Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions

Răzvan V. Chereji, Denis Tolkunov, George Locke, and Alexandre V. Morozov

Statistical mechanics of nucleosome ordering by chromatin structure-induced two-body interactions

Răzvan V. Chereji,[1] Denis Tolkunov,[1, 2] George Locke,[1] and Alexandre V. Morozov[1, 2, *]

[1]*Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854-8019*
[2]*BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854-8019*

One-dimensional arrays of nucleosomes (DNA-bound histone octamers separated by stretches of linker DNA) fold into higher-order chromatin structures which ultimately make up eukaryotic chromosomes. Chromatin structure formation leads to $10 - 11$ base pair (bp) discretization of linker lengths caused by the smaller free energy cost of packaging nucleosomes into regular chromatin fibers if their rotational setting (defined by the DNA helical twist) is conserved. We describe nucleosome positions along the fiber using a thermodynamic model of finite-size particles with both intrinsic histone-DNA interactions and an effective two-body potential. We infer one- and two-body energies directly from high-throughput maps of nucleosome positions. We show that chromatin structure explains *in vitro* and *in vivo* nucleosome ordering in transcribed regions, and plays a leading role in establishing well-known $10 - 11$ bp genome-wide periodicity of nucleosome positions.

In living cells, eukaryotic DNA is found in a compact, multi-scale chromatin state [1]. The fundamental unit of chromatin is a nucleosome: 147 bp of DNA wrapped around a histone octamer [2]. In addition to its primary function of DNA compaction, chromatin modulates DNA accessibility to transcription factors and other molecular machines in response to external signals, exerting a profound influence on numerous DNA-mediated biological processes such as gene transcription, DNA repair, and replication [3].

Equilibrium thermodynamic models that account for intrinsic histone-DNA sequence preferences and nearest-neighbor steric exclusion have been used to predict nucleosome positions and formation energies [4–6]. However, structural regularity of the chromatin fiber imposes additional constraints, leading to discretization of linker lengths between neighboring nucleosomes with the $10-11$ bp periodicity of the DNA double helix [7, 8]. The discretization is required to avoid steric clashes caused by the nucleosome rotating around the linker DNA axis as the linker is extended [9], and more generally to minimize the free energy costs associated with maintaining a regular pattern of protein-protein and protein-DNA contacts in the chromatin fiber [8]. Indeed, adding a short DNA segment to the linker will result in a rotation of the nucleosome with respect to the rest of the fiber, disrupting its periodic structure. This additional twist has to be compensated unless the segment is $10 - 11$ bp in length, bringing the nucleosome into an equivalent rotational position.

Large-scale maps of *in vivo* and *in vitro* nucleosome positions in yeast reveal nucleosome-depleted regions (NDRs) in the vicinity of transcription start and termination sites (TSS and TTS) [5, 10, 11]. In these experiments, chromatin is digested with micrococcal nuclease to obtain mononucleosome core particles, and the mononucleosomal DNA is purified and either sequenced or hybridized to microarrays [12]. 5' NDRs play a key role in gene regulation [10]. NDRs are also observed *in vitro*, where they are defined by poly(dA:dT) tracts and other nucleosome-disfavoring sequences. Surprisingly, there are no oscillations in nucleosome occupancy around *in vitro* NDRs and, on average, just a $\sim 25\%$ depletion of the occupancy over 5' NDRs compared with the genome-wide mean [5, 11] (bp occupancy is defined as its probability to be nucleosome-covered). This is true even if genomic DNA from *S.cerevisiae* is combined with purified histones in a 1:1 mass ratio, yielding a maximum nucleosome occupancy of 0.82 which is close to the *in vivo* value [11]. This behavior is in sharp contrast with *in vivo* chromatin in which the action of transcription factors, chromatin remodeling enzymes and components of transcriptional machinery results in well-positioned genic nucleosomes and highly pronounced 5' NDRs ($\sim 70\%$ depletion on average with respect to the mean) [5, 10]. Because occupancy oscillations are a generic feature of one-dimensional liquids of finite-size particles in the vicinity of potential barriers and wells [13], the absence of such oscillations *in vitro* and shallow NDRs strongly suggest that sequence-specific histone-DNA interaction energies are on average comparable to $k_B T$. Consequently, nucleosome-positioning and disfavoring sequences are expected to play a minor role in establishing *in vivo* localization of genic nucleosomes.

Here we focus on how nucleosome positions are affected by effective two-body interactions imposed on neighboring particles by regular chromatin structure. We map a three-dimensional chromatin fiber onto a system of non-overlapping particles of length $a = 147$ bp with both histone-DNA and short-range nearest-neighbor interactions. The particles are confined to a one-dimensional lattice of length L. We develop a theory in which the interaction (that reflects linker discretization) is deduced exactly from the two-particle distribution, even in the
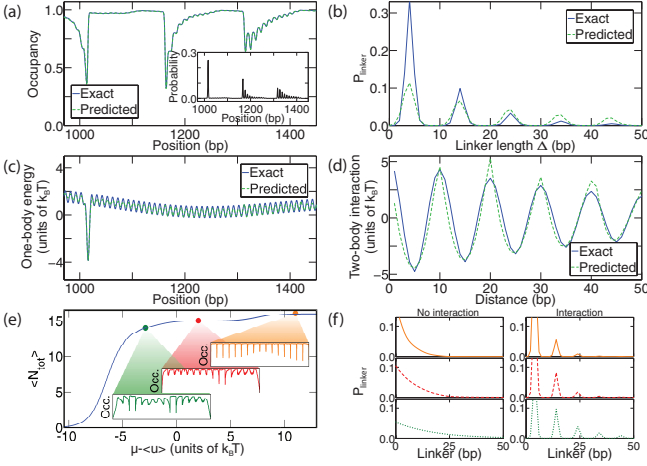
FIG. 1: (Color online) A model with 10 bp oscillations in both one-body and two-body energies. The two-body interaction is $\Phi(x) = A\cos\left(\frac{2\pi}{10}x\right)e^{-x/b}$, where $A = 5\ k_BT$ and $b = 50$ bp. For the one-body potential, 10 bp oscillations with the 0.5 k_BT amplitude were superimposed onto a smooth energy profile with two $-5\ k_BT$ potential wells separated by 1000 bp. DNA length of 2416 bp was chosen to be able to position 16 nucleosomes with 151 bp repeat length. The occupancy profile (a), the linker length distribution (b), the one-body energy (c), and the two-body interaction (d): exact (solid blue line) and predicted (dashed green line). $\mu - \langle u \rangle = -1\ k_BT$ in (a)-(d). Inset of (a): the probability of starting a nucleosome at a given bp. (e) Average number of nucleosomes $\langle N_{tot} \rangle$ vs. $\mu - \langle u \rangle$. Insets: Occupancy profiles corresponding to three different chemical potentials. (f) Linker length distributions for three values of $\langle N_{tot} \rangle$ shown as points in (e), with and without two-body interactions.

presence of $10-11$ bp periodic one-body energies related to the rotational positioning of the nucleosome [10, 11].

Let $u(k)$ be the external potential energy of a particle that occupies positions k through $k + a - 1$ on the DNA, and let $\Phi(k,l)$ be the two-body interaction between a pair of nearest-neighbor particles with starting positions k and l, respectively. Here $u(k)$ describes intrinsic histone-DNA interactions, whereas $\Phi(k,l)$ accounts for the effects of chromatin structure. The grand-canonical partition function is given by

$$Z = 1 + \sum_{N=1}^{N_{max}} \langle J|(zw)^{N-1}z|J\rangle = 1 + \langle J|(I - zw)^{-1}z|J\rangle, \tag{1}$$

where N_{max} is the maximum number of particles that can be positioned on L bp, I is the identity matrix, $|j\rangle$ is a unit vector of dimension $L - a + 1$ with 1 at position j, and $|J\rangle = \sum_{j=1}^{L-a+1} |j\rangle$. In matrix notation, $\langle k|z|l\rangle = e^{\beta[\mu-u(k)]}\delta_{k,l}$ and $\langle k|w|l\rangle = e^{-\beta\Phi(k,l)}\Theta(l - k)$, where μ is the chemical potential, $\delta_{k,l}$ is the Kronecker delta, β is the inverse temperature, and Θ is the Heaviside step function.

The one-particle and nearest-neighbor pair distribution functions are:

$$n(i) = \frac{1}{Z}\langle J|(I - zw)^{-1}|i\rangle\langle i|z|i\rangle\langle i|(I - wz)^{-1}|J\rangle, \tag{2}$$

$$\overline{n}_2(i,j) = \frac{1}{Z}\langle J|(I - zw)^{-1}|i\rangle\langle i|zwz|j\rangle\langle j|(I - wz)^{-1}|J\rangle. \tag{3}$$

Note that for $0 < j - i < 2a$, $\overline{n}_2(i,j) = n_2(i,j)$, where n_2 is the ordinary two-particle distribution function. Defining two matrices, $\langle i|N|j\rangle = n(i)\delta_{i,j}$ and $\langle i|N_2|j\rangle = \overline{n}_2(i,j)$, we rewrite the partition function as

$$Z = \frac{1}{1 - \langle J|(I - N_2 N^{-1})N|J\rangle}. \tag{4}$$

By inverting Eqs. (2) and (3) we obtain the exact expressions for one- and two-body energies [14, 15]:

$$-\beta[u(k) - \mu] = \ln\left(\frac{\langle J|I - N_2 N^{-1}|k\rangle\langle k|N|k\rangle\langle k|I - N^{-1}N_2|J\rangle}{1 - \langle J|I - N_2 N^{-1}N|J\rangle}\right), \tag{5}$$

$$-\beta\Phi(k,l) = \ln\left(\frac{\langle k|N^{-1}N_2 N^{-1}|l\rangle\left[1 - \langle J|I - N_2 N^{-1}N|J\rangle\right]}{\langle k|I - N^{-1}N_2|J\rangle\langle J|I - N_2 N^{-1}|l\rangle}\right). \tag{6}$$

Note that if the two-body interactions are neglected, Eq. (5) reduces to [6]

$$e^{-\beta[u(i)-\mu]} = \frac{n(i)}{1 - O(i) + n(i)} \prod_{j=i}^{i+a-1} \frac{1 - O(j) + n(j)}{1 - O(j)}, \tag{7}$$

where $O(i)$ is the nucleosome occupancy of bp i [$O(i) = \sum_{j=i-a+1}^{i} n(j)$].

If one-body energies u and two-body interactions Φ are known, Eqs. (2) and (3) allow us to construct particle distributions n and $\overline{n}_2$ exactly. Conversely, we can use Eqs. (5) and (6) to find u and Φ from one- and two-particle distributions. However, the two-particle distribution is not directly measured in current high-

throughput experiments, in which chromatin from many cells is mixed together before mononucleosomes are isolated and sequenced. In other words, it is not known which particular genome a given nucleosome comes from. This is irrelevant for n but may present a problem for $\overline{n}_2$, which requires two-nucleosome configurations. Nonetheless, we can build a model for $\overline{n}_2$ which allows us to approximate the two-body interaction.

Let $g(i, j)$ be the pair distribution $n_2(i, j)/[n(i)n(j)]$. Without one-body energies, the system is homogeneous and g is a function of only the relative distance between the nucleosomes: $g(i, j) \equiv g(j - i)$. In this case Eq. (6) reduces to

$$-\beta\Phi(i, j) = \ln[g(j - i)] + \alpha(j - i) + \ln C \quad (8)$$

for arbitrary interactions Φ [16]. The constants C and α can be determined from the asymptotic condition $\lim_{(j-i)\to\infty} \Phi(i, j) = 0$. However, position-dependent one-body energies break translational invariance of the pair distribution g. Assuming that Φ is translationally invariant, we introduce $P_{\text{linker}}(\Delta) = \langle g(i, i + \Delta + a)\rangle_i$ and approximate Φ as

$$-\beta\Phi(i, j) \approx \ln[P_{\text{linker}}(j - (i + a))] + \alpha(j - i) + \ln C. \quad (9)$$

This step is reminiscent of replacing the ensemble average with the time average in statistical mechanics. Our numerical tests show this to be an excellent approximation, even if one- and two-body energies are comparable in magnitude, making the system strongly inhomogeneous.

Experimental nucleosome positioning data sets consist of the histogram of the number of nucleosomes starting at each genomic bp i. We preprocess the data by removing all counts of height 1 from the histogram, and smoothing the remaining profile with a $\sigma = 2$ Gaussian kernel. Next, we compute $n(i)$ by rescaling the smoothed profile so that the maximum occupancy for each chromosome is 1. Finally, we identify all local maxima on the n profile and assume that they mark prevalent nucleosome positions. For each maximum at bp i we find subsequent maxima at positions $i + 146 < j_1 < j_2 < j_3 < \ldots$ in the 50 bp window. To each pair of maxima $(i, j_1), (i, j_2), \ldots$ we assign the probability that they represent neighboring nucleosomes: $n(i)n(j_1), n(i)[1 - n(j_1)]n(j_2), \ldots$ By summing over all initial positions i and normalizing, we obtain the linker length probability which gives us an empirical estimate of P_{linker}.

Fig. 1 demonstrates our procedure in a model system, with preprocessing and rescaling steps skipped since the simulated n profile is noise-free and already properly normalized. Specifically, we use local maxima in the nucleosome starting probability profile [inset of Fig. 1(a)] to obtain P_{linker} [Fig. 1(b)]. Fig. 1(d) shows that the two-body interaction can be reconstructed using Eq. (9), even in the presence of one-body energies with the same periodicity. The reconstruction is facilitated by the presence
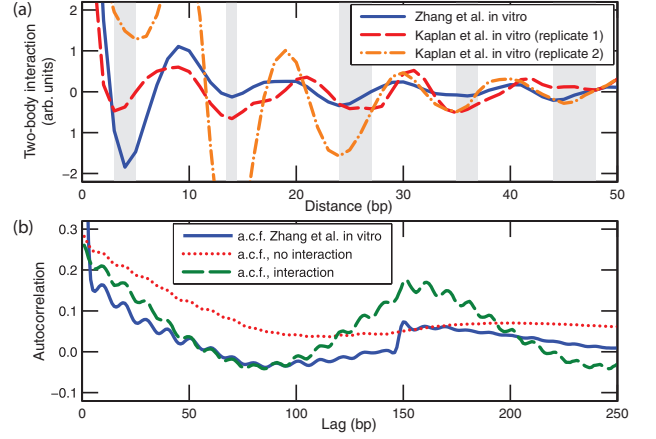

FIG. 2: (Color online) (a) Two-body interaction Φ inferred from *in vitro* maps of nucleosome positions [5, 11]. Grey bars indicate consensus positions of the minima. (b) Autocorrelation of nucleosome starting positions in one of the *in vitro* data sets [11], and of starting positions predicted using sequence-specific one-body energies from the "spatially resolved" model [6], with and without Φ. The two-body potential is from Fig. 1, consistent with the minima of Φ observed in (a). The one-body energies have $\sigma = 0.23\ k_B T$. To account for the limited size of the *in vitro* data set, model output was degraded by randomly removing 1% of predicted nucleosome probabilities.

of potential wells or barriers in the one-body energy profile that are strong enough to create non-uniform density of nearby nucleosomes. To find the one-body energies, we substitute predicted Φ into Eq. (2), which we solve numerically for z [Fig. 1(c)]. Nucleosome occupancies inferred from predicted u and Φ are virtually identical to the exact profile [Fig. 1(a)].

As the chemical potential is increased, nucleosomes undergo a transition in which their average number goes up in a step-like fashion [Fig. 1(e)] [17]. In contrast to the $\Phi = 0$ case in which linkers are distributed exponentially, two-body interactions lead to the pronounced discretization of linker lengths [Fig. 1(f)]. The first minimum of Φ becomes more dominant as the number of nucleosomes increases, leading to a well-positioned array with 4 bp-long linkers.

We now use Eq. (9) to predict nearest-neighbor interactions from genome-wide nucleosome maps [Fig. 2(a)]. We find that despite significant experiment-to-experiment variations, all two-body potentials have minima within $1 - 2$ bp of $5 + 10m$ bp, $m = 0, 1, \ldots$ [18]. Surprisingly, there are substantial differences between two Kaplan *et al.* [5] *in vitro* replicates, with one replicate exhibiting higher values of Φ due to the pronounced depletion of nucleosomes separated by < 10 bp. Apparently, chromatin structure can undergo subtle uncontrolled changes from experiment to experiment.

Two-body interactions are reflected in the autocorrelation of nucleosome starting positions [Fig. 2(b)].
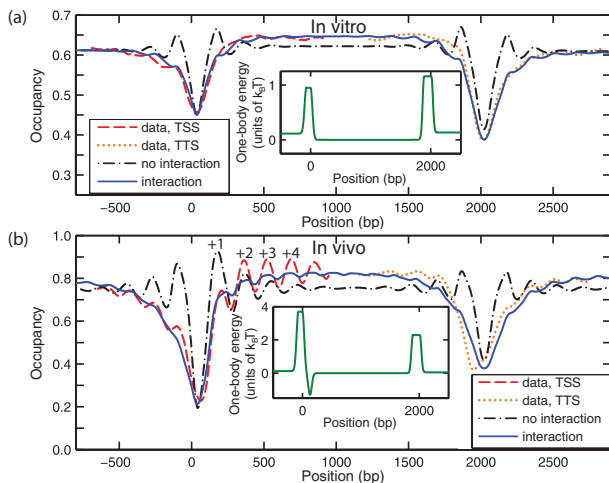
FIG. 3: (Color online) A minimal model of nucleosome ordering in genic regions. (a) Dashed red and dotted orange lines: average nucleosome occupancy *in vitro* around TSS and TTS [11]. Solid blue and dash-dot black lines: model predictions with and without Φ from Fig. 1. Both models have the average occupancy of 0.60 (less than the maximum possible occupancy of 0.82 because some histone octamers are not DNA-bound). Inset: one-body energy landscape with barrier heights, widths and shapes adjusted to reproduce observed NDRs. (b) Same as (a), for *in vivo* nucleosomes (YPD medium) [19]. Φ is from Fig. 1 with $A = 7$ $k_B T$. The log-intensities from the microarray were exponentiated and normalized separately for each gene, yielding the average occupancy of 0.70 which was also used in the models.

The oscillations in the autocorrelation function are suppressed when nucleosome positions are predicted using a sequence-specific model which neglects two-body interactions [6]. This "spatially resolved" model assigns mono- and dinucleotide energies independently at each position within the nucleosomal site and is thus capable of capturing the $10 - 11$ bp periodicity of one-body interactions. We find that the autocorrelation function is much closer to experiment if the two-body potential is included into the model [Fig. 2(b)].

Two-body interactions are also essential for reconstructing nucleosome occupancy profiles over transcribed regions [Fig. 3]. Sequence-specific energy barriers over NDRs must be low *in vitro* to account for the lack of occupancy oscillations induced by steric exclusion at 1:1 DNA:histone mass ratio [11]. Even with the low barriers shown in Fig. 3(a), the interaction-free model yields an oscillatory profile which is not observed in the data. The oscillations are suppressed by the two-body potential, and the resulting profile increases towards the center of the gene, in contrast with the pure steric exclusion scenario in which nucleosomes adjacent to the barriers are always the most localized [13]. This behavior is also observed *in vivo* where the +2 nucleosome is higher than the +1 nucleosome [Fig. 3(b)]. The *in vivo* barriers are

more pronounced to account for additional nucleosome depletion in the NDRs due to effects other than intrinsic histone-DNA interactions. Finally, in agreement with a previous hypothesis [11], a potential well is added to localize the +1 nucleosome *in vivo*. The well makes the TSS profile asymmetric with respect to the center of the NDR [compare to the more symmetric TTS profile in Fig. 3(b)].

In summary, our study is the first to show that short-range two-body interactions induced by chromatin fiber formation play a major role in genome-wide nucleosome ordering. We demonstrate that large-scale mononucleosome maps contain evidence of the two-body potential. This potential is more important than intrinsic histone-DNA interactions for predicting $10 - 11$ bp periodicity in genome-wide nucleosome positions, and for understanding nucleosome occupancy in transcribed regions. Clearly, two-body interactions should be an integral part of genome-wide models of nucleosome occupancy. Our study also underscores the need for future experiments focused on multi-nucleosome distributions, which can be analyzed using our exact theory [Eqs. (5) and (6)].

* Corresponding author: morozov@physics.rutgers.edu
[1] G. Felsenfeld and M. Groudine, Nature **421**, 448 (2003).
[2] T. J. Richmond and C. A. Davey, Nature **423**, 145 (2003).
[3] B. Li et al., Cell **128**, 707 (2007).
[4] A. V. Morozov et al., Nucleic Acids Res. **37**, 4707 (2009).
[5] N. Kaplan et al., Nature **458**, 362 (2009).
[6] G. Locke et al., Proc. Natl. Acad. Sci. USA **107**, 20998 (2010).
[7] F. Strauss and A. Prunell, EMBO J **2**, 51 (1983).
[8] J. Widom, Proc. Natl. Acad. Sci. USA **89**, 1095 (1992).
[9] L. E. Ulanovsky and E. N. Trifonov, *Biomolecular Stereodynamics III* (Adenine Press, 1986), pp. 35–44.
[10] T. N. Mavrich et al., Nature **453**, 358 (2008).
[11] Y. Zhang et al., Nature Struct. Mol. Biol. **16**, 847 (2009).
[12] D. Tolkunov and A. Morozov, Adv. Protein Chem. Struct. Biol. **79**, 1 (2010).
[13] R. D. Kornberg and L. Stryer, Nucleic Acids Res. **16**, 6677 (1988).
[14] J. K. Percus, J. Stat. Phys. **15**, 505 (1976).
[15] J. K. Percus, J. Phys. Condens. Matter **1**, 2911 (1989).
[16] F. Gürsey, in *Mathematical Proceedings of the Cambridge Philosophical Society* (Cambridge University Press, 1950), vol. 46, pp. 182–194.
[17] D. J. Schwab et al., Phys. Rev. Lett. **100**, 228105 (2008).
[18] J. P. Wang et al., PLoS Comput. Biol. **4**, e1000175 (2008).
[19] K. A. Zawadzki et al., Mol. Biol. Cell **20**, 3503 (2009).