Editors' Suggestion

Translational nucleosome positioning: A computational study

J. Neipel,[1,2,3] G. Brandani,[4] and H. Schiessel[3]

[1]*Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany*
[2]*Faculty of Physics, Ludwig-Maximilians-Universität München, 80333 München, Germany*
[3]*Instituut-Lorentz, Universiteit Leiden, Postbus 9506, 2300 RA Leiden, The Netherlands*
[4]*Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan*

(Received 19 August 2019; accepted 25 November 2019; published 10 February 2020)

About three-quarters of eukaryotic DNA is wrapped into nucleosomes; DNA spools with a protein core. The affinity of a given DNA stretch to be incorporated into a nucleosome is known to depend on the base-pair sequence-dependent geometry and elasticity of the DNA double helix. This causes the rotational and translational positioning of nucleosomes. In this study we ask the question whether the latter can be predicted by a simple coarse-grained DNA model with sequence-dependent elasticity, the rigid base-pair model. Whereas this model is known to be rather robust in predicting rotational nucleosome positioning, we show that the translational positioning is a rather subtle effect that is dominated by the guanine-cytosine content dependence of entropy rather than energy. A correct qualitative prediction within the rigid base-pair framework can only be achieved by assuming that DNA elasticity effectively changes on complexation into the nucleosome complex. With that extra assumption we arrive at a model which gives an excellent quantitative agreement to experimental *in vitro* nucleosome maps, under the additional assumption that nucleosomes equilibrate their positions only locally.

DOI: 10.1103/PhysRevE.101.022405

I. INTRODUCTION

DNA in eukaryotic cells is compacted with the help of proteins into the hierarchical chromatin complex. Details of the higher levels are not yet well understood, even though there is currently rapid progress [1]. The first level of complexation, however, is known in great detail. It consists of the basic repeated structure, the nucleosome, involving a short stretch of DNA, 147 base pairs (bp) in length, wrapped in 1-3/4 turns around a cylindrical aggregate of eight histone proteins. This results in a disk-shaped particle with a diameter of 11 nm and a height of 6 nm [2]. A short stretch of DNA, called the linker, connects to the next such protein spool.

DNA is a rather stiff molecule with a persistence length of about 150 bp or 50 nm. Therefore, wrapping DNA into a nucleosome is rather costly as it involves about one persistence length to be bent nearly two turns. This huge bending cost is compensated by the binding of the DNA backbones to the histone octamer at 14 binding sites [2]. The bending cost shows a strong dependence on the sequence-dependent geometry and elasticity of the involved DNA stretch, whereas the binding occurs mainly to the DNA backbones whose chemistry does not depend on sequence. This suggests that the affinity of a given DNA sequence to be part of a nucleosome reflects mainly the ease with which the DNA is wrapped into a nucleosome. This allows for mechanical cues to be written along DNA molecules, telling nucleosomes where to sit and where not to sit, sometimes called the "nucleosome positioning code" [3] (for earlier versions of this idea see, e.g., Refs. [4] and [5]).

The nonuniform, sequence-dependent positioning of nucleosomes along genomic DNA can be clearly observed by reconstituting nucleosomes from their pure components, DNA and histone proteins, via salt dialysis and then producing nucleosome maps using genomewide assays that extract DNA stretches which were stably wrapped into nucleosomes (see, e.g., Ref. [6]). A typical quantity to be determined is the nucleosome occupancy at each base-pair position which is the probability that the corresponding base pair is covered by a nucleosome. One finds two types of nucleosome positioning along DNA: rotational and translational positioning [7]. Rotational positioning reflects the fact that a given DNA stretch is typically not intrinsically straight due to the intrinsic geometries of the involved base-pair steps. This results in a local preference for the nucleosome to sit in a certain orientation on the DNA, leading to sets of positions 10 bp apart, reflecting the DNA helical repeat. Translational positioning is caused by DNA stretches that have overall a higher affinity for nucleosomes. This is known to correlate well with their GC content, i.e., the fraction of nucleotides in a DNA sequence that are guanine (G) or cytosine (C) [8–12]; the physics underlying this sequence preference is the subject of the current study. Examples for such translational mechanical cues are nucleosome-depleted regions before transcription start sites in yeast facilitating transcription initiation [6,12], mechanically encoded retention of a small fraction of nucleosomes in human sperm cells allowing transmission of paternal epigenetic information [13] or the positioning of six million nucleosomes around nucleosome-inhibiting barriers in human somatic cells [11].

Of importance is also the fact that histone octamers can spontaneously "slide" along DNA [14] and therefore sample different positions, allowing for (a rather slow) equilibration of nucleosomes, at least locally. Two mechanisms have been suggested, both are based on thermally induced defects inside the nucleosome: single base-pair twist defects (a missing

 ©2020 American Physical Society

or an extra base pair) [15–17] and 10-bp bulges [18,19]. New simulation studies [20,21] strongly suggest that both mechanisms can be at play and that it depends on the underlying base-pair sequence which one is preferred for a given DNA stretch. *In vivo* there are in addition chromatin remodellers at work that use adenosine triphosphate to move nucleosomes along DNA. New experiments [22–24] and simulations [25] suggest that at least some of them induce twist defects in the nucleosome. Chromatin remodellers might help nucleosomes to equilibrate their locations along DNA [26], but they might also perturb the intrinsically preferred positioning of nucleosomes, together with other proteins that compete for DNA target sites [10].

There is wide range of questions that need to be answered when considering nucleosome positioning and the underlying physical mechanisms. Restricting ourselves to the *in vitro* situation (neglecting the more complex *in vivo* case), we may ask: Is it the base pairs' shapes or stiffnesses or both that underlie translational and rotational positioning? Can nucleosomes reconstituted on genomic DNA be described as an equilibrium ensemble or are the "sliding" mechanisms typically too slow to allow for global equilibration? Are nucleosome models based on a DNA representation with local deformation energies like the rigid base-pair model [27] (the latter is used in Refs. [12,28–43]) sufficient to predict translational and rotational nucleosome positioning?

We have used a series of simulations [38,40] and analytical approaches [42] to study some of these questions; also a probabilistic model [12,43] informed by a simulation [38] is used. All our models are based on the rigid base-pair model for the DNA double helix and we force this model into a shape that resembles conformations of nucleosomal DNA. These kind of models are particular successful in predicting the rotational positioning of nucleosomes [38,41,42] as they recover the well-known nucleosome base-pair step preferences [3,5], e.g., the preferences of GC steps for locations where the DNA major groove faces inward. We were able to explain these sequence rules as a consequence of the requirement of sequence continuity (e.g., after a GC step must follow a step that starts with C), often going against the sequence preferences of individual base-pair steps [42]. Importantly, we found that rotational sequence preferences are mainly caused by the intrinsic shape of base-pair steps, whereas differences in softness can be neglected [42] (also a different computational model came to this conclusion [44]). In addition, using these models we demonstrated that rotational positioning cues can be put freely even on top of genes as the genetic code is degenerate [38,43].

Our model showed also close agreement with the translational positioning of nucleosomes [12]. We have, however, not yet performed a critical analysis of the underlying physical mechanism that causes the sequence preferences of nucleosomes for GC-rich DNA. It is the purpose of the current study to perform this analysis. We demonstrate that translational positioning of nucleosomes, at least within the rigid base-pair model, is a rather subtle effect related to the overall higher softness of GC-rich sequences. Surprisingly, however, if we consistently account for the entropy cost in the free DNA, our previous models cannot capture the observed GC preference. However, we find that it is indeed possible to come to a consistent framework to quantitatively predict translational nucleosome positioning using the rigid base-pair model. To achieve this, we introduce a multiharmonic model which uses the two distinct sets of parameters for nucleosomal and free DNA, respectively obtained from protein-DNA cocrystals and all-atom molecular dynamics (MD) simulation. This is based on the recognition that a single harmonic approximation to the DNA may not be accurate enough to represent diverse conditions such as those of free and nucleosomal DNA. Instead, each DNA region is parametrized using a set of the DNA conformations obtained under similar conditions.

The paper is organized as follows: In the next section we introduce our coarse-grained nucleosome model together with a Markov-chain approach to calculate the sequence affinities of this model. In Sec. III we identify the origin of the GC content dependence of this model. This insight requires to account for the entropy change that results from wrapping DNA into a nucleosome, undermining the GC content dependence of past models, see Sec. IV. By comparing the predictions of different models, we show that only the new multiharmonic model can explain the nucleosome preference for sequences with high GC content. In Sec. V we study the possibility that the GC preferences of real nucleosomes might actually reflect their dislike for poly-dA:dT sequences which are particularly stiff, an effect that cannot be accounted for by our DNA model with local energies. After showing that there is a genuine preference of real nucleosomes for GC-rich DNA, we compare experimental observations with the predictions of our new model, showing that reaching a quantitative agreement requires taking into account the nonequilibrium features of reconstituted nucleosomes. We finish with a Conclusions section.

II. MODEL

We employ the same nucleosome model as in our previous work [37–41]. We stress, however, that the results of this study shed also light on other models that employ the same DNA model, namely Refs. [28–36]. In our (and these other models) the DNA molecule is represented by the rigid base-pair model [27] which treats each base pair as a rigid body, the spatial position and orientation of which are described by six (three translational and three rotational) degrees of freedom. It assumes only nearest-neighbour interactions with a quadratic deformation energy between successive base pairs [27]. The elastic energy of an $N+1$-bp-long chain is then given by

$$\mathcal{H}_{\rm pot}(\{\mathbf{x}^a\}) = \sum_{a=1}^{N} \sum_{i,j} \left(x_i^a - x_{0,i}^a\right) K_{ij}^a \left(x_j^a - x_{0,j}^a\right), \qquad (1)$$

where K_{ij}^a is the sequence-dependent stiffness matrix and $\mathbf{x}_0^a$ is the equilibrium conformation of the base-pair step state $\mathbf{x}^a$, $a = 1, \ldots, N$, defined as

$$\mathbf{x}^a = \left(v_1^a, v_2^a, v_3^a, u_1^a, u_2^a, u_3^a\right)^T. \qquad (2)$$

The stiffness matrices and equilibrium conformations can be derived from DNA-protein cocrystals [27] or from all-atom MD simulations of DNA oligomers [45]. Also a hybrid parametrization [46] has been proposed with the shape parameters taken from the cocrystals and the stiffnesses from the simulations. We used this hybrid approach in Refs. [37–41].

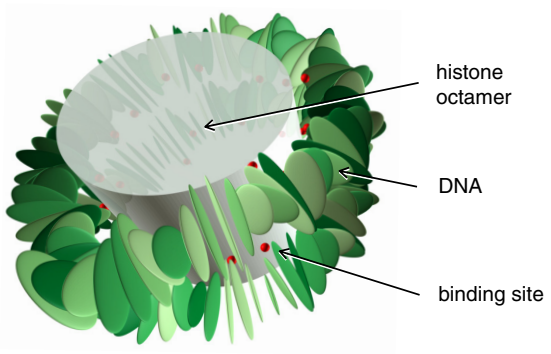

FIG. 1. Nucleosome model. Each rigid plate represents a base pair, the locations of the constraints (corresponding to bound phosphates) are shown by beads, two per binding site. The cylinder is a rough representation of the protein core but is not simulated explicitly (except through the binding sites).

In our nucleosome model a 147-bp-long DNA molecule is forced into a superhelix through a set of 28 constraints that represent the 14 binding sites to the histone octamer (see Fig. 1) and which were extracted from the nucleosome crystal structure 1kx5 [47] without introducing free parameters [38] (later we will also consider how the results are affected by using different reference crystal structures). These constraints correspond to bound phosphates in the DNA backbones and are represented in our model by a special treatment for the corresponding base-pair steps. This is necessary because the rigid base pair model does not contain the phosphates explicitly. We have shown [38] that the location of a given phosphate can be predicted with high accuracy from the positions and orientations of the base pairs connected to it. Specifically, a given phosphate lies very close to the midplane of the corresponding base-pair step. We therefore model bound phosphates by imposing fixed midframes for all the base-pair steps closest to such phosphates, 28 in total (two per binding site). Whereas their positions are rather insensitive to the used crystal structure (here 1kx5 [47]), their orientations are affected by the DNA sequence. To remove this base-pair sequence-dependent bias, we performed a prerelaxation step of the structure with a homogeneous DNA sequence where we allowed for a free rotation of the bound midframes, before we fixed them in their preferred orientations (see Ref. [38] for more details).

This model is used here for two applications. One is to calculate the average elastic energy for a given 147-bp-long sequence by producing random samples of nucleosome conformations using the standard Metropolis algorithm. Each Monte Carlo move consists of a local conformational move of a randomly picked base pair. Base pairs next to bound phosphates are not moved individually but as a pair such that the rotation and translation of one base pair determines that of the other, keeping the midframe fixed (as detailed in Ref. [38]). The second application of this nucleosome model is to build a probabilistic model that allows to determine the free energy, and therefore the positioning probability along the genome, to a very good approximation, as detailed further below.

For this purpose, we perform a Monte Carlo simulation with moves in configuration as well as sequence space. This mutation Monte Carlo (MMC) algorithm [38] can be used to simulate a nucleosome which is in equilibrium in sequence as well as configuration space. The probability of finding a certain sequence in a certain configuration along a very long trajectory of a MMC simulation is proportional to the Boltzmann weight of this state. Hence, we are able to determine the free-energy difference of two 147-bp sequences S and S' by comparing the probabilities of the sequences irrespective of their configurations:

$$\frac{P(S)}{P(S')} = \frac{\int_{x\in\Omega_{\rm nuc}} dx e^{-\beta E(x,S)}}{\int_{x\in\Omega_{\rm nuc}} dx e^{-\beta E(x,S')}} = e^{-\beta[F_{\rm nuc}(S)-F_{\rm nuc}(S')]}, \quad (3)$$

where S and S' are nucleosome-bound DNA sequences, x are DNA configurations, $\Omega_{\rm nuc}$ is the set of possible conformations that leave the bound midframes fixed, $\beta = 1/k_BT$, E is the energy from the rigid base-pair model, and $F_{\rm nuc}$ is the mechanical free energy of nucleosome-bound DNA.

$P(S)$ cannot be determined from a trajectory of an MMC simulation for all possible sequences as there are far too many, namely $4^{147} \approx 10^{88}$. However, the probability for shorter stretches, namely all dinucleotides or trinucleotides, at all positions in the model nucleosome can be determined to high precision. Following Tompitak *et al.* [48] these probabilities can be used to approximate conditional probabilities. For instance, the probability that a nucleosome bound sequence features nucleotide S_n at position n, given nucleotides S_1 to S_{n-1}, can be approximated by

$$P(S_n|S_{n-1}\wedge S_{n-2}\wedge ...\wedge S_1) \approx P(S_n|S_{n-1}\wedge S_{n-2}). \quad (4)$$

This next-nearest-neighbor approximation can then be used to estimate the probability $P(S)$ for a given sequence $S = \{S_1, S_2, \ldots, S_N\}$ with $N = 147$:

$$\begin{aligned} P(S_N\wedge ...\wedge S_1) &= P(S_2\wedge S_1)\prod_{i=3}^{N} P(S_i|S_{i-1}\wedge ...S_1) \\ &\approx P(S_2\wedge S_1)\prod_{i=3}^{N} P(S_i|S_{i-1}\wedge S_{i-2}) \\ &= P(S_2\wedge S_1)\prod_{i=3}^{N} \frac{P(S_i\wedge S_{i-1}\wedge S_{i-2})}{P(S_{i-1}\wedge S_{i-2})}. \end{aligned} \quad (5)$$

Note that this contains no information about free-energy differences in the unbound DNA states (which are purely entropic), what is usually referred to as linker or free DNA. Therefore, the free-energy difference in Eq. (3) is only informative for nucleosome positioning if the system is dominated by the energetic wrapping costs, so that entropic differences in the unbound states can be neglected. One might hope that this is indeed the case, as the substantial DNA bending inside nucleosomes amounts to large energy costs, and in the past we also employed this approximation [48]. However, there are certain aspects of nucleosome positioning where the entropy of the unbound DNA is important, as we demonstrate in Secs. III and IV. Given that two sequences A and B compete for a nucleosome, the probabilities of finding the nucleosome bound to A and B depends on the free-energy difference

between the two states of the entire system, not just the nucleosome-bound sequence:

$$\begin{aligned} k_B T \ln &\left[\frac{P(\text{Nuc. bound to } A)}{P(\text{Nuc. bound to } B)}\right] \\ &= F_{\text{sys}}(\text{Nuc. bound to } B) - F_{\text{sys}}(\text{Nuc. bound to } A) \\ &= [F_{\text{nuc}}(B) + F_{\text{free}}(A)] - [F_{\text{nuc}}(A) + F_{\text{free}}(B)] \\ &= [F_{\text{nuc}}(B) - F_{\text{nuc}}(A)] - [F_{\text{free}}(B) - F_{\text{free}}(A)] \\ &= \Delta F_{\text{nuc}} - \Delta F_{\text{free}}, \end{aligned} \tag{6}$$

where F_{nuc} and F_{free} are the sequence-dependent free energies of nucleosome-bound and unconstrained (free) DNA, respectively. In other words, a genomic sequence needs to be favorable in the nucleosome-bound state and unfavorable in the unbound state to compete for nucleosomes with high affinity.

In our model, making use of Eq. (5), the free energy of nucleosome-bound DNA for a sequence S is given by:

$$\begin{aligned} F_{\text{nuc}}(S) &= -k_B T \log P(S) \\ &\approx -k_B T \log P(S_2 \wedge S_1) \prod_{i=3}^{N} \frac{P(S_i \wedge S_{i-1} \wedge S_{i-2})}{P(S_{i-1} \wedge S_{i-2})}, \end{aligned} \tag{7}$$

with the probabilities obtained from the MMC simulation. The sequence-dependent free energy of the free DNA can be in principle determined with the very same MMC algorithm we used for parametrizing the free energy of nucleosomal DNA; however, as shown in Appendix B, it also can be computed analytically. For a sequence S of length $N+1$, the free energy of the free DNA is

$$F_{\text{free}}(S) = \frac{k_B T}{2} \sum_{a=1}^{N} \ln \det \mathbf{K}^a, \tag{8}$$

where $\mathbf{K}^a$ is the stiffness matrix of the base-pair step starting at base pair a. Therefore the free energy of free DNA depends solely on the stiffness of the DNA base steps, and it is purely entropic in origin. The energy of unconstrained DNA is given by the equipartition theorem. Hence, it solely depends on the number of degrees of freedom, i.e., the number of base pairs, and the temperature, but not the DNA sequence. In Appendix B we also derive the probabilities of individual base-pair steps along unconstrained DNA. Depending on the parametrization, these probabilities can display strong sequence dependencies and the entropy of free DNA might thus represent an important contribution to nucleosome positioning. This will be discussed in Sec. IV.

Our model is very similar to those used by Segal and coworkers [3,6,49], except that in their case the sequence-dependent probabilities entering in the expression of the free energy were estimated directly from experiments, whereas in our case they are estimated either analytically (for free DNA) or from the MMC simulations (for nucleosomal DNA) based on the rigid base-pair model. The advantage of our approach is that it enables us to clearly discriminate the energetic versus the entropic contributions to the nucleosome positioning along the genome, a main focus of this article. A second important aim is to investigate how the model predictions, and the agreement with experiments, depend on the DNA parametrization. As mentioned earlier, in the rigid base-pair model the stiffness matrices $\mathbf{K}$ may be determined either from a database of DNA-protein cocrystals [27] or from all-atom MD simulations [45]. A critical issue with the choice of DNA parametrization is that the rigid base-pair model is only a harmonic approximation to the true DNA Hamiltonian. In principle we should choose the parameters that provide the best agreement with experiments, but it is not clear whether a single set of parameters can represent well both unconstrained and nucleosomal DNA at the same time. In fact, one could expect that the MD parameters should be more appropriate for modeling unconstrained DNA, since it corresponds to the condition of the MD simulations from which parameters are derived, whereas the crystal parameters should be more appropriate for nucleosomal DNA, since these parameters were also obtained from DNA-protein complexes. Given these considerations, we decided to evaluate the predictions and the agreement with experimental *in vitro* nucleosome occupancies for four main different models:

(i) The crystal model, where both the equilibrium and stiffness parameters are derived from DNA-protein cocrystals [27].

(ii) The hybrid model, where the equilibrium base-pair step parameters are obtained from DNA-protein cocrystals, whereas the stiffness parameters come from MD simulations. This model was used many times in the past [37–41,46], and it behaves very similarly to a pure MD parametrization (which is not discussed here).

(iii) The multiharmonic model, where the nucleosomal DNA is modeled using the crystal parameters (both equilibrium and stiffness), whereas the unconstrained DNA is modeled using the MD parameters (here only stiffness enters in the free energy). This model is motivated by the realization that DNA is highly nonlinear and that it may be preferable to have two distinct harmonic approximations for nucleosomal and free DNA.

(iv) The ET model introduced in Ref. [12], where ET denotes Eslami-Mossallam-Tompitak, because it was the first to combine the MMC simulation approach by Eslami-Mossallam [38] with the approximation for the probability by Tompitak [12] [Eq. (5)]. This model uses the same DNA parameters as in the crystal model. However, the ET model relied on the assumption that the free energy of nucleosome formation is dominated by the bending energy penalty, whereas in the other three models presented here we also take into account the entropy of the free DNA [Eq. (6)].

III. ORIGIN OF GC CONTENT DEPENDENCE OF THE ET MODEL

The ET model has been used to predict the average nucleosome occupancy (the probability that a given base pair is covered by a nucleosome) around transcription start sites (TSS) in various organisms [12]. This theoretical "average nucleosome occupancy" is determined from the genomewide average energy landscape around TSSs, whereas in experimental studies the average logarithm of nucleosome occupancy is computed. After averaging the energy landscapes over several thousand sequences, only sequence features

remain that generally inhibit or favor nucleosome positions, independent of the rotational phase. Tompitak *et al.* [12] already noted that the average GC content curves look very similar to their predicted nucleosome occupancy, but only when using a purely crystallographic parametrization [27]. However, the physics underlying the GC content dependence of the model was not investigated and it was not explained why the crystallographic stiffnesses seem more reliable in this particular context, whereas in all other studies of our nucleosome model the stiffness matrices were derived from MD simulations [45]. In the following, we will thoroughly study the predictions of the ET model and the sequence dependence of nucleosome occupancy in yeast in order to answer these questions.

The ET model [12] is based on the assumption that energy dominates over entropy in determining the sequence dependence of nucleosome affinity, because DNA is strongly deformed inside nucleosomes. In order to test this assumption, Tompitak *et al.* compared the free-energy differences between various sequences along chromosome I from yeast to the corresponding average energy differences [48]. The free energies were derived from the MMC-based Markov model, whereas the average energies were determined with Monte Carlo simulations of the coarse-grained nucleosome model in configuration space only. Both simulations were performed at $T = 50$ K and used a hybrid parametrization [46], combining the equilibrium configurations from crystal structures with the stiffness matrices from MD simulations. In a different paper the AT-rich dinucleotide frequencies in the nucleosome were determined at different temperatures, which showed hardly any temperature dependence [40]. The confidence that energy is also dominating *in vitro* affinities comes from the original MMC study, where the average energies of tetramers were compared to experimental nucleosome affinities for a few sequences [38]. In all three cases, however, the hybrid parametrization was used, which does not yield the GC content dependence observed *in vitro*. Hence, it has never been tested whether the GC content dependence observed in Ref. [12] is in agreement with the assumptions of the ET model.

Given that the stiffnesses $\mathbf{K}^a$ in both parametrizations are of the same order of magnitude, one might expect that the dominance of energy is independent of the choice of parameters. However, a closer look at the diagonal entries of $\mathbf{K}^a$ reveals that the crystallographic parameters span a broader range than the MD parameters, especially for roll and tilt, which are particularly relevant for DNA bending inside nucleosomes [see Figs. 2(b) and 2(c)]. On average, GC-rich dinucleotides are less stiff than AT-rich dinucleotides which can be seen particularly well at the determinant of the stiffness matrix, Fig. 2(a). Interestingly, this trend is not found for roll, which is the dominant mode of bending in the nucleosome. Hence, it is not clear, whether the low total stiffness of GC dinucleotides actually makes them energetically more favorable.

To answer this question, we computed the energy and the free energy of 144 000 random sequences. The energy is calculated from the coarse-grained nucleosome model [38] (shown in Fig. 1), whereas the free energy is determined from our ET model, see Eq. (7). All simulations in this section are performed at 100 K, the same temperature we used in some of our earlier simulations, e.g., the ones in Ref. [38]. Compared to physiological temperatures, this underestimates the entropic contribution to the free energy. In order to properly test the GC content dependence, the number of GCs in each sequence was fixed such that there are 1000 sequences for each possible GC content between $\frac{2}{147}$ and $\frac{145}{147}$. In Figs. 2(d) and 2(g) the free energies of all these sequences are plotted vs. the average energy. Free as well as average energies have been shifted such that both energies average to zero for sequences with a GC content of 50%. Among sequences with similar GC content, differences in free energy are largely due to differences in energy, independent of parametrization [see Fig. 2(d) and 2(g)]. When using the hybrid parametrization, this dominance of energy is also found for sequences with very different GC content, Fig. 2(g). When using the crystallographic stiffnesses, however, entropy becomes relevant [Fig. 2(d)].

In fact, it can now be seen clearly that the GC content dependence observed in Ref. [12] (which uses crystallographic parameters) is mainly an entropic effect. The low total stiffness does not make GC-rich sequences energetically favorable, but it allows them to assume a larger number of configurations than AT-rich sequences [Fig. 2(f)]. For large differences in GC content, the entropic contribution to the free-energy difference is four times larger than the maximal energy difference among the 144 000 sequences. This is particularly remarkable, because all simulations were performed at a low temperature of 100 K. On the other hand, hardly any entropic contribution to the GC content dependence is observed when using MD stiffnesses, whereas the energetic contribution makes GC-rich sequences unfavorable [Fig. 2(i)]. Interestingly, the entropy of the ET model is largely independent of temperature, see Figs. 2(e) and 2(h). Furthermore, it seems largely independent of the overall DNA configuration as the entropy of unconstrained DNA follows very similar slopes, see Figs. 2(f) and 2(i).

In Appendix A we present analytical results for simplified versions of the nucleosome model to understand better the competition between energy and entropy in such systems. It also demonstrates that the independence of entropy with respect to temperature results from the harmonic potential. The weak dependence of entropy on the overall DNA configuration will be discussed in the next section.

IV. GC CONTENT DEPENDENCE OF CONSISTENT MECHANICAL MODELS

The results in the previous section demonstrate that, especially for the crystallographic DNA parametrization, the entropy of nucleosomal DNA depends on the sequence, violating the assumption of the dominance of energy underlying the ET model [12]. Therefore, only the complete crystal, hybrid, and multiharmonic models can be considered consistent mechanical models of nucleosome positioning. In the following, we study the sequence dependencies of these models.

As an illustration we simulate the coarse-grained nucleosome model, Fig. 1, at 300 K but with extra free DNA attached to the nucleosomal DNA at both ends, namely 50 bp each. The resulting nucleotide distributions are shown for the hybrid

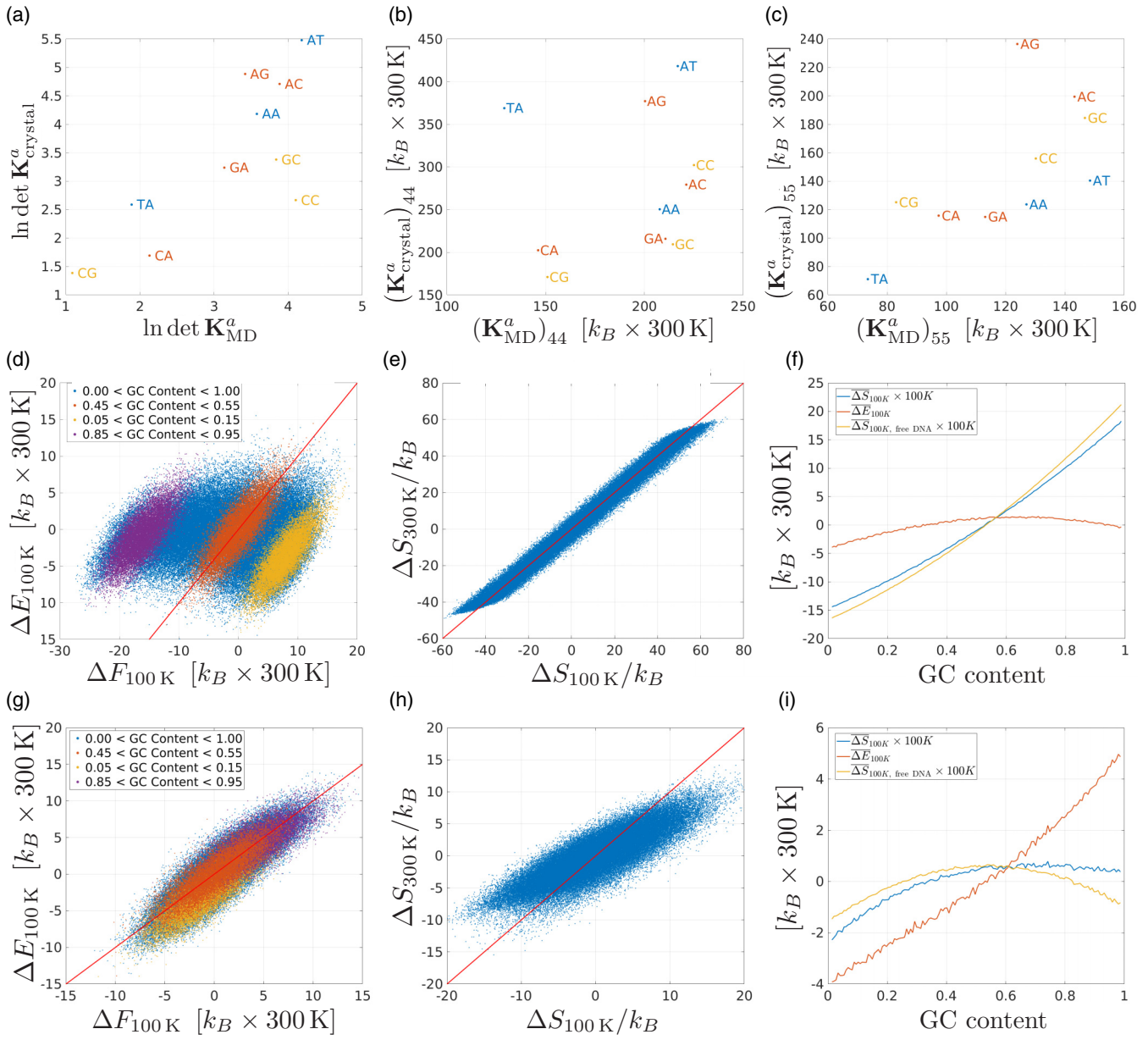

FIG. 2. GC content dependence of energy and entropy: (a) Determinant of stiffness matrix $\mathbf{K}^a$ from rigid base-pair model, derived from molecular dynamics simulations (MD) and protein-DNA crystal structures (crystal) for all distinct base-pair steps. In the crystallographic parametrization GC-rich base pairs (yellow) tend to be less stiff than AT-rich steps (blue). (b) Roll-roll and (c) tilt-tilt stiffnesses vary more strongly in the crystallographic parametrization. [(d)–(f)] Comparison of entropy and energy in nucleosomal [(d)–(f)] and unbound (f) DNA using parameters from protein-DNA cocrystals [27] demonstrates an entropy-dominated GC content dependence. Specifically, (d) average energy E from Monte Carlo simulations vs. free energy F from ET model for 144 000 random sequences with equally distributed GC content. Energies and free energies are given relative to the average values at 50% GC content. All simulations were performed at 100 K. The $x = y$ line is depicted in red. (e) Entropy of the same sequences as in (d) using the free energy from MMC simulations at 100 K and 300 K and the energy from Monte Carlo simulations at 100 K. Values are given relative to the average entropy of all sequences. The $x = y$ is drawn in red illustrating the temperature independence of entropy. (f) Average entropy S of nucleosomal as well as free DNA and energy E of nucleosomal DNA for sequences with identical GC content from (d) and (e), relative to average over all sequences. [(g)–(i)] Same as in (d)–(f) but using base-pair step shape parameters from protein-DNA crystal structures and stiffnesses from MD simulations (hybrid parametrization [46]).

parametrization in Fig. 3(a) and for the crystallographic parametrization in Fig. 3(b). Specifically is depicted the GC content (and the corresponding AT content) as a function of the base-pair position. For both cases we find inside the nucleosome characteristic 10-bp oscillations, whereas the GC content in the free DNA is constant. The oscillations are well known from experiments [3,5,6] and computer simulations [32,38,40]. They reflect the fact that sequences rich in AT content where the minor groove faces inward and in GC content where it faces outward lead on average to an intrinsically

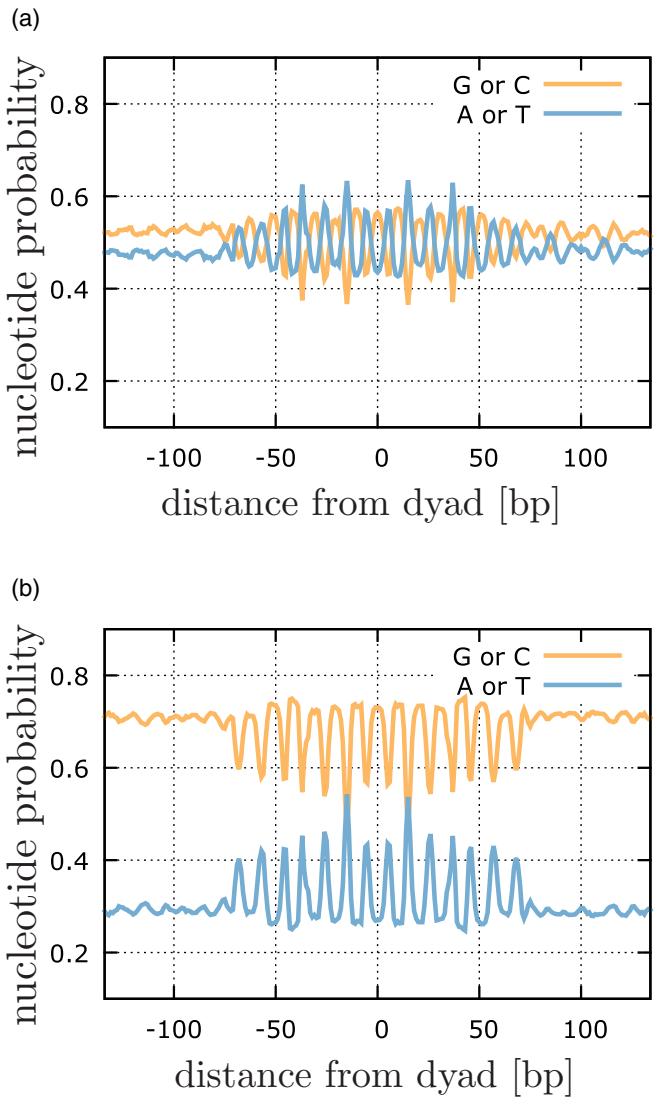

FIG. 3. MMC simulation of the nucleosomal model in Fig. 1 but with extra 50 bp of free DNA attached at each end. (a) Hybrid parametrization and (b) crystallographic parametrization. Both simulations were performed at 300 K.

bent DNA double helix with that preferred rotational setting [42].

Note the dramatic difference in GC content (for both free and wrapped DNA) between the two parameter sets. Whereas the hybrid parametrization yields only a small preference for GC-rich DNA, Fig. 3(a), the crystallographic parameters show a strong preference, Fig. 3(b). Moreover, the GC content of wrapped DNA for positions in between binding sites and of free DNA is almost the same. Only close to the binding sites do the probabilities drop. This reflects the fact that in our nucleosome model it is entropically unfavorable for the softer GC-rich base pairs to be close to the fixed midplanes that represent the binding sites, see Fig. 1.

In the following we will not rely on the MMC simulations for the free DNA since it it possible to compute the base-pair step probabilities analytically, as shown in Appendix B. The resulting dinucleotide probabilities are plotted in Figs. 4(c) and 4(f). In this plot we depict the probabilities of all 10 independent dinucleotides, colored according to their GC content. Adding up the corresponding probabilities leads back to the values found in the MMC simulations for the free DNA stretches, see Fig. 3.

For comparing sequence preferences of free and nucleosomal DNA it is useful to compute the average dinucleotide probabilities of nucleosomal DNA in 10-bp intervals to remove the strong 10-bp undulations inside nucleosomal DNA (seen, e.g., in Fig. 3). Figures 4(a) and 4(d) show these probabilities for two different types of parametrization, both at 100 K. The corresponding probabilities at 300 K are depicted in Figs. 4(b) and 4(e). As can be seen for both parametrizations, with increasing temperature the dinucleotide probabilities in nucleosomal DNA converge to the values found in free DNA [e.g., for the crystal case compare Figs. 4(a) to 4(c)]. This clearly suggests that the average dinucleotide probabilities are in general dominated by entropy and that the sequence dependence of entropy is configuration independent.

After having recognized the potential importance of the entropy in determining the sequence preferences of nucleosomes, we now come back to our original question: Can simple theoretical models predict the observed nucleosome preference for sequences with high GC content? In Fig. 5 we show the relationship between GC content and the free energy of nucleosome assembly ($\Delta F_{\rm nuc} - \Delta F_{\rm free}$) for the 144 000 random sequences used in Fig. 2. In Figs. 5(a), 5(b) and 5(c), we respectively compare the predictions from the complete crystal, hybrid, and multiharmonic models. We focus first on the blue curves, which correspond to the results obtained using the 1kx5 crystal structure as reference (as done in all past models). As can be seen, both pure crystallographic and hybrid DNA parameters predict an increase in free energy with increasing GC content, in disagreement with the experimental observations. However, this similar behavior has two distinct origins. In the crystallographic case, the energy of nucleosomal DNA shows little GC dependence, and the nucleosome preference for sequences with low GC content is due to the larger loss of entropy on the free DNA [Fig. 2(f)]. On the other hand, in the hybrid case, entropy does not display a large GC dependence, but the preference for sequences with low GC content is still present due to their lower energy within the nucleosome [Fig. 2(i)].

The only theoretical model that is able to correctly capture the high GC-content preference is the multiharmonic model. Here the unbound DNA is modeled using the hybrid (or, equivalently, MD) parameters, so that the entropy of free DNA shows little sequence dependence [Fig. 2(i)], whereas nucleosomal DNA is modeled using the crystallographic parameters, so that GC-rich sequences are favored due to their higher entropy [Fig. 2(f)]. This result suggests that indeed it may be preferable to model unconstrained and nucleosomal DNA using two distinct sets of DNA parameters, possibly reflecting the existence of nonlinearities that cannot be captured by a single harmonic approximation. In this sense the ET model [12], which ignores the entropy of the free DNA and also correctly predicts the GC preference, may be considered as a consistent approximation to the multiharmonic model.

However, our predictions may have been also influenced by some specific aspects of our models. In particular, one

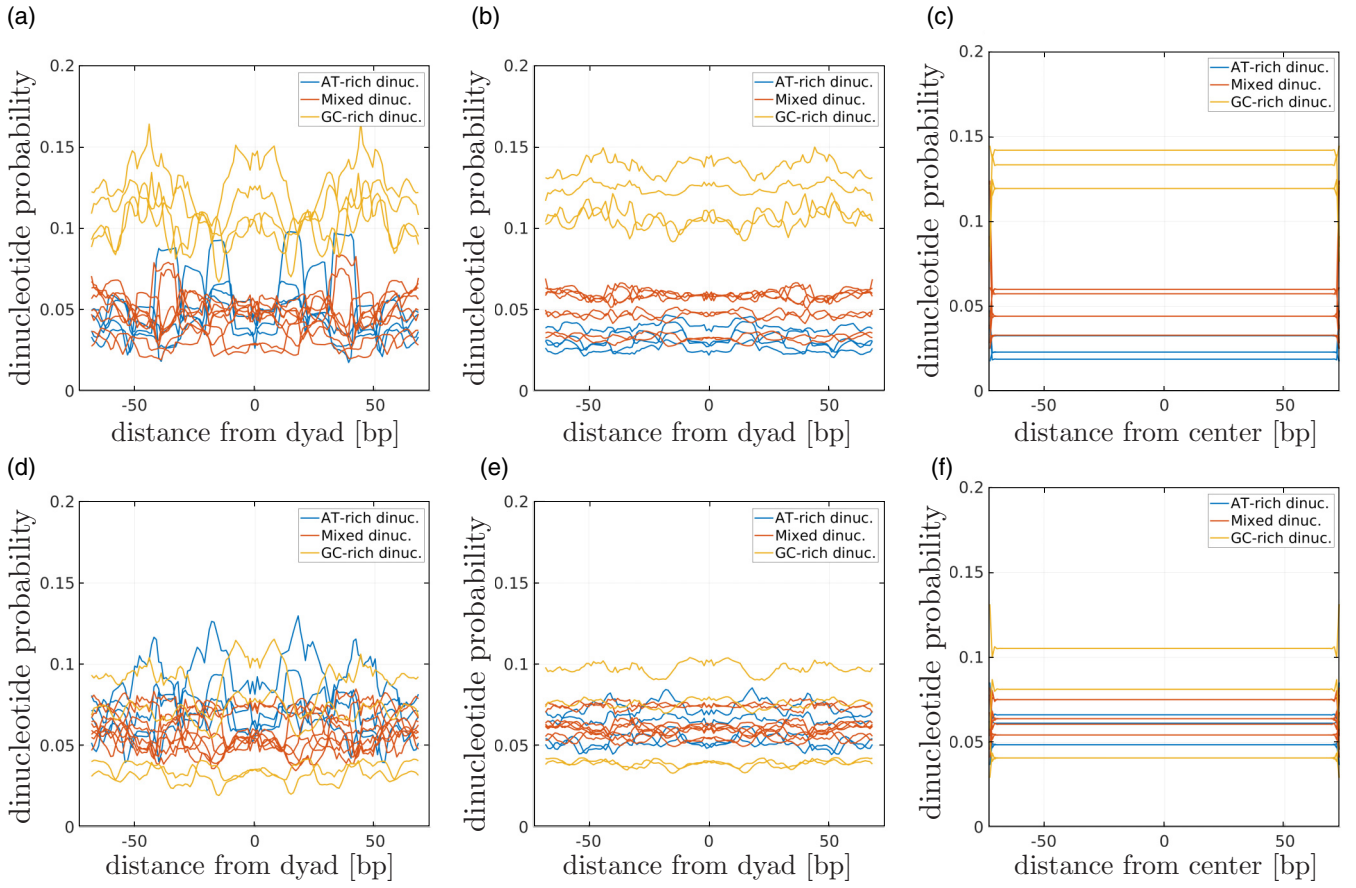

FIG. 4. Dinucleotide probabilities in nucleosomal and free DNA converge at high temperatures. The nucleosomal probabilities were determined with the MMC algorithm applied to the prerelaxed nucleosome structure of Ref. [38]. The simulations were performed at 100 K and 300 K using the crystal parameters or the hybrid parametrization. Specifically, (a) $T = 100$ K, crystal parametrization; (b) $T = 300$ K, crystal parametrization; (d) $T = 100$ K, hybrid parametrization; and (e) $T = 300$ K, hybrid parametrization. To facilitate the comparison with unconstrained DNA shown in (c) for crystal parametrization and in (f) for hybrid parametrization, the average probabilities in (a), (b), (d), and (e) were computed in 10-bp intervals. The dinucleotide probabilities in unconstrained DNA were computed using Eq. (B4).

possibility might be that the specific DNA conformation used in our nucleosome model [38] is unrealistic, as it is based on a single nucleosome crystal structure using a single DNA sequence [47]. We attempted to make this configuration less sequence specific by implementing a prerelaxation step before fixing the midplanes that mimic the binding sites [38].

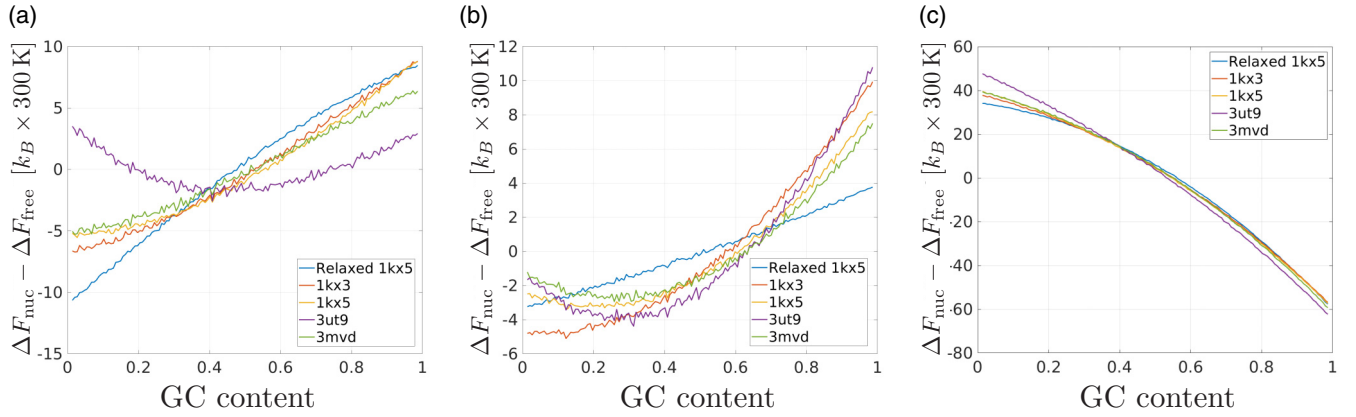

FIG. 5. GC content dependence of relative free energy [Eq. (6)] for different nucleosome structures at 300 K for (a) crystallographic, (b) hybrid, and (c) multiharmonic parameters. The relaxed 1kx5 structure is the structure used in Refs. [12,38,40], which is based on the 1kx5 crystal structure and was adapted using the prerelaxation procedure described in Ref. [38]. The other structures are the unrelaxed rigid base-pair fits of nucleosome crystal structures from the protein data base (PDB). The PDB IDs are given in the legend.

Nevertheless, this procedure might have led to constraints on our nucleosomal DNA that happen to be energetically unfavorable for GC-rich sequences. Therefore, we performed MMC simulations of four different structures, derived from nucleosome crystal structures without further prerelaxation (PDB: 1kx5 [47], 1kx3 [47], 3ut9 [50], and 3mvd [51]). Subsequently, the resulting di- and trinucleotide probabilities were used to compute the free energy of nucleosome binding for the same 144,000 sequences used in Fig. 2. However, our results were largely unchanged from those obtained with the original 1kx5 reference. For most structures, the crystal and hybrid models predict that GC-rich sequences are unfavorable for nucleosome binding (see Fig. 5). Only for one nucleosome crystal structures (PDB: 3ut9) a significant energetic contribution balances the entropic penalty of high GC content. On the other hand, the multiharmonic model consistently predicts higher nucleosome affinity for sequences with high GC content.

V. THE *IN VITRO* GC PREFERENCE IS INDEPENDENT OF POLY-dA:dT SEQUENCES BUT LIKELY INFLUENCED BY NONEQUILIBRIUM EFFECTS

In this section, we address the question whether translational nucleosome positioning *in vitro* is affected by phenomena our nucleosome model cannot capture, namely nonequilibrium effects and the peculiar properties of poly-dA:dT tracts. To answer this question, we take a closer look at the genome of baker's yeast (*Saccharomyces cerevisiae*), relating the nucleosome map produced by Kaplan *et al.* [6] to GC content and the presence of poly-dA:dT sequences. We are aware of the fact that this is a single study using micrococcal nuclease. While similar results have been obtained in a very similar independent study [52], the main problem with MNase is its significant sequence specificity independent from the presence of nucleosomes [53]. Notably, the MNase preference for digesting exposed AT-rich sequences [54] may partially account for the observed enrichment in GC-content within nucleosomal DNA [9,55]. However, we note that similar results have been also obtained with a different enzyme [56]. In addition, although the chemical *in vivo* nucleosome map obtained by Widom and coworkers does not highlight an enrichment in GC content in nucleosomal vs. linker DNA [55], it does show a significant correlation between the probabilities of the individual stable nucleosomes and the GC content of the underlying 147-bp nucleosomal DNA ($\rho = 0.20$, $p \approx 0$); such a correlation is still present after excluding nucleosomes containing poly-dA:dT tracts of four or more base pairs ($\rho = 0.045$, $p = 0.0004$). The consistency of these findings makes us confident that the considered MNase study can serve as a representative example of *in vitro* nucleosome maps. Furthermore, the GC content dependence of nucleosome occupancy (*in vivo*) has been observed in cells as different as human sperm [13] and archaea [57,58].

First we consider the possibility that the preference of nucleosomes for GC-rich DNA might actually reflect to a large extent their dislike for poly-dA:dT sequences (i.e., AAAA... or TTTT...). It is known that poly-dA:dT tracts have unusual properties, making them resistant to be incorporated into nucleosomes [59]. This is a cooperative sequence effect that cannot be captured by the rigid base-pair model [27] as in this model the stiffness and geometry of any given base-pair step is independent of the rest of the sequence. It is worthwhile to mention that a less coarse-grained DNA model, *cgDNA*, a rigid base model, predicts that poly-dA:dT sequences are exceptionally stiff [60]. This suggests that our model cannot capture certain aspects that are inherent to real DNA mechanics. This failure might be important as it is well known that poly-dA:dT tracts are more abundant in eukaryotic genomes than expected by chance, in particular around TSSs [61], see, e.g., the peak in the frequency of poly-dA:dT sequences in yeast in Fig. 6(i). Hence the question arises whether the nucleosome depletion around TSSs and other GC-poor regions in the genome is a result of poly-dA:dT tracts instead of GC content.

To address this question, the entire genome was scanned for poly-dA:dT tracts (defined here as a series of only A or only T nucleotides longer than 4 bp). Every given base pair which was closer than 147 bp to at least one A or T of a poly-dA:dT sequence longer than 4 bp was marked as possibly poly-dA:dT affected. Subsequently, the GC content dependence of the nucleosome occupancy was analyzed, using either the entire genome or only the positions which are not affected by poly-dA:dT tracts. This was done by binning $2 \times 146 + 1$ bp DNA stretches according to their GC content with each bin corresponding to a 2% GC content interval. For these stretches the average *in vitro* and *in silico* occupancy was computed. In Figs. 6(a) and 6(b) the statistics of the nucleosome occupancy for each GC content bin are plotted. The blue solid curves corresponds to the entire yeast genome, whereas for the red solid curve only poly-dA:dT-free intervals were taken into account. For each bin and each subset of the genome, three values are plotted: the average, the 10th percentile, and the 90th percentile (dashed lines). The *in vitro* data, Fig. 6(a), clearly indicate that the observed GC content dependence is independent of poly-dA:dT sequences. Furthermore, this is not an effect of some other special sequences, as the slopes of average and percentile values are nearly identical. The same is also true for the multiharmonic model, Fig. 6(b). However, the GC content dependence turns out to be very different from the *in vitro* data. While the *in vitro* occupancy saturates for high GC content, the slope of the *in silico* occupancy even increases slightly for high GC content.

A possible explanation for this effect may be that the nucleosomes are not able to try out all possible sequences on the genome during one experiment as nucleosome repositioning along DNA, mentioned in the Introduction, is a very slow process. Hence, a given interval on the genome only competes with nearby intervals for nucleosomes, because nucleosomes will not move far along the genome after they are bound to DNA. In other words, the system is only locally at thermodynamic equilibrium. To test this hypothesis, we assume that the nucleosomes were uniformly distributed at the beginning of the experiment. As nucleosome exchange over long distances is negligible, this initial distribution fixes the nucleosome distribution on large scales. We implement this by normalizing the probability in 1 kbp intervals. In each interval the number of nucleosomes is assumed to be of order one, such that excluded volume effects are negligible. The

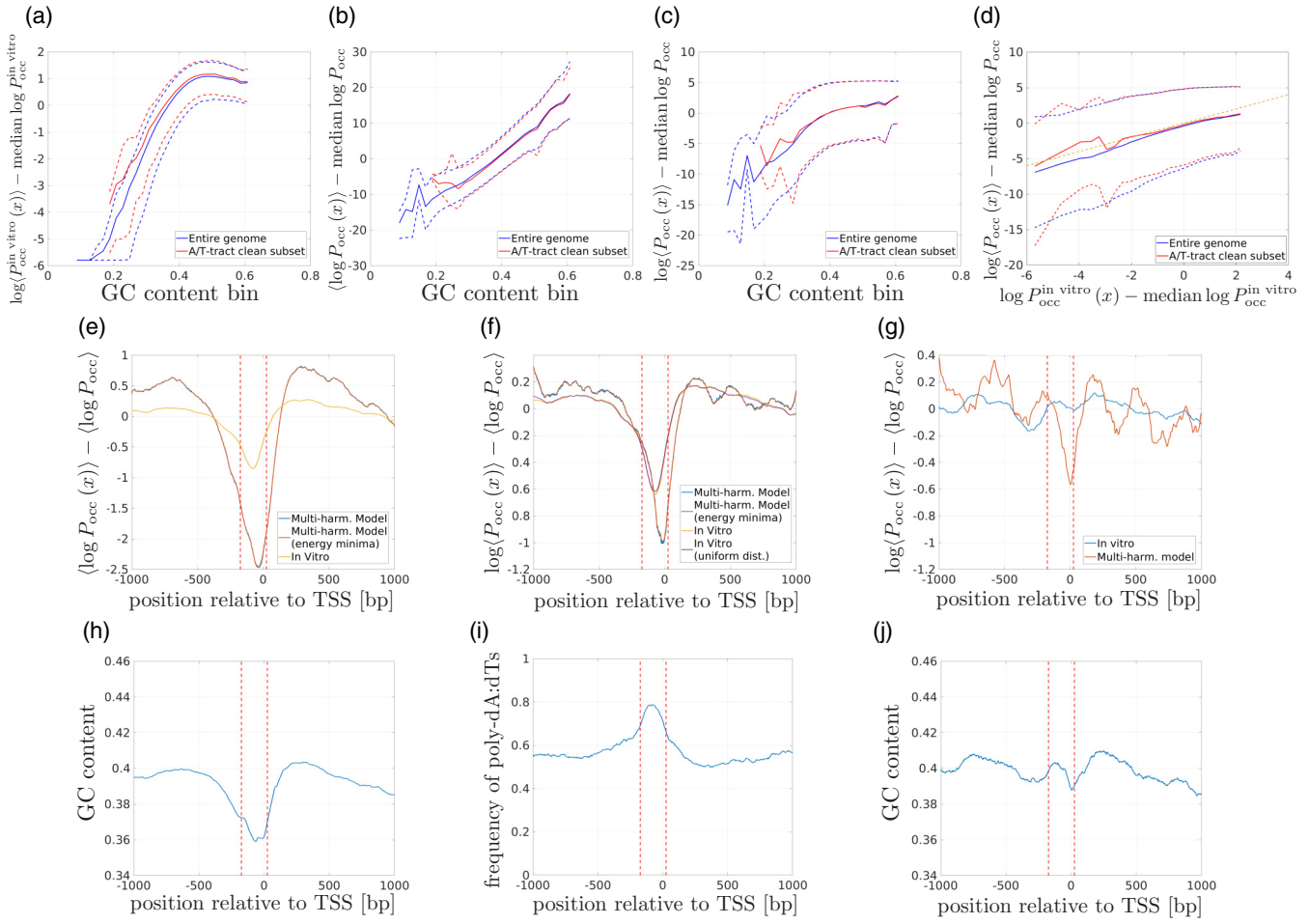

FIG. 6. The interplay of poly-dA:dT sequences and GC content in *in vitro* and *in silico* nucleosome positioning. *In vitro* data are taken from Kaplan *et al.* [6]. The *in silico* data correspond to the predictions of the multiharmonic model. (a) Average (solid line), 10th percentile, and 90th percentile (dashed lines) of *in vitro* nucleosome occupancy in yeast as a function of GC content. For the blue curves the statistics of the entire *S. cerevisiae* genome is taken into account, whereas for the red curves positions that are closer than 147 bp to a poly-dA:dT sequence (>4bp) are excluded. (b) Same as (a) but using the prediction of the multiharmonic model. (c) Same as (b) but after normalizing the predicted nucleosome binding probabilities in 1 kbp intervals. (d) Same as in (c), but after binning the sequences according to the *in vitro* occupancy instead of GC content. The $x = y$ line is indicated by a dashed yellow line. Plots (e), (f), (h), and (i) present averages over all TSSs in *S. cerevisiae*, whereas plots (g) and (j) show averages only for TSSs which do not contain a poly-dA:dT tract in the 200-bp interval indicated by the red dashed lines. (e) Average logarithm of nucleosome occupancy *in vitro* and *in silico* (at 300 K). [(f) and (g)] Logarithm of average occupancy *in vitro* and *in silico* (at 300 K) with the occupancy for each gene normalized to 1 in the 2000-bp interval (except for the curve labeled "in vitro"), see text for details. [(h) and (j)] Average GC content in 147-bp intervals. (i) Fraction of 147-bp intervals that contain at least one poly-dA:dT sequence.

resulting prediction of the multiharmonic model, Fig. 6(c), yields indeed saturation similar to the *in vitro* data, Fig. 6(a). Note that the different behavior at very low and very high GC content does not have to be meaningful, as they are the result of only a few sequences.

Figure 6(d) demonstrates that the combination of local normalization and multiharmonic parameters enables us to predict the average nucleosome occupancy over several orders of magnitudes. At the same time, we also note a large span in *in silico* occupancy for sequences with similar *in vitro* occupancy [see the curves for the 90th and 10th percentiles in Fig. 6(d)]. Hence, we cannot exclude that there are further aspects of translational nucleosome positioning which our model cannot resolve. However, also the limited number of sequencing reads as well as the nucleosome-independent sequence specificity of MNase certainly play a role here.

A frequently studied aspect of nucleosome maps is the average occupancy around TSSs, which sheds further light on the GC dependence of nucleosome occupancy. As observed in several studies, the TSSs of *S. cerevisiae* are nucleosome depleted on the genomewide average, *in vitro* [Figs. 6(e) and 6(f)] as well as *in vivo* (see, e.g., Ref. [6]). Also here we can ask the question whether this is related to a drop in GC content, Fig. 6(h), or a peak in the fraction of poly-dA:dT sequences, Fig. 6(i), or both. We therefore looked at the subset of TSSs which do not contain poly-dA:dT sequences longer than 4 bp in a 200-bp interval close to the TSS, see Figs. 6(g) and 6(j) (the interval is demarcated by red lines). Because such sequences are so abundant around TSSs, only a few hundred TSSs remain, leading to poor statistics. Furthermore, this procedure selects for sequences with relatively high GC content around the TSS [Fig. 6(j)]. In doing so, we lose the

nucleosome depletion region usually observed around TSSs. Nucleosome occupancy [Fig. 6(g)] and GC content [Fig. 6(j)], both averaged over this subset of TSSs, do not show a clear signal. Hence, a separate analysis of GC content and poly-dA:dT dependence of nucleosome occupancy—as done for the whole genome above—is not possible for TSSs.

However, we can test the predictions of the multiharmonic model for the TSSs of *S. cerevisiae*. When computing the average logarithm of the predicted nucleosome occupancy around TSSs across all genes [Fig. 6(e)], we observe that our model correctly predicts a nucleosome-depleted region. However, it overestimates the effect, as we observed before for the genomewide GC content dependence [Fig. 6(b)]. We also show a slightly modified version of this curve where the nucleosome free energy is given by the minimum free energy of all position up to 5 bp away. This means we assume that nucleosomes can slide over a short distance to always be rotationally optimally positioned. This modification has negligible effects on the occupancy curve.

Next, we once again implemented the idea that nucleosomes are only locally in thermal equilibrium. To do so, we normalize the occupancies for each TSS in the corresponding 2000-bp interval to 1. We then average across genes and finally take the logarithm. Remarkably, the normalization step has hardly any effect for the *in vitro* data: The curve where we did not normalize the occupancies is nearly identical to the one where we employed the normalization, see Fig. 6(f). This is a strong indication that the reconstituted chromatin sample is indeed not equilibrated. Applying the same approach to the multiharmonic model gives an occupancy curve that is much closer to the *in vitro* data than what we found in Fig. 6(e). Accounting for short range sliding up to 5 bp to the lowest free-energy state leads to a smoothening of the curve [Fig. 6(f)].

Remarkably, our model predicts three maxima to the right of the TSS, the second being 250 bp apart from the first maximum and the third 550 bp. These maxima cannot be seen in the *in vitro* curve. Also they are hardly visible when using the previous averaging procedure, Fig. 6(e). We speculate that these might be real mechanical cues on the DNA molecules that are present downstream of a substantial fraction of the TSSs. However, it should also be noted that actual *in vivo* nucleosome densities are very high so that the average spacing between nucleosomes is much smaller than suggested by these peaks. *In vivo* occupancies show well-defined peaks with a regular spacing of about 160 bp, see, e.g., Fig. 4(a) in Ref. [55]. These peaks reflect the statistical ordering close to a boundary (here caused by the dip in GC content at TSSs), an effect already proposed in Ref. [62]. In fact, short-enough genes show a crystal-like configuration between their TSSs and transcription termination sites, which also act as nucleosome barriers [63]. The interesting question remains whether a subset of genes shows deviations from this statistical ordering as a result of the mechanical cues found in Fig. 6(f). We plan to address this question in a future study.

VI. CONCLUSIONS

In this paper we asked the question how the sequence-dependent elasticity and geometry of the DNA double helix causes the translational positioning of nucleosomes. We found that when modeling DNA with the rigid base-pair model, a model with a local harmonic elastic energy, translational positioning is a rather subtle effect that in our nucleosome model is predominantly caused by entropy. The overall softer GC-rich steps of the crystallographic parameter set are entropically preferred for nucleosomal DNA which would be in accordance with experimental observations. However, a full model needs also to account for the entropy change when free DNA is wrapped into nucleosomes which is entropically more costly for the softer GC-rich DNA. As a result, if the model employs the same set of parameters for both nucleosomal and free DNA, then it predicts the wrong GC dependence. We checked that this failure of the model is not the result of the specific nucleosome geometry assumed in our model or of the peculiarities of poly-dA:dT tracts.

We introduce a multiharmonic model which uses different parametrizations for free and complexed DNA, namely parameters that have been extracted from these two DNA "states." Specifically, free DNA is parametrized by MD simulations of DNA oligomers (which shows a weak GC dependence) and wrapped DNA by the crystal parameters extracted from protein-DNA crystals. Using in addition the fact that genomic nucleosome maps are typically only locally equilibrated [see Figs. 6(a) and 6(f) for baker's yeast] one finds a good quantitative agreement between the model and the data [see Figs. 6(d) and 6(f)], with a nucleosome preference for high GC content which, however, saturates at large values.

This study therefore suggests that the rigid base-pair model can indeed be used to predict translational nucleosome positioning (and not only rotational positioning [38], which reflects mainly the elastic energy cost due to DNA shape [42]). However, some caution is necessary as the entropy of free DNA needs to be taken into account which, in the most straightforward implementation, inverts the GC dependence. The fact that we need to use a mixed parametrization to arrive at the proper GC dependence points to some shortcomings of the rigid base-pair model. One problem might be that the model is strictly local and therefore cannot describe the stiffening of poly-dA:dT tracts. However, our analysis suggested that this is not the mechanism behind the experimental GC dependence. Another problem might be that the model is harmonic, i.e., one assumes that the elastic energy of each base-pair step varies quadratically with the deviation from the preferred geometry. Furthermore, the different electrostatic environment in the protein-bound state due to counter-ion release and the low dielectric protein core might simply yield a very different but still harmonic potential than the one obtained from MD simulations of bare DNA. This should be investigated in the future. The fact that we need to use a different parametrization for the strongly deformed complexed DNA suggests that this assumption breaks down at least partly, namely when looking at entropy, whereas other sequence-dependent effects, including rotational positioning [38], asymmetric nucleosome breathing [41], and force-induced unwrapping [37], are rather robustly described by the model in Fig. 1. Finally, since the GC dependence of the MD parametrization used for the free DNA is weak, a simplified approach that does not account for the free DNA at all, as done in Ref. [12], might be sufficient.

ACKNOWLEDGMENTS

We thank Johannes Nübler and Lennart de Bruin for discussions. J.N. acknowledges support by the Erasmus+ programme of the European Union and the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation).

APPENDIX A: ON THE INTERPLAY OF ENTROPY AND ENERGY IN SIMPLIFIED ONE-DIMENSIONAL NUCLEOSOME TOY MODELS

We study here various versions of systems consisting of one-dimensional chains of masses connected by harmonic springs. These Rouse model-like systems can be solved analytically and allow us to understand better the effect of the spring stiffnesses on energy and entropy. The surprising dominance of entropy in determining the GC dependence of our nucleosome model can then be better interpreted. First we look just at a series of springs without any constraints ("free DNA"), then we look at systems where the ends of the system are fixed (either rigidly or via extra springs), and, finally, we mimic our nucleosome model rather closely by introducing fixed midplanes.

1. Without constraints

The "free DNA" version of the one-dimensional chain is shown in Fig. 7. Its Hamiltonian is given by:

$$\mathcal{H}_1(\{q_i\},\{p_i\}) = \sum_{i=1}^{N} \frac{p_i^2}{2m} + \sum_{i=2}^{N} \frac{k}{2}(q_i - q_{i-1} - l)^2, \quad \text{(A1)}$$

where q_i is the position, p_i is the momentum of particle i, and k denotes the stiffness of the Hookean springs of length l and N the number of particles. The partition function is then given by

$$\mathcal{Z}_1(T,V,N) = \int \prod_{i=1}^{N} dq_i dp_i \exp[-\beta \mathcal{H}_1(\{q_j\},\{p_j\}]], \quad \text{(A2)}$$

where V is the (one-dimensional) volume to which the chain is constrained. This integral is a product of simple Gaussian integrals leading to

$$\mathcal{Z}_1(T,V,N) = \left(\frac{2\pi m}{\beta}\right)^{\frac{N}{2}} \left(\frac{2\pi}{\beta k}\right)^{\frac{N-1}{2}} V. \quad \text{(A3)}$$

From this, we can calculate the free energy:

$$F_1(T,V,N,k) = -\frac{\ln \mathcal{Z}_1(T,V,N)}{\beta} = \frac{N-1}{2\beta} \ln k + F_1(T,V,N,k=1). \quad \text{(A4)}$$

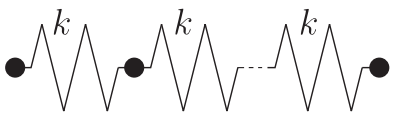

FIG. 7. One-dimensional chain without constraints.

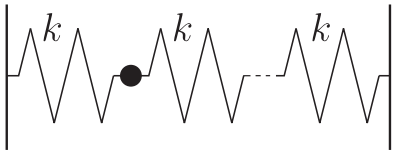

FIG. 8. One-dimensional chain with its ends rigidly constrained.

The energy is simply given by the equipartition theorem:

$$\overline{E_1} = \frac{2N-1}{2} k_B T. \quad \text{(A5)}$$

As expected, the k dependence of the free energy of this system is caused only by the temperature-independent entropy:

$$S_1 = \frac{\overline{E_1} - F_1}{T} = -k_B \frac{N-1}{2} \ln k + S_1(k=1). \quad \text{(A6)}$$

2. With rigid constraints

Next, we study the impact of constraints ("bound DNA"). Here we fix the ends of the chain, see Fig. 8, and compute the partition function of the remaining chain with $N_2 = N - 2$ particles. We redefine the positions of particles as $q_i \to q_i - il$. The Hamiltonian is then given by

$$\mathcal{H}_2(\{q_i\},\{p_i\}) = \sum_{i=1}^{N_2} \frac{p_i^2}{2m} + \frac{k}{2} \sum_{i=2}^{N_2} (q_i - q_{i-1})^2 + \frac{k}{2}\left[q_1^2 + (L - (N_2+1)l - q_{N_2})^2\right]. \quad \text{(A7)}$$

After a longer calculations we arrive at the exact partition function:

$$\mathcal{Z}_2(T,V,N_2) = \left(\frac{2\pi m}{\beta}\right)^{\frac{N_2}{2}} \left(\frac{2\pi}{\beta k}\right)^{\frac{N_2}{2}} \frac{1}{\sqrt{N_2+1}} \times \exp\left\{-\frac{\beta k [L - (N_2+1)l]^2}{2(N_2+1)}\right\}. \quad \text{(A8)}$$

This yields an additional term in the k dependence of the free energy compared to the unbound case discussed in the previous section [see Eq. (A4)]:

$$F_2(k) = \frac{N_2}{2\beta} \ln k + \frac{[L - (N_2+1)l]^2}{2(N_2+1)}(k-1) + F_2(1). \quad \text{(A9)}$$

This term corresponds to an energetic contribution:

$$\overline{E_2}(k) = -\frac{\partial_\beta \mathcal{Z}_2}{\mathcal{Z}_2} = N_2 k_B T + \frac{[L - (N_2+1)l]^2 k}{2(N_2+1)}. \quad \text{(A10)}$$

Note that the k dependence of the energy vanishes if we choose L to be equal to the chain's equilibrium length $(N_2 + 1)l$. The k dependence of entropy behaves as if we had removed one particle from the free DNA:

$$S_2(k) = \frac{\overline{E_2}(k) - F_2(k)}{T} = -k_B \frac{N_2}{2} \ln k + S_2(1) = -k_B \frac{N-2}{2} \ln k + S_2(1). \quad \text{(A11)}$$

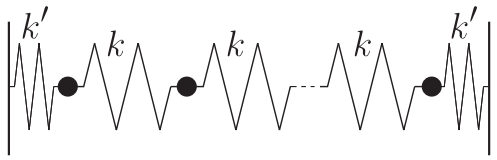

FIG. 9. One-dimensional chain with flexible constraints.

With the results from this and the previous section, we can now determine the k dependence of the likelihood of "binding":

$$\Delta F = F_2 - F_1 = -\frac{1}{2\beta}\ln k + \frac{[L-(N-1)l]^2}{2(N-1)}(k-1) + F_2(1) - F_1(1). \tag{A12}$$

For $L \neq (N-1)l$, i.e., when the imposed end-to-end distance causes the compression or stretching of the chain of springs, there are two regimes of k dependence: For large k, energy dominates yielding a positive k dependence of ΔF and thus stiffer springs are less likely to bind. For small k, entropy dominates (diverging as $\ln k$), making softer chains less likely to bind. The boundary between these two regimes depends on temperature and the degree of compression or stretching. From $\partial \Delta F/\partial k = 0$ follows:

$$k_0 = \frac{N-1}{\beta[L-(N-1)l]^2}. \tag{A13}$$

For $k = k_0$ there is a global minimum of ΔF. The sequences which are most likely to bind are those with a k close to k_0. If k_0 is smaller than the typical stiffness of an ensemble of sequences, then the ensemble will show a positive correlation of free energy and stiffness (energy dominated case). If k_0 is larger, then one finds an anticorrelation (entropy dominated case).

The latter case corresponds to our nucleosome model which leads to the problematic prediction that nucleosomes tend to avoid the softer GC-rich DNA. This suggests that a successful nucleosome model should impose stronger deformations on the complexed DNA. However, using variants of our model, based on several crystal structures and with the prerelaxation step removed, did not resolve the issue, see Fig. 5.

3. With flexible constraints

There might be a conceptual problem when comparing the free energies of the free and the bound state, using rigid constraints for the latter. As we fix the first and the last particle of the chain, the entropic penalty of binding is infinite. This makes the validity of Eq. (A12) questionable. We overcome this here by binding the ends to springs with stiffness k' and zero equilibrium length, Fig. 9. This ensures that the difference in free energy of the bound and the unbound state will not be infinite and an analysis of the k dependence is valid. In the end, we can compute the limit for $k' \to \infty$. As we show here we recover and, thus, validate Eq. (A12). Let us again start from the Hamiltonian, now with the positions redefined as $q_i \to q_i - il + l$:

$$\mathcal{H}_3(\{q_i\},\{p_i\}) = \sum_{i=1}^{N}\frac{p_i^2}{2m} + \frac{k}{2}\sum_{i=2}^{N}(q_i - q_{i-1})^2 + \frac{k'}{2}\{q_1^2 + [L-(N-1)l - q_N]^2\}. \tag{A14}$$

The partition function of this system follows from a longer calculation (similar to the previous section) to be

$$\mathcal{Z}_3 = \left(\frac{2\pi m}{\beta}\right)^{\frac{N}{2}}\left(\frac{2\pi}{\beta k}\right)^{\frac{N}{2}} \times\sqrt{\frac{k^2}{k'[2k+(N-1)k']}}\exp\left\{-\frac{\beta k' k[L-(N-1)l]^2}{2[2k+(N-1)k']}\right\}, \tag{A15}$$

from which we find the free energy

$$F_3(k) = \frac{N-2}{2\beta}\ln k + \frac{k'[L-(N-1)l]^2}{2[2k+(N-1)k']}(k-1) + \ln\left[\frac{2k+(N-1)k'}{2+(N-1)k'}\right] + F_3(1). \tag{A16}$$

Performing the limit $k' \to \infty$, we recover indeed the result from the previous section:

$$\lim_{k'\to\infty}[F_3(k) - F_3(1)] = F_2(k) - F_2(1). \tag{A17}$$

4. With fixed midplanes

In our nucleosome model we do not fix the positions of base pairs but of certain midplanes to mimic the bound phosphates. Therefore we study here entropy and energy of the corresponding one-dimensional system. Let us begin with two fixed midplanes at positions x_1 and x_2. We assume n_1 beads to the left of x_1, n_2 beads between x_1 and x_2 and n_3 beads to the right of x_2. The fixed midplanes allow us to eliminate four degrees of freedom in the Hamiltonian:

$$\frac{q_{n_1} + q_{n_1+1}}{2} = x_1, \quad \frac{q_{n_1+n_2} + q_{n_1+n_2+1}}{2} = x_2 \tag{A18}$$

and

$$p_{n_1} = -p_{n_1+1}, \quad p_{n_1+n_2} = -p_{n_1+n_2+1}. \tag{A19}$$

The kinetic part of the Hamiltonian is then given by

$$\Rightarrow \mathcal{H}_{4,\,\mathrm{kin}} = {\sum_{i=1}^{n_1+n_2+n_3}}' \frac{p_i^2}{2m} + \frac{p_{n_1}^2}{m} + \frac{p_{n_1+n_2}^2}{m}, \tag{A20}$$

where the sum, Σ', skips $i = n_1$, n_1+1, n_1+n_2, and n_1+n_2+1. The potential part reads

$$\mathcal{H}_{4,\,\mathrm{pot}} = {\sum_{i=1}^{n_1+n_2+n_3-1}}' \frac{k}{2}(q_{i+1} - q_i - l)^2 + 2k\left(x_1 - q_{n_1} - \frac{l}{2}\right)^2 + 2k\left(x_2 - q_{n_1+n_2} - \frac{l}{2}\right)^2 + \frac{k}{2}\left(q_{n_1+2} + q_{n_1} - 2x_1 - l\right)^2 + \frac{k}{2}\left(q_{n_1+n_2+2} + q_{n_1+n_2} - 2x_2 - l\right)^2, \tag{A21}$$

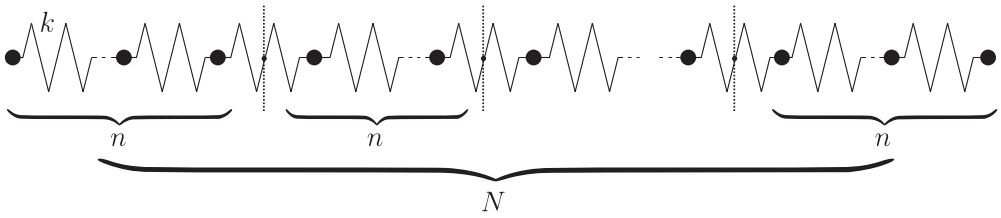

FIG. 10. One-dimensional chain with $N-1$ fixed midplanes.

with the sum, Σ', defined as above. The system behaves as if we had removed particles n_1+1 and n_1+n_2+1, equipped particles n_1 and n_1+n_2 with half the mass and attached them to x_1 and x_2 via springs with stiffness $4k$. As in the previous sections one can calculate the potential part of the partition function,

$$\mathcal{Z}_{4,\,\mathrm{pot}} = \left(\frac{2\pi}{\beta k}\right)^{\frac{n_1+n_2+n_3-2}{2}} \sqrt{\frac{1}{16n_2-8}} \times \exp\left\{-\frac{\beta k[x_2-x_1-(n_1-1)l]^2}{2n_2-1}\right\}, \quad \text{(A22)}$$

from which we calculate the free energy

$$F_4(k) = \frac{n_1+n_2+n_3-2}{2\beta}\ln k + \frac{(x_2-x_1)^2}{2n_2-1}(k-1) + F_4(1). \quad \text{(A23)}$$

We find a similar k dependence of the energy as for the rigid constraints, Eq. (A9), only that the effective number of particles (between the constraints) is half-integer, $n_2-1/2$, instead of n_2+1.

In our nucleosome model, we fix 28 midplanes. Thus, let us generalize the system to a chain of $N_{\mathrm{tot}}=\sum_{i=1}^{N} n_i = Nn$ particles, see Fig. 10. We constrain the midplanes x_i between subchains n_{i+1} and n_i such that $x_{i+1}-x_i=\Delta x$. The particles 1 and Nn are unconstrained, whereas each internal subchain is bound to particles constrained by midplanes. After a longer calculation we find the free energy of the form

$$F_5(k) = \frac{Nn-(N-1)}{2\beta}\ln k + \frac{1}{2}CD(N,n)(k-1) + F_5(k=1), \quad \text{(A24)}$$

with

$$C = \frac{(\Delta x)^2}{n-1}. \quad \text{(A25)}$$

Unfortunately, we did not find a closed analytical form for $D(N,n)$ which can be written as a matrix product.

With this, we can calculate the stiffness k_0 for which $F(k)$ is minimal, because entropic [first term in Eq. (A24)] and energetic contributions [second term in Eq. (A24)] to the k dependence are equal:

$$k_0 = \frac{N-1}{D(N,n)}\frac{k_BT}{C}. \quad \text{(A26)}$$

As for the fixed particles [Eq. (A13)], k_0 is proportional to k_BT/C, although the proportionality factor depends on N and n in a nontrivial manner. Calculating $D(N,n)$ for $3\leqslant N\leqslant 40$ and $2\leqslant n<40$, we find that $D(N,n)$ is of order $N-2$, see Fig. 11. This yields once again $k_0\approx k_BT/C$.

APPENDIX B: THE DINUCLEOTIDE PROBABILITIES IN UNCONSTRAINED DNA

We calculate here the dinucleotide probabilities of free DNA. The Hamiltonian of the rigid base-pair model is given by Eq. (1). The partition function of an $N+1$-bp-long free DNA chain is then given by

$$\begin{aligned}\mathcal{Z}_{\mathrm{free}} &= \int\prod_{a=1}^{N} d\mathbf{x}^a \exp\left[-\frac{\beta}{2}\left(\mathbf{x}^a-\mathbf{x}_0^a\right)^T\mathbf{K}^{\mathbf{a}}\left(\mathbf{x}^a-\mathbf{x}_0^a\right)\right]\\ &= \int\prod_{a=1}^{N} d\mathbf{x}^a \exp\left[-\frac{\beta}{2}(\mathbf{x}^a)^T\mathbf{K}^{\mathbf{a}}\mathbf{x}^a\right]\\ &= \prod_{a=1}^{N}\sqrt{\frac{64\pi^6k_B^6T^6}{\det\mathbf{K}^a}}. \end{aligned} \quad \text{(B1)}$$

Note that the first step contains an approximation as we set the Jacobian determinant for the transformation from canonical momenta to configuration-independent momenta as defined in Ref. [64] to 1; we checked that even for the strongly deformed nucleosomal DNA geometry this quantity takes values close to 1 (namely between 0.95 and 1). From this we obtain the

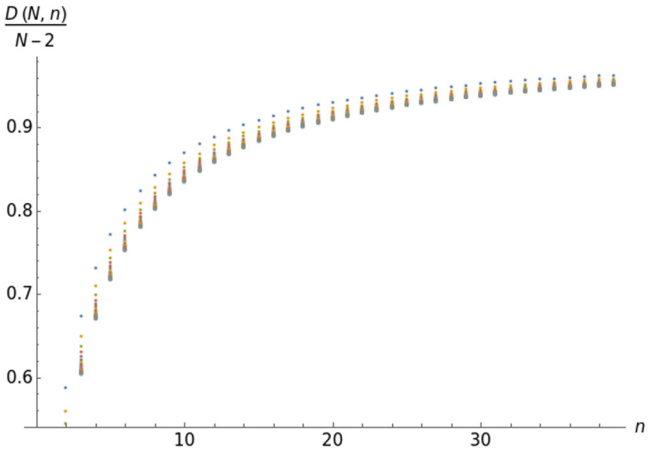

FIG. 11. $D(N,n)$ for $3\leqslant N\leqslant 40$.

sequence-dependent free energy of unconstrained DNA:

$$F_{\text{free}}(\{\mathbf{K}^a\}) - F_{\text{free}}(\mathbf{K}=\mathbf{1}) = \frac{k_B T}{2}\sum_{a=1}^{N} \ln\det \mathbf{K}^a = -T\sum_{a=1}^{N} S_{\text{free}}(s_{a-1}, s_a), \quad \text{(B2)}$$

where S is the stiffness-dependent entropy contribution of the dinucleotide (s_{a-1}, s_a).

This allows us to compute the dinucleotide probabilities in a free DNA molecule with the transfer matrix method used in Refs. [42] and [46]. The transfer matrix $\mathbf{M}$ is defined in the nucleotide basis $B = \{|A\rangle, |C\rangle, |G\rangle, |T\rangle\}$:

$$\langle n\,|\,\mathbf{M}\,|\,m\rangle \equiv \exp[-\beta F_{\text{free}}(n,m)] = \exp[S_{\text{free}}(n,m)/k_B] = \frac{1}{\sqrt{\det \mathbf{K}(n,m)}}, \quad \text{(B3)}$$

which corresponds to the Boltzmann weight of the dinucleotide (n, m). To compute the probability P_i of this dinucleotide at position i, we need to take the Boltzmann weight of all possible neighboring sequences into account and compare it to the Boltzmann weight of all possible sequences:

$$P_i(a,b) = \frac{\sum_{n_0,n_N} \langle n_0\,|\,\mathbf{M}^i\,|\,a\rangle\langle a\,|\,\mathbf{M}\,|\,b\rangle\langle b\,|\,\mathbf{M}^{N-i-1}\,|\,n_N\rangle}{\sum_{n_0,n_N}\langle n_0\,|\,\mathbf{M}^N\,|\,n_N\rangle}. \quad \text{(B4)}$$

The configuration, and therefore the transfer matrix, is not position dependent, in contrast to the nucleosome. Still, the dinucleotide probabilities are position dependent, because for dinucleotides at the 5′ or 3′ end the entropy of possible 3′ or 5′ neighbors does not matter, in contrast to dinucleotides in the middle of the DNA molecule. However, these boundary effects are only relevant to dinucleotides in the immediate vicinity of the 5′ or 3′ end, see Figs. 4(c) and 4(f).

[1] S. Sazer and H. Schiessel, Traffic **19**, 87 (2018).
[2] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Nature **389**, 251 (1997).
[3] E. Segal, Y. N. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, Nature **442**, 772 (2006).
[4] E. N. Trifonov and J. L. Sussman, Proc. Natl. Acad. Sci. USA **77**, 3816 (1980).
[5] S. C. Satchwell, H. R. Drew, and A. A. Travers, J. Mol. Biol. **191**, 659 (1986).
[6] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. Leproust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, Nature **458**, 362 (2009).
[7] P. T. Lowary and J. Widom, Proc. Natl. Acad. Sci. USA **94**, 1183 (1997).
[8] D. Tillo and T. R. Hughes, BMC Bioinform. **10**, 442 (2009).
[9] G. Locke, D. Tolkunov, Z. Moqtaderi, K. Struhl, and A. V. Morozov, Proc. Natl. Acad. Sci. USA **107**, 20998 (2010).
[10] K. Struhl and E. Segal, Nat. Struct. Mol. Biol. **20**, 267 (2013).
[11] G. Drillon, B. Audit, F. Argoul, and A. Arneodo, BMC Genom. **17**, 526 (2016).
[12] M. Tompitak, C. Vaillant, and H. Schiessel, Biophys. J. **112**, 505 (2017).
[13] T. Vavouri and B. Lehner, PLOS Genet. **7**, e1002036 (2011).
[14] G. Meersseman, S. Pennings, and E. M. Bradbury, EMBO J. **11**, 2951 (1992).
[15] I. M. Kulić and H. Schiessel, Phys. Rev. Lett. **91**, 148103 (2003).
[16] F. Mohammad-Rafiee, I. M. Kulić, and H. Schiessel, J. Mol. Biol. **344**, 47 (2004).
[17] G. B. Brandani, T. Niina, C. Tan, and S. Takada, Nucleic Acids Res. **46**, 2788 (2018).
[18] H. Schiessel, J. Widom, R. F. Bruinsma, and W. M. Gelbart, Phys. Rev. Lett. **86**, 4414 (2001).
[19] I. M. Kulić and H. Schiessel, Biophys. J. **84**, 3197 (2003).
[20] J. Lequieu, D. C. Schwartz, and J. J. de Pablo, Proc. Natl. Acad. Sci. USA **114**, E9197 (2017).
[21] T. Niina, G. B. Brandani, C. Tan, and S. Takada, PLOS Comput. Biol. **13**, e1005880 (2017).
[22] J. Winger, I. M. Nodelman, R. F. Levendosky, and G. D. Bowman, eLife **7**, e34100 (2018).
[23] M. Li, X. Xia, Y. Tian, Q. Jia, X. Liu, Y. Lu, M. Li, X. Li, and Z. Chen, Nature **567**, 409 (2019).
[24] A. Sabantsev, R. F. Levendosky, X. Zhuang, G. D. Bowman, and S. Deindl, Nat. Commun. **10**, 1720 (2019).
[25] G. B. Brandani and S. Takada, PLOS Comput. Biol. **14**, e1006512 (2018).
[26] E. Segal and J. Widom, Trends Genet. **25**, 335 (2009).
[27] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, Proc. Natl. Acad. Sci. USA **95**, 11163 (1998).
[28] C. Anselmi, G. Bocchinfuso, P. D. Santis, M. Savino, and A. Scipioni, Biophys. J. **79**, 601 (2000).
[29] M. Y. Tolstorukov, A. V. Colasanti, D. M. McCandlish, W. K. Olson, and V. B. Zhurkin, J. Mol. Biol. **371**, 725 (2007).
[30] C. Vaillant, B. Audit, and A. Arneodo, Phys. Rev. Lett. **99**, 218103 (2007).
[31] S. Balasubramanian, F. Xu, and W. K. Olson, Biophys. J. **96**, 2245 (2009).
[32] A. V. Morozov, K. Fortney, D. A. Gaykalova, V. M. Studitsky, J. Widom, and E. D. Siggia, Nucleic Acids Res. **37**, 4707 (2009).
[33] N. B. Becker and R. Everaers, Structure **17**, 579 (2009).
[34] A. Fathizadeh, A. B. Besya, M. R. Ejtehadi, and H. Schiessel, Eur. Phys. J. E **36**, 21 (2013).
[35] T. Držata, N. Špačková, P. Jurečka, M. Zgarbová, S. Šponer, and F. Lankaš, Nucleic Acids Res. **42**, 7383 (2014).
[36] D. Norouzi and F. Mohammad-Rafiee, J. Biomol. Struct. Dyn. **32**, 104 (2014).
[37] L. de Bruin, M. Tompitak, B. Eslami-Mossallam, and H. Schiessel, J. Phys. Chem. B **120**, 5855 (2016).
[38] B. Eslami-Mossallam, R. D. Schram, M. Tompitak, J. van Noort, and H. Schiessel, PLOS One **11**, e0156905 (2016).
[39] M. Tompitak, L. de Bruin, B. Eslami-Mossallam, and H. Schiessel, Phys. Rev. E **95**, 052402 (2017).
[40] J. A. J. Wondergem, H. Schiessel, and M. Tompitak, J. Chem. Phys. **147**, 174101 (2017).

[41] J. Culkin, L. de Bruin, M. Tompitak, R. Phillips, and H. Schiessel, Eur. Phys. J. E **40**, 106 (2017).
[42] M. Zuiddam, R. Everaers, and H. Schiessel, Phys. Rev. E **96**, 052412 (2017).
[43] M. Zuiddam and H. Schiessel, Phys. Rev. E **99**, 012422 (2019).
[44] G. S. Freeman, J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. J. de Pablo, Phys. Rev. Lett. **113**, 168101 (2014).
[45] F. Lankaš, J. Sponer, J. Langowski, and T. E. Cheatham III, Biophys. J. **85**, 2872 (2003).
[46] N. B. Becker, L. Wolff, and R. Everaers, Nucleic Acids Res. **34**, 5638 (2006).
[47] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, J. Mol. Biol. **319**, 1097 (2002).
[48] M. Tompitak, G. T. Barkema, and H. Schiessel, BMC Bioinform. **18**, 157 (2017).
[49] Y. Field, N. Kaplan, Y. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal, PLOS Comput. Biol. **4**, e1000216 (2008).
[50] E. Y. D. Chua, D. Vasudevan, G. E. Davey, B. Wu, and C. A. Davey, Nucleic Acids Res. **40**, 6338 (2012).
[51] R. D. Makde, J. R. England, H. P. Yennawar, and S. Tan, Nature **467**, 562 (2010).
[52] Z. Zhang, C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh, Science **332**, 977 (2011).
[53] H. Chung, I. Dunkel, F. Heise, C. Linke, S. Krobitsch, A. E. Ehrenhofer-Murray, S. R. Sperling, and M. Vingron, PLOS One **5**, e15754 (2010).
[54] W. Hörz and W. Altenburger, Nucleic Acids Res. **9**, 2643 (1981).
[55] K. Brogaard, L. Xi, J. P. Wang, and J. Widom, Nature **486**, 496 (2012).
[56] J. Allan, R. M Fraser, T. Owen-Hughes, and D. Keszenman-Pereyra, J. Mol. Biol. **417**, 152 (2012).
[57] R. Ammar, D. Torti, K. Tsui, M. Gebbia, T. Durbic, G. D. Bader, G. Giaever, and C. Nislow, Elife **1**, e00078 (2012).
[58] H. Maruyama, J. C. Harwood, K. M. Moore, K. Paszkiewicz, S. C. Durley, H. Fukushima, H. Atomi, K. Takeyasu, and N. A. Kent, EMBO Rep. **14**, 711 (2013).
[59] E. Segal and J. Widom, Curr. Opin. Struct. Biol. **19**, 65 (2009).
[60] J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, and J. H. Maddocks, J. Chem. Theory Comput. **13**, 1539 (2017).
[61] T. E. Haran and U. Mohanty, Q. Rev. Biophys **42**, 41 (2009).
[62] R. D. Kornberg and L. Stryer, Nucleic Acids Res. **16**, 6677 (1988).
[63] G. Chevereau, L. Palmeira, C. Thermes, A. Arneodo, and C. Vaillant, Phys. Rev. Lett. **103**, 188103 (2009).
[64] O. Gonzalez and J. H. Maddocks, Theor. Chem. Acc. **106**, 76 (2001).