Experiments Confirm the Influence of Genome Long-Range Correlations on Nucleosome Positioning

C. Vaillant, B. Audit, and A. Arneodo

Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS, ENS-Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France

(Received 31 May 2007; published 21 November 2007)

From the statistical analysis of nucleosome positioning data for chromosome III of *S. cerevisiae,* we demonstrate that long-range correlations (LRC) in the genomic sequence strongly influence the organization of nucleosomes. We present a physical explanation of how LRC may significantly condition the overall formation and positioning of nucleosomes including the nucleosome-free regions observed at gene promoters. From grand canonical Monte Carlo simulations based upon a simple sequence-dependent nucleosome model, we show that LRC induce a patchy nucleosome occupancy landscape with alternation of ''crystal-like'' phases of confined regularly spaced nucleosomes and ''fluidlike'' phases of rather diluted nonpositioned nucleosomes.

DOI: 10.1103/PhysRevLett.99.218103 PACS numbers: 87.15.Cc, 87.14.Gg, 87.15.Aa, 87.16.Sr

Because the spatial organization of DNA plays a major role in regulating nuclear metabolisms including transcription and replication [1], it is natural to wonder to which extent the genomic sequence codes for the structure and dynamics of chromatin. In eukaryotes, for instance, the first level of organization is the packaging of DNA in arrays of nucleosomes consisting in the wrapping of 146 bp of DNA around a protein octamer [2]. Many experiments have shown that the local mechanical properties of the DNA double helix depend on the nucleotide content and are likely to be sensed by binding proteins [3]. Hence, the energy cost for nucleosome formation presents a sequence-dependent contribution which may influence the overall formation and positioning of nucleosomes [4] and in turn proteins access to their target sites [5]. In this setting, identifying the underlying mechanisms that control the linear organization of nucleosomes appears as a challenging task; even more so, we would like to determine to which extent this organization has been imprinted in DNA sequence during evolution.

It is known that some genomic sequences presenting a 10 bp periodicity in the distribution of di- or trinucleotides (e.g., AA/TT) show (*in vivo*) higher affinity for nucleosomes [6]. The periodic positioning of these motifs might contribute in a coherent manner to a global spontaneous curvature of DNA that would favor its wrapping on the histone surface [7]. This explains why genomic 10 bp periodicity has been considered as a typical nucleosomal positioning signal for years [8]. However, additional *in vitro* experiments have revealed that more than 95% of genomic sequences present binding affinity for core histones that do not differ from random fragments, the strongest positioning sequences being artificial ones [9]. To this day, even though some phenomenological models provide satisfactory agreement between theory and experiments [10], a properly detailed model of nucleosome remains an unachieved goal. At larger scales, *in vivo* and *in vitro* studies have shown that chromatin can locally present a well-ordered linear organization with a well-defined nucleosomal repeat which turns out to be tissue and species specific [11]. Yet, the biological and physical basis of such periodic ordering and the role of the underlying DNA sequence are still unclear.

As an alternative to the 10 bp periodicity nucleosome positioning signal that fails to account for the vast majority of bulk nucleosomes, a recent work has shown that in eukaryotic genomes, sequence motifs that favor the local bending of the double helix present long-range correlations (LRC) that are related to the scale-invariance properties of the DNA sequences [12]. When analyzing the intrinsic bending profiles generated using the experimental ''Pnuc'' coding table established from nucleosome positioning data [13], two LRC regimes were identified. In the 10–200 bp range, LRC are observed for eukaryote sequences as quantified by a Hurst exponent value $H \sim 0.6$ (but not for eubacterial sequences that are found uncorrelated with $H = 0.5$), as the signature of the nucleosome structure. These LRC were shown to favor the formation of small (a few hundreds bp) 2D DNA loops and in turn the propensity of eukaryotic DNA to interact with histones to form nucleosomes [14]. Over larger distances (from 200 to $\sim$5000 bp), stronger LRC with $H = 0.8$ are observed in all sequences [12,15] and are likely to play a role in the condensation of the nucleosomal string into the 30 nm fiber.

Recently, Yuan *et al.* [16] used the tiled microarray approach to experimentally measure nucleosome density $\rho(s)$ along chromosome III of *S. cerevisiae*. As shown in Figs. 1(a) and 1(c), the nucleosome occupancy landscape (OL), defined as $Y(s) = \ln[\rho(s)]$, displays a rather disordered behavior at the chromosomal level despite the presence of some local periodic oscillations (of mean period $l^* \sim 160 \pm 10$ bp) corresponding to regular arrays of nucleosomes. Yuan *et al.* have estimated that a majority of the nucleosomes are preferentially positioned on well-defined genomic positions along the chromosome. From a fine

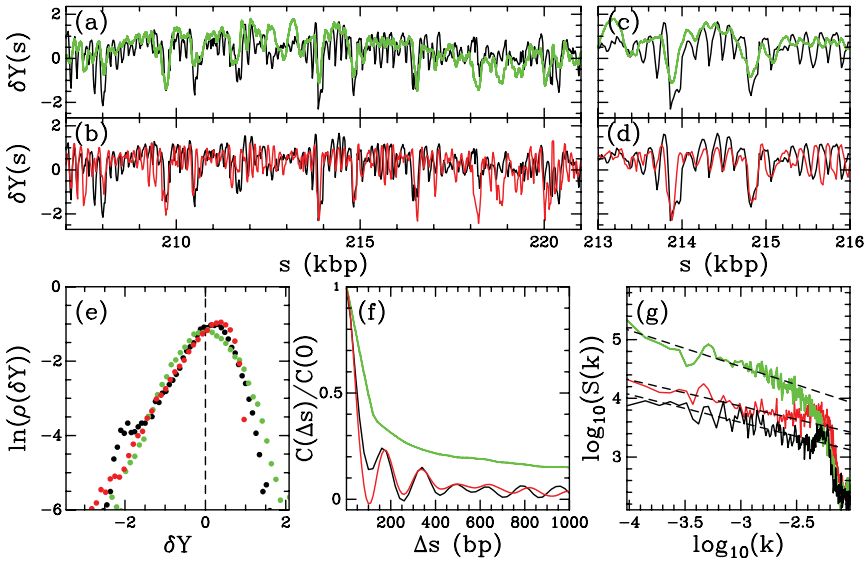

FIG. 1 (color). Statistical analysis and modeling of Yuan *et al.* nucleosome positioning data [16]. (a)–(d) The experimental OL: $\delta Y(s) = Y(s) - \bar{Y}$ (black), the theoretical negative energy landscape (EL): $-\beta\delta F_{P\mathrm{nuc}}(s, 125)$ (green) for a wrapping length $l = 125$ bp, and the numerical Monte Carlo OL (MCOL): $\delta \ln\rho(s)$ (red) obtained from Monte Carlo simulations (see the text). (e) Pdfs, (f) two-point correlation function $C(\Delta s)$, (g) power spectrum of $\delta Y(s)$ (black), $-\beta\delta F_{p\mathrm{nuc}}(s, 125)$ (green), and $\delta \ln\rho(s)$ (red). In (g) the dashed lines correspond, from top to bottom, to power-law scaling exponents $\nu = 0.60$, 0.45, and 0.48 corresponding to $H = 0.80$, 0.73, and 0.74 LRC properties, respectively.

mapping of a large set of yeast promoters, they have confirmed that promoters are essentially depleted in nucleosomes and bordered by patterns of well-positioned nucleosomes [16]. Segal *et al.* [17] and Ioshikhes *et al.* [18] have further analyzed the Yuan *et al.* data and have drawn the conclusion that a large set of well-defined nucleosome positions could effectively be related to a "10 bp periodic genomic nucleosomal pattern." Using probabilistic models that take into account the matching of their pattern to the sequence and the steric hindrance between nucleosomes, they predict a nucleosome OL that correlates well with the data, especially in regions of well-positioned nucleosomes. However, the nucleosomal 160 bp short-range ordering is not the only characteristic feature of the experimental OL. As shown in Fig. 1(g), the corresponding power spectrum displays a power-law decay $S(k) \propto k^{-\nu}$, with exponent $\nu = 0.48 \pm 0.02$ that is likely to be a direct consequence of the presence of LRC in the DNA sequence. Here, our goal is to show that the LRC observed in the sequence are not only recovered in the nucleosome OL but that they globally provide a convincing understanding of the observed alternation of patterns of well-ordered nucleosomes confined near energy barriers and of regions of more diluted nonpositioned nucleosomes.

As reported in Fig. 1, when performing statistical analysis of the OL fluctuations $\delta Y(s) = Y(s) - \bar{Y}$, we observe rather peculiar features. In Fig. 1(e), we show that the OL fluctuation pdf is nonsymmetric with an exponential-like tail for low occupancy values. In Fig. 1(f), the autocorrelation function $C(\Delta s) = \overline{\delta Y(s)\delta Y(s + \Delta s)}$ clearly confirms the presence of a small-scale periodic arrangement of nucleosomes with a characteristic spacing $l^* = 160 \pm 10$ bp in good agreement with the well-known 165 bp value for yeast [11]. But this correlation function also reveals that this "nucleosomal periodicity" statistically appears as a modulation of a dominant slow-decaying component which mainly characterizes the large-scale disordered OL fluctuations. $C(\Delta s) \propto s^{-\alpha}$ decays as a power-law with an exponent $\alpha = 2 - 2H = 0.52 \pm 0.02$, in good agreement with the $H = 0.8$ value characterizing the LRC in yeast sequences [12]. This result is quite consistent with the value found for the spectral exponent in Fig. 1(g), $\nu = 2H - 1 = 0.48$. Altogether these results demonstrate that the spatial organization of nucleosomes is long-range correlated with characteristics similar to the LRC imprinted in the sequence.

Here, we aim at providing some understanding of the above observations using a phenomenological approach based on simple physical arguments. When focusing on the dynamical assembly of histone octamers along the DNA chain, chromatin can be reasonably modeled by a fluid of rods of finite extension l (the DNA wrapping length around the octamer), binding and moving in an external potential $F(s)$ (the effective nucleosome formation potential) and interacting [potential $v(s, s')$] along a 1D substrate (the DNA chain) [19,20]. Within the grand canonical formalism, considering that the fluid is in contact with a thermal bath (at reciprocal temperature β) and a histone reservoir (at chemical potential μ), the thermodynamical equilibrium properties of the system are described by the grand partition function:

$$\Xi = \sum_{N=0}^{N_{\max}} \frac{1}{N!} \int D[\mathbf{s}^{(N)}] \exp\{-\beta[V(\mathbf{s}^{(N)}) - \mu N]\}, \quad (1)$$

where $V(\mathbf{s}^{(N)}) = \sum_1^N F(s_k) + 1/2\sum_{j\neq k} v(s_k, s_j)$ is the total potential energy of the N-rod system. From Eq. (1), we obtain the nucleosome density profile as $\rho(s) = -\beta^{-1}\partial \ln\Xi/\partial F(s)$. For monodisperse hard rods on a uniform potential, this is the well-known Tonks gas [21]. In the case of a nonuniform external potential of interest here, the problem was solved by Percus [22] who derived a functional relationship between the residual chemical potential $\mu - F(s)$ and the hard rod density $\rho(s)$.

To compute the free EL for the formation of one nucleosome at a given position along DNA, we will assume that (i) DNA is an unshearable elastic rod whose conformations are described by the local angles $\Omega_1(s)$ (tilt), $\Omega_2(s)$ (roll), $\Omega_3(s)$ (twist) and (ii) the DNA chain along the nucleosome at position s is constrained to form an ideal superhelix of radius $R = 4.19$ nm and pitch $P = 2.59$ nm [2] over a total length l, which forces the distribution of angular deformations $[\Omega_i^{\mathrm{nuc}}(u)]_{i=1,2,3}$, $u = s, \ldots, s + l$. Within linear elasticity approximation, the energy cost for nucleosome formation is written as

$$\beta F(s, l) = \int_s^{s+l} \sum_{i=1}^{3} \frac{A_i}{2} [\Omega_i^{\mathrm{nuc}}(u) - \Omega_{\mathrm{oi}}(u)]^2 du, \quad (2)$$

where A_1, A_2, and A_3 are the stiffnesses associated to the tilt, roll, and twist deformations around their intrinsic values Ω_{o1}, Ω_{o2}, and Ω_{o3}, respectively. Consistent with our previous works [12,14], we shall make use of the Pnuc structural bending table [13] which is a trinucleotide roll coding table (Ω_{o2}), with zero tilt ($\Omega_{o1} = 0$) and constant twist ($\Omega_{o3} = 2\pi/10.5$). As the values of this bending table were arbitrarily assigned between 0 and $\pi/18$, we will perform the following affine rescaling $\Omega^*_{o2} = \gamma\Omega_{o2} - \eta$ with $\gamma = 0.4$, $\eta = 0.06$, in order to get a comparable range of EL fluctuations as in the experiments.

In Fig. 1(a), the numerical EL along the yeast chromosome III obtained for a wrapping length $l = 125$ bp is shown. In a single nucleosome system, the occupancy probability of a site s is given by $\rho(s) \propto \exp[-\beta F(s, l)]$. Thus, when comparing the negative EL fluctuations $-\beta\delta F(s, 125)$ of such a dilute system with the Yuan *et al.* OL fluctuations $\delta Y(s)$, we observe in Figs. 1(a) and 1(c) a surprisingly strong similarity. In particular, we see that the strongest nucleosome depleted regions are well predicted by high energy barriers that manifest as heavy exponential tail in the EL and OL fluctuations pdfs. Note that we now observe that the autocorrelation function decays as a power law with exponent $\alpha = 0.40 \pm 0.02$ [Fig. 1(f)] consistently with the power-law exponent $\nu = 0.60 \pm 0.02$ of the power spectrum [Fig. 1(g)]. These sequence-induced $H = 0.8$ LRC properties are in good agreement with the corresponding exponents found for the experimental OL except for the fact that $C(\Delta s)$ does not display any periodic modulation.

Being more than one nucleosome, interactions between neighboring nucleosomes have to be considered. We assume here that these interactions reduce to steric hindrance, modeled by a hard core potential of size l between the 1D rods. To compute the rods OL of this high-density system, we now turn to Grand Canonical Monte Carlo simulations [23]. As shown in Figs. 1(b) and 1(d), this MCOL nicely matches the experimental OL since it now accounts for both the disordered nonpositioning component but also for the small-scale periodic positioning patterns. On top of a better concordance of the MCOL and OL fluctuation pdfs [Fig. 1(e)], we also recover the additional harmonic modulation with period $l^* \sim 170$ bp in $C(\Delta s)$ [Fig. 1(f)]. Furthermore, the power-law exponents of both autocorrelation function [Fig. 1(f)] and power spectrum [Fig. 1(g)], namely, $\alpha = 0.54 \pm 0.02$ and $\nu = 0.45 \pm 0.02$, are sligthly different from those of the EL but in better agreement with those of the experimental OL. Note that these results are obtained for a hard rod size of 125 bp and a chemical potential μ fixing the average density of about one nucleosome per 250 bp in good agreement with recent experimental bulk nucleosome density estimates [24]. In future work, we plan to refine our model by (i) taking into account a spatial short-range attraction between neighboring nucleosomes [20,25] and (ii) considering a fluctuating wrapping length [19,26]. This should further improve the matching to the experimental results and confirm the major role of LRC genomic sequences in nucleosome positioning.

When focusing on a small region of chromosome III [Fig. 1(c) and 1(d)], we realize that, whereas the EL does account for the overall shape and, in particular, for the two depleted regions, it clearly fails to reproduce the periodic patterns in-between these two energy barriers. But when considering a higher nucleosome density in this EL [Fig. 1(d)], the MCOL reproduces remarkably well the experimental data. This is a nice illustration of a very simple mechanism of nucleosome ordering already proposed by Kornberg and Stryer [27]. When investigating 1D hard rod distribution in the vicinity of a boundary (i.e., a fixed exclusion zone), we obtain, for a sufficiently dense system, a damped oscillating density pattern, i.e., a statistical short-range ordering due to the interplay between boundary confinement and rod-rod excluded volume interaction. Realistic boundaries can be induced by stably bound proteins on specific sites [28], fixed nucleosomes on strong positioning sequences, or alternatively an unfavorable sequence inducing a depletion zone. Let us point out that this confinement-induced nucleosome ordering can be enhanced when two high barriers are found at a sufficiently short distance from each other, as illustrated in Fig. 1(d). Actually there exists a functional relationship between the EL and the hard rods OL [22]. Ordering is a function of confinement which itself is controlled by the EL topography and the chemical potential. At low density the confinement is weak and nucleosomes distribute everywhere according to EL fluctuations. Our Monte Carlo simulations confirm that when the nucleosome density is increased, ordering progressively appears around the highest energy barriers leading to an overall organization with "crystallike" phases coexisting with "fluidlike" phases where ordering is lost. As shown by visual inspection in Fig. 1(a) and by means of correlation analysis in Fig. 2(a), the experimental OL displays such a patchy landscape with regions of regular nucleosome ordering as characterized by a modulated ($l^* \sim 170$ bp) power-law

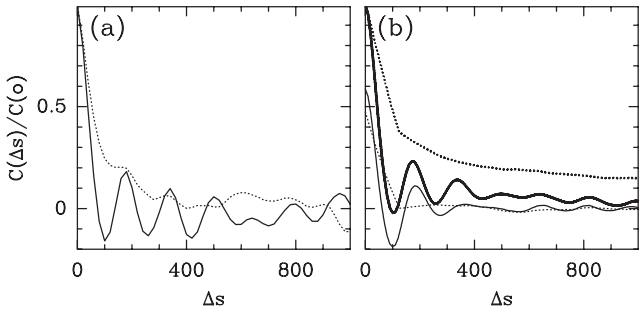

FIG. 2. (a) Autocorrelation function of the experimental OL computed in regions of 5000 bp: average over 5 regions with nonpositioned nucleosomes (dotted line) and 5 regions with well-positioned nucleosomes (solid line). (b) $C(\Delta s)/C_{\text{yeast}}(0)$ of the theoretical EL (dotted lines) and MCOL (solid lines) for the chromosome III of yeast (thick lines) and for an artificial random uncorrelated ($H = 0.5$) sequence (thin lines).

decay of $C(\Delta s)$, and regions with no ordering where $C(\Delta s)$ no longer display any oscillatory component.

We can now address the influence of LRC by comparing the theoretical results for yeast and those for uncorrelated sequences ($H = 0.5$). The difference is striking: as shown in Fig. 2(b), for the same chemical potential, the MCOL of uncorrelated sequences reveals a weaker short-range ordering with a larger nucleosomal period $l^* \sim 180$ bp. Indeed ''crystal-like'' phases are shorter, less ordered, and with a larger nucleosomal repeat. More importantly, large-scale OL does not present any correlation, similarly to the EL. This is an evidence that at this intermediate average nucleosomal density, the OL fluctuations drastically depend at all scales on the underlying EL and consequently on the presence or absence of genomic LRC. Indeed, genomic LRC lead to energy landscapes that fluctuate more smoothly over short distances but with a larger amplitude over large distances. In other words, LRC lead to energy landscapes with wider and deeper energy ''valleys'' separated by higher energy barriers. Hence, as compared to uncorrelated sequences, LRC not only promote the formation of nucleosomes [14] but they also enhance their confinement into larger ''arrays'' of better positioned and regularly spaced nucleosomes separated by regions of nonpositioned nucleosomes.

In conclusion, the results reported in this work corroborate that the positioning of nucleosomes is fundamentally influenced by the sequence over rather long distances. In particular, the data of Yuan *et al.* [16] confirm that the LRC in the sequence disorder contribute to the overall arrangement of the nucleosomes into crystal-like phases of confined regularly spaced nucleosomes and fluidlike phases of rather nonpositioned nucleosomes. Furthermore, these LRC provide a very attractive interpretation of the nucleosome-free regions observed at gene promoters [29] in terms of very high energy barriers that are absent in EL generated from uncorrelated sequences.

We are very grateful to P. St-Jean for a careful reading of our manuscript. This work was supported by ACI IMPBio 2004, the Conseil Regional Rhônes-Alpes (Emergence 2005), and the Agence Nationale de la Recherche (No. PCV06135105).

[1] G. S. Stein *et al.*, Trends Cell Biol. **13**, 584 (2003).
[2] T. J. Richmond and C. A. Davey, Nature (London) **423**, 145 (2003).
[3] P. J. Hagerman, Annu. Rev. Biophys. Biophys. Chem. **17**, 265 (1988); D. M. Crothers, T. E. Haran, and J. G. Nadeau, J. Biol. Chem. **265**, 7093 (1990).
[4] O. J. Rando and K. Ahmad, Curr. Opin. Cell Biol. **19**, 250 (2007).
[5] X. Liu, Genome Res. **16**, 1517 (2006).
[6] S. C. Satchwell, H. R. Drew, and A. A. Travers, J. Mol. Biol. **191**, 659 (1986).
[7] E. N. Trifonov, Nucleic Acids Res. **8**, 4041 (1980).
[8] I. P. Ioshikhes *et al.*, J. Mol. Biol. **262**, 129 (1996); J. Widom, Q. Rev. Biophys. **34**, 269 (2001); A. B. Cohanim, Y. Kashi, and E. N. Trifonov, J. Biomol. Struct. Dyn. **22**, 687 (2005).
[9] A. Thaström *et al.*, J. Mol. Biol. **288**, 213 (1999).
[10] C. Anselmi *et al.*, Biophys. J. **79**, 601 (2000).
[11] C. L. Woodcock *et al.*, Chromosome Res. **14**, 17 (2006).
[12] B. Audit *et al.*, Phys. Rev. Lett. **86**, 2471 (2001); J. Mol. Biol. **316**, 903 (2002).
[13] D. S. Goodsell and R. E. Dickerson, Nucleic Acids Res. **22**, 5497 (1994).
[14] C. Vaillant, B. Audit, and A. Arneodo, Phys. Rev. Lett. **95**, 068101 (2005); C. Vaillant *et al.*, Eur. Phys. J. E **19**, 263 (2006).
[15] J. Moukhtar *et al.*, Phys. Rev. Lett. **98**, 178101 (2007).
[16] G.-C. Yuan *et al.*, Science **309**, 626 (2005).
[17] E. Segal *et al.*, Nature (London) **442**, 772 (2006).
[18] I. P. Ioshikhes *et al.*, Nat. Genet. **38**, 1210 (2006).
[19] T. Chou, Europhys. Lett. **62**, 753 (2003).
[20] F. J. Solis *et al.*, Biophys. J. **87**, 3372 (2004); Biochemistry **46**, 5623 (2007).
[21] L. Tonks, Phys. Rev. **50**, 955 (1936).
[22] J. K. Percus, J. Stat. Phys. **15**, 505 (1976).
[23] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Elsevier Academic Press, Amsterdam, 1986).
[24] W. Tong *et al.*, J. Mol. Biol. **361**, 813 (2006).
[25] F. Mühlbacher, H. Schiessel, and C. Holm, Phys. Rev. E **74**, 031919 (2006).
[26] F. Montel *et al.*, Biophys. J. **93**, 566 (2007).
[27] R. D. Kornberg and L. Stryer, Nucleic Acids Res. **16**, 6677 (1988).
[28] R. H. Pusaria, V. Vinayachandran, and P. Bhargava, FEBS J. **274**, 2396 (2007).
[29] V. Miele *et al.* (to be published).