doi:10.1006/jmbi.2001.5363 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **316**, 903–918

JMB

Long-range Correlations between DNA Bending Sites: Relation to the Structure and Dynamics of Nucleosomes

Benjamin Audit[1], Cedric Vaillant[1], Alain Arneodo[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, CNRS, Laboratoire associé à l'Université Pierre et Marie Curie, Avenue de la Terrasse, 91198 Gif-sur-Yvette France*

**Corresponding author*

It has been established that the precise positioning of nucleosomes on genomic DNA can be achieved, at least for a minority of them, through sequence-dependent processes. However, to what extent DNA sequences play a role in the positioning of the major part of nucleosomes is still debated. The aim of the present study is to examine to what extent long-range correlations (LRC) are related to the presence of nucleosomes. 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. The exploration of a number of eukaryotic and bacterial genomes 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. Here, we focus on the existence of LRC in the small-scale regime (10-200 bp) which are actually observed in eukaryotic genomes, in contrast to their absence in eubacterial genomes. Analysis of viral DNA genomes shows that, like their host's genomes, eukaryotic viruses present LRC but eubacterial viruses do not. There is one exception for genomes of poxviruses (*Vaccinia* and *Melamoplus sanguinipes*) which do not replicate in the cell nucleus and do not exhibit LRC. No small-scale LRC are detected in the genomes of all examined RNA viruses, with the exception of retroviruses. These results together with the observation of LRC between particular sequence motifs known to participate in the formation of nucleosomes (e.g. AA dinucleotides) strongly suggest that the 10-200 bp LRC are a signature of the sequence-dependence of nucleosome positioning. Finally, we discuss possible interpretations of these LRC in terms of the physical mechanisms that might govern the positioning and the dynamics of the nucleosomes along the DNA chain through cooperative processes.

Keywords: chromatin; nucleosome positioning; long-range correlations; fractals; scale-invariance

Introduction

The genomic DNA of eukaryotic cells is tightly packaged into nucleosomes which constitute the basic units of chromatin. Each nucleosome consists of almost two turns of DNA wrapped around an octamer of core histone proteins.[1] An additional fragment of DNA associated with a linker histone separates successive nucleosomes which are disposed as beads-on-a-string along the DNA.[2] This motif is further organized into successive higher-order structures that involve large-scale ranges and reach a full extent of condensation in metaphase chromosomes. Although these structures are still

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

Abbreviations used: LRC, long-range correlations; WT, wavelet transform; WTMM, wavelet transform modulus maxima; FARIMA, fractional auto-regressive integrated moving average process; ss, single-stranded; ds, double-stranded.

E-mail address of the corresponding author: thermes@cgm.cnrs-gif.fr

0022-2836/02/040903–16 $35.00/0

poorly understood, it is established that they play crucial roles in gene regulation and cell division.[3,4] Since the discovery of naturally bent DNA,[5] it has been proposed that certain nucleotide motifs help DNA folding around the core histones, suggesting that the nucleosome packaging can be facilitated by proper DNA sequences.[6–8] Moreover, there is an increasing number of examples of nucleosomes positioned at specific sites where they constitute essential elements of the regulation of gene expression, and this precise positioning is also likely to be encoded in the DNA sequence. However, can we consider that apart from these definite regions, the majority of genomic sequences evolved to favour a compact arrangement of the DNA molecule? In other words, are the bulk of the nucleosome arrays that one observes in the cell nucleus, organized *via* sequence dependent processes? This question has been investigated over many years and has received controversial answers.

Among the results indicating that nucleosome formation is facilitated by DNA sequences, analyses of aligned fragments of nucleosomal DNA have exhibited dinucleotide motifs presenting a 10 bp periodicity.[9–17] The periodic positioning of these motifs contributes in a coherent manner to a global curvature of DNA which amplifies the affinity for the histone octamer and therefore favours the wrapping of DNA on the histone surface.[18,19] This is for example the case for the AA dinucleotide which confers anisotropic flexibility to the DNA double helix. Studies focused on short DNA regulatory regions have pointed to the importance of sequence-directed nucleosome positioning. These mainly involved the LTR promoter of the mouse mammary tumour virus, the *Xenopus* 5 S rRNA genes and several genes of *Tetrahymena* (for a review, see Zlakanova *et al*.[4]). *In vitro* chromatin reconstitution experiments performed with 2-3 kb vertebrate DNA fragments have revealed that nucleosomes tend to form highly ordered physiologically spaced arrays on some of these fragments but not on others, in agreement with *in vivo* localization of nucleosomes on the corresponding DNA regions.[20] Other studies of the native SV40 chromatin have shown that in the late region of the SV40 genome, nucleosomes tend to form in segments that contain high concentrations of the 10 bp periodic trinucleotides (non-T)(A/T)G and to avoid regions where the concentration of this motif is low.[16] Overall, the result from long series of *in vitro* and *in vivo* studies is that DNA sequences can contribute to their own packaging process.

However, other results have shed a different light on this conclusion. Competitive nucleosome reconstitution assays performed with genomic and random synthetic DNA fragments have shown that the vast majority of the genomic sequences (>95 %) present binding affinities for core histones that do not differ significantly from those of the random fragments.[21] This led the authors to conclude that for most nucleosomes, the DNA sequence does not play an important role in their positioning. In a synthetic view, they propose that for one part, a minority of nucleosome packaging and positioning signals are concentrated into a small subset of the genome where they contribute to position the nucleosomes in preferential locations. For the other part, the majority of signals would be sparsely but rather uniformly distributed along regions that represent more than 95 % of the genome, where they poorly contribute to their own packaging at the level of individual nucleosomes. Does this vast set of signals which apparently do not differ significantly from random sequences present any biological significance? It is possible that they simply constitute an ''incoherent noise'' with no biological role. Alternatively, it can be conjectured that acting ''collectively'' or ''coherently'', they may favour the formation of entire arrays of nucleosomes. These ''noisy'' signals would then play a significant role in nucleosome formation and positioning.

Along the lines of this last hypothesis, how could such a mechanism be achieved? More precisely, how could the distribution of these sites, which are apparently dispersed at random along the DNA, facilitate a collective organisation of nucleosomes, leading ultimately to their arrangement into regular arrays. Such distributions would indeed differ from the 10 bp periodic patterns previously worked out by Fourier and correlation function analyses that result in the tight binding of nucleosomes (larger than those of the random fragments). A recent work has shown that in eukaryotic genomes, sequence motifs that favour the formation and positioning of nucleosomes present particular correlation properties.[22] These correlations, which are referred to as long-range correlations (LRC), extend over large distances and are related to the scale-invariant properties of the DNA sequences. The LRC are identified and quantified using a space-scale analysis performed by the so-called wavelet transform.[23] Here, our aim is to use this technique to determine to what extent the LRC observed in the 10 to $\approx$200 nucleotides range, can be considered as a signature of the presence of nucleosomes. An overview of the basic theoretical concepts necessary to understand the study of the scale-invariance properties of DNA sequences with the wavelet-based methodology is first presented as a reference guide. We then proceed to an extensive study of the distributions of various nucleosome packaging and positioning motifs in a number of eukaryotic and bacterial genomes. The results show that for all the motifs examined, these LRC are observed in eukaryotic genomes and are absent from eubacterial genomes. We discuss the mechanisms by which these correlations could facilitate the positioning and the mobility of the nucleosomes, and their possible roles in the structure and dynamics of chromatin.

Theoretical Concepts and Methodology: Wavelet Analysis of Long-range Correlations in DNA Sequences

In order to illustrate the concept of long-range power-law correlations (LRC) and its relationship to scale invariance properties, we will first use artificial sequences that mimic the distribution of purines and pyrimidines along a DNA sequence. We will then explain how to quantify LRC from the measurement of the Hurst exponent, H.

Long-range correlations and scale invariance properties of symbolic sequences

Let us consider two artificially built sequences. The first one is an uncorrelated sequence drawn with equal probability for each of the four nucleotides. Using a method based on a two-valued fractional auto-regressive integrated moving average process, FARIMA,[24] we built a second sequence under the constraint that its purine (or equivalently its pyrimidine) positions be long-range correlated (characterized by a Hurst exponent $H = 0.9$, see later).

In Figure 1(a) and (b), we use a bar-code representation of the purine content along the first 200 bp of the two sequences. These sequences can readily be distinguished by visual inspection. Stretches of black meaning stretches of purines (white = pyrimidines) are clearly wider for the correlated sequence than for the uncorrelated one. This qualitative difference is the signature of what we will refer to as persistence.[25] When the positions of purines are positively correlated, if there is a purine at a position, the probability of having a purine at the following position is enhanced; LRC mean that this enhancement also depends on the presence of purines at positions on the sequence over arbitrarily large distances. To illustrate this particular structural property induced by these correlations and its relation to scale invariance, 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 amounts to doing a coarse graining of the sequence, replacing each box by a purine or a pyrimidine according to the purine content in that box (Figure 1(c) and (d), for $w = 32$ bp and (e) and (f), for $w = 512$ bp). When comparing these results to Figure 1(a) and (b), an important feature emerges: the coarse-grained sequences are statistically indistinguishable from the original sequence. 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 some coarse graining. For the correlated sequence, scale invariance results in purines being correlated in the same manner as boxes containing an excess of purines and this, whatever the box width†. In the same manner, for the uncorrelated sequence, scale invariance results in the absence of correlations at all scales. It is important to note that persistence does not mean that there is much smaller variation of the purine concentration along the correlated sequence than along the uncorrelated one. 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. To summarize, LRC in DNA sequences are likely to result from processes that structure objects of size w along the genome in the same statistical manner whatever the scale of observation w.

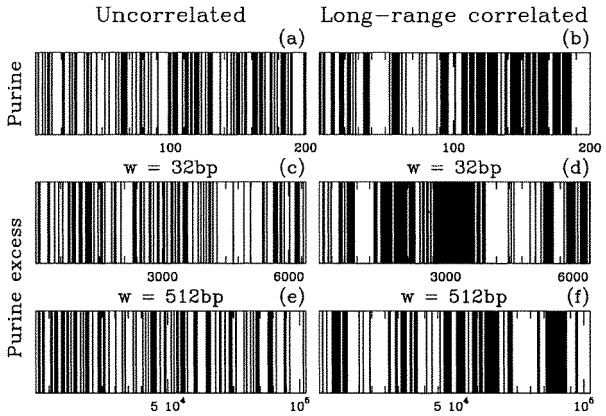

Figure 1. Bar code representation of the purine/pyrimidine content for two artificial DNA sequences (both are 262, 144 bp long). (a), (c) and (e) Analysis of a sequence generated by uncorrelated trials with equal probability for the four nucleotides; (b), (d) and (f) analysis of a sequence obtained under 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 with equal probability for the four nucleotides. 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) (and (e) and (f)), the sequence has been divided into non-overlapping boxes of size 32 bp (512 bp) and we measure the purine content in each box. If this concentration is $\geqslant 50\,\%$, a black bar of width 32 bp (512 bp) is drawn at the corresponding position. For the six pictures, the abscissa range is 200 bars wide to ensure that the visual effect obtained is not due to bars of different sizes. Going from (e) to (c) to (a) (and (f) to (d) to (b)) is equivalent to zooming in the uncorrelated (long-range correlated) sequence with a black and white 200 pixels camera.

† It is important to note that LRC cannot be constructed with a Markov model of finite size memory.[26] 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 indistinguishable from an uncorrelated one as soon as the size of the box $w \gg m$.

Quantification of long-range correlations in symbolic sequences

Let us now perform a quantitative analysis of the fluctuations of purine concentration along the same two sequences as in Figure 1. Figure 2(a)-(d) display the purine concentration calculated for the same box sizes as those used in Figure 1(c)-(f), respectively. For the uncorrelated sequence, the amplitude of the fluctuations around the mean value of 0.5 clearly decreases as the window width increases from $w_1 = 32$ bp (Figure 2(a)) to $w_2 = 512$ bp (Figure 2(c)). In the case of the long-range correlated sequence, this phenomenon is also visible but to a much smaller extent. To distinguish the two sequences, we can take advantage of this difference and 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 $\sigma(w)$ is thus of the form:

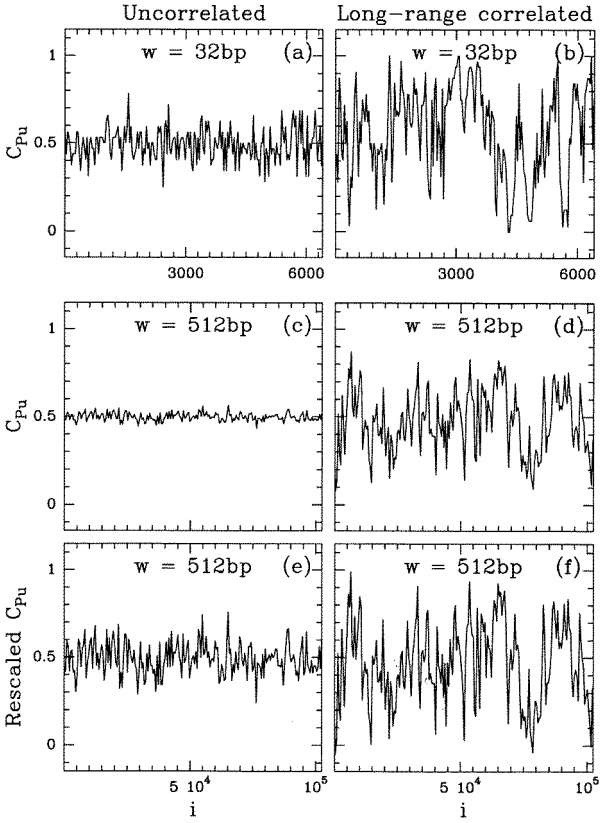

Figure 2. Fluctuations of the purine content C_{Pu} within non-overlapping boxes of size w as a function of the box position i for the same sequences as in Figure 1. In (a) and (b), $w_1 = 32$ bp; in (c) and (d), $w_2 = 512$ bp. In (e) and (f), C_{Pu} computed with $w_2 = 512$ bp, is rescaled according to equation (2) i.e. $0.5 + (C_{Pu} - 0.5)(w_1/w_2)^{H-1}$, with (e) $H = 0.5$ and (f) $H = 0.9$. For the six pictures, the abscissa range is 200 box-wide so that all curves present the same number of points.

$$\sigma(w) = \sigma(1)/\sqrt{w} \quad (1)$$

Similarly, for sequences possessing scale-invariance properties in general, the standard deviation is:

$$\sigma_H(w) = \sigma_H(1) w^{H-1} \quad (2)$$

where H is the Hurst exponent.[25,27] Note that $H = 0.5$ for the uncorrelated sequence. As a visual check of this power-law behaviour of the root-mean square (r.m.s.) fluctuations of purine concentration, we have plotted in Figure 2(e) and (f), the purine concentration computed for the box size $w_2 = 512$ bp after rescaling the fluctuations by $(w_1/w_2)^{H-1}$. Once rescaled using the appropriate Hurst exponent value, the purine concentration fluctuations obtained in Figure 2(e) ($H = 0.5$) and (f) ($H = 0.9$) are statistically indistinguishable from the corresponding fluctuations obtained with boxes of smaller size in Figure 2(a) and (b), respectively.

The quantitative characterization of scale-invariant properties is a straightforward consequence of equation (2). Taking the logarithm of equation (2), one gets:

$$\log_{10}(w\sigma_H(w)) = H \log_{10} w + \log_{10}(\sigma_H(1)) \quad (3)$$

So, when plotting $\log_{10}(w\sigma_H(w))$ as a function of $\log_{10}w$, the fact that all the data points fall on a linear curve enables us to diagnose scale invariance properties. Measuring the slope of this straight line gives an estimate of H. Figure 3(a) illustrates 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 use throughout this work for the analysis of genomic sequences. By plotting $\Lambda(w) = \log_{10}(w\sigma_H(w)) - 0.6 \log_{10}w$ *versus* $\log_{10}w$, we select $H = 0.6$ as the Hurst exponent value of reference for a horizontal linear scaling behaviour. The straight lines corresponding to $H = 0.5$ and $H = 0.8$ are drawn to guide the eye, allowing visual diagnostics.

Let us emphasize that quantifying scale-invariant properties *via* equation (2) also amounts to characterizing the behaviour of the correlation function C_H. Actually, equation (2) implies that C_H decreases as a power-law with exponent $2H-2$ of the distance n between nucleotides:[28–30]

$$C_H(n) \underset{m \to +\infty}{\propto} n^{2H-2} \quad (4)$$

This characteristic behaviour is drastically different from the much faster exponential decrease observed for instance in the Markov model[26] and justifies the terminology long-range correlations. The larger H (<1), the slower the correlation function decreases and, in that sense, the stronger the LRC. Let us point out that the Hurst exponent H describes how fluctuations or correlations are modified when increasing size or distance but does

not provide an estimate of the amplitude or intensity (equations (3) and (4)).

To conclude, it is fundamental to stress the necessity of using the wavelet analysis to investigate the scale invariance properties of DNA sequences. (i) The mosaic structure of genomic DNA has dramatic consequences on the standard deviation calculations proposed in equation (2) and thus leads to severe bias in the estimate of H.[31] A way to overcome this difficulty is to extract fluctuations by means of oscillating boxes, i.e. wavelets, instead of simple boxes.[32] (ii) Describing the scale-invariance properties of genomic DNA by means of a single exponent, namely the Hurst exponent H, amounts to assuming implicitly that the sequence is monofractal, as opposed to multifractal for which an infinite number of exponents are required to fully describe scale-invariant characteristics. As originally established by Arneodo *et al.*,[23] the so-called wavelet transform modulus maxima (WTMM) method is a very efficient technique for performing such a diagnostic analysis. The results reported here have been obtained using the WTMM method, and the monofractal nature of all the genomic sequences considered has been explicitely checked. This justifies *a posteriori* that we will focus our study and discussion on the estimate of the Hurst exponent H, using the r.m.s. of the wavelet coefficients:

$$\sigma_{WT}(w) \sim w^H$$

Results

Wavelet analysis of the scale-invariance properties of genomic sequences

We report the results of a wavelet-based statistical analysis of the scale-invariance properties of genomic sequences that belong to eukaryotic, eubacterial, and archaeal genomes, as well as sequences of DNA and RNA viruses. For the sake of simplicity, we will systematically report the results using the representation of the estimate of the Hurst exponent H described in Theoretical Concepts and Methodology (Figure 3(b)).

† The probability density functions (pdfs) of wavelet coefficient values of the DNA sequence of the yeast chromosome 1 ($L = 230{,}209$ bp), computed at different scales using the A coding, are very well approximated by Gaussian distributions in the small-scale regime, 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 trinucleotide codings used here (data not shown).

Saccharomyces cerevisiae

To set up the general framework of our study, we will first investigate the genome of *S. cerevisiae* which will allow us to perform a scaling analysis over a wide range of scales extending from tens to thousands of nucleotides. When looking at the global estimate of the standard deviation of the wavelet transform (WT) coefficients $\sigma_{WT}(w)$ obtained for each of the 16 yeast chromosomes, using the A mononucleotide coding rule, it has been observed that all present superimposable behaviour with the same characteristic scale $w_C = 200$ bp, that separates two different scaling regimes (Figure 4(a)).[22] At small scales, $10 < w < 200$, LRC are characterized by a mean Hurst exponent value $H = 0.57(\pm 0.03)$, which is significantly larger than the theoretical prediction $H = 1/2$ for uncorrelated sequences. At large scales, $200 < w < 5000$ bp, stronger LRC with $H = 0.82(\pm 0.02)$ become dominant with a cutoff around 10 kbp (a number by no means accurate) above which uncorrelated behaviour is observed.

In Figure 4(b) are reported the results of some tests 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 indistinguishable as well as the data for the G and C codings (this justifies that in the following, we will present results corresponding to the average over the A and T codings, A(+T), and over the G and C codings, G(+C)). While each of these mononucleotide codings displays a characteristic scale ($w_C \approx 200$ bp) that separates two scaling regimes, there is however some difference between them. This difference arises mainly in the small-scale regime ($10 < w < 200$ bp) 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†

We have further performed a comparative wavelet analysis of the yeast DNA sequences using the Pnuc coding rule that provides an estimate of DNA roll-angle values from which one determines the bending profile of the axis of the double helix (see Materials and Methods). This analysis 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 (Figure 4(b)). To ensure that these observations are not simply due to a ''recoding'' of the DNA sequences, but rather to the proper values of Pnuc, we have randomly changed the Pnuc table to a new table obtained using a Gaussian distribution of the same mean, variance and symmetries as the original table. This results in the vanishing of the observed LRC ($H = 0.50(\pm 0.01)$) at all scales $w < 1000$ bp, which strongly suggests that these LRC are likely to reflect the structural information included in the Pnuc table (Figure 4(b)). Additional evidence that

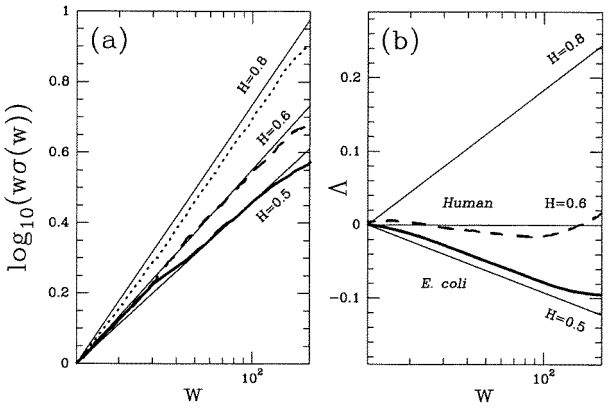

Figure 3. Scale-invariance analysis of purine concentration fluctuations. (a) $\mathrm{Log}_{10}(w\sigma_H(w))$ *versus* $\log_{10}w$ for artificial sequences of length 10 kbp with Hurst exponents $H = 0.5$ (continuous line), 0.6 (broken line) and 0.8 (dotted line), respectively. (b) $\Lambda(w) = \log_{10}(w\sigma_H(w)) - 0.6\ \log_{10}w$ *versus* $\log_{10}w$ for the genome sequence of *E. coli* (continuous line) and the human chromosome 21 (broken line). The straight lines correspond to uncorrelated ($H = 0.5$) and long-range correlated ($H = 0.6$ and $H = 0.8$) sequences ; the curves in (a) and (b) have been vertically shifted to start from the same value of $\Lambda(10)$.

the Pnuc trinucleotide coding is not a trivial recoding of the DNA sequences is provided by the data obtained when using the DNase table of curvature (see Materials and Methods). One notices a significant weakening of the LRC exponent observed in the large-scale regime ($H \approx 0.6$ with the DNase coding instead of $H \approx 0.8$ with the Pnuc coding, Figure 4(b)). We will see in the following that a significant weakening is also observed with DNA sequences of other eukaryotic genomes when using the DNase coding. To strengthen our interpretation of the observed LRC in terms of structural constraints, we have performed the wavelet analysis of DNA sequences using dinucleotide codings which are known to contribute to the intrinsic bending and flexibility properties of the DNA double helix. We show in Figure 4(c) the results obtained with the AA (=TT) coding when averaging over the 16 yeast chromosomes. When comparing to the results obtained with the $A_{\mathrm{iso}}(=T_{\mathrm{iso}})$ 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 $A_{\mathrm{iso}}(=T_{\mathrm{iso}})$ coding, while the AA(=TT) coding accounts for a major part of the LRC observed with the A and Pnuc codings.

We have extended this statistical analysis of DNA sequences to various eukaryotic, eubacterial and archaebacterial genomes. A general observation is the existence of a characteristic scale $w_C = 100\text{-}200$ bp that separates two different monofractal scaling regimes whatever the coding rule used to digitize the DNA sequences. In the large-scale regime ($200 < w < 1000$ bp), when using the Pnuc coding rule as well as the four elementary mononucleotide coding rules, strong LRC ($H \approx 0.8(\pm 0.1)$) are systematically observed in most DNA sequences whatever the organism, the kingdom and the coding or non-coding nature of the sequence under study. The biological meaning of

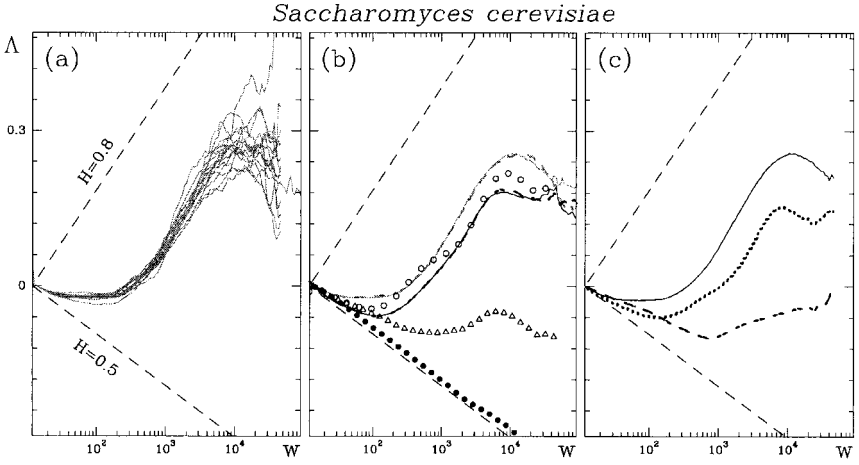

Figure 4. Global estimate of the r.m.s. of WT coefficients of the 16 chromosomes of *S. cerevisiae*. $\Lambda(w) = \log_{10}\sigma_{WT}(w) - 0.6\ \log_{10}w$ is plotted *versus* $\log_{10}w$. (a) Comparative analysis of the 16 chromosomes when using the A mononucleotide coding. (b) Comparative analysis of the A (grey continuous line), T (grey broken line), G (black continuous line) and C (black broken line) mononucleotide codings, with the Pnuc (circles) and DNase (triangles) trinucleotide codings ; the black dots correspond to a randomly shuffled Pnuc table (see the text). (c) Comparative analysis of the A(+T) mononucleotide coding (continuous line) with the AA(=TT) dinucleotide coding (dotted line) and the $A_{iso}(=T_{iso})$ mononucleotide coding (broken line). In (b) and (c), the results correspond to averaging over the 16 chromosomes. For the coding rules see Materials and Methods. The analysing wavelet is the first derivative of the Gaussian function.[23]

Table 1. Values of the Hurst exponent H in the small-scale regime

	Pnuc	DNase	AA (=TT)	A_{iso} (=T_{iso})	A (+T)	GG (=CC)	G_{iso} (=C_{iso})	G (+C)
H. sapiens	0.67(4)	0.59(3)	0.64(2)	0.55(1)	0.61(3)	n.a.	0.54(2)	0.59(3)
D. rerio	0.61(5)	0.58(3)	0.58(5)	0.58(3)	0.60(3)	n.a.	0.56(3)	0.57(6)
D. melanogaster	0.62(3)	0.56(3)	0.60(5)	0.59(2)	0.63(5)	n.a.	0.56(1)	0.61(6)
C. elegans	0.66(6)	0.59(5)	0.63(5)	056(2)	0.59(5)	n.a.	0.57(4)	0.62(7)
A. thaliana	0.60(7)	0.55(2)	0.58(5)	0.55(1)	0.60(3)	n.a.	0.54(2)	0.58(7)
S. cerevisiae	0.54(1)	0.54(2)	0.54(1)	0.55(1)	0.57(3)	n.a.	0.52(1)	0.53(3)
Herpesviruses	0.57(1)	0.52(1)	n.a.	0.53(1)	0.53(1)	0.57(2)	0.53(1)	0.59(1)
Adenoviruses	0.57(1)	0.53(2)	0.55(4)	0.54(2)	0.53(2)	n.a.	0.52(4)	0.54(2)
M. sanguinipes	0.51(1)	0.49(2)	0.50(2)	0.51(3)	0.49(2)	n.a.	0.49(2)	0.51(1)
Vaccinia	0.51(2)	0.49(1)	0.51(2)	0.50(2)	0.48(4)	n.a.	0.51(1)	0.48(1)
ssRNAp	0.53(1)	0.52(2)	0.51(1)	0.53(1)	0.50(1)	n.a.	0.53(2)	0.49(2)
dsRNA	0.55(1)	0.49(2)	n.a.	0.51(1)	0.50(3)	0.53(3)	0.53(3)	0.51(1)
Spumaretrovirus	0.62(2)	0.50(2)	0.53(1)	0.51(2)	0.49(1)	n.a.	0.62(4)	0.54(2)
Retroviruses	0.57(3)	0.50(1)	0.57(1)	0.51(2)	0.53(2)	0.61(1)	0.56(1)	0.54(1)
B. subtilis	0.51(2)	0.49(1)	0.50(2)	0.51(1)	0.52(1)	n.a.	0.50(2)	0.51(1)
C. trachomatis	0.51(1)	0.51(2)	0.52(1)	0.52(1)	0.46(2)	n.a.	0.51(1)	0.49(1)
E. coli	0.49(2)	0.48(2)	0.50(2)	0.50(1)	0.51(1)	0.50(1)	0.50(1)	0.50(1)
H. pylori	0.51(5)	0.51(2)	0.52(4)	0.52(1)	0.52(4)	n.a.	0.50(3)	0.55(3)
M. pneumoniae	0.51(4)	0.51(3)	0.52(3)	0.52(1)	0.52(2)	0.52(3)	0.52(3)	0.52(1)
Synechocystis	0.51(1)	0.47(2)	0.50(1)	0.50(1)	0.50(1)	0.49(2)	0.49(1)	0.51(1)
T. maritima	0.52(2)	0.49(1)	0.52(3)	0.51(1)	0.51(2)	0.52(2)	0.51(1)	0.50(2)
T. pallidum	0.54(4)	0.51(1)	0.50(1)	0.51(1)	0.50(2)	0.49(3)	0.53(1)	0.50(1)
13 other eubacteria	0.52(2)	0.50(2)	0.51(1)	0.50(1)	0.51(1)	0.50(1)	0.51(1)	0.51(1)
A. aeolicus	0.57(4)	0.51(2)	0.57(3)	0.51(1)	0.54(3)	0.52(2)	0.51(1)	0.53(2)
B. burgdorferi	0.55(1)	0.51(1)	0.53(1)	0.51(1)	0.50(1)	n.a.	0.52(1)	0.53(1)
Buchnera sp.	0.59(2)	0.52(1)	0.57(2)	0.52(1)	0.50(1)	n.a.	0.54(2)	0.53(4)
C. jejuni	0.55(4)	0.50(4)	0.55(4)	0.51(1)	0.54(2)	n.a.	0.52(3)	0.54(3)
R. prowazekii	0.54(3)	0.50(2)	0.54(2)	0.52(1)	0.52(2)	n.a.	0.52(2)	0.51(3)
T4	0.50(1)	0.50(1)	0.50(1)	0.52(1)	0.49(1)	n.a.	0.53(2)	0.47(1)
SPBc2	0.49(1)	0.50(3)	0.50(1)	0.52(1)	0.51(1)	n.a.	0.51(1)	0.50(1)
A. pernix	0.53(1)	0.51(1)	n.a.	0.50(1)	0.48(3)	0.50(1)	0.51(1)	0.58(2)
A. fulgidus	0.54(5)	0.51(2)	0.52(5)	0.50(1)	0.50(3)	0.50(2)	0.52(1)	0.58(2)
M. jannaschii	0.56(5)	0.52(3)	0.55(3)	0.51(1)	0.55(4)	n.a.	0.51(2)	0.53(2)
P. horikoshii	0.52(4)	0.51(2)	0.53(4)	0.51(1)	0.53(1)	0.50(2)	0.51(1)	0.52(2)
S. solfataricus	0.54(3)	0.50(1)	0.53(2)	0.51(1)	0.53(2)	n.a.	0.51(1)	0.55(2)
T. acidophilum	0.56(3)	0.50(2)	0.52(5)	0.50(1)	0.54(3)	0.52(3)	0.51(1)	0.51(2)

H is estimated using the wavelet-based method (see the text). The numbers correspond to a linear regression fit of $\log_{10}(\sigma_{WT}(w))$ *versus* $\log_{10}w$, in the 10(20)-100 bp range. The error terms ($\times 10^2$) in parentheses, 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. n.a., non-attributed, when the density of the mono-, di- or trinucleotide under consideration is too small to allow the quantification of LRC.

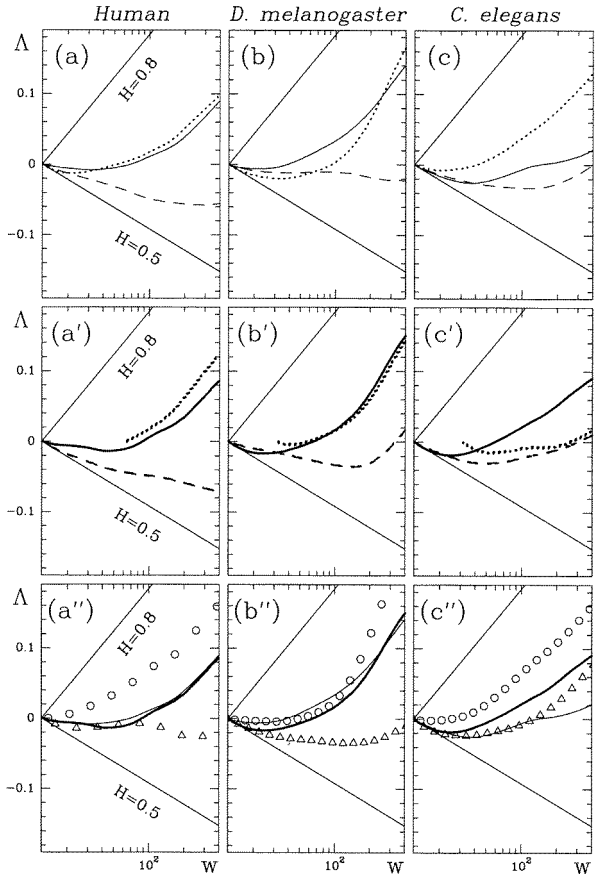

Figure 5. Global estimate of the r.m.s. of WT coefficients of the human chromosome 21 (a)-(a″), *D. melanogaster* (b)-(b″) and *C. elegans* (c)-(c″) genomes. Coordinates and analysing wavelet as in Figure 4. (a)-(c): A(+T) mononucleotide coding (continuous line); AA(=TT) dinucleotide coding (dotted line); $A_{iso}(=T_{iso})$ mononucleotide coding (broken line). (a′)-(c′) G(+C) mononucleotide coding (black continuous line); GG(=CC) dinucleotide coding (black dotted line); $G_{iso}(=C_{iso})$ mononucleotide coding (black broken line). (a″)-(c″) Pnuc coding (circles); DNase coding (triangles); A(+T) mononucleotide coding (continuous line); G(+C) mononucleotide coding (black continuous line).

this long-range correlated large-scale regime is the subject of current research. Here we will concentrate on the small-scale regime ($10 < w < 200$ 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. The data corresponding to different mono-, di- and trinucleotide coding rules will be compared on the range of scales $10 < w < 400$ bp. When the curve corresponding to some coding (e.g. GG coding) will be missing or cut at very small scales, this will reflect that the density of the mono-, di- or trinucleotide under consideration (e.g. GG) is too small for the investigation of the LRC properties over this range of scales to make any sense.

Wavelet analysis of LRC in the small-scale regime

Other eukaryotic genomes

In Figure 5 and Table 1 are reported the results of a wavelet transform analysis of various eukaryotic genomes in the small-scale regime. As a first general observation, there exist significant LRC in every examined eukaryotic DNA sequence when using the Pnuc coding rule (Figure 5(a″)-(c″)). For example, one gets the following estimates of the Hurst exponent H: 0.67(±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 but they are systematically weaker than those identified with the Pnuc coding rule namely $H = 0.59(\pm 0.03)$ (human chromosome 21), 0.56(±0.03) (*D. melanogaster*), $H = 0.59(\pm 0.05)$ (*C. elegans*) and 0.55(±0.02) (*A. thaliana*). In contrast to the above observation, the estimates obtained for the yeast genome are similar: $H = 0.54(\pm 0.01)$ with the Pnuc coding and $H = 0.54(\pm 0.02)$ with the DNase coding.

Another rather general observation is that the data obtained with the Pnuc coding yield an estimate of the strength H of the LRC which is larger than, or similar to, the values obtained with the mononucleotide codings (Figure 5(a″), (c″), Table 1 and data not shown). We also note that the AA(=TT) dinucleotide coding reproduces quite well the Pnuc data (Figure 5(a), (c), and Table 1). The A_{iso} $(=T_{iso})$ coding in Figure 5(a)-(c) 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 Figure 5(a′)-(c′) that the $G_{iso}(=C_{iso})$ coding does not participate to a large extent in the LRC revealed by the Pnuc coding. When the density of GG(=CC) dinucleotides allows us to investigate LRC, these LRC are generally similar to the ones observed with the Pnuc coding (Figure 5(a′), (b′) and data not shown). As a final observation, the main features recognized in the results illustrated in Figure 5 are quite characteristic of the other studied eukaryotic genomes (Table 1 and data not shown).

Eubacterial genomes

We have analysed the scale-invariance properties of complete eubacterial genomes that belong to the following groups: proteobacteria, Gram-positive, spirochaetes, cyanobacteria, thermotogales and chlamydiae. In Figure 6 are illustrated the data for some selected complete eubacterial genomes from these various groups. The visualized range of scales is the same as for the eukaryotic sequences in Figure 5. In eubacterial genomes, the character-

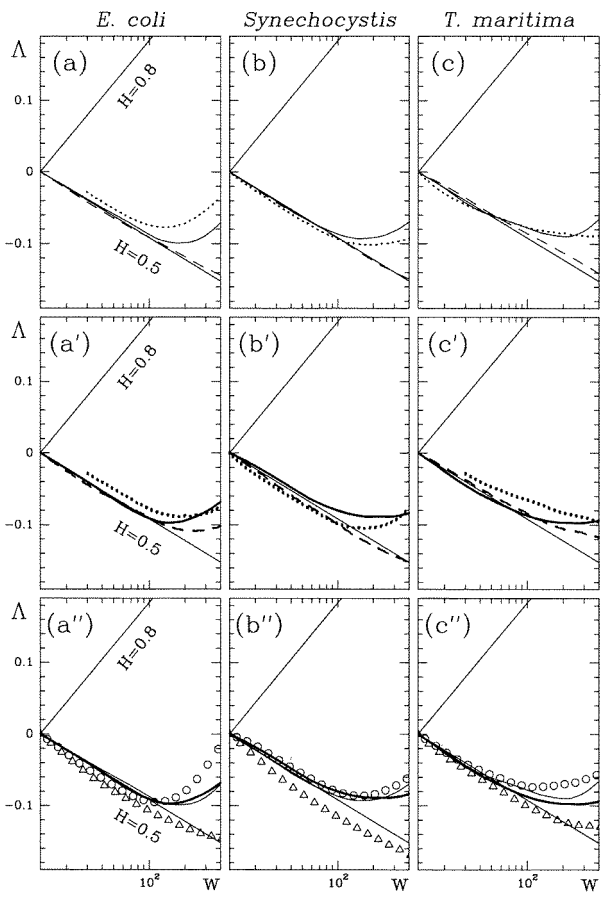

Figure 6. Global estimate of the r.m.s. of WT coefficients of *E. coli* (a)-(a″), *Synechocystis* sp. (b)-(b″) and *T. maritima* (c)-(c″). Coordinates, curve lines and analysing wavelet as in Figure 5.

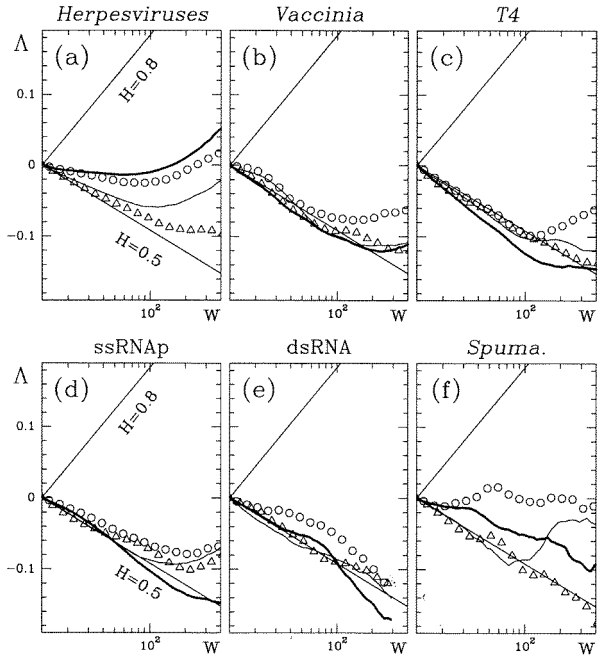

Figure 7. Global estimate of the r.m.s. of WT coefficients of viral genomes (for genome references, see Materials and Methods). (a) Average over seven genomes of Herpesviruses; (b) *Vaccinia* virus; (c) Bacteriophage *T4*; (d) average over 20 genomes of positive strand ssRNA viruses; (e) average over four genomes of dsRNA viruses; (f) genome of Spumaretrovirus. Coordinates, curve lines and analysing wavelet as in Figure 5(a″)-(c″).

istic scale 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. w_C is more likely about 200 bp. 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 with the 100-150 bp long DNA regions which are wrapped around histone proteins to form the eukaryotic nucleosomes.[1]

The main observation when examining the data in Figure 6 is that whatever the rules used to code the eubacterial DNA sequences, one does not observe any evidence of the existence of LRC. As one can see quantitatively in Table 1, the estimates of the Hurst exponent H all fall in the range $0.48 < H < 0.52$ and therefore cannot be distinguished from the canonical value $H = 1/2$ for uncorrelated sequences. Both Pnuc and DNase codings (as well as the AA(=TT) coding) yield quantitative estimates of $H = 0.50(\pm 0.02)$ similar to those obtained with the four mononucleotide codings. This observation stands for 24 eubacterial genomes among 29 examined (Table 1). Among the five exceptions, two genomes exhibit rather strong LRC with the Pnuc coding, namely *Aquifex aeolicus* ($H = 0.57(\pm 0.04)$) and *Buchnera* sp. ($H = 0.59$ (± 0.02)). The three other genomes present weaker detectable LRC (e.g. *Borrelia burgdorferi*, $H = 0.55$ (± 0.01), *Campylobacter jejuni*, $H = 0.55(\pm 0.04)$ and *Rickettsia prowazekii*, $H = 0.54(\pm 0.03)$). In addition, a questionable case is observed with *Treponema pallidum* for which $H = 0.54(\pm 0.04)$ (Pnuc coding), but which has no LRC with the AA(=TT) coding (Table 1).

Viral DNA genomes

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[33] and Adenoviruses.[34] Small-scale LRC are clearly identified in these genomes as shown in Figure 7(a) and Table 1. For these genomes, LRC are clearly detected with the Pnuc coding rule but not with the DNase coding rule (for Herpesviruses, $H = 0.57(\pm 0.01)$ and $0.52(\pm 0.02)$ for the Pnuc and DNase codings, respectively). Note that the LRC observed with Pnuc are quite comparable to the ones exhibited by the AA(=TT) and/or GG(=CC)

dinucleotide codings ($H = 0.57(\pm 0.02)$ for the Herpesviruses with the latter coding), which is similar to what has been previously observed for eukaryotic genomes. We have also investigated the existence of LRC in poxviruses. As shown in Figure 7(b) and Table 1, the genomes of *Melanoplus sanguinipes* and *Vaccinia* exhibit H values very close to 1/2, indicating the absence of LRC. 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 (Figure 7(c) and Table 1). Indeed, none of the considered mono-, di-, or trinucleotide codings exhibit any evidence for LRC (H values do not deviate significantly from $H = 1/2$). These results show that prokaryotic viral sequences present no LRC in the DNA texts as well as in the DNA bending profiles, as previously observed for their hosts genomes.

Viral RNA genomes

We have examined several classes of single-stranded and double-stranded RNA genomes (Figure 7 and Table 1). The results reveal the absence of LRC in most genomes. The estimates of H do not deviate significantly from $H = 0.5$ for the ssRNA, and this for all codings (Figure 7(d)). For the dsRNA, similar estimates are obtained except for the Pnuc coding, which exhibits weak LRC (Figure 7(e)) (note that the Pnuc table has no structural significance for RNA molecules). We have also examined, but separately, the case of retroviruses, since the retroviral genomes are inserted as double-stranded DNA in their host genomes. The results obtained for the Spumaretrovirus and for several distantly related retroviruses clearly show that, contrary to the other RNA viral genomes, these retroviral sequences exhibit LRC with the Pnuc coding (Figure 7(f)) as in their host genomic sequences (which contrasts with their absence with the DNase coding). Significant LRC are also observed with the AA(=TT) and GG(=CC) dinucleotide codings (Table 1).

Archaebacterial genomes

We have examined the complete genomes of euryarchaeota identified to contain histones (*Thermoplasma acidophilum, Methanococcus jannaschi, Pyrococcus horikoshi, archaeoglobus fulgidus*) and of two crenarcheota which did not have histones (*Aeropyrum pernix* and *Sulfolobus solfataricus*) (see Sandman & Reeve[35] and references therein) for the presence of LRC. As shown in Table 1, 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 $A_{iso}(=T_{iso})$. For example, for the *M. jannaschii* 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 contrasting situation is observed with the G(+C) coding, which presents strong LRC ($H = 0.58$ (± 0.02)) but no LRC with GG(=CC) ($H = 0.50(\pm 0.01)$). These observations differ significantly from what has been observed with the eukaryotic genomes, for which the LRC obtained with GG(=CC) are mostly comparable to the LRC evidenced with the G(+C) coding (Table 1). 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. We note that similar results are obtained with the genome of the euryarchaeota *A. fulgidus* (Table 1).

Discussion

A regime of long-range correlations (10 bp to 200 bp) specific to eukaryotic sequences

Here, we have analysed genomic sequences in the new perspective to show long-range correlations (LRC) related to structural properties of the DNA molecule involved in the processes of chromatin packaging. This suggested 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 nucleosome packaging and nucleosome positioning. The analyses were carried out with the bending profiles obtained by coding the DNA sequences with these structural motifs. We have performed genome-wide studies of LRC in an overview of organisms belonging to the three kingdoms, eukaryotes, eubacteria, and archaebacteria, as well as in DNA and RNA viral genomes. The studies were performed systematically in parallel with the analysis of the DNA texts and they allowed us to show 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 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 in all the organisms that we have analysed in the three kingdoms. In contrast to this robust ''stability'' of the large-scale regime, 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 uncorrelated sequences characterized by $H = 0.5$. A totally different situation emerges from eukaryotic genomes. These exhibit H values significantly larger than 0.5 that reveal the existence of LRC. Furthermore, the presence of LRC in eukaryotic genomes and their absence from 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 di- and trinucleotides.

The 10-200 bp LRC regime is related to the existence of nucleosomes

Several works have determined sequence-dependent preferences for the bending of the DNA double helix around the core histones. Sequence analyses based on the alignment of nucleosomal DNA allowed us to set up a table (Pnuc) of the bending values (roll angles) associated with all trinucleotides,[9,36] that we used to establish a ''bending profile'' of the DNA sequence (see Materials and Methods). These profiles were examined with the WWTM method to search for the presence of LRC.[23] It must be noted that the determination of sequence-dependent DNA bending sites is somehow controversial and that there exist various coding tables that represent different ways of modelling the DNA local bending or flexibility. Models have been derived from different types of experimental observations, like for instance gel-electrophoresis, X-rays, NMR, DNase I and the nucleosome-positioning model from which is derived the Pnuc table.[37] The resulting values are not always consistent but can be considered as providing significant elements of the DNA structure that depend on the situation under analysis. One can thus consider that all coding tables generally predict rather correctly the DNA curvature, but that they present differences that are related to the experimental technique they are derived from. In this respect, the Pnuc coding is the most suitable table of bending values when one wants to study the nucleosomal curvature. This table is derived from the statistical analysis of the sequences of nucleosomal DNA fragments. This allowed us to determine the sequence-dependent preferences that correlate with the rotational orientation of the DNA at the surface of the histone octamer.[9] The DNase I compilation is based on a completely different rationale. This enzyme binds to DNA without particular sequence specificity and bends it towards the major groove. The cutting rate is then used to estimate the DNA rigidity.[38] The comparison of the Pnuc and DNase tables shows that the relative positions of the trinucleotides are different (the linear correlation coefficient is 0.53).[39] However, several trinucleotides, like for instance AAA/TTT, which play a major role in nucleosome binding, occupy similar positions in both tables. How can these similarities be reconciled with the fact that the two tables lead to strongly different LRCs? First, we observe that several motifs (several trinucleotides containing the TA/TA step like ATA/TAT or the CA/TG step, like CAG/CTG) occupy very different relative positions in the tables.[39] These properties might be related to the fact that the TA/TA and CA/TG steps can adopt dinstinct modes of bending, depending on the protein they are bound to.[40] Second, it is worth noting that the differences between the two tables are clearly sufficient to produce significantly different DNA bending profiles as seen in Figure 8 (let us point out that this does not necessarily imply that these two profiles present different LRCs). Taken together, these observations allow us to consider that the small but discriminating differences between the Pnuc and DNase tables sustain the specificity of the LRC-based diagnosis of the nucleosomal structure.

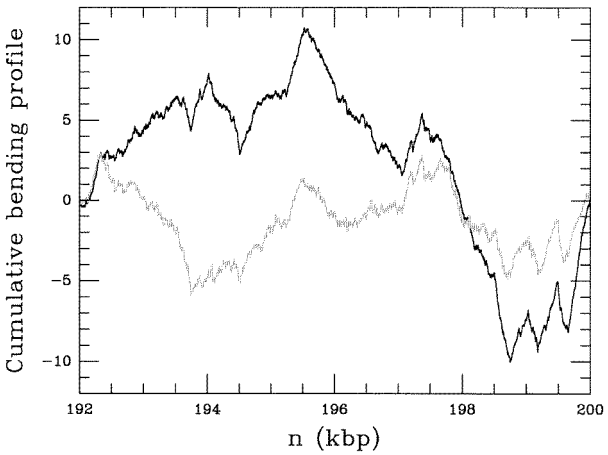

Figure 8. Cumulative bending profiles for a human DNA fragment (chromosome 21, positions 192 kb to 200 kb). Abscissa is the position on the sequence; the curves are cumulative representations of the PNuc (black) and DNase (grey) codings (in order to facilitate the comparison, the mean drift of the curves has been eliminated). Note that the two structural codings display very different fluctuations.

We observed that for most eukaryotic sequences, the bending profiles obtained with the Pnuc table presented H values similar to, or larger than those obtained with the DNA texts; on the contrary, the H values corresponding to the DNase table were significantly smaller (Figure 5 and Table 1). For eubacterial sequences, in the 10-200 bp range, the H values obtained with both the Pnuc and the DNase coding tables were similar to those of uncorrelated sequences (Figure 6 and Table 1). We also measured the H values of bending profiles obtained with other bending tables. In all cases tested, they led to values that were smaller than those obtained with the Pnuc table and larger than those obtained with the DNase table (data not

shown). The main conclusion from these results is that LRC do exist in the 10-200 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 weakly detected by the DNase coding, the other types of codings of the DNA curvature leading to intermediate situations.

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 ≈150 bp DNA regions that are wrapped around the core histones to form the eukaryotic nucleosomes. In prokaryotes, the absence of such structures is in agreement with the values of $H = 0.5$ observed for the majority of eubacterial genomes. In this respect, how can we interpret the observation of H values larger than 0.5 for a few bacterial genomes? We favour the hypothesis that the interpretation of LRCs in terms of DNA structural constraints also stands for these particular bacterial genomes. The fact that the Pnuc table reveals significant LRC in these few bacterial genomes suggests that some sequence motifs important for the nucleosomal organisation of eukaryotic genomes are also important for this particular bacterial DNA packaging process(es). Among these eubacterial exceptions, *A. aeolicus* is an hyperthermophilic bacterium which is most deeply branched in the bacterial phylogenetic tree. Interestingly, if we except the case of *C. jejuni*, we observe that all the other genomes that present LRCs (e.g. *B. burgdorferi*, *Buchnera* sp., *R. prowazekii* and *T. pallidum*) correspond to obligate intracellular bacteria, and that there is no other obligate intracellular organism among the 29 examined eubacteria. Further studies will be required to determine what features of these particular bacteria are related to the presence of LRCs in their genomic sequences.

The nucleosomal hypothesis can further 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 nucleotides that are not part of a dinucleotide AA, namely A_{iso} (all A bases 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 mononucleotide 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. Consistent results were obtained with other eukaryotic genomes although the GG distributions could not always be examined due to lack of abundance of this dinucleotide. Conversely, the $A_{\mathrm{iso}}(=T_{\mathrm{iso}})$ as well as the $G_{\mathrm{iso}}(=C_{\mathrm{iso}})$ codings generally revealed significantly weaker LRC (Figure 5 and Table 1). Examination of Figure 5 shows significant differences between the LRC detected in the examined genomes. For example, the Pnuc coding reveals LRC which are similar to those of the mononucleotide codings A(+T) and G(+C) for *D. melanogaster* (b″), but on the opposite, Pnuc leads to LRC larger than those of the mononucleotide codings for human (a″) and *C. elegans* (c″). This property is not simply due to the AA dinucleotide, since the LRC associated to AA are larger than those of the mononucleotide codings for (c) *C. elegans* but not for (a) human and (b) *D. melanogaster*. Taken together, these observations allow us to consider that the optimal parameter sets for nucleosome positioning can be considered as genome-dependent. They could for instance depend on the compositional biases of the genomes.

The hypothesis that these LRC are associated with the presence of nucleosomes can be further tested by searching for LRC in viral 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[41]. 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[33] and Adenoviruses.[34] We also examined the genomes of poxviruses. These are the only animal viruses that replicate in the cytoplasm, which suggests that their genomic DNA molecule is not expected to form nucleosomes. The results (Figure 7(a) and (b); and Table 1) 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) with 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 7(c) and Table 1). Overall, these results are in remarkable agreement with our hypothesis.

To end up with this overview of complete genomes, we have also examined the sequences of viral single-strand and double-strand RNA genomes. In line with 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.[42] As shown in Figure 7(d)-(f) and Table 1, the results demonstrate that the examined RNA genomes do not deviate significantly from uncorrelated sequences, except for the retroviral genomes, which again strongly sustains our hypothesis.

Histones are known to exist not only in most eukaryotes but also in a class of archaebacteria, the

euryarchaeota.[35] Furthermore, histone packaging of DNA has apparently imposed similar constraints on the genomes of both types of organisms to direct nucleosome positioning, involving for example the AA(=TT) dinucleotides.[43] These various observations prompted us to examine the complete genomes of euryarchaeota identified to contain histones (*T. acidophilum*, *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. The results obtained with the archaebacteria (see Table 1) can be summarized as follows. Among euryarchaeotic genomes, *T. acidophilum* and *M. jannaschii* present significant LRC with the AA(=TT) and the Pnuc codings, at the opposite of the two crenarchaeotic genomes available at this time. For these two euryarchaeota, this is consistent with the observation that archaeal nucleosome packaging involves sequence regularities similar to those of eukaryotic nucleosomes.[44,45] However, the observation of LRC between G(+C) nucleotides, and simultaneously of no LRC between GG(=CC) dinucleotides reveals a new type of correlation which is unprecedented in all eukaryotic and eubacterial genomes examined here. This property displayed by the two crenarcheota (*A. pernix* and *S. solfataricus*) as well as one euryarchaeota (*A. fulgidus*), reveals that LRC in archaebacterial genomic sequences present specific features that remain to be investigated.

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 mononucleotides (DNA text), then any arrangement of nucleotides (words) should also present similar LRC. The analyses with the various bending tables actually demonstrate the opposite, since: (i) 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); (ii) this is further evidenced by the fact that the A nucleotides exhibit strong LRC when they belong to the AA dinucleotide subgroup, but to a much lesser extent when they belong to the ''isolated A'' subgroup; (iii) finally, this is also strengthened by the analysis of the DNA profiles generated with a coding table obtained by the shuffling of the Pnuc table, and which leads to a total vanishing of LRC (Figure 4(b)). These observations demonstrate that the LRC observed between bending sites are not a trivial consequence of the existence of LRC between single nucleotides. On the other hand, the latter should rather be considered as resulting from LRC between bending sites. This does not mean that the Pnuc table allows the exact and unique evaluation of the ''words'' which are long-range-correlated in all DNA sequences. However, this characterization of the DNA bending sites resulting from the analysis of nucleosomal DNA provides the coding which, among those used here, most efficiently detects LRC in genomic sequences. Along this line, we observe that the analysis of the *C. elegans* genome with the AA(=TT) dinucleotide coding reveals larger H values than with the A(+T) coding (Figure 5(c)), indicating that the contribution of AA dinucleotides to the formation of nucleosomes is particularly important in this genome. 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* compared to *S. cerevisiae*.[46]

At this point, we must emphasize that there exists a fundamental difference between the Fourier analysis, which is based on the periodic properties of genomic sequences, and the present analysis based on scale-invariant properties (see Theoretical Concepts and Methodology). Both analyses contribute to identify statistical properties resulting from the nucleosomal organisation. However, each of them detects a particular type of property. In this respect, they should not be considered as opposed, but rather as constituting complementary approaches.

What mechanisms underly LRC in genome sequences?

Although the analysis of LRC in genome sequences is still at an early stage, we can tentatively identify the basis 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. 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 DNA. Actually, it has been estimated that 95% of bulk genomic DNA sequences have an affinity for the histone octamer similar to that of random sequences[21] (this number is likely to be an overestimation, since it includes particular motifs that impair nucleosome formation like TGGA repeats[47]). The remaining 5% of sequences present affinities that are significantly larger than average.[21] How sparsely distributed these specific regions are in genomic DNA 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.[46] Indeed, one cannot exclude the possibility that the rather well positioned nucleosomes are 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.[48] Since a large

proportion (about 95 %) of genomic DNA has a free energy for nucleosome formation that differs little 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.

We propose that, in contrast to the tight histone binding obtained with an adequate periodic distribution of bending sites, LRC would facilitate the positioning of the histone core throughout a major part of the genome. If one considers the translational positioning of nucleosomes as a mechanism of diffusion along the DNA chain and if we assume that, once the histone core is bound to DNA, the distribution of bending sites has a direct consequence on this diffusion process, then an analogy can be made with an already known category of diffusion phenomena. In such phenomena known as ''abnormal diffusion'', the r.m.s. distance covered after a given number of steps is larger than in a classical Brownian motion. In the present case, an analogous situation could be achieved because the nucleosome positioning sites on the DNA present ''correlation properties'' similar to those involved in ''super-diffusive'' processes. LRC would allow nucleosome mobility along DNA to proceed with an average displacement (after a given number of elementary steps) larger than with uncorrelated sequences, in the same manner as LRC induced larger stretches of black (or white) in Figure 1(b), (d), (f), as compared to Figure 1(a), (c), (e). The persistent nature of this scale-invariant spatial organization of bending sites would be selected in order to favour the overall dynamic of compaction of nucleosomes by enabling them to explore larger segments of DNA. In other words, nucleosomes would require less energy for similar amplitude of displacements. Persistence therefore offers some understanding of the modest free-energy of nucleosome formation observed for most DNA sequences, which also facilitates the translational mobility and thus the propensity of nucleosomes to be dynamical structures. Such properties could then favour an optimal compromise between DNA compaction and accessibility constraints. These hypotheses constitute new directions for the study of the effects of LRC on the structural, mechanical and dynamical properties of DNA in chromatin.

Materials and Methods

Coding rules

The A, G, T and C mononucleotide coding rules are defined by putting 1 at the considered nucleotide positions and 0 at the other positions. These mononucleotide codings allow us to study how A, G, T and C are distributed along the DNA sequence. We define the binary coding rules A_{iso}, G_{iso}, T_{iso} and C_{iso} by 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. $A_{\mathrm{iso}}(=T_{\mathrm{iso}})$ and $G_{\mathrm{iso}}(=C_{\mathrm{iso}})$ code with 1 both the isolated A and T on the one hand or both the isolated G and C on the other hand. The AA, GG, TT and CC dinucleotide coding rules consist of 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. The two trinucleotide coding rules given by the Pnuc and DNase tables are obtained, respectively, from Ref.[36] and Ref.[49] The former is deduced from experimentally determined nucleosome positioning.[10] The latter is based on sensitivity of DNA fragments to DNase I.[39] The Pnuc and DNase trinucleotide coding rules are defined by coding the nucleotide n_i at position i by the numerical value given by either one of these tables for the trinucleotide (n_{i-1}, n_i, n_{i+1}). A complete coding of DNA sequences is achieved by repeating this operation for all the positions i from 2 to $L-1$ (L is the length of the considered sequence).

Data sets

All genomes, chromosomes and contigs were downloaded 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 analysed: *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, seven genomes were chosen in the subfamilies of alphaherpesviruse AJ004801, AF030027, X14112; betaherpesviruses, X17403; gammaherpesviruses, AF005370, V01555; and one unclassified, AB049735. For the adenoviruses, three genomes were analysed: human adenovirus type 5, M73260; ovine adenovirus isolate 287, U40839; turkey adenovirus 3, AF074946. For the poxviruses, two genomes were studied: *Vaccinia* virus (strain *Tan Tan*), AF095689; *Melanoplus sanguinipes*, AF063866. For positive 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, four 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 seven retrovirus genera: lentiviruses, L07625; spumaviruses, U21247; mammalian type B retroviruses, M15122; mammalian type C retroviruses, M23385; Aavian type C retroviruses, J02342; D-type retroviruses, M12349; BLV-HTLV retroviruses, K02120.

The following complete bacterial and virus genomes were analysed and included in Table 1: *Aquifex aeolicus*, AE000657; *Bacillus subtilis*, AL009126; *Borrelia burgdorferi*, AE000783; *Buchnera* sp., BA000003; *Campylobacter jejuni*, AL111168; *Chlamydia trachomatis*, AE001273; *Escherichia coli*, U00096; *Helicobacter pylori* 26695, AE000511; *Mycoplasma pneumoniae*, U00089; *Rickettsia prowazekii*, AJ235269; *Synechocystis* sp. PCC6803, AB001339; *Thermotoga maritima*, AE000512; *Treponema pallidum*, AE000520; bacteriophage T4, AF158101; bacteriophage SPBc2, AF020713; the group of 13 eubacterial genomes included in Table 1 is: *Bacillus halodurans*, BA000004; *Chlamydia pneumoniae* AR39, AE002161; *Chlamydia muridarum*, AE002160; *Deinococcus radiodurans*, AE000513 and

AE001825; *Haemophilus influenzae* Rd, L42023; *Helicobacter pylori* J99, AE001439; *Mycobacterium tuberculosis*, AL123456; *Mycoplasma genitalium*, L43967; *Neisseria meningitidis* Z2491, AL157959; *Pseudomonas aeruginosa*, AE004091; *Ureaplasma urealyticum*, AF222894; *Vibrio cholerae*, AE003852 and AE003853; *Xylella fastidiosa*, AE003849; three other genomes were analysed but were not included in Table 1: *Chlamydia pneumoniae* CWL029, AE001363; *Chlamydophila pneumoniae* J138, BA000008, *Neisseria meningitidis* MC58, AE002098.

Archaebacterial genomes: *Thermoplasma acidophilum*, AL139299; *Methanococcus jannaschii*, L77117; *Pyrococcus horikoshii* (NCBI); *Archaeoglobus fulgidus*, AE000782; *Aeropyrum pernix* (NCBI); *Sulfolobus solfataricus* (directly from CBR http://www.cbr.nrc.ca/).

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 2000 (CNRS).

References

1. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature,* **389**, 251-260.
2. van Holde, K. (1989). *Chromatin*, Springer, New York.
3. Orphanides, G. & Reinberg, D. (2000). RNA polymerase II elongation through chromatin. *Nature,* **407**, 471-475.
4. Zlatanova, J., Caiafa, P. & Van Holde, K. (2000). Linker histone binding and displacement: versatile mechanism for transcriptional regulation. *FASEB J.* **14**, 1697-1704.
5. Marini, J. C., Levene, S. D., Crothers, D. M. & Englund, P. T. (1983). A bent helix in kinetoplast DNA. *Cold Spring Harbor Symp. Quant. Biol.* **47**, 279-283.
6. Simpson, R. T. & Stafford, D. W. (1983). Structural features of a phased nucleosome core particle. *Proc. Natl Acad. Sci. USA,* **80**, 51-55.
7. Widom, J. (1984). DNA bending and kinking. *Nature,* **309**, 312-313.
8. Travers, A. A. & Klug, A. (1987). The bending of DNA in nucleosomes and its wider implications. *Phil. Trans. Roy. Soc. ser B,* **317**, 537-561.
9. Drew, H. R. & Travers, A. A. (1985). DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* **186**, 773-790.
10. Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659-675.
11. Bina, M. (1994). Periodicity of dinucleotides in nucleosomes derived from simian virus 40 chromatin. *J. Mol. Biol.* **235**, 198-208.
12. Muyldermans, S. & Travers, A. A. (1994). DNA sequence organization in chromatosomes. *J. Mol. Biol.* **235**, 855-870.
13. Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* **262**, 129-139.
14. Widlund, H. R., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P. E. *et al*. (1997). Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.* **267**, 807-817.
15. Herzel, H., Weiss, O. & Trifonov, E. N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics,* **15**, 187-193.
16. Stein, A. & Bina, M. (1999). A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucl. Acids Res.* **27**, 848-853.
17. Thaström, A., Lowary, P. T., Widlund, H. R., Cao, H., Kubista, M. & Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* **288**, 213-229.
18. Trifonov, E. N. & Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA,* **77**, 3816-3820.
19. Shrader, T. E. & Crothers, D. M. (1989). Artificial nucleosome positioning sequences. *Proc. Natl Acad. Sci. USA,* **86**, 7418-7422.
20. Liu, K., Sandgren, E. P., Palmiter, R. D. & Stein, A. (1995). Rat growth hormone gene introns stimulate nucleosome alignment *in vitro* and in transgenic mice. *Proc. Natl Acad. Sci. USA,* **92**, 7724-7728.
21. Lowary, P. T. & Widom, J. (1997). Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc. Natl Acad. Sci. USA,* **94**, 1183-1188.
22. Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J. F. & Arneodo, A. (2001). Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys. Rev. Letters,* **86**, 2471-2474.
23. Arneodo, A., d'Aubenton-Carafa, Y., Bacry, E., Graves, P. V., Muzy, J.-F. & Thermes, C. (1996). Wavelet based fractal analysis of DNA sequences. *Physica D,* **96**, 291-320.
24. Hosking, J. R. M. (1981). Fractional differencing. *Biometrika,* **68**, 165-176.
25. Mandelbrot, B. B. (1982). *The Fractal Geometry of Nature*, Freeman & Co., San Francisco.
26. Weir, B. S. (1990). Editor of *Genetic Data Analysis, Methods for Discrete Population Genetic Data*, pp. 237-240, Sinauer Associates Inc. Publishers, Sunderland, MA.
27. Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Ossadnik, S. M., Peng, C. K. & Simons, M. (1993). Fractal landscapes in biological systems. *Fractals,* **1**, 283-301.
28. Li, W. & Kaneko, K. (1992). Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Letters,* **17**, 655-660.
29. Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. & Stanley, H. E. (1992). Long-range correlations in nucleotide sequences. *Nature,* **356**, 168-170.
30. Voss, R. F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Letters,* **68**, 3805-3808.
31. Karlin, S. & Brendel, V. (1993). Patchiness and correlations in DNA sequences. *Science,* **259**, 677-680.
32. Arneodo, A., Bacry, E., Graves, P. V. & Muzy, J. F. (1995). Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Letters,* **74**, 3293-3296.

33. Deshmane, S. L. & Fraser, N. W. (1989). During latency, herpes simplex virus type 1 DNA is associated with nucleosomes in a chromatin structure. *J. Virol.* **63**, 943-947.
34. Marcus-Sekura, C. J. & Carter, B. J. (1983). Chromatin-like structure of adeno-associated virus DNA in infected cells. *J. Virol.* **48**, 79-87.
35. Sandman, K. & Reeve, J. N. (2000). Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. *Arch. Microbiol.* **173**, 165-169.
36. Goodsell, D. S. & Dickerson, R. E. (1994). Bending and curvature calculations in *B*-DNA. *Nucl. Acids Res.* **22**, 5497-5503.
37. Munteanu, M. G., Vlahovicek, K., Parthasarathy, S., Simon, I. & Pongor, S. (1998). Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena. *Trends Biochem. Sci.* **23**, 341-347.
38. Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995). Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.* **14**, 1812-1818.
39. Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995). Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. *J. Biomol. Struct. Dynam.* **13**, 309-317.
40. El Hassan, M. A. & Calladine, C. R. (1998). Two distinct modes of protein-induced bending in DNA. *J. Mol. Biol.* **282**, 331-343.
41. Challberg, M. D. & Kelly, T. J. (1989). Animal virus DNA replication. *Annu. Rev. Biochem.* **58**, 671-717.
42. Verdin, E., Paras, P., Jr & Van Lint, C. (1993). Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation. *EMBO J.* **12**, 3249-3259.
43. Bailey, K. A., Pereira, S. L., Widom, J. & Reeve, J. N. (2000). Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. *J. Mol. Biol.* **303**, 25-34.
44. Pereira, S. L., Grayling, R. A., Lurz, R. & Reeve, J. N. (1997). Archaeal nucleosomes. *Proc. Natl Acad. Sci. USA,* **94**, 12633-12637.
45. Decanniere, K., Babu, A. M., Sandman, K., Reeve, J. N. & Heinemann, U. (2000). Crystal structures of recombinant histones HMfA and HMfB from the hyperthermophilic archaeon *Methanothermus fervidus*. *J. Mol. Biol.* **303**, 35-47.
46. Widom, J. (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.* **259**, 579-588.
47. Cao, H., Widlund, H. R., Simonsson, T. & Kubista, M. (1998). TGGA repeats impair nucleosome formation. *J. Mol. Biol.* **281**, 253-260.
48. Liu, K. & Stein, A. (1997). DNA sequence encodes information for nucleosome array formation. *J. Mol. Biol.* **270**, 559-573.
49. Gabrielian, A. & Pongor, S. (1996). Correlation of intrinsic DNA curvature with DNA property periodicity. *FEBS Letters,* **393**, 65-68.

Edited by T. Richmond

(Received 25 July 2001; received in revised form 12 December 2001; accepted 14 December 2001)