.01	0.53 ±0.02	0.55 ±0.04	0.54 ±0.02	0.53 ±0.02	n.a.	0.52 ±0.04	0.54 ±0.02
Melanoplus *sanguinipes*	0.51 ±0.01	0.49 ±0.02	0.50 ±0.02	0.51 ±0.03	0.49 ±0.02	n.a.	0.49 ±0.02	0.51 ±0.01
Vaccinia virus	0.51 ±0.02	0.49 ±0.01	0.51 ±0.02	0.50 ±0.02	0.48 ±0.04	n.a.	0.51 ±0.01	0.48 ±0.01
Positive-strand *ssRNA viruses*	0.53 ±0.01	0.52 ±0.02	0.51 ±0.01	0.53 ±0.01	0.50 ±0.01	n.a.	0.53 ±0.02	0.49 ±0.02
dsRNA viruses	0.55 ±0.01	0.49 ±0.02	n.a.	0.51 ±0.01	0.50 ±0.03	0.53 ±0.03	0.53 ±0.03	0.51 ±0.01

Table I. Continued

	PNuc	DNase	AA $(= TT)$	A_{iso} $(= T_{iso})$	A $(+T)$	GG $(= CC)$	G_{iso} $(= C_{iso})$	G $(+C)$
Human Spumaretrovirus	0.62 0.22	0.50 ±0.02	0.53 ±0.01	0.51 ±0.02	0.49 ±0.01	n.a.	0.62 ±0.04	0.54 ±0.02
Retroviridae	0.57 ±0.03	0.50 ±0.01	0.57 ±0.01	0.51 ±0.02	0.53 ±0.02	0.61 ±0.01	0.56 ±0.01	0.54 ±0.01
Escherichia coli	0.49 ±0.02	0.48 ±0.02	0.50 ±0.02	0.50 ±0.01	0.51 ±0.01	0.50 ±0.01	0.50 ±0.01	0.50 ±0.01
Rickettsia prowazekii	0.54 ±0.03	0.50 ±0.02	0.54 ±0.02	0.52 ±0.01	0.52 ±0.02	n.a.	0.52 ±0.02	0.51 ±0.03
Helicobacter pylori 26695	0.51 ±0.05	0.51 ±0.02	0.52 ±0.04	0.52 ±0.01	0.52 ±0.04	n.a.	0.50 ±0.03	0.55 ±0.03
Chlamydia trachomatis	0.51 ±0.01	0.51 ±0.02	0.52 ±0.01	0.52 ±0.01	0.46 ±0.02	n.a.	0.51 ±0.01	0.49 ±0.01
Treponema pallidum	0.54 ±0.04	0.51 ±0.01	0.50 ±0.01	0.51 ±0.01	0.50 ±0.02	0.49 ±0.03	0.53 ±0.01	0.50 ±0.01
Mycoplasma pneumoniae	0.51 ±0.04	0.51 ±0.03	0.52 ±0.03	0.52 ±0.01	0.52 ±0.02	0.52 ±0.03	0.52 ±0.03	0.52 ±0.01
Bacillus subtilis	0.51 ±0.02	0.49 ±0.01	0.50 ±0.02	0.51 ±0.01	0.52 ±0.01	n.a.	0.50 ±0.02	0.51 ±0.01
Synechocystis PCC6803	0.51 ±0.01	0.47 ±0.02	0.50 ±0.01	0.50 ±0.01	0.50 ±0.01	0.49 ±0.02	0.49 ±0.01	0.51 ±0.01
Thermotoga maritima	0.52 ±0.02	0.49 ±0.01	0.52 ±0.03	0.51 ±0.01	0.51 ±0.02	0.52 ±0.02	0.51 ±0.01	0.50 ±0.02
Aquifex aeolicus	0.57 ±0.04	0.51 ±0.02	0.57 ±0.03	0.51 ±0.01	0.54 ±0.03	0.52 ±0.02	0.51 ±0.01	0.53 ±0.02
Bacteriophage *T4*	0.50 ±0.01	0.50 ±0.01	0.50 ±0.01	0.52 ±0.01	0.49 ±0.01	n.a.	0.53 ±0.02	0.47 ±0.01
Bacteriophage *SPBc2*	0.49 ±0.01	0.50 ±0.03	0.50 ±0.01	0.52 ±0.01	0.51 ±0.01	n.a.	0.51 ±0.01	0.50 ±0.01
Thermoplasma acidophilum	0.56 ±0.03	0.50 ±0.02	0.52 ±0.05	0.50 ±0.01	0.54 ±0.03	0.52 ±0.03	0.51 ±0.01	0.51 ±0.02
Methanococcus jannaschii	0.56 ±0.05	0.52 ±0.03	0.55 ±0.03	0.51 ±0.01	0.55 ±0.04	n.a.	0.51 ±0.02	0.53 ±0.02
Pyrococcus horikoshii	0.52 ±0.04	0.51 ±0.02	0.53 ±0.04	0.51 ±0.01	0.53 ±0.01	0.50 ±0.02	0.51 ±0.01	0.52 ±0.02
Archaeoglobus fulgidus	0.54 ±0.05	0.51 ±0.02	0.52 ±0.05	0.50 ±0.01	0.50 ±0.03	0.50 ±0.02	0.52 ±0.01	0.58 ±0.02
Aeropyrum pernix	0.53 ±0.01	0.51 ±0.01	n.a.	0.50 ±0.01	0.48 ±0.03	0.50 ±0.01	0.51 ±0.01	0.58 ±0.02
Sulfolobus solfataricus	0.54 ±0.03	0.50 ±0.01	0.53 ±0.02	0.51 ±0.01	0.53 ±0.02	n.a.	0.51 ±0.01	0.55 ±0.02

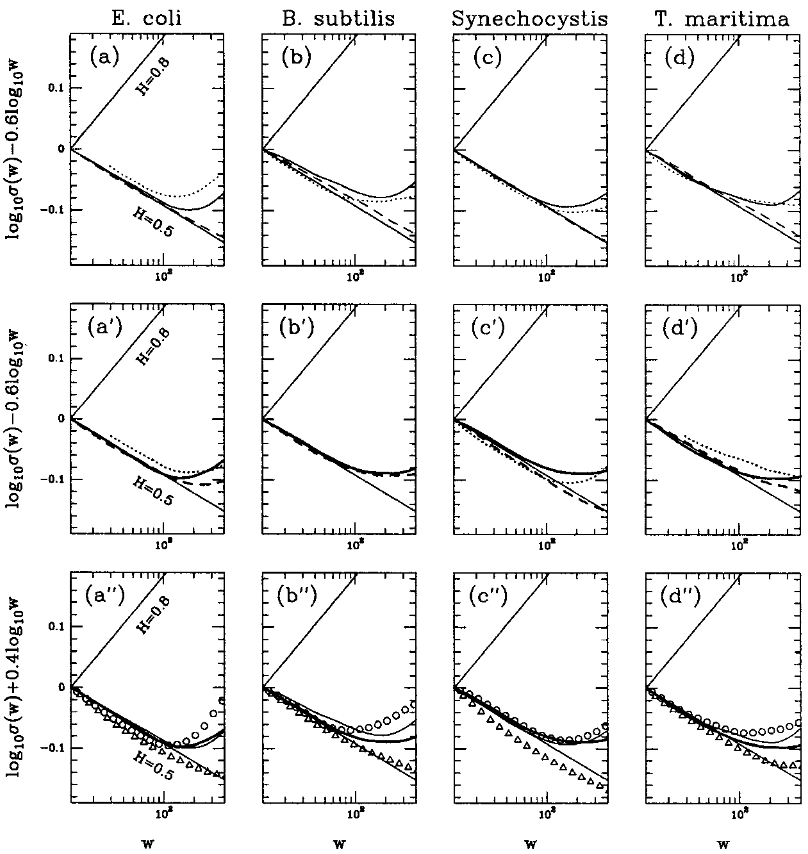

Figure 11. Global estimate of the r.m.s. of WT coefficients of *Escherichia coli* (a-a″), *Bacillus subtilis* (b-b″), *Synechocystis sp. PCC 6803* (c-c″) and *Thermotoga maritima* (d-d″): $\log_{10}\sigma_{WT}(\mathrm{w}) - 0.6\log_{10}\mathrm{w}$ is plotted versus $\log_{10}\mathrm{w}$. The analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). The various curves correspond to the same mono-, di- and tri-nucleotide coding rules as in Figure 10.

Finally, let us mention that two eubacterial genomic sequences (among 29 examined sequences) exhibit rather strong LRC with the ‘Pnuc’ coding (Table I) namely *Buchnera sp.* ($H = 0.59 \pm 0.02$) and *Aquifex aeolicus sp.* ($H = 0.57 \pm 0.04$). In three cases, weaker detectable LRC are identified like for *Rickettsia prowazekii* ($H = 0.54 \pm 0.03$) as reported in Table I.

3.4. VIRAL DNA GENOMES

Most dsDNA eukaryotic viruses replicate in the cell nucleus of their host in which their genomic DNA molecules associate to the host histones to form nucleosomes

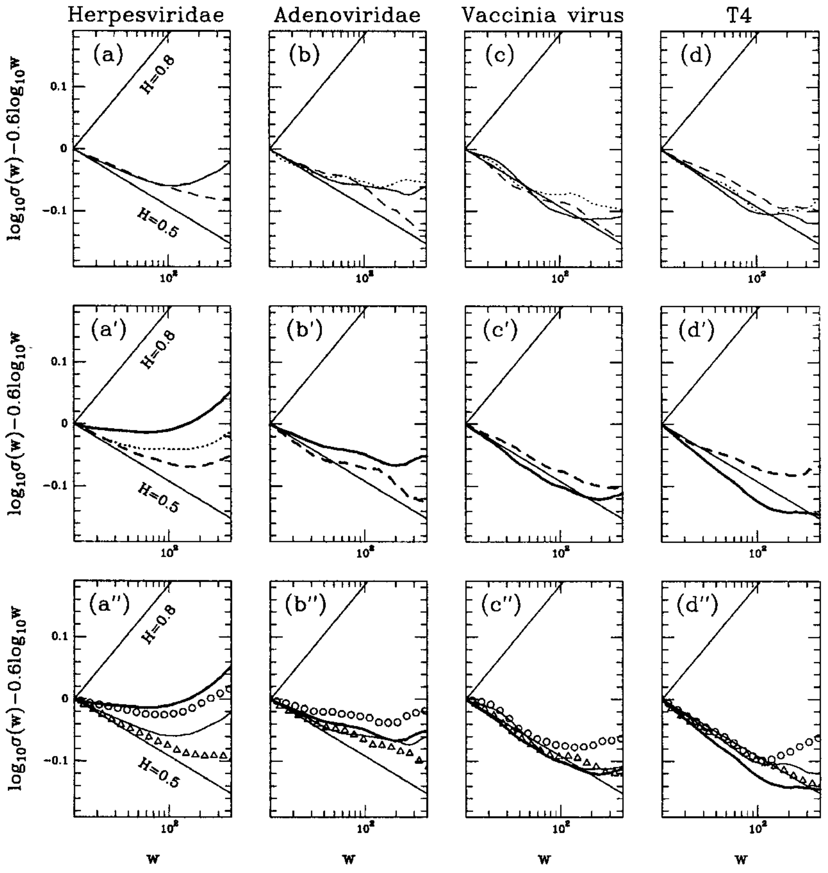

Figure 12. Global estimate of the r.m.s. of WT coefficients of viral DNA genomes: average over 7 complete genomes of *Herpesviridae* (a-a″), average over 3 complete genomes of *Adenoviridae* (b-b″), *Vaccinia virus* (c-c″) and bacteriophage *T4* (d-d″). $\log_{10}\sigma_{WT}(\mathrm{w})-0.6\log_{10}\mathrm{w}$ is plotted versus $\log_{10}\mathrm{w}$. The analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). The various curves correspond to the same mono-, di- and tri-nucleotide coding rules as in Figure 10.

(for a review see [135]). In the line of our previous observations, it can then be expected than the genome sequences of eukaryotic viruses present LRC in the small-scale range. We have performed the wavelet based statistical analysis of a number of dsDNA eukaryotic viruses which are known to form nucleosomes in the cell nucleus, namely *Herpesviruses* [136] and *Adenoviruses* [137]. Small-scale LRC are clearly detected in these genomes as shown in Figures 12(a-a″) and 12(b-b″). In this case, LRC are clearly detected with the ‘Pnuc’ coding rule ($H = 0.57 \pm 0.01$) but not with the ‘DNase’ coding rule ($H = 0.52 \pm 0.02$) as illustrated in Figures 12(a″) and 12(b″). Note that as illustrated in Figures 12(a’) and 12(b'), the LRC observed with the 'Pnuc' coding are quite comparable to the ones exhibited by the 'AA' (= 'TT') and 'GG' (='CC') dinucleotide codings as previously observed for eukaryotic genomes. In particular, one gets the following estimates $H = 0.56 \pm 0.01$ for *Adenoviridae* with the former coding and $H = 0.57 \pm 0.02$ for *Herpesviridae* with the latter coding. We have also investigated the existence of LRC in *Poxviruses*. The *Poxviridae* constitute the only family of animal viruses whose genome does not replicate in the cell nucleus, suggesting that genomic DNA should not associate with the host cell's histones. Indeed, these exhibit H values very close to $1/2$ (see Table I) in the small-scale range as illustrated in Figure 12(c-c$''$) for *Vaccinia virus*. In the case of the prokaryotic DNA viruses, no LRC are detected in the $10-200$ bp range, as examplified by the *T4* and the SPBc2 bacteriophages in Figure 12(d-d$''$) and Table I. Note that no one of the considered mono-, di- or tri-nucleotide codings do exhibit any evidence for LRC; the reported estimates fot the corresponding H values in Table I do not deviate significantly from the value $H = 1/2$ for uncorrelated sequences. These results show that prokaryotic viral sequences present DNA texts, as well as DNA bending profiles identical to those exhibited by their hosts genomes (Figure 11).

3.5. VIRAL RNA GENOMES

We have further extended our wavelet-based analysis to viral RNA genomes. The bending profiles based on the Pnuc and DNases tables present no relevance for single- and double-stranded RNA molecules. However, the analysis has been carried out for RNA genome sequences as for DNA genomes in order to allow numerical comparisons. We have examined several classes of single-stranded plus and minus (data not shown) RNA genomes , as well as double-stranded RNA genomes. The results reveal the absence (or very weak) LRC in these sequences as shown in Figures 13(a-a$''$) and 13(b-b$''$) respectively. The numerical estimates of H in Table I corroborate the visual estimates one can perform by a direct look at the curves in Figures 13; these do not significantly deviate from the $H = 0.5$ straight-line for ssRNAp and this for all codings. For dsRNA, similar estimates are obtained except for the 'Pnuc' coding which exhibits weak LRC with $H = 0.55 \pm 0.01$ (reminds that the Pnuc table has no structural significance for dsRNA molecules). We have also examined, but separately, the case of retroviruses since the retroviral genomes are inserted, as double-stranded DNA, in the host genome. We see respectively in Figure 13(c-c$''$) and Figure 13(d-d$''$) the results obtained for *Spumaretrovirus* and for a group of several distantly related retroviruses. It appears clearly that on the contrary to the other RNA viral genomes, the retroviral sequences exhibit LRC with $H = 0.57 \pm 0.03$ in the $10 - 100$ bp range for the 'Pnuc' coding (which contrasts with their total absence for the 'DNase' coding) as in their host genomic sequences. Significant LRC are also observed with mono- and di-nucleotide codings, although to a less extent with the 'A' (+ 'T') and 'AA' (= 'TT') codings.

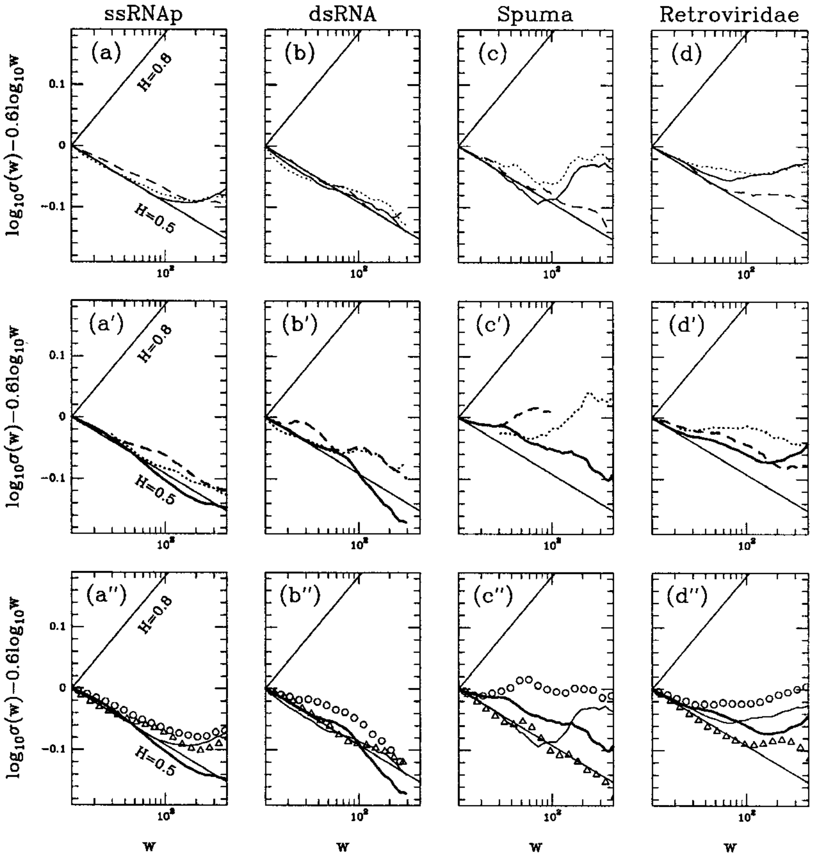

Figure 13. Global estimate of the r.m.s. of WT coefficients of viral RNA genomes: average over 20 complete genomes of positive strand ssRNA viruses (a-a″), average over 4 complete genomes of double strand RNA viruses (b-b″), complete genome of *Spumaretrovirus* (c-c″) and average over 6 complete genomes of retroviruses (d-d″). $\log_{10}\sigma_{WT}(\mathrm{w}) - 0.6\log_{10}\mathrm{w}$ is plotted versus $\log_{10}\mathrm{w}$. The analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). The various curves correspond to the same mono-, di- and tri-nucleotide coding rules as in Figure 10.

3.6. ARCHAEBACTERIAL GENOMES

Histones are known to exist not only in most eukaryotes but also in euryarchaeota, a class of the prokaryotic domain (for a review see [60]). Despite the large difference in euryarchaeotic and eukaryotic genome sizes, it appears that apparently similar architectural motifs function to package DNA in both types of organisms, the archaeal nucleosomes being constituted by histone tetramers [58, 138]. Furthermore, histone packaging of DNA has apparently imposed similar constraints on the genomes of both types of organisms to direct nucleosomes positioning, involving for

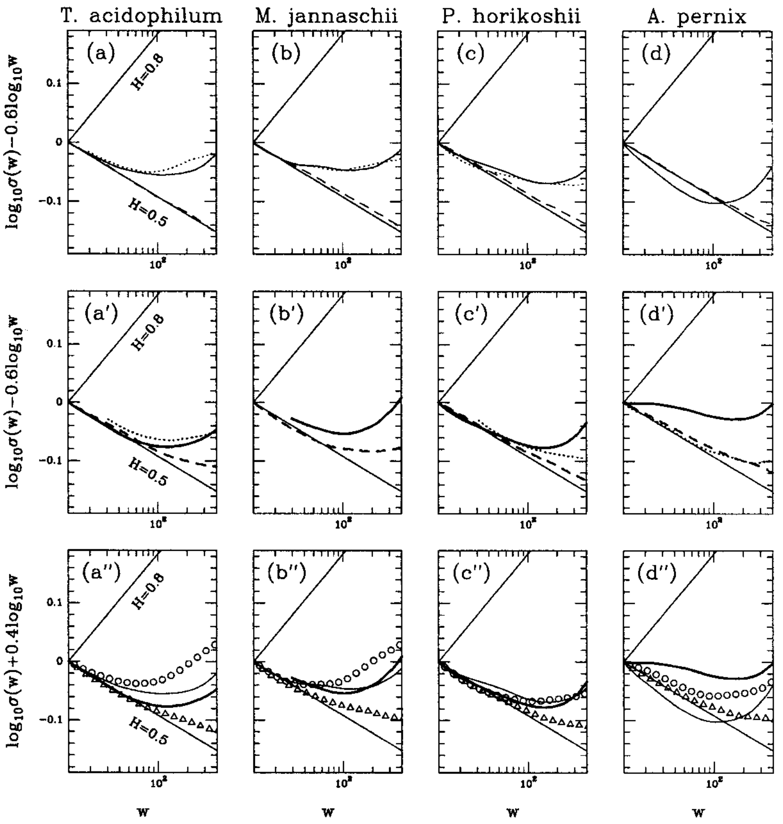

Figure 14. Global estimate of the r.m.s. of WT coefficients of Archaebacterial complete genomes: *Thermoplasma acidophilum* (a-a″), *Methanococcus jannaschii* (b-b″), *Pyrococcus horikoshii* (c-c″) and *Aeropyrum pernix* (d-d″). $\log_{10}\sigma_{WT}(\text{w}) - 0.6\log_{10}\text{w}$ is plotted versus $\log_{10}\text{w}$. The analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). The various curves correspond to the same mono-, di- and tri-nucleotide coding rules as in Figure 10.

example the AA (=TT) dinucleotides [40]. These various observations prompted us to examine the complete genomes of euryarchaeota identified to contain histones (*M. jannaschii*, *P. horikoshii*, *A. fulgidus*) for the presence of LRC in the DNA text, as well as in the 'Pnuc' and dinucleotide bending codings. We have also examined the sequences of one euryarchaeota *T. acidophilum* and of two crenarchaeota (*A. pernix*, *S. solfataricus*) which do not have histones.

As observed in Figure 14(a-c), mild LRC are detected with the 'A' (+ 'T') mononucleotide coding in the genomes of *T. acidophilum*, *M. jannashii* and *P. horikoshii*. Similar LRC are observed with the 'AA' (= 'TT') dinucleotide coding, as well as with the 'Pnuc' coding, but not with the 'A_{iso} (= 'T_{iso}') (see Table I).

For example, for the *M. jannashii* genome, $H = 0.55 \pm 0.04$ with the 'A' (+ 'T') coding, $H = 0.55 \pm 0.03$ with the 'AA' (= 'TT') coding and $H = 0.56 \pm 0.05$ with the 'Pnuc' coding. On the contrary, the mononucleotide 'G' (+ 'C') as well as the 'GG' (= 'CC') and 'G_{iso}' (= 'C_{iso}') codings present a total absence of LRC. For the genome of *A. pernix*, a contrasted situation is observed with the 'G' (+ 'C') coding which presents strong LRC ($H = 0.58 \pm 0.02$) but no LRC with 'GG' (= 'CC') ($H = 0.50 \pm 0.01$). This result differs from what we have observed with the eukaryotic genomes, for which the LRC obtained with 'GG' (= 'CC') are always comparable to the LRC evidenced with the 'G' (+ 'C') coding as seen in Figure 10(a'-d'). On the other hand, we observe little LRC with the 'Pnuc' coding in the *A. pernix* genome ($H = 0.53 \pm 0.01$) which is consistent with the absence of LRC with the 'AA' (= 'TT') and 'GG' (= 'CC') codings. The observation of LRC with the 'PNuc' and 'AA' (= 'TT') codings in euryarchaeotic genomes that contain histones is consistent with the observation that archaeal nucleosome packaging involves sequence regularities similar to those of eukaryotic nucleosomes. However, similar small-scale LRC are also observed in the genome of *T. acidophilum*, which does not contain histones. In addition, the observation of LRC between 'G' (+ 'C') nucleotides, and simultaneously of no LRC between 'GG' (= 'CC') dinucleotides (in *A. pernix*, *S. Solfataricus* and *A. fulgidus*) reveals a new type of correlations which is unprecedented in all eukaryotic and eubacterial genomes examined in this work. Together, these particularities indicate that small-scale LRC in archaebacterial genomic sequences present specific features that remain to be investigated.

3.7. HUMAN INTRONS AND EXONS

We report in Figure 15, the results concerning the human introns and coding exons. As shown in Figure 15(a-a″), LRC are observed for intronic sequences with the 'Pnuc' coding ($H = 0.68 \pm 0.02$) and mainly originate from the LRC induced by the distribution of the dinucleotides AA and TT ($H = 0.64 \pm 0.01$). These data are in remarkable agreement with the results reported in Figure 10(a-a″) for the human chromosome 21 which corroborates the fact that intronic sequences present LRC properties strongly similar to those of intergenic regions (about 80 % of the human genome corresponds to intergenic regions).

We have reproduced this analysis for human coding exons in Figure 15(b-b″). On the contrary to the conclusions of the pioneering statistical analysis of DNA sequences [61, 62, 63, 64, 69, 75, 99], there exist LRC in the human exonic sequences when one considers the 'Pnuc' coding. These LRC are not as strong as in the intronic sequences but they are characterized by an average Hurst value $H = 0.55 \pm 0.01$, a value which is significantly larger than the theoretical prediction $H = 1/2$ for uncorrelated sequences. We have shown in a previous work [123] that the strength H of the LRC observed in both the human introns and exons definitely increases when increasing the (G+C) content of the sequence under study.

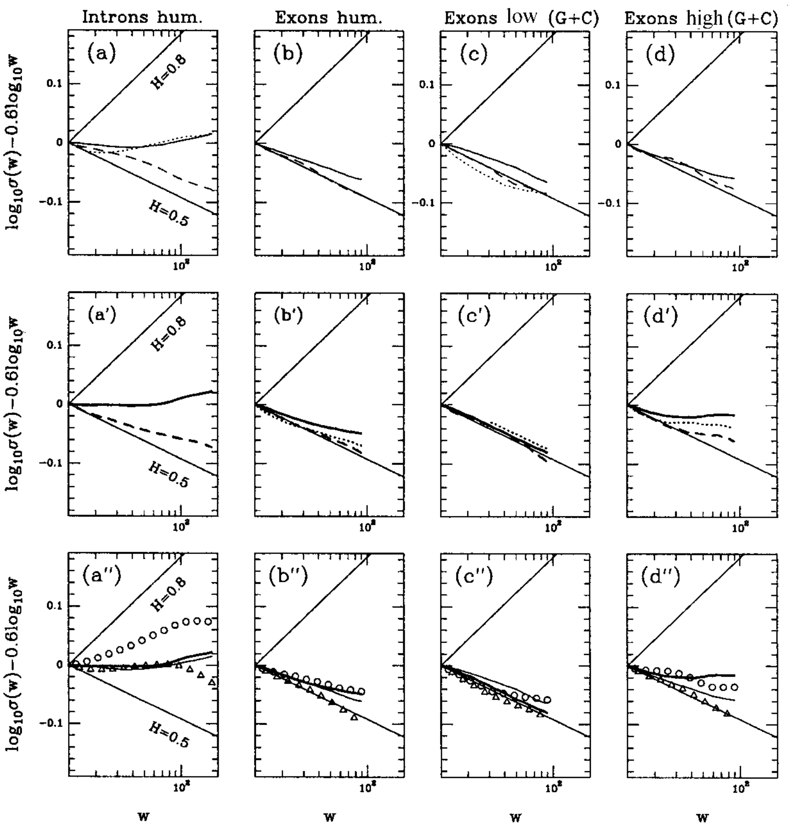

Figure 15. Global estimate of the r.m.s. of WT coefficients of human introns and exons: average over 2184 introns of length larger than 800 bp (a-a″); average over 226 exons of length larger than 600 bp (b-b″); average over 82 exons of length larger than 600 bp and with a G+C content less than 50% (c-c″); average over 73 exons of length larger than 600 bp and with a G+C content larger than 60% (d-d″). $\log_{10}\sigma_{WT}(w) - 0.6\log_{10}w$ is plotted versus $\log_{10}w$. The analyzing wavelet is the first-derivative of the Gaussian function $g^{(1)}$ (Figure 6(d)). The various curves correspond to the same mono-, di- and tri-nucleotide coding rules as in Figure 10.

If one concentrates our WT analysis on the subset of human exons with a G+C content larger than 60% as shown in Figure 15(d″), we observe LRC with the ‘Pnuc’ coding which are slightly larger than those obtained for the exons with a G+C content smaller than 50% (Figure 15(c″)). This effect is strongly enhanced if one considers the ‘G’ (+ ‘C’) and ‘GG’ (= ‘CC’) codings which lead to the Hurst exponent values $H = 0.58 \pm 0.03$ and $H = 0.56 \pm 0.03$ respectively. For exons with a low G+C content, we see in Figure 15(c’), a total absence of LRC when using the same codings. Let us point out that very weak LRC are evidenced in human exons with the ‘A’ (+ ‘T’) and ‘AA’ (= ‘TT’) codings (Figure 15(b-d)). The understanding of the effect of the (G+C) content on the observed LRC is likely to provide new insight into the mechanisms that govern the wrapping of DNA around histones to form nucleosomes. Work in this direction is in current progress.

4. Discussion

4.1. A SMALL-SCALE LRC REGIME IS OBSERVED IN EUKARYOTIC SEQUENCES

The existence and the significance of long-range correlations in genome nucleotide sequences is a long-debated problem which has been examined in a number of previous works on a large variety of DNA sequences [61, 62, 64, 65, 66, 67, 69, 70, 71, 99, 102, 115, 121]. In these studies, nucleotide sequences have been searched for correlations between the individual nucleotides of these sequences (A, G, T, C), or between particular ‘characters’ which can be encoded with these nucleotides (A or G, T or C, etc.). The observed correlations were mainly interpreted in the context of particular DNA informational contents like the coding/non-coding nature of DNA segments (genes, exons) [61, 62, 63, 64, 65, 66, 67, 69, 70, 71, 75, 80, 81, 83, 85, 86, 92, 99], or the presence of particular regularities resulting from the duplication-mutation events associated to genome dynamic [61, 75, 93, 94, 95, 96, 97, 98]. In the present work, we have enlarged the search for LRC in two different directions. First, we have performed systematic genome-wide studies of LRC in complete genomes over scale ranges which were as largely extended as possible (up to thousands of base-pairs) and this in an overview of organisms belonging to the three kingdoms, eukaryotes, eubacteria, archaebacteria, as well as in DNA and RNA viral genomes. Second, we have analyzed these genomes in a new perspective: our aim was to evidence LRC related to structural properties of the DNA molecule [139] involved in the processes of chromatin packaging associated to the various mechanisms of gene expression, and during the successive stages of the cell cycle. This implied not to search genome sequences only for correlations between ‘one-character’ motifs (DNA text), but rather between DNA segments or ‘words’ known to be associated to the structural properties of the DNA double helix. The analyses were performed with the bending profiles obtained by coding the DNA sequences with these structural motifs. They were carried out in parallel with the study of the DNA text and they allowed us to evidence the existence of LRC between the structure-associated DNA words, as well as between mononucleotides.

Among the various properties exhibited by these LRC, the first outstanding feature is the monofractal structure of the signals which allowed us to characterize the LRC in a defined scale range by a single Hurst exponent (see Theoretical Concepts and Methodology). This led us to reveal in all the genomes examined, the existence of a characteristic scale of about $100 - 200$ bp that separates two different regimes of correlations. A first regime spans over a range of about $10 - 200$ bp, which we refer to as the small-scale regime. The second regime (large-scale) extends from about 200 bp to much larger scales depending on the size of the sequence under study. As a general trait, the large-scale regime always presents very large values of H, in general $H > 0.75$, and this with very few exceptions, in all the organisms that we have analyzed in all three kingdoms. On the opposite to this robust 'stability' of the large-scale regime across the diversity of genomes, the small-scale regime presents two different ranges of H values that depend on the class of organisms. Indeed, in the small-scale regime, the eubacterial genomes cannot in general be significantly distinguished from non-correlated sequences characterized by $H = 0.5$. A totally different situation emerges from eukaryotic genomes. These exhibit H values significantly larger than $1/2$, that reveal the existence of LRC. Furthermore, the presence of LRC in eukaryotic genomes and their absence in eubacterial genomes are common features observed with both types of sequence codings, i.e. the coding with single nucleotides, as well as the 'structural coding' with trinucleotides.

4.2. TO WHAT MECHANISMS OF THE EUKARYOTIC CELLS ARE RELATED THE SMALL-SCALE LRC ?

We conjectured that the biological processes that might require the presence of LRC in eukaryotic genomes are related to the structure and the dynamics of the DNA molecule in chromatin. To test for this possibility, we searched for correlations between DNA sequence-dependent motifs that were likely to play a role in the structure of chromatin. In eukaryotes, the first level of compaction of the DNA molecule consists in the solenoidal folding of the DNA molecule around the histone octamer, which is favoured by the distribution of DNA bending sites allowing a proper rotational orientation of the double helix relatively to the histone protein surface [7, 8]. Accordingly, previous works have determined sequence-dependent preferences for the bending of the DNA double helix around the core histones [18, 19, 20, 21, 140, 141]. These allowed to set up a table of the bending values (roll angles) associated to all tri-nucleotides [51], the PNuc table that we used to establish a 'bending profile' of the DNA sequence (see Materials and Methods). These profiles were then examined with the wavelet-transform modulus maxima method (WWTM) to search for the presence of LRC [1, 2, 102, 115] (see Theoretical Concepts and Methodology). As a control, we systematically analyzed the bending profiles obtained with the alternative DNase table [125] based on the cutting of DNA by the DNase I enzyme. It does not present sequence specificity but rather depends on the bending of DNA by the DNase I protein which differs from the bending by histones.

Since these various determinations led to differents sets of bending values, it was of great interest to compare the results of the wavelet-based analyses of the corresponding bending profiles. We observed that for most eukaryotic sequences, the bending profiles obtained with the Pnuc table (nucleosomal DNA) presented H values similar to, or larger than those obtained with the DNA texts; on the opposite, the H values corresponding to the DNase table were significantly smaller (see

Table I). For eubacterial sequences, in the small-scale range, the H values obtained with both the 'Pnuc' and the 'DNase' coding tables were similar to those of uncorrelated sequences. We also measured the H values of bending profiles obtained with other bending tables. In all cases tested, they led to values which were smaller than those obtained with the Pnuc table and generally larger than those obtained with the DNase table (data not shown). The main conclusion from these results is that LRC do exist in the $\sim 10 - 100$ bp range between DNA bending sites in eukaryotic sequences and that these LRC are mostly 'extracted' from the sequence by the 'Pnuc' coding. On the opposite, LRC are poorly detected by the 'DNase' coding, as to some extent by the other types of codings of the DNA curvature.

Taken together, the studies of the eukaryotic and eubacterial genomes strongly suggest that small-scale LRC are related to particular distributions of bending sites in the $\sim$ 150 bp DNA regions which are wrapped around the core histones to form the eukaryotic nucleosomes. This hypothesis can be tested by examining the LRC between individual DNA bending sites that contribute in large part to the bending of nucleosomal DNA, like for instance the AA and GG dinucleotides. We thus examined the LRC in bending profiles obtained with the dinucleotides AA, and compared them to the profiles obtained with the A's nucleotides that are not part of a dinucleotide AA ('A_{iso}') (all A's belong to one and only one of these two subsets). Similar analyses were carried out with GG and G_{iso}. The results show that both AA (= TT) and GG (= CC) dinucleotides do present strong LRC. Furthermore, these are in general close to, or larger than the values measured with the corresponding mono-nucleotide codings. For example, in the case of the human chromosome 21, $H = 0.64 \pm 0.02$ with the 'AA' (= 'TT') coding and $H = 0.61 \pm 0.03$ with the 'A' (+ 'T') coding; similarly $H = 0.59 \pm 0.03$ with the 'G' (+ 'C') coding. Consistant results were obtained with other eukaryotic genomes although the GG distributions could not always be examined by lack of abundance of this di-nucleotide (Figure 10 and Table I). On the opposite, the 'A_{iso}' (= 'T_{iso}') as well as the 'G_{iso}' (= 'C_{iso}') codings revealed significantly weaker LRC, respectively $H = 0.55 \pm 0.01$ and $H = 0.54 \pm 0.02$.

The hypothesis that these LRC are associated to the presence of nucleosomes can be further tested by searching for LRC in viral genomes. The wavelet based analysis of dsDNA eukaryotic viral genomes was thus performed for a number of viruses whose genomic DNA is known to form nucleosomes in the cell nucleus, namely *Herpesviruses* [136] and *Adenoviruses* [137]. We also examined the genomes of *Poxviruses*. These are the only animal viruses that replicate in the cytoplasm, which implicates that their genomic DNA molecule is not expected to form nucleosomes. The results clearly reveal that all the examined viral genomes exhibit the presence of LRC when using the 'Pnuc' coding table (e.g., $H = 0.57 \pm 0.01$ for our set of *Herpesviruses*) to the exception of the *Poxviridae* (e.g., $H = 0.51 \pm 0.02$ for *Vaccinia virus*). In parallel, we also examined the genomes of eubacterial DNA viruses which showed a total absence of LRC (Figure 12(d-d$''$) and Table I). Overall, these results are in remarkable agreement with our hypothesis.

To end up with this overview of complete genomes, we also examined the sequences of viral single strand and double strand RNA genomes. In the line of our hypothesis, these are not expected to exhibit LRC except in the case of the retroviruses since their replication cycle includes the insertion of the double stranded DNA copy of the viral genome into the host genome. This copy of viral DNA is then associated with the host histones to form nucleosomes [142]. As shown in Figure 13, the results of these analyses demonstrate that the examined RNA genomes do not deviate significantly from uncorrelated sequences, except for the retroviral genomes, which again strongly sustains our hypothesis.

Previous work has established the presence of LRC in coding sequences, in particular in human coding exons which present a high (G+C) content [123]. To conclude this study of LRC in biological sequences, we systematically reexamined this question with the various types of codings used here. We see in Figure 15(a-a') that, as expected, the LRC exhibited by the human intronic sequences display similar LRC than the overall genomic sequences for all coding rules (mononucleotides, dinucleotides, Pnuc and DNase). We also see in Figure 15(b') that overall protein coding exons display moderate LRC when using the 'Pnuc' coding. We also confirm that larger LRC are observed for high (G+C) containing exons but interestingly, these are restrained to G and C mononucleotides ($H = 0.58 \pm 0.03$) and GG (= CC) dinucleotides ($H = 0.56 \pm 0.03$), and to the bending profile obtained with the 'Pnuc' coding ($H = 0.55 \pm 0.02$) as shown in Figure 15(d-d'). This observation can be paralleled to the presence of LRC in viral DNA sequences that are more pronounced with the GG (=CC) dinucleotides than with the AA (=TT) dinucleotides. This particularity might be related to the constraints exerted on the evolution of these sequences by their protein coding contents. This result extends the possibility that the formation of nucleosomes in exonic regions can involve sequence patterns that differ from those of intronic regions, as already suggested by Baldi and collaborators [22].

4.3. DO LRC BETWEEN DNA BENDING SITES RESULT FROM A SIMPLE RECODING OF THE DNA TEXT ?

An important point concerns the possibility that the LRC between DNA bending sites might be a trivial observation. In effect, one might argue that since LRC exist between all mono-nucleotides (DNA text), then any arrangement of nucleotides (words) should present as well similar LRC. The analyses with the various bending tables demonstrate that, on the contrary, the choice of particular words can reveal strong LRC, as evidenced with the 'Pnuc' coding rule, while other types of coding do not (e.g., 'DNase' coding). This is further evidenced by the fact that the A nucleotides exhibit strong LRC when they belong to the AA di-nucleotide subgroup, but to a much lesser extent when they belong to the 'isolated A′s' subgroup. Finally, this is also strengthened by the analysis of the DNA profiles obtained with a modified Pnuc coding table. This table is obtained by the shuffling of the Pnuc table and leads to a total vanishing of LRC (Figure 9(b)). These essential results demonstrate that the LRC observed between bending sites are not a trivial consequence of the existence of LRC between single nucleotides. On the opposite, we can propose that the latter should rather be considered as resulting from LRC between bending sites. This does not mean that the Pnuc table allows the exact evaluation of the 'words' which are long-range correlated in all DNA sequences. However, this characterisation of the DNA bending sites issued from the analysis of nucleosomal DNA provides the coding which, among others, most efficiently detects LRC in genomic sequences. Along this line, we notice that the analysis of the *C. elegans* genome with the 'AA" (= 'TT') dinucleotide coding reveals larger H values than with the 'A' (+ 'T') mononucleotide coding (Figure 10(c)). This suggests that the contribution of AA dinucleotides to the formation of nucleosomes is increased in *C. elegans* as compared to other eukaryotic genomes (*Human*, *D. melanogaster*, *A. thaliana*, *S. cerevisiae*). This result can be paralleled with a previous work which showed with the Fourier transform technique, that the spectral component corresponding to AA (= TT) at the 10.2 bp periodicity is strongly enhanced in *C. elegans* comparatively to *S. cerevisiae* [23].

4.4. WHAT MECHANISMS UNDERLY LRC IN GENOME SEQUENCES?

Although the analysis of LRC in genome sequences is still at an early stage, we can tentatively put the grounds of such mechanisms. The perfectly well established structure of nucleosomes dictates that the DNA sequence provides a proper rotational orientation of the double helix around the core histone. It is admitted that among the sequences that favour the formation of nucleosomes, those which contribute significantly to their positioning display a characteristic periodicity of about 10.2 bp, like for example the dinucleotides AA (=TT) and GG (=CC) which are known to play a major role in the intrinsic bending and flexibility properties of the DNA double helix [20, 23, 25, 26, 37, 43, 44, 45, 46, 47, 48]. Actually, it has been estimated that only a small fraction of about 5% of the genome presents this periodic sequence-directed nucleosome positioning properties, that are larger than in the bulk genomic sequences. How sparsely are distributed these specific regions in genomic DNA ? This is still an open question. Periodic signals have been found in coding and non-coding sequences and are not restricted to particular regions as promoters [23]. Indeed, one cannot exclude the possibility that the rather well positioned nucleosomes be concentrated in vast regions leading to the formation of somehow distinct chromatin structures which may facilitate DNA function in a chromatin context, i.e. the functioning of particular genes or loci [24]. Since a large proportion (about 95%) of genomic DNA has a free energy for nucleosome formation that little differs from that of random DNA, one may be tempted to conclude that the DNA sequence has no appreciable influence on nucleosome formation for the vast majority of them. This is probably true as far as nucleosome positioning is concerned. What our analysis strongly suggests is that the LRC observed at small scales ($<$ 200 bp) in eukaryotic genomes are mainly devoted to the formation of the solenoidal supercoils of DNA in nucleosomes. We thus propose that, on the contrary to the tight histone binding obtained with an adequate periodic distribution of bending sites, LRC would allow the major part of the genome to facilitate the left-handed superhelical wrapping of DNA, whatever the positioning of the histone core. The fact that bending sites are long-range correlated means that these sites are more likely spatially distributed on a persistent Cantor set structure as sketched in Figures 1(b), 1(d) and 1(f). This observation brings into light the possibility that the mechanisms underlying the interactions between DNA and histones to form nucleosomes are multi-scale phenomena that involve the interplay of all scales up to $100 - 200$ bp. The presence of LRC between bending sites might not only reflect some mechanical and structural ability of DNA to wrap around histones, but also some propensity of the nucleosomes to be dynamical structures that could favour an optimal compromise between DNA compaction and accessibility constraints. During processes like replication or transcription, the entire length of nucleosomal DNA is exposed (although not necessarily all at once) to the polymerases. Processes of 'site exposure' more rapid than the characteristic time for nucleosome sliding has been presented as an attractive model for the initial binding of regulatory proteins to nucleosomal target sites [31, 143]. The observed LRC between bending sites might play a role in the dynamical DNA peeling off the histone octamer surface as well as in the mechanisms by which the polymerases progress through nucleosomes. In this context, we propose that the LRC would facilitate the translational mobility (sliding) of the nucleosomes [144, 145, 146, 147, 148, 149, 150]. If one considers this mobility as a diffusion mechanism along the DNA molecule, we can assume that the long-range correlated distribution of bending sites exerts a direct effect on this diffusion process. This effect effect might either increase the diffusion coefficient or lead to abnormal diffusion process in which the average rms distance covered after a given number of steps is larger than in classical Brownian motion. LRC between bending sites would thus allow larger nucleosome displacement by 'super-diffusive' processes. This property is reminiscent of the larger black and white segments induced by the presence of LRC in Figures 1(b), 1(d) and 1(f) compared to Figures 1(a), 1(c) and 1(e). The persistent nature of the scale invariant organisation of bending sites would favour the overall dynamic of nucleosomes by allowing them to explore larger DNA fragment. It would also offer an understanding of the modest free energy of nucleosome formation observed for most DNA sequences.

Following this vision of LRC associated to the superhelical states of DNA, we can enlarge the interpretation of our results to the LRC observed at large scales ($>$ 200 bp). This presence of strong large-scale LRC might favour the formation of large solenoidal supercoils that would contribute to the supercoiling of chromatin. In eukaryotes, the LRC observed in the large-scale regime would favour the regular supercoiling of interphase chromatin by condensin [56, 151, 152, 153]. We suggest that to some extent, this mechanism can be paralleled to the way small-scale LRC favour the supercoiling of nucleosomal DNA around the core histone. Along this line, the LRC observed in the large-scale bacterial genomes would, similarly as for eukaryotes, facilitate the supercoiling of DNA to achieve the condensation-decondensation processes of chromatin. These hypotheses constitute new directions for the study of the large-scale LRC and are currently under investigation.

5. Materials and Methods

5.1. CODING DNA SEQUENCES FOR STRUCTURAL ANALYSIS

To apply numerical methods to a DNA sequence $\{n_i\}$ consisting of four nucleotides A, G, T and C, one needs to map the corresponding text on a digital sequence $\{u_i\}$. In previous works [102, 115, 123], we have mainly used the three independent binary mapping rules based on identifying two by two the four bases [72, 109]. For example, the purine-pyrimidine distinction rule amounts to code the purines (A or G) by 1 and the pyrimidines (C or T) by -1. These binary codings have proved to be very convenient to convert DNA sequences into 'DNA walks' using u_i as an incremental variable, the graph of the DNA walks being defined by the cumulative variable $f(k) = \sum_{i=1}^{k} u_i$ [64, 102, 109]. Let us point out that the Hurst exponent H defined in equations (2) and (8) actually characterizes the global regularity properties of the fractal landscape of the graph $f(k)$ of the considered DNA walk [102].

In this work, we use different mapping rules based on the identification of mono-, di- or tri-nucleotides. These codings are actually inspired from the binary coding method extensively used by Voss [65, 67] and which consists in decomposing the nucleotide sequence into four sequences corresponding to A, G, T or C, coding with 1 at the considered nucleotide positions and 0 at the other positions. To investigate the scale-invariance properties of the fluctuations of the local bendability/bending distribution of DNA, we use specific trinucleotide codings which are no longer binary codings since they consist in using the numerical values provided by the Pnuc [51] and DNase [125] tables respectively.

5.1.1. *Mononucleotide Coding Rules*

As mentioned just above, one can define the 'A', 'G', 'T' and 'C' coding rules, by putting 1 at the considered nucleotide positions and 0 at the other positions. These mononucleotide codings allow us to study the way that each nucleotide A, G, T and C is distributed along the DNA sequence.

With the specific goal to compare the statistical distribution of some isolated nucleotides, e.g., the adenines that are not part of a dinucleotide AA, to the overall distribution of the considered nucleotide, e.g. all adenines, we define the binary coding rules 'A_{iso}', 'G_{iso}', 'T_{iso}' and 'C_{iso}'. These rules consist in coding by 1 at the considered nucleotide positions provided the two nearest neighbour nucleotides be different from the considered nucleotide and 0 at the other positions. To improve statistical convergence, we mainly report results obtained with the 'A_{iso}' (='T_{iso}') and 'G_{iso}' (= 'C_{iso}') codings which respectively amount to code with 1 both the isolated A and T on the one hand or both the isolated G and C on the other hand.

5.1.2. *Dinucleotide Coding Rules*

As alternative mapping rules, one can process DNA sequences looking for the respective distributions of each one of the sixteen dinucleotides. In the present study, we report results obtained with particular dinucleotides which are known to participate to the positioning and formation of nucleosomes [26]. The 'AA', 'GG', 'TT' and 'CC' coding rules consist in coding with 1 at the considered nucleotide positions provided at least one of the nearest neighbour nucleotides be the same nucleotide and 0 at the other positions. Note that very much like for the isolated mononucleotide coding rules and for the same reasons, most of the results reported in this work correspond to using the 'AA' (= 'TT') and 'GG' (= 'CC') dinucleotide codings.

5.1.3. *Trinucleotide Coding Rules*

In the context of the present study which is mainly devoted to extracting the structural information which is encoded in the primary DNA sequences, we consider the two trinucleotide coding rules given by the Pnuc and DNase tables reported respectively in [51] and [125]. As previously emphasized, these tables are likely to provide pertinent codings of the local bending and flexibility properties of the DNA double helix. The former is deduced from experimentally determined nucleosome positioning [43]. The later is based on sensitivity of DNA fragments to the enzyme DNase I [52, 154]. The 'Pnuc' and 'DNase' trinucleotide coding rules are thus defined by coding the nucleotide n_i at position i by the numerical value given by either one of these tables for the trinucleotide defined by this nucleotide and its two nearest neighbours, i.e., the triplet (n_{i-1}, n_i, n_{i+1}). A complete coding of DNA sequence is achieved by repeating this operation for all the positions i from 2 to $L - 1$, where L is the overall length of the sequence.

5.2. DATA SETS

All genomes, chromosomes and contigs were down-loaded using either one of the facilities offered at EBI (`http://www.ebi.ac.uk`) or at NCBI (`http://ncbi.nlm.nih.gov`). The following sequences were analyzed, *Homo sapiens* chromosome 21 (from NCBI); *Danio rerio*, AF112374; *Drosophila melanogaster*, AE002602; *Caenorhabditis elegans* chromosome 1 (from NCBI); *Arabidopsis thaliana* chromosome 2, AE002093; *Saccharomyces cerevisiae* 16 chromosomes: U00091, Y13136, Y13137, Y13138, Z71257, Y13139, Y13140, U00094, Y13134, X59720, Z71256, U00092, D50617, Y13135, U00093, Z47047. For the family of *Herpesviruses*, 7 genomes were chosen in the subfamilies of *Alphaherpesviruses*, AJ004801, AF030027, X14112; *Betaherpesviruses*, X17403; *Gammaherpesviruses*, AF005370, V01555; and one unclassified, AB049735. For the *Adenoviruses*, 3 genomes were analyzed: *human adenovirus type 5*, M73260; *ovine adenovirus isolate 287*, U40839; *turkey adenovirus 3*, AF074946. For the *Poxviruses*, 2 genomes were studied: *Vaccinia virus (strain Tan Tan)*, AF095689; *Melanoplus sanguinipes*, AF063866. For positive single strand ssRNA viruses, 20 genomes were studied (all pairs presented less than 50% identity): AF022937, D86371, M87512, M95169, Y10237, U15146, Y07862, X97251, U05771, U38304, U27495, AF029248, M12294, AF039204, AF046869, AF056575, AF094612, X04129, M31182, Y18420. For double strand RNA viruses, 4 genomes were chosen in the families of *Totiviruses*, L13218, AF039080; *Hypoviruses*, AF082191; and *Cystoviruses*, AF226851. For *retroviruses*, one genome was chosen in each of the 7 retrovirus genera: *Lentiviruses*, L07625; *Spumaviruses*, U21247; *Mammalian type B* retroviruses, M15122; *Mammalian type C* retroviruses, M23385; *Avian type C* retroviruses, J02342; *D-type* retroviruses, M12349; *BLV-HTLV* retroviruses, K02120.

The following complete bacterial genomes and virus were analysed: *Escherichia coli*, U00096; *Rickettsia prowazekii*, AJ235269; *Helicobacter pylori*, AE000511; *Chlamydia trachomatis*, AE001273; *Treponema pallidum*, AE000520; *Mycoplasma pneumoniae*, U00089; *Bacillus subtilis*, AL009126; *Synechocystis sp.* PCC6803, AB001339; *Thermotoga maritima*, AE000512; *Aquifex aeolicus*, AE000657; bacteriophage λ, J02459; bacteriophage *T4*, AF158101; bacteriophage *SPBc2*, AF020713; *Thermoplasma acidophilum*, AL139299; *Methanococcus jannashii*, L77117; *Pyroccocus horikoshii* (NCBI); *Archaeoglobus fulgidus*, AE000782; *Aeropyrum pernix*, (NCBI); *Sulfolobus solfataricus* (directly from CBR `http://www.cbr.nrc.ca/`).

For human exons and introns, we also present the results obtained after averaging over several sequences. We extracted all human exons and introns from the EMBL database release 57 and only kept a subset of sequences with a minimum length of 600 bp for exons and 800 bp for introns and a maximum identity of 60%. Moreover, only intron sequences starting with GT and finishing with AG were kept. In the same manner, exon sequences were selected such that they were immediately preceded by AG and followed by GT dinucleotides and with a clearly defined phase (at least one stop codon in exactly two phases). These last criteria avoid selecting the large non coding exons which can be found at both ends of gene sequences and thus guaranties that we work on coding exons only. Overall 226 exons and 2184 introns were retained by this procedure.

Acknowledgements

This research was supported by the GIP GREG (project ‘Motifs dans les Séquences’) and by the Ministère de l’Education Nationale, de l’Enseignement Supérieur, de la

Recherche et de l'Insertion Professionnelle ACC-SV (project 'Génétique et Environnement') and the Action Bioinformatique. B. Audit is currently supported by a Marie Curie Fellowship of the European Community programme 'Improving Human Research Potential and the Socio-economic Knowledge Base' under contract number HPMF-CT-2001-01321.

References

1. Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J.-F. and Arneodo, A.: Long-range Correlations in Genomic DNA: A Signature of the Nucleosomal Structure, *Phys. Rev. Lett.* **86**(2001), 2471–2474.
2. Audit, B., Vaillant, C., Arneodo, A., d'Aubenton-Carafa, Y. and Thermes, C.: Long-Range Correlations between DNA Bending Sites: Relation to the Structure and Dynamics of Nucleosomes, *J. Mol. Biol.* **316**(2002), 903–918.
3. Kornberg, R.D.: Structure of Chromatin, *Annu. Rev. Biochem.* **46** (1977), 931–954.
4. Klug, A., Rhodes, D., Smith, J., Finch, T.J. and Thomas, J.O.: A Low Resolution Structure for the Histone Core of the Nucleosome, *Nature* **287** (1980), 509–516.
5. Richmond, T.J., Finch, J.T., Rusliton, B., Rhodes, D. and Klug, A.: Structure of the Nucleosome Core Particle at 7Å Resolution, *Nature* **311** (1984), 532–537.
6. Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J.: Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution, *Nature* **389**, 251–260 (1997).
7. van Holde, K.E.: *Chromatin*, Springer, New York, 1989.
8. Wolffe, A.P.: *Chromatin Structure and Function*, Academic Press, London, 1995.
9. Finch, J.T. and Klug, A.: Solenoidal Model for Superstructure in Chromatin, *Proc. Natl. Acad. Sci. USA* **73** (1976), 1897–1901.
10. Widom, J. and Klug, A.: Structure of the 300Å Chromatin Filament: X-Ray Diffraction from Oriented Samples, *Cell* **43** (1985), 207–213.
11. Yang, G., Leuba, S.H., Bustamante, C., Zlatanova, J. and van Holde, K.: Role of Linker Histones in Extended Chromatin Fibre Structure, *Nat. Struct. Biol.* **1** (1994), 761–763.
12. Furrer, P., Bednar, J., Dubochet, J., Hamiche, A. and Prunell, A.: DNA at the Entry-Exit of the Nucleosome observed by Cryoelectron Microscopy, *J. P. Struct. Biol.* **114** (1995), 177–183.
13. Woodcock, C.L. and Horowitz, R.A.: Electron Microscopic Imaging of Chromatin with Nucleosome Resolution, *Methods Cell Biol.* **53** (1998), 167–186.
14. Cui, Y. and Bustamante, C.: Pulling a Single Chromatin Fiber reveals the Forces that Maintain its Higher-Order Structure, *Proc. Natl. Acad. Sci. USA* **57** (2000), 127–132.
15. Travers, A.A.: DNA Conformation and Protein Binding, *Annu. Rev. Biochem.* **58** (1989), 427–452.
16. Yao, J., Lowary, P.T. and Widom, J.: Direct Detection of Linker DNA Bending in Defined-Length Oligomers of Chromatin, *Proc. Natl. Acad. Sci. USA* **87** (1990), 7603–7607.
17. van Holde, K. and Zlatanova, J.: What determines the Folding of the Chromatin Fiber ? *Proc. Natl. Acad. Sci. USA* **93** (1996), 10548–10555.
18. Widom, J.: Structure, Dynamics and Function of Chromatin *in vitro*, *Annu. Rev. Biophys. Biomol. Struct.* **27** (1998), 285–327.
19. Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M. and Scipioni, A.: Dual Role of DNA Intrinsic Curvature and Flexibility in Determining Nucleosome Stability, *J. Mol. Biol.* **286** (1999), 1293–1301.
20. Trifonov, E.N. and Sussman, J.L.: The Pitch of Chromatin DNA is reflected in its Nucleotide Sequence, *Proc. Natl. Acad. Sci. USA* **77** (1980), 3816–3820.

21. Drew, H.R. and Travers, A.A.: DNA Bending and its Relation to Nucleosome Positioning, *J. Mol. Biol.* **186** (1985), 773–790.
22. Baldi, P., Brunak, S., Chauvin, Y. and Krogh, A.: Naturally occurring Nucleosome Positioning Signals in Human Exons and Introns, *J. Mol. Biol.* **263** (1996), 503–510.
23. Widom, J.: Short-Range Order in Two Eukaryotic Genomes: Relation to Chromosome Structure, *J. Mol. Biol.* **259** (1996), 579–588.
24. Liu, K., and Stein, A.: DNA Sequence Encodes Information for Nucleosome Array Formation, *J. Mol. Biol.* **270** (1997), 559–573.
25. Herzel, H., Weiss, O. and Trifonov, E.N.: 10-11bp Periodicities in Complete Genomes Reflect Protein Structure and DNA Folding, *Bioinformatics* **15** (1999), 187–193.
26. Thaström, A., Lowary, P.T., Widlund, H.R., Cao, H., Kubista, M. and Widom, J.: Sequence Motifs and Free Energies of Selected Natural and Non-Natural Nucleosome Positioning DNA Sequences, *J. Mol. Biol.* **288** (1999), 213–219.
27. Simpson, R.T.: Nucleosome Positioning: Occurence, Mechanisms and Functional Consequences, *Prog. Nucl. Acid. Res.* **40** (1991), 143–184.
28. Grunstein, M., Durrin, L.K., Mann, R.K., Fisher-Adams, G. and Johnson, L.M.: Histones: Regulators of Transcription in Yeast, In: S. McKnight and K. Yamamoto (eds.), *Transcriptional Regulation,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992, pp. 1295–1315.
29. Wolffe, A.P.: Transcription: In Tune with the Histones, *Cell* **77** (1994), 13–16.
30. Felsenfeld, G.: Chromatin Unfolds, *Cell* **86** (1996), 13–19.
31. Protacio, R.U., Polach, K.J. and Widom, J.: Coupled-Enzymatic Assays for the Rate and Mechanism of DNA Site Exposure in a Nucleosome, *J. Mol. Biol.* **274** (1997), 708–721.
32. Sudarsanam, P. and Winston, F.: The Swi/Snf Family Nucleosome-Remodeling Complexes and Transcriptional Control, *Trends Genet.* **16** (2000), 345–351.
33. Suto, R.K., Clarkson, M.J., Tremethick, D.J. and Luger, K.: Crystal Structure of a Nucleosome Core Particle Containing the Variant Histone H2A.Z, *Nat. Struct. Biol.* **7** (2000), 1121–1124.
34. Wu, J. and Grunstein, M.: 25 Years after the Nucleosome Model: Chromatin Modifications, *Trends Biochem. Sci.* **25** (2000), 619–623.
35. Romero, D., Martinez-Salazar, J., Ortiz, E., Rodriguez, C. and Valencia-Morales, E.: Repeated Sequences in Bacterial Chromosomes and Plasmids: A Glimpse from Sequenced Genomes, *Res. Microbiol.* **150** (1999), 735–743.
36. Gaut, B.S., Le Thierry d'Ennequin, M., Peek, A.S. and Sawkins, M.C.: Maize as a Model for the Evolution of Plant Nuclear Genomes, *Proc. Natl. Acad. Sci. USA* **97** (2000), 7008–7015.
37. Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. and Trifonov, E.N.: Nucleosome DNA Sequence Pattern revealed by Multiple Alignment of Experimentally Mapped Sequences, *J. Mol. Biol.* **262** (1996), 129–139.
38. Trifonov, E.N.: 3-, 10.5-, 200- and 400-Base Periodicities in Genome Sequences, *Physica A* **249** (1998), 511–516.
39. Ioshikhes, I., Trifonov, E. and Zhang, M.: Periodical Distribution Factor Sites in Promoter Regions and Connection with Chromatin Structure, *Proc. Natl. Acad. Sci. USA* **96** (1999), 2891–2895.
40. Bailey, K.A., Pereira, S.L., Widom, J. and Reeve, J.N.: Archaeal Histone Selection of Nucleosome Positioning Sequences and the Procaryotic Origin of Histone-Dependent Genome Evolution, *J. Mol. Biol.* **303** (2000), 25–34.
41. Zhurkin, V.B.: Periodicity in DNA Primary Structure is Defined by Secondary Structure of the Coded Protein, *Nucl. Acids Res.* **9** (1981), 1963–1971.
42. Creighton, T.E.: *Proteins: Structure and Molecular Properties*, Freeman, New York, 1993.
43. Satchwell, S.C., Drew, H.R. and Travers, A.A.: Sequence Periodicities in Chicken Nucleosome core DNA, *J. Mol. Biol.* **191** (1986), 659–675.

44. Ioshikhes, I., Bolshoy, A. and Trifonov, E.N.: Preferred Positions of AA and TT Dinucleotides in Aligned Nucleosomal DNA Sequences, *J. Biomol. Struct. Dynam.* **9** (1992), 1111–1117.

45. Bina, M.: Periodicity of Dinucleotides in Nucleosomes Derived from Simian Virus 40 Chromatin, *J. Mol. Biol.* **235** (1994), 198–208.

46. Staffelbach, H., Koller, T. and Burks, C.: DNA Structure Patterns and Nucleosome Positioning, *J. Biomol. Struct. Dynam.* **12** (1994), 301–325.

47. Bolshoy, A.: CC Dinucleotides Contribute to the Bending of DNA in Chromatin, *Nature Struct. Biol.* **2** (1995), 446–448.

48. Stein, A. and Bina, M.: A Signal Encoded in Vertebrate DNA that Influences Nucleosome Positioning and Alignment, *Nucl. Acids Res.* **27** (1999), 848–853.

49. Trifonov, E.N.: Hidden Segmentation of Protein Sequences: Structural Connection with DNA, In: A. Pullman et al. (eds.), *Modelling of Biomolecular Structures and Mechanisms*, Kluwer, Dordrecht, 1995, pp. 473–479.

50. Trifonov, E.N. and Mengeritsky, G.: Bent DNA in Chromatin versus Force Curved DNA, In: W.K. Olson, M.H. Sarma, R.H. Sarma and M. Sundaraligam (eds.), *Structure and Expression*, Adenine Press, Shenectady, NY, 1988, pp. 159–167.

51. Goodsell, D.S. and Dickerson, R.E.: Bending and Curvature Calculations in B-DNA, *Nucl. Acids Res.* **22** (1994), 5497–5503.

52. Brukner, I., Sanchez, R., Suck, D. and Pongor, S.: Sequence-Dependent Bending Propensity of DNA as revealed by DNase I: Parameters for Trinucleotides, *The EMBO Journal* **14** (1995), 1812–1818.

53. Trifonov, E.N.: Segmented Structure of Separate and Transposable DNA and RNA Elements as Suggested by Their Size Distributions, *J. Biomol. Struct. Dynam.* **14** (1997), 449–457.

54. Vologodskii, A. and Cozzarelli, N.: Conformational and Thermodynamic Properties of Supercoiled DNA, *Annu. Rev. Biophys. Biomol. Struct.* **23** (1994), 609–643.

55. Sawitzke, J. and Austin, S.: Suppression of Chromosome Segregation Defects of *Escherichia coli* Muk Mutants by Mutations in Topoisomerase I, *Proc. Natl. Acad. Sci. USA* **97** (2000), 1671–1676.

56. Holmes, V. and Cozzarelli, N.: Closing the Ring: Links between SMC Proteins and Chromosome Partitioning, Condensation and Supercoiling, *Proc. Natl. Acad. Sci. USA* **97** (2000), 1322–1324.

57. Lopez-Garcia, P. and Forterre, P.: DNA Topology in Hyperthermophilic Archae: Reference States and Their Variation with Growth Phase, Growth Temperature, and Temperature Stresses, *Mol. Microbiol.* **23** (1997), 1267–1279.

58. Decanniere, K., Babu, A., Sandman, K., Reeve, J.N. and Heinemann, U.: Crystal Structures of Recombinant Histones HMfA and HMfB from the Hyperthermophilic Archaeon Methanothermus Fervidus, *J. Mol. Biol.* **303** (2000), 35–47.

59. Musgrave, D., Forterre, P. and Slesarev, A.: Negative Constrained DNA Supercoiling in Archaeal Nucleosomes, *Mol. Microbiol.* **35** (2000), 341–349.

60. Sandman, K. and Reeve, J.N.: Structure and Functional Relationships of Archaeal and Eukaryal Histones and Nucleosomes, *Microbiol.* **173** (2000), 165–169.

61. Li, W.: Generating Non Trivial Long-Range Correlations and $1/f$ Spectra by Replication and Mutation, *Int. J. Bifurc. Chaos* **2** (1992), 137–154.

62. Li, W. and Kaneko, K.: Long-Range Correlation and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence, *Europhys. Lett.* **17** (1992), 655–660.

63. Li, W. and Kaneko, K.: DNA Correlations, *Nature* **360** (1992), 635–636.

64. Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H.E.: Long-Range Correlations in Nucleotide Sequences, *Nature* **356** (1992), 168–170.

65. Voss, R.F.: Evolution of Long-Range Fractal Correlations and $1/f$ Noise in DNA Base Sequences, *Phys. Rev. Lett.* **68** (1992), 3805–3808.

66. Voss, R.F.: Long-Range Power-Law Correlations in DNA (reply), *Phys. Rev. Lett.* **71** (1993), 1777–1777.
67. Voss, R.F.: Long-Range Fractal Correlations in DNA Introns and Exons, *Fractals* **2** (1994), 1–6.
68. Borštnik, B., Pumpernik, D. and Lukman, D.: Analysis of Apparent $1/f^{\alpha}$ Spectrum in DNA Sequences, *Europhys. Lett.* **23** (1993), 389–394.
69. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M., Sciortino, F. and Stanley, H.E.: Long-Range Power-Law Correlations in DNA (comment), *Phys. Rev. Lett.* **71** (1993), 1776–1776.
70. Azbel', M.Y.: Universality in a DNA Statistical Structure, *Phys. Rev. Lett.* **75** (1995), 168–171.
71. Herzel, H. and Grosse, I.: Measuring Correlations in Symbol Sequence, *Physica A* **216** (1995), 518–542.
72. Gates, M.A.: A Simple Way to Look at DNA, *J. Theor. Biol.* **119** (1986), 319–328.
73. Jeffrey, H.J.: Chaos Game Representation of Gene Structure, *Nucl. Acids Res.* **18** (1990), 2163–2170.
74. Berthelsen, C.L., Glazier, J.A. and Skolnick, M.H.: Global Fractal Dimension of Human DNA Sequences treated as Pseudorandom Walks, *Phys. Rev. A* **45** (1992), 8902–8913.
75. Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Ossadnik, S.M., Peng, C.-K. and Simons, M.: Fractal Landscapes in Biological Systems, *Fractals* **1** (1993), 283–301.
76. Solovyev, V.V., Korolev, S.V. and Lim, H.A.: A New Approach for the Classification of the Functional Regions of DNA Sequences based on Fractal Representation, *Int. J. Gen. Res.* **1** (1993), 109–128.
77. Li, W.: Mutual Information Function versus Correlation Functions, *J. Stat. Phys.* **60** (1990), 823–837.
78. Herzel, H. and Grosse, I.: Correlations in DNA Sequences: The Role of Protein Coding Segments, *Phys. Rev. E* **55** (1997), 800–810.
79. Nee, S.: Uncorrelated DNA Walks, *Nature* **357** (1992), 450–450.
80. Prabhu, V.V. and Claverie, J.-M.: Correlations in Intronless DNA, *Nature* **357** (1992), 782–782.
81. Munson, P.J., Taylor, R.C. and Michaels, G.S.: DNA Correlations, *Nature* **360** (1992), 635–636.
82. Karlin, S. and Brendel, V.: Patchiness and Correlations in DNA Sequences, *Science* **259** (1993), 677–679.
83. Chatzidimitriou-Dreismann, C.A. and Larhammar, D.: Long-Range Correlations in DNA, *Nature* **361** (1993), 212–213.
84. Larhammar, D. and Chatzidimitriou-Dreismann, C.A.: Biological Origins of Long-Range Correlations and Compositional Variations in DNA, *Nucl. Acids Res.* **21** (1993), 5167–5170.
85. Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H.E.: Linguistic Features of Noncoding Sequences, *Phys. Rev. Lett.* **73** (1994), 3169–3172.
86. Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H.E.: Systematic Analysis of Coding and Noncoding DNA Sequences using Methods of Statistical Linguistics, *Phys. Rev. E* **52** (1995), 2939–2950.
87. Havlin, S., Buldyrev, S.V., Goldberger, A.L., Mantegna, R.N., Peng, C.-K., Simons, M. and Stanley, H.E.: Statistical and Linguistic Features of DNA Sequences, *Fractals* **3** (1995), 269–284.
88. Herzel, H., Ebeling W. and Schmitt, A.O.: Entropies of Biosequences: The Role of Repeats, *Phys. Rev. E* **50** (1994), 5061–5071.
89. Bernaola-Galván, P., Román-Roldán, R. and Oliver, J.L.: Compositional Segmentation and Long-Range Fractal Correlations in DNA Sequences, *Phys. Rev. E* **53** (1996), 5181–5189.

90. Li, W.: The Measure of Compositional Heterogeneity in DNA Sequences is related to Measures of Complexity, *Complexity* **3** (1997), 33–37.
91. Román-Roldán, R., Bernaola-Galván, P. and Oliver, J.L.: Sequence Compositional Complexity of DNA through an Entropic Segmentation Method, *Phys. Rev. Lett.* **80** (1998), 1344–1347.
92. Borštnik, B.: The Character of the Correlations in DNA Sequences, *Int. J. of Quantum Chemistry* **52** (1994), 457–463.
93. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H.E.: Generalized Lévy-Walk Model for DNA Nucleotide Sequences, *Phys. Rev. E* **47** (1993), 4514–4523.
94. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Stanley, H.E., Stanley, M.H.R. Simons, M.: Fractal Landscapes and Molecular Evolution: Modeling the Myosin Heavy Chain Gene Family, *Biophys. J.* **65** (1993), 2673–2679.
95. Li, W., Marr, T.G. and Kaneko, K.: Understanding Long-Range Correlations in DNA Sequences, *Physica D* **75** (1994), 392–416.
96. Li, W.: The Study of Correlation Structure of DNA Sequences: A Critical Review, *Comp. Chem.* **21** (1997), 257–272.
97. Dokholyan, N.V., Buldyrev, S.V., Havlin, S. and Stanley, H.E.: Model of Unequal Chromosomal Crossing over in DNA Sequences, *Physica A* **249** (1998), 594–599.
98. Provata, A.: Random Aggregation Models for the Formation and Evolution of Coding and Non-Coding DNA, *Physica A* **264** (1999), 570–580.
99. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K. and Stanley, H.E.: Fractals in Biology and Medecine: from DNA to Heartbeat, In: A. Bunde and S. Havlin (eds.), *Fractals in Science*, Springer-Verlag, Berlin, 1994, pp. 49–87.
100. Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Simons, M. and Stanley, H.E.: Finite-Size Effects on Long-Range Correlations: Implications for Analysing DNA Sequences, *Phys. Rev. E* **47** (1993), 3730–3733.
101. Berthelsen, C.L., Glazier, J.A. and Raghavachari, S.: Effective Multifractal Spectrum of a Random Walk, *Phys. Rev. E* **49** (1994), 1860–1864.
102. Arneodo, A., d'Aubenton-Carafa, Y., Bacry, E., Graves, P.V., Muzy, J.-F. and Thermes, C.: Wavelet based Fractal Analysis of DNA Sequences, *Physica D* **96** (1996), 291–320.
103. Gardiner, K.: Base Composition and Gene Distribution: Critical Patterns in Mammalian Genome Organization, *Trends Genet.* **12** (1996), 519–524.
104. Barakat, A., Matassi, G. and Bernardi, G.: Distribution of Genes in the Genome of *Arabidopsis thaliana* and its Implications for the Genome Organization of Plants, *Proc. Natl. Acad. Sci. USA* **95** (1998), 10044–10049.
105. Li, W., Stolovitzky, G., Bernaola-Galván, P. and Oliver, J.L.: Compositional Heterogeneity within, and Uniformity between, DNA Sequences of Yeast Chromosomes, *Genome Research* **8** (1998), 916–928.
106. Bernardi, G.: Isochores and the Evolutionary Genomics of Vertebrates, *Gene* **241** (2000), 3–17.
107. Viswanathan, G.M., Buldyrev, S.V., Havlin, S. and Stanley, H.E.: Long-Range Correlation Measures for Quantifying Patchiness: Deviations from Uniform Power-Law Scaling in Genomic DNA, *Physica A* **249** (1998), 581–586.
108. Peng, C.-K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E. and Goldberger, A.L.: Mosaic Organization of DNA Nucleotides, *Phys. Rev. E* **49** (1994), 1685–1689.
109. Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E. Peng, C.-K., Simons, M. and Stanley, H.E.: Long-Range Correlation Properties of Coding and Noncoding DNA Sequences: GenBank Analysis, *Phys. Rev. E* **51** (1995), 5084–5091.
110. Daubechies, I.: *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
111. Meyer, Y. (ed.): *Wavelets and their Applications*, Springer, Berlin, 1992.

112. Meyer, Y. and Roques, S. (eds.): *Progress in Wavelets Analysis and Applications*, Editions frontières, Gif-sur-Yvette, 1993.

113. Arneodo, A., Argoul, F., Bacry, E., Elezgaray, J. and Muzy, J.-F.: *Ondelettes Multifractales et Turbulences: de l'ADN aux croissances cristallines*, Diderot Editeur, Arts et Sciences, Paris, 1995.

114. Mallat, S.: *A Wavelet Tour of Signal Processing*, Academic Press, New York, 1998.

115. Arneodo, A., Bacry, E., Graves, P.V. and Muzy, J.-F.: Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis, *Phys. Rev. Lett.* **74** (1995), 3293–3296.

116. Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C. and Marcourt, L.: Fourier and Wavelet Transform Analysis, a Tool for Visualizing Regular Patterns in DNA Sequences, *J. Theor. Biol.* **206** (2000), 323–326.

117. Holschneider, M.: On the Wavelet Transform of Fractal Objects, *J. Stat. Phys.* **50** (1988), 963–993.

118. Arneodo, A., Grasseau, G. and Holschneider, M.: Wavelet Transform of Multifractals, *Phys. Rev. Lett.* **61** (1988), 2281–2284.

119. Muzy, J.-F., Bacry, E. and Arneodo, A.: The Multifractal Formalism Revisited with Wavelets, *Int. J. Bifurc. Chaos* **4** (1994), 245–302.

120. Arneodo, A., Bacry, E. and Muzy, J.-F.: The Thermodynamics of Fractals Revisited with Wavelets, *Physica A* **213** (1995), 232–275.

121. Arneodo, A., d'Aubenton-Carafa, Y., Audit, B., Bacry, E., Muzy, J.-F. and Thermes, C.: What can we Learn with Wavelets about DNA Sequences? *Physica A* **249** (1998), 439–448.

122. Muzy, J.-F., Bacry, E. and Arneodo, A.: Wavelets and Mulifractal Formalism for Singular Signals: Application to Turbulence Data, *Phys. Rev. Lett.* **(1991) 67**, 3515–3518.

123. Arneodo, A., d'Aubenton-Carafa, Y., Audit, B., Bacry, E., Muzy, J.-F. and Thermes, C.: Nucleotide Composition Effects on the Long-Range Correlations in Human Genes, *Eur. Phys. J. B* **1** (1998), 259–263.

124. Yeramian, E.: Gene and the Physics of the DNA Double-Helix, *Gene* **255** (2000), 139–150.

125. Gabrielian, A. and Pongor, S.: Correlation of Intrinsic DNA Curvature with DNA Property Periodicity, *FEBS Letters* **393** (1996), 65–68.

126. Taqqu, M.S., Teverovsky, V. and Willinger, W.: Estimator for Long-Range Dependence: An Empirical Study, *Fractals* **3** (1995), 785–798.

127. Audit, B., Bacry, E., Muzy, J.-F. and Arneodo, A.: Wavelet-Based Estimators of Scaling Behavior, *IEEE Trans. Info. Theory* **48** (2000), 2938–2954.

128. Mandelbrot, B.B.: *The Fractal Geometry of Nature*, Freeman and Co., San Franscisco, 1982.

129. Peitgen, H.-O. and Saupe, D. (eds.): *The Science of Fractal Images*, Springer Verlag, New York, 1987.

130. Weir, B.S.: *Genetic Data Analysis, Methods for Discrete Population Genetic Data*, Sinauer Associates Inc. Publishers, Sunderland, Massachusets, 1990.

131. Torresani, B.: *Analyse Continue par Ondelettes*, Editions de Physique, Les Ulis, 1998.

132. Goupillaud, P., Grossmann, A. and Morlet, J.: Cycle-Octave and Related Transforms in Seismic Signal Analysis, *Geoexploration* **23** (1984), 85–102.

133. Grossmann, A. and Morlet, J.: Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape, *S.I.A.M. J. of Math. Anal.* **15** (1984), 723–736.

134. Grossmann, A. and Morlet, J.: In: L. Streit (ed.), *Mathematics and Physics, Lectures on Recent Results*, World Scientific, Singapore, 1985.

135. Kelly, J.: Animal Virus Replication, *Annu. Rev. Biochem.* **58** (1989), 671–717.

136. Deshmane, S.L. and Fraser, N.W.: During Latency, Herpes Simplex Virus Type 1 DNA is associated with Nucleosomes in a Chromatin Structure, *J. Virol.* **63** (1989), 943–947.

137. Marcus-Sekura, C.J. and Carter, B.J.: Chromatin-Like Structure of Adeno-Associated Virus DNA in Infected Cells, *J. Virol.* **48** (1983), 79–87.

138. Pereira, S.L., Grayling, R.A., Lurz, R. and Reeve, J.N.: Archael Nucleosomes, *Proc. Natl. Acad. Sci. USA* **94** (1997), 12633–12637.

139. Grosberg, A., Rabin, Y., Havlin, S. and Neer, A.: Crumpled Globule Model of the Three-Dimensional Structure of DNA, *Europhys. Lett.* **23** (1993), 373–378.

140. Gabrielian, A., Simoncsits, A. and Pongor, S.: Distribution of Bending Propensity in DNA Sequences, *FEBS Letters* **393** (1996), 124–130.

141. Lowary, P.T. and Widom, J.: Nucleosome Packaging and Nucleosome Positioning of Genomic DNA, *Proc. Natl. Acad. Sci. USA* **94** (1997), 1183–1188.

142. Meisterernst, M., Horikoshi, M. and Roeder, R.G.: Chromatin Disruption in the Promoter of Human Immunodeficiency Virus Type 1 during Transcriptional Activation, *EMBO J.* **12** (1993), 3249–3259.

143. Polach, K.J. and Widom, J.: Mechanism of Protein Access to Specific DNA Sequences in Chromatin: A Dynamic Equilibrium Model for Gene Regulation, *J. Mol. Biol.* **254** (1995), 130–149.

144. Spadafora, C., Oudet, P. and Chambon, P.: Rearrangement of Chromatin Structure Induced by Increasing Ionic Strength and Temperature, *Eur. J. Biochem.* **100** (1979), 225–235.

145. Pennings, S., Meersseman, G. and Bradbury, E.M.: Mobility of Positioned Nucleosomes on 5 S rDNA, *J. Mol. Biol.* **220** (1991), 101–110.

146. Meersseman, G., Pennings, S. and Bradbury, E.M.: Mobile Nucleosomes – A General Behavior, *EMBO J.* **11** (1992), 2951–2959.

147. Ura, K., Hayes, J.J. and Wolffe, A.P.: A Positive Role for Nucleosome Mobility in the Transcriptional Activity of Chromatin Templates: Restriction by Linker Histones, *EMBO J.* **14** (1995), 3752–3765.

148. Hamiche, A., Sandaltzopoulos, R., Gdula, D.A. and Wu, C.: ATP-Dependent Histone Octamer Sliding Mediated by the Chromatin Remodeling Complex NURF, *Cell* **97** (1999), 833–842.

149. Langst, G., Bonte, E.J., Corona, D.F. and Becker, P.B.: Nucleosome Movement by CHRAC and ISWI without Disruption or Trans-Displacement of the Histone Octamer, *Cell* **97** (1999), 843–852.

150. Whitehouse, I., Flaus, A., Cairns, B.R., White, M., Workman, J.L. and Owen-Hughes, T.: Nucleosome Mobilization Catalysed by the Yeast SWI/SNF Complex, *Nature* **400** (1999), 784–787.

151. Hirano, T. and Mitchison, T.J.: A Heterodimeric Coiled-Coil Protein Required for Mitotic Chromosome Condensation *in vitro*, *Cell* **79** (1994), 449–458.

152. Kimura, K. and Hirano, T.: ATP-Dependent Positive Supercoiling of DNA by 13S Condensin: A Biochemical Implication for Chromosome Condensation, *Cell* **90** (1997), 625–634.

153. Kimura, K., Rybenkov, V.V., Crisona, N.J., Hirano, T. and Cozzarelli, N.R.: 13S Condensin Actively Reconfigures DNA by Introducing Global Positive Writhe: Implications for Chromosome Condensation, *Cell* **98** (1999), 239–248.

154. Brukner, I., Sanchez, R., Suck, D. and Pongor, S.: Trinucleotide Models for DNA Bending Propensity: Comparison of Models Based on DNase I Digestion and Nucleosome Packaging Data, *J. Biomol. Struct. Dynam.* **13** (1995), 309–317.