Physics behind the mechanical nucleosome positioning code

Martijn Zuiddam
Institute Lorentz for Theoretical Physics, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands

Ralf Everaers
Univ Lyon, ENS de Lyon, Univ Claude Bernard Lyon 1, CNRS, Laboratoire de Physique and Centre Blaise Pascal, F-69342 Lyon, France

Helmut Schiessel
Institute Lorentz for Theoretical Physics, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
(Received 11 July 2017; published 28 November 2017)

The positions along DNA molecules of nucleosomes, the most abundant DNA-protein complexes in cells, are influenced by the sequence-dependent DNA mechanics and geometry. This leads to the "nucleosome positioning code", a preference of nucleosomes for certain sequence motives. Here we introduce a simplified model of the nucleosome where a coarse-grained DNA molecule is frozen into an idealized superhelical shape. We calculate the exact sequence preferences of our nucleosome model and find it to reproduce qualitatively all the main features known to influence nucleosome positions. Moreover, using well-controlled approximations to this model allows us to come to a detailed understanding of the physics behind the sequence preferences of nucleosomes.

DOI: 10.1103/PhysRevE.96.052412

I. INTRODUCTION

The DNA double helix carries, in addition to the classical genetic information (the genes encoding for the proteins), a mechanical layer of information. This is possible because the mechanical properties of DNA depend on the underlying sequence of base pairs (bp). Certain combinations of letters (especially bp steps) are softer than others and some cause intrinsic bends on the DNA molecule [1]. So unlike in a book where the stiffness of the paper does not depend on the text printed, DNA elasticity and geometry is intimately linked to the text it carries.

Possibly the most important biological consequence of sequence-dependent DNA mechanics is its impact on the positioning of DNA spools, called nucleosomes (Fig. 1). The core of each spool is a cylinder composed of eight histone proteins, and it is wrapped by a DNA stretch of 147-bp length. A short stretch of unbound DNA, the linker DNA, connects to the next protein spool. It is known from the nucleosome crystal structure [2] that the DNA is bound to the protein core at 14 locations where the minor groove of the DNA double helix faces the cylinder. This defines the binding path, a left-handed superhelix of one and three quarter turns.

This structure makes nucleosomes ideal "readers" of mechanical cues. First, the length that is wrapped in a nucleosome is about one persistence length, 50 nm. It follows that the bending energy is much larger (about 60 times [3]) than the thermal energy. Thus, even a small change in the wrapped bp sequence is expected to have a strong effect on the nucleosome affinity. Second, as the binding to the histone octamer occurs mostly with the two backbones of the DNA double helix, there is no direct readout of the sequence but instead the nucleosome affinity results from the elasticity and geometry of the involved DNA stretch.

It is indeed known from various experiments that nucleosomes have sequence preferences [4–6]. High affinity sequences show certain motifs along the wrapped DNA. This "nucleosome positioning code" is typically formulated in terms of bp steps or, looking along one strand, dinucleotides: most importantly, the probability of finding GC steps (nucleotide G followed by nucleotide C) peaks at positions where the major groove faces the protein cylinder (every 10th bp), whereas TT, AA, and TA are all in phase and have their peaks in between where the minor groove faces the cylinder [see Fig. 1(a)].

Over evolutionary time scales mechanical signals have evolved along genomes. Examples are nucleosome depleted regions at transcription start sites in yeast facilitating transcription initiation [6,7], mechanically encoded retention of a small fraction of nucleosomes in human sperm cells allowing transmission of paternal epigenetic information [7,8], or the positioning of six million nucleosomes around nucleosome inhibiting barriers in human somatic cells [9].

However, what is still missing is a deeper understanding of the physics underlying nucleosome positioning rules. An example, mentioned in Ref. [10], are the positions where GC steps typically occur in high-affinity sequences. These correspond to positions that GC steps dislike the most. Even more remarkably, of *all* 16-bp steps it is the GC step that is energetically most costly at these positions.

A first step toward understanding the nucleosome positioning rules is using coarse-grained DNA models with sequence-dependent elasticity and force them into shapes that resemble the wrapped DNA portions in nucleosomes. Several such models use the so-called rigid base pair model [11,12], in which the conformation of a DNA molecule is described by the positions and orientations of its base pairs that are modeled as rigid objects. These nucleosome models have been used to predict nucleosome stability and positioning [13–20], forces and torques on the wrapped DNA [21], nucleosome mobility along DNA [22], and the response of nucleosomes to external forces [23,24]. One recent study [10,25] specifically addresses the question whether such models can predict the above-mentioned rules of the nucleosome positioning code. This was achieved by introducing the mutation Monte Carlo method, which mixes conformational and sequence moves.

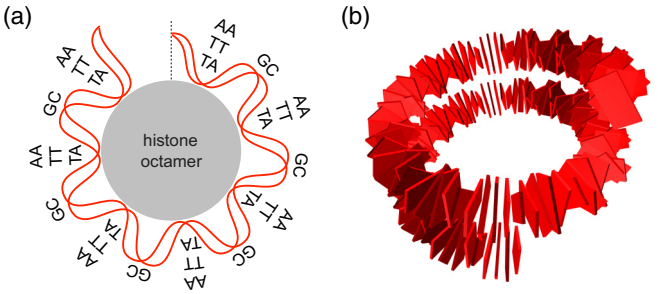

FIG. 1. (a) Schematic view of one half of the symmetric nucleosome; the vertical dashed line indicates the dyad axis. Key dinucleotides that position the nucleosome along DNA are indicated with their location inside the nucleosome constituting the "nucleosome positioning code". (b) Visual representation of our model for nucleosomal DNA. Base pairs represented by rigid plates are frozen in an idealized superhelical shape.

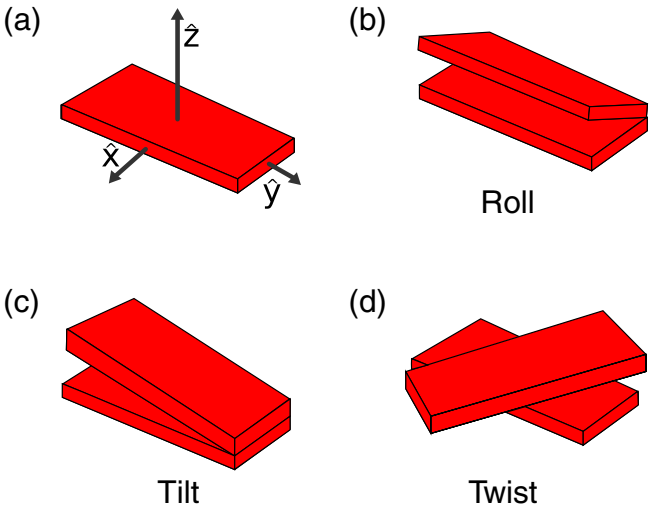

FIG. 2. The rotational degrees of freedom between neighboring bp in the rigid base pair model. Each base pair has a coordinate system (a), which can be used to describe the relative orientation between two plates. In our model we account only for energy contributions from (b) roll and (c) tilt but neglect contributions from (d) twist. Also the translational degrees are not considered.

This method automatically produces the sequence preferences along the wrapped DNA, and it was indeed found that it reproduces the nucleosome positioning rules. However, the model is still far too complex to really come to a clear interpretation of how the rules result from the underlying elasticity and geometry of the DNA.

Here we overcome this complexity by reducing the model to its bare essentials: we consider a piece of DNA that is forcibly curved and idealize the shape by placing it on a superhelical path [Fig. 1(b)]. Assuming such an idealized shape (as done in Refs. [13,15–17]) instead of trying to imitate details of the crystal structure (as done in Refs. [10,14,18–20,22,23,25]) makes our model analytically tractable and allows us to pinpoint the dominant contributions that underlie the positioning code. Moreover, we freeze the model into this configuration, unlike in some models where the base pairs are free to move with respect to others (at some energy cost) [10,17,19,20,22,23,25]. Variants of our approach are in principle applicable to any model that freezes the DNA into a fixed configuration like it is done in Refs. [13–16,18].

The goal of this work is not to come up with yet another tool for nucleosome positioning. Based on the more complete model [10], we were able to build a probabilistic model that is as fast as the model introduced here and is very successful in predicting nucleosome positioning [7]. The goal of the current work is instead to come to a deep understanding of the positioning rules. For instance, we will be able to explain what cause GC steps to "favor" the most costly positions on the wrapped DNA. To achieve this, an analytical approach as presented here is indispensable.

In Sec. II we introduce our model. In Sec. III we explain how it can be solved using transfer matrices. This is followed by two sections that develop approximations that allow us to come to a detailed understanding of the nucleosome positioning rules: in Sec. IV we take a limited number of neighbors around the given base pair step into account to derive upper and lower bounds for the probabilities of its occurrence, and in Sec. V we introduce the average neighbor energy approximation, an effective approximation for interpreting nucleosome positioning rules. The exact dinucleotide probabilities, approximations to them, and an interpretation of the rules is presented in Sec. VI, and a conclusion is provided in the final section.

II. MODEL

The model is a simplified version of our computational model that we introduced earlier to predict the nucleosome positioning rules [10]. This full model was extensively tested to produce not only those rules but also to predict the proper response of nucleosomes with various sequences to external forces [23] and to reproduce nucleosome positioning maps of various organisms around transcription start sites [7]. We also used the model to explain sequence dependence in nucleosome breathing [20] and the outcome of SELEX experiments with nucleosomes and DNA rings [25]. That earlier model is based on the rigid base pair model [11,12], a coarse-grained representation of the DNA double helix that treats the base pairs as rigid plates. Neighboring plates differ by six degrees of freedom called shift, slide, rise, roll, tilt, and twist. The rotational degrees of freedom, roll, tilt, and twist, are shown in Fig. 2. In the earlier model we accounted for the shape of the DNA in the nucleosome crystal structure by modeling the various binding sites between the histone octamer and the DNA, we allowed for the structural relaxation of the DNA, and we took into account the energetic contributions from all degrees of freedom between the base pairs, including the cross-terms between them. However, such a model is too complex to interpret the precise nature by which the nucleosome sequence preferences come about.

To arrive at a model that can be solved analytically and that can be interpreted in a straightforward manner we make the following simplifications. We force our DNA model into an ideal superhelix to mimic the bending of the DNA inside a nucleosome, neglecting the nonuniform bending of the nucleosomal DNA observed in its crystal structure [2]. As

the general nucleosome positioning rules hold all along the wrapped part [5], we expect that this simplification does not affect the rules whose origin we aim to understand here. As we were able to rationalize the findings in our earlier nucleosome model just based on the overall bending shape of the DNA [10], we neglect here the possibility of structural relaxation of the DNA molecule. Moreover, as we arrived earlier [10] at a qualitative understanding of the positioning rules just based on two degrees of freedom, roll and tilt, we only account for the energetic contributions from these two degrees. Especially, we assume here the DNA to be unshearable, as, e.g., also done by Vaillant *et al.* [15]. This means that translational degrees of freedom are not accounted for, even though it is known that slide plays a role in the roll-and-slide mechanism in real nucleosomes [14]. However, since this mechanism was not observed in our full model, it should not be included in its simplified version. In principle, it is straightforward to account for the third rotational degree of freedom, twist, and the coupling between all the three rotational degrees. In the main text, however, we do not account for these contributions, even though they are not strictly negligible. The reasons are that they do not change qualitatively the nucleosome positioning rules and that with the remaining two degrees of freedom, roll and tilt, the interpretation of the exact results proves to be challenging enough. For completeness, however, we provide results when also twist and all rotational cross terms are accounted for in Appendix A.

The rigid base pair model assumes only nearest-neighbor interactions and places a quadratic deformation energy between successive base pairs with bp-step-dependent stiffnesses and intrinsically preferred configurations. We use in the following the hybrid parametrization, where the intrinsic values are derived from protein-DNA crystals and the stiffnesses from atomistic molecular simulations [26]; see Table I for a list of the parameters for roll and tilt.

To calculate the difference between the preferred and the *actual* configuration, we need to formally define the shape of our superhelix. We consider a superhelix with pitch P and radius R (similar to Morozov *et al.*, Ref. [17]):

$$\vec{r}(s) = [R\cos(s/R_{\text{eff}}), R\sin(s/R_{\text{eff}}), -(P/2\pi R_{\text{eff}})s], \quad (1)$$

with $R_{\text{eff}} = \sqrt{R^2 + (P/2\pi)^2}$. The set of Frenet-Serret vectors at position s on the superhelix are given by

$$[\hat{t}(s),\hat{n}(s),\hat{b}(s)] = \left[\frac{d\vec{r}}{ds}, \frac{d\vec{t}}{ds}\Big/\left|\frac{d\vec{t}}{ds}\right|, \vec{t}\times\vec{n}\right], \quad (2)$$

where $\hat{t}$ is the tangent unit vector, $\hat{n}$ the principal normal unit vector, and $\hat{b}$ the binormal unit vector.

The rotational orientation of a base pair plate, compared to the origin, can be described using the three orthonormal vectors $\hat{x},\hat{y},\hat{z}$; see Fig. 2(a). We place the double helical shape of the DNA on the superhelix by defining the orthonormal vectors with respect to the Frenet-Serret vectors, such that the double helix revolves (twists) righthandedly around the superhelix:

$$\begin{aligned}[\hat{x}(p),\hat{y}(p),\hat{z}(p)] &= [\hat{n}(p)\cos(\theta p+\phi) - \hat{b}(p)\sin(\theta p+\phi),\\ &\quad -\hat{n}(p)\sin(\theta p+\phi) - \hat{b}(p)\cos(\theta p+\phi),\hat{t}(p)],\end{aligned} \quad (3)$$

with $p(s) = s(L-1)/(2\pi R_{\text{eff}}\alpha) + 1/2$ the position of the dinucleotide (right in between the two plates), where α denotes the number of superhelical turns and L the number of base pairs wrapped around the nucleosome. The constants θ and ϕ determine how much the double helix is twisted, and which positions correspond to maximum and minimum roll and tilt. To reflect the fact that the minor groove is facing toward the histone octamer every 10 bp's, we set $\theta = 2\pi/10$. The phase ϕ is set to $-147\pi/10$ such that the bp at the central position between dinucleotide steps 73 and 74 corresponds to the position of maximal roll, in accordance with the fact that at that position the major groove faces the histone octamer. This is also the place where the tilt changes sign from negative to positive values.

The convention we use to calculate the roll, tilt, and twist degrees of freedom from the orientation of the plates has been well-explained in the literature [27] and will not be discussed here. We will provide the (numerical) results of this method, as well as a short explanation of the values. Using $P = 25.9$ Å, $R = 41.9$ Å, $\alpha = 1.84$, and $L = 147$ [17], we find expressions for the angles q_p^i, $i \in \{\text{roll,tilt,twist}\}$ given by

$$\begin{aligned}\left[q_p^{\text{roll}}, q_p^{\text{tilt}}, q_p^{\text{twist}}\right] &= [\Gamma\cos(2\pi p/10 - 147\pi/10), \Gamma\sin(2\pi p/10\\ &\quad - 147\pi/10), q^{\text{twist}}],\end{aligned} \quad (4)$$

with $\Gamma \approx 0.0796$ rad and $q^{\text{twist}} \approx 2\pi/10.17$ rad. These values can be rationalized the following way. Our superhelix has constant curvature, and as a result, a constant angle between each dinucleotide pair, to which roll and tilt make equal contributions [17]. This angle is given by $\arccos\{\vec{t}\,[s(p)]\cdot\vec{t}\,[s(p+1)]\} \approx 0.0788$, which is a great approximation for our value of Γ. The twist q^{twist} we report is lower than the value for θ we defined. While roll and tilt are 10-bp periodic, the twist corresponds to a periodicity of 10.17. This discrepancy between these values reflects the fact that the path is superhelical; a twist of $2\pi/10$ rad would lead to a planar ring instead.

TABLE I. Parametrization used to calculate the dinucleotide energy, Eqs. (5) to (7). The symbols q and Q denote the intrinsic value and the stiffness of roll or tilt.

	q^{roll} [rad]	Q^{roll} [$\frac{k_B T_r}{\text{rad}^2}$]	q^{tilt} [rad]	Q^{tilt} [$\frac{k_B T_r}{\text{rad}^2}$]
AA	0.012410451	126.98464	−0.024820902	207.73324
AT	0.019409417	148.42141	0	216.86174
AC	0.012372536	143.15931	−0.0017675051	221.16218
AG	0.079562987	123.91326	−0.030057128	200.28179
TA	0.058653564	73.527282	0	129.10674
TT	0.012410451	126.98464	0.024820902	207.73324
TC	0.03372236	113.06128	0.026622916	210.62471
TG	0.083496463	97.396194	−0.0088826025	146.17762
CA	0.083496463	97.396194	0.0088826025	146.17762
CT	0.079562987	123.91326	0.030057128	200.28179
CC	0.063703201	130.1586	0.0017695334	225.01953
CG	0.095824007	83.019248	0	150.88272
GA	0.03372236	113.06128	−0.026622916	210.62471
GT	0.012372536	143.15931	0.0017675051	221.16218
GC	0.0053117746	146.67053	0	214.38125
GG	0.063703201	130.1586	−0.0017695334	225.01953

As mentioned before, we only account for two degrees of freedom and also neglect cross terms between them. Hence, the energy of placing a dinucleotide step $a,b \in \{A,T,C,G\}$ at position p is the sum of the roll and tilt energies:

$$E_p(a,b) = E_p^{\text{roll}}(a,b) + E_p^{\text{tilt}}(a,b), \tag{5}$$

with

$$E_p^{\text{roll}}(a,b) \equiv \tfrac{1}{2} Q^{\text{roll}}(a,b)\left[q_p^{\text{roll}} - \bar{q}^{\text{roll}}(a,b)\right]^2, \tag{6}$$

and

$$E_p^{\text{tilt}}(a,b) \equiv \tfrac{1}{2} Q^{\text{tilt}}(a,b)\left[q_p^{\text{tilt}} - \bar{q}^{\text{tilt}}(a,b)\right]^2. \tag{7}$$

The bp-step-dependent stiffnesses in the roll and tilt degrees of freedom are given by $Q^{\text{roll}}(a,b)$ and $Q^{\text{tilt}}(a,b)$ and the corresponding intrinsic values by $\bar{q}^{\text{roll}}(a,b)$ and $\bar{q}^{\text{tilt}}(a,b)$.

III. DINUCLEOTIDE PROBABILITIES

Here we calculate the dinucleotide probability distribution along our nucleosome model. Base pair steps are the mechanical units in our model and also the experimentally observed nucleosome sequence preferences are typically formulated in terms of dinucleotides [1,5]. We therefore aim to obtain the probability of having nucleotides a and b at dinucleotide position p on the DNA molecule of length $L = 147$. The nucleotides are numbered from 1 to L, such that the pth dinucleotide position contains nucleotides p and $p+1$. The probability does not merely depend on the energy stored between a and b. These bases are connected to other bases as well. To find the probability we need to sum over all possible DNA strands containing a and b at position p, and divide by the partition sum. Therefore, the probability is given by

$$P_p(a,b) = \frac{\sum_{\substack{n_1,\ldots,n_L \\ n_p=a,n_{p+1}=b}} \exp\left[-\beta \sum_{i=1}^{L-1} E_i(n_i,n_{i+1})\right]}{\sum_{n_1,\ldots,n_L} \exp\left[-\beta \sum_{i=1}^{L-1} E_i(n_i,n_{i+1})\right]}, \tag{8}$$

where we sum over all possible states $n_i \in \{A,T,C,G\}$, with β the inverse temperature. The probability given by Eq. (8) corresponds to the case where the nucleosomal DNA sequence mutates freely. This is distinct from the scenario where various DNA stretches compete for nucleosomes, as it is typically the case in experiments such as Refs. [4–6]. Then also entropic effects play a role (e.g., softer bp steps prefer to reside outside nucleosomes for entropic reasons). However, our model is also a reasonable approximation to this case since this system is energy-dominated for physiological temperatures (and lower). In this study, we therefore consider only energies but neglect entropic contributions associated with conformational degrees of freedom.

This type of probabilities can be evaluated using *transfer matrices*. Transfer matrix formalisms have been used both in the context of calculating dinucleotide probabilities for a single nucleosome and evaluating many-nucleosome systems [5,17,26,28] (see Ref. [29] for an overview).

We define the position-dependent transfer matrix T_i in the basis $B = \{|A\rangle, |T\rangle, |C\rangle, |G\rangle\}$ such that

$$\langle n|T_i|m\rangle \equiv \exp[-\beta E_i(n,m)], \tag{9}$$

with $|n\rangle, |m\rangle \in B$. This allows us to rewrite the probability as

$$P_p(a,b) = \frac{\sum_{n_1,n_L} \langle n_1|\, T_1 \ldots T_{p-1}\, |a\rangle\, \langle a|T_p|b\rangle\, \langle b|\, T_{p+1}\ldots T_{L-1}\, |n_L\rangle}{\sum_{n_1,n_L} \langle n_1|\, T_1 \ldots T_{p-1} T_p T_{p+1} \ldots T_{L-1}\, |n_L\rangle}. \tag{10}$$

Finding this probability involves multiplying $L-1 = 146$ four-by-four transfer matrices in the nominator and denominator.

While this quantity is easy to calculate, the sheer number of terms makes it hard to determine which terms influence the probability most and which terms can be neglected. It seems reasonable that bases at positions far away from position p are not as important to the probability as its close neighbors, e.g., at positions $p+1$ and $p-1$. In the next section we will show this by quantifying the effect that far-away bases can possibly have on the probability.

IV. BOUNDS OF DINUCLEOTIDE PROBABILITIES

Here we show how much the probability $P_p(a,b)$ can be affected by the energies of nucleotides some steps away from the position p. In the following we quantify the effect by calculating kth-order bounds of the probability, which we obtain using only the energies of k bases to the left and k bases to the right of the dinucleotide at position p. We assume that all the "unused" bases either try to make the probability $P_p(a,b)$ as high or as low as possible. This is done by substituting all terms related to the unused bases on the left by $\langle x_k|$, and the terms related to unused bases on the right by $|y_k\rangle$. The probability for $k \geqslant 1$ is then given by

$$P_p(a,b) = \frac{\langle x_k| \prod_{i=p-k}^{p-1} T_i\, |a\rangle\, \langle a|T_p|b\rangle\, \langle b| \prod_{j=p+1}^{p+k} T_j\, |y_k\rangle}{\langle x_k| \prod_{i=p-k}^{p+k} T_i\, |y_k\rangle}, \tag{11}$$

with

$$\langle x_k| \equiv \frac{1}{c_k} \sum_n \langle n|\, T_1 T_2 \ldots T_{p-k-2} T_{p-k-1} \tag{12}$$

and

$$|y_k\rangle \equiv \frac{1}{d_k} \sum_n T_{p+k+1} T_{p+k+2} \ldots T_{L-2} T_{L-1}\, |n\rangle\,, \tag{13}$$

where c_k and d_k are normalization constants such that $|\langle x_k|x_k\rangle| = 1$ and $|\langle y_k|y_k\rangle| = 1$. Note that $\langle x_k|$ and $|y_k\rangle$ implicitly depend on p.

To find the kth-order bounds on the probability, we assume that we know nothing about $\langle x_k|$ or $|y_k\rangle$ other than that they represent physically possible states. We formally define the kth-order upper and lower bound by taking the maximum and minimum of Eq. (11), where we let $\langle x_k|$ and $|y_k\rangle$ run over all their possible states. Because the transfer matrix contains Boltzmann weights only, all entries in the transfer matrices T_i are positive. From this it follows that $|x_k\rangle = \sum_{n\in\{A,T,C,G\}} x_{n,k}\, |n\rangle$ and $|y_k\rangle = \sum_{n\in\{A,T,C,G\}} y_{n,k}\, |n\rangle$, with $0 < x_{n,k} \leqslant 1$ and $0 < y_{n,k} \leqslant 1$. These equations are equivalent to the quantum mechanical representation of *mixed states*. The

probabilities to encounter the four possible bases k positions to the left and right of dinucleotide a,b are weighted by $x_{n,k}$ and $y_{n,k}$, parameters that depend on the energy costs of bases farther away.

It turns out that one finds the minimally and maximally possible value of the probability when $|x_k\rangle$ and $|y_k\rangle$ are *pure states*, states from the basis $B = \{|A\rangle, |T\rangle, |C\rangle, |G\rangle\}$. Pure states correspond to *exactly* knowing which bases are present k bases to the left and to the right of the dinucleotide a,b. (Strictly speaking, this happens only when the energy costs of encountering the other possible bases are infinitely high. In other words, this is a limiting case.)

Since, as we prove below, the minimally and maximally possible value of the probability is found when $|x_k\rangle$ and $|y_k\rangle$ are pure states, one can compute the kth-order upper and lower bounds of the probability, $P^{(k)}_{\max,p}(a,b)$ and $P^{(k)}_{\min,p}(a,b)$, by simply evaluating the probability for all 16 possible combinations of pure states. This leads to the expressions

$$P^{(k)}_{\max,p}(a,b) = \max_{|x_k^*\rangle,|y_k^*\rangle\in B} \frac{\langle x_k^*| \prod_{i=p-k}^{p-1} T_i\, |a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j\, |y_k^*\rangle}{\langle x_k^*| \prod_{i=p-k}^{p+k} T_i\, |y_k^*\rangle} \tag{14}$$

and

$$P^{(k)}_{\min,p}(a,b) = \min_{|x_k^*\rangle,|y_k^*\rangle\in B} \frac{\langle x_k^*| \prod_{i=p-k}^{p-1} T_i\, |a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j\, |y_k^*\rangle}{\langle x_k^*| \prod_{i=p-k}^{p+k} T_i\, |y_k^*\rangle}. \tag{15}$$

We prove now the expression for the kth-order upper bound of the probability (the proof for the lower bound can be obtained analogously). We substitute $|x_k\rangle = \sum_{n\in\{B\}} x_{n,k}\,|n\rangle$ and $|y_k\rangle = \sum_{m\in\{B\}} y_{m,k}\,|m\rangle$ into Eq. (11). To prove Eq. (14), we need to show that one finds the largest possible value for the probability when $x_{n,k}$ and $y_{m,k}$ are zero for all n,m except for one value of n and m. For convenience, we define $\bar{\mathcal{T}}_{nm} \equiv \langle n| \prod_{i=p-k}^{p-1} T_i\, |a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j\, |m\rangle$ and $\mathcal{T}_{nm} \equiv \langle n| \prod_{i=p-k}^{p+k} T_i\, |m\rangle$ for $n,m \in B$. The probability can then be stated as

$$P_p(a,b) = \frac{\sum_{n\in B}\sum_{m\in B} x_{n,k}\bar{\mathcal{T}}_{nm}y_{m,k}}{\sum_{n\in B}\sum_{m\in B} x_{n,k}\mathcal{T}_{nm}y_{m,k}}. \tag{16}$$

Without loss of generality, we assume that

$$\frac{\bar{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}} = \min\left(\frac{\bar{\mathcal{T}}_{AA}}{\mathcal{T}_{AA}}, \frac{\bar{\mathcal{T}}_{AT}}{\mathcal{T}_{AT}}, \ldots, \frac{\bar{\mathcal{T}}_{GG}}{\mathcal{T}_{GG}}\right) \tag{17}$$

holds for some $i,j \in B$, which does not have to be unique. We evaluate the sign of the derivative of $P_p(a,b)$ with respect to $x_{i,k}y_{j,k}$:

$$\frac{\partial P_p(a,b)}{\partial(x_{i,k}y_{j,k})} = \frac{\sum_{n\in B}\sum_{m\in B} x_{m,k}\mathcal{T}_{nm}y_{n,k}\mathcal{T}_{ij}\left(\frac{\bar{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}} - \frac{\bar{\mathcal{T}}_{nm}}{\mathcal{T}_{nm}}\right)}{\left(\sum_{n\in B}\sum_{m\in B} x_{n,k}\mathcal{T}_{nm}y_{m,k}\right)^2} \leqslant 0. \tag{18}$$

The less-than-or-equal-to sign follows from the fact that $\bar{\mathcal{T}}_{nm}$, $\mathcal{T}_{nm}$, $x_{m,k}$, and $y_{n,k}$ are nonnegative for all n,m and from Eq. (17). Because the derivative is nonpositive, the probability is nonincreasing as a function of $x_{i,k}y_{j,k}$, thus a maximum can be found when $x_{i,k}y_{j,k}$ is minimal, i.e., in the limit of $x_{i,k}y_{j,k} \to 0$. Now we have "eliminated" one combination of variables: $x_{i,k}y_{j,k}$ and the corresponding ratio $\frac{\bar{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}}$ from Eq. (16) [this can be checked by inserting $x_{i,k}y_{j,k} = 0$ into Eq. (16)]. This process can be performed iteratively until only one combination of variables is left. Now we assume, again without loss of generality, that this final combination is $x_{r,k}y_{s,k}$ for some $r,s \in B$. The probability is now independent of these variables:

$$P^{(k)}_{\max,p}(a,b) = \frac{x_{r,k}\bar{\mathcal{T}}_{rs}y_{s,k}}{x_{r,k}\mathcal{T}_{rs}y_{s,k}} = \frac{\bar{\mathcal{T}}_{rs}}{\mathcal{T}_{rs}}. \tag{19}$$

This does not mean we can freely assign a number to $x_{r,k}y_{s,k}$. Recall that $|x_k\rangle$ and $|y_k\rangle$ are unit vectors. Since $x_{m,k}y_{n,k} \to 0$ for all $m \neq r, n \neq s$, it is required that $x_{r,k} \to 1$ and $y_{s,k} \to 1$, and $x_{m,k} \to 0$ and $y_{n,k} \to 0$ for all $m \neq r, n \neq s$. Therefore, we find the kth upper bound of the probability when $|x_k\rangle$ and $|y_k\rangle$ are pure states from the basis B, as we stated in Eq. (14).

For the zeroth-order bounds, where no neighbors are taken into account, a similar result holds. This can be obtained in the same manner as Eqs. (14) and (15), therefore no proof is provided. These bounds are given by

$$P^{(0)}_{\max,p}(a,b) = \max_{|x_0^*\rangle,|y_0^*\rangle\in B} \frac{\langle x_0^*|a\rangle \langle a|T_p|b\rangle \langle b|y_0^*\rangle}{\langle x_0^*|\, T_p\, |y_0^*\rangle} = 1 \tag{20}$$

and

$$P^{(0)}_{\min,p}(a,b) = \min_{|x_0^*\rangle,|y_0^*\rangle\in B} \frac{\langle x_k^*|a\rangle \langle a|T_p|b\rangle \langle b|y_k^*\rangle}{\langle x_k^*|\, T_p\, |y_k^*\rangle} = 0. \tag{21}$$

These bounds are 1 and 0 because $\min_{|x_0^*\rangle,|y_0^*\rangle\in B} \langle x_k^*|a\rangle = 0$ and $\max_{|x_0^*\rangle,|y_0^*\rangle\in B} \langle x_k^*|a\rangle = 1$. This shows that one needs to take at least one neighbor into account to obtain non-trivial results.

Furthermore, one can show that the bounds on the probability get sharper at higher order, i.e., increasing k:

$$P^{(k)}_{\max,p}(a,b) \geqslant P^{(k+1)}_{\max,p}(a,b) \geqslant P_p(a,b), \tag{22}$$

$$P_p(a,b) \geqslant P^{(k+1)}_{\min,p}(a,b) \geqslant P^{(k)}_{\min,p}(a,b). \tag{23}$$

An intuitive explanation is that adding the information on more and more bases to our calculation should lead to sharper bounds on the probability. It is straightforward to prove. Consider Eq. (14) for the $(l+1)$th order upper bound (such that $k = l+1$) with its two maximizing pure states $\langle x_{l+1}^*| = \langle n|$ and $|y_{l+1}^*\rangle = |m\rangle$. This bound is smaller or equal to the upper bound for $k = l$ for the following reason: one finds exactly the same expression as above if one inserts into Eq. (14)

the states $\langle x_l^*| = \langle n|\, T_{p-l}$ and $|y_l^*\rangle = T_{p+l}\,|m\rangle$. We find the lth-order upper bound if $\langle x_l^*|$ and $|y_l^*\rangle$ are pure states. Using other values, i.e., $\langle n|\, T_{p-l}$ and $T_{p+l}\,|n\rangle$, can only result in probabilities equal to or lower than this lth-order maximum. Therefore, the $(l+1)$th-order upper bound cannot be higher than the lth-order upper bound. The same reasoning holds for the lower bounds.

V. THE AVERAGE NEIGHBOR ENERGY APPROXIMATION

The method of finding bounds on the probability in the previous section allows us to quantify how much nucleotides a given number of steps away from a given position can affect the dinucleotide preferences at that position. By comparing the results of the bounds on the probability at different orders, we will show in the next section that long-range interactions are unimportant. On the other hand, we will also find that a purely local picture where the probability of a dinucleotide is determined only by its own elastic properties is not predictive. Even the first-order bounds on the probability that take the nearest neighbors into account are too far apart to confine sufficiently the position-dependent variations of the probabilities. It is the difference between the second-order upper and lower bounds that is much smaller than these variations. This demonstrates that only a limited number of neighbors determines the nucleosome positioning rules.

Here we further expand on this idea by showing that, for our model at room temperature, the probability of finding a dinucleotide at a given position p mostly depends on only two parameters: the energy of the dinucleotide at position p, and the sum of the averages of the energies of their possible neighbors at positions $p+1$ and $p-1$. Looking at these two parameters allows us to interpret the base pair step preferences in our nucleosome model. We will call the corresponding approximation the *average neighbor energy approximation*. This approximation will be used later, not to calculate probabilities but to give a physical interpretation of our findings from the exact treatment.

Since the first-order bounds on the probability are not good enough to confine the dinucleotide preferences, it may seem counter-intuitive to use only the nearest neighbors. This can be explained by the fact that the upper and lower bounds on the probability take extreme scenarios into account where the neglected nucleotides have the highest possible impact on the probability, whereas the actual system does not behave as extremely.

We introduce now the approximated probability that we indicate by a superscript (e) as follows: $P_p^{(e)}(a,b)$. Using the notation

$$\langle f(x)\rangle_x = \frac{1}{4}\sum_{x\in\{A,T,C,G\}} f(x), \tag{24}$$

$$\langle g(x,y)\rangle_{x,y} = \frac{1}{16}\sum_{x,y\in\{A,T,C,G\}} g(x,y), \tag{25}$$

we define the average neighbor energy approximation of the probability as

$$P_p^{(e)}(a,b) \equiv \frac{\exp\left[-\beta\langle E_{p-1}(n_{p-1},a)\rangle_{n_{p-1}}\right]\exp[-\beta E_p(a,b)]\exp\left[-\beta\langle E_{p+1}(b,n_{p+2})\rangle_{n_{p+2}}\right]}{\sum_{n_p,n_{p+1}}\exp\left[-\beta\langle E_{p-1}(n_{p-1},n_p)\rangle_{n_{p-1}}\right]\exp[-\beta E_p(n_p,n_{p+1})]\exp\left[-\beta\langle E_{p+1}(n_{p+1},n_{p+2})\rangle_{n_{p+2}}\right]}. \tag{26}$$

Note that this approximation depends on $E_p(a,b)$, the energy of the dinucleotide step ab at position p, and on $\langle E_{p-1}(n_{p-1},a)\rangle_{n_{p-1}} + \langle E_{p+1}(b,n_{p+2})\rangle_{n_{p+2}}$, an average of the energies of possible nearest neighbors of ab. We have calculated the error introduced by using the average neighbor energy approximation and found it not to be larger than 3.5 percent at any position for any dinucleotide; see Appendix B.

Next we provide an explanation why this approximation works so well for our model. Our strategy is to bring the approximated probability, Eq. (26), and the full probability, Eq. (8), into a similar form. Comparison of the two similar expressions allows then to explain the nature of this approximation that is otherwise not straightforward to see. We start by rewriting the approximation such that it resembles more the exact probability [Eq. (8)]:

$$P_p^{(e)}(a,b) = \frac{\sum_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}: \\ n_p=a,n_{p+1}=b}} \exp\left[-\beta\sum_{i=1}^{L-1}\langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\right]}{\sum_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}}} \exp\left[-\beta\sum_{i=1}^{L-1}\langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\right]}. \tag{27}$$

The hats above n_{p-1} and n_{p+2} denote that these variables are not to be summed over. The nominator factorizes in three terms: (1) a sum of terms where each term depends explicitly on at least one of the variables n_1 to n_{p-2}, (2) a sum of terms where each term depends explicitly on at least one of the variables n_{p+3} to n_L, and terms independent of those variables. The first and second factors cancel out with the exact same expressions in the denominator leading back to Eq. (26).

We will now make the exact probability [Eq. (8)] look more like the approximation in the form of Eq. (27). By substituting the function $C_p(i,j)$, defined as

$$C_p(m,o) \equiv \frac{\frac{1}{4}\sum_n \exp[-\beta E_{p+1}(m,n) - \beta E_{p+2}(n,o)]}{\exp[-\beta\langle E_{p+1}(m,n) + E_{p+2}(n,o)\rangle_n]}, \tag{28}$$

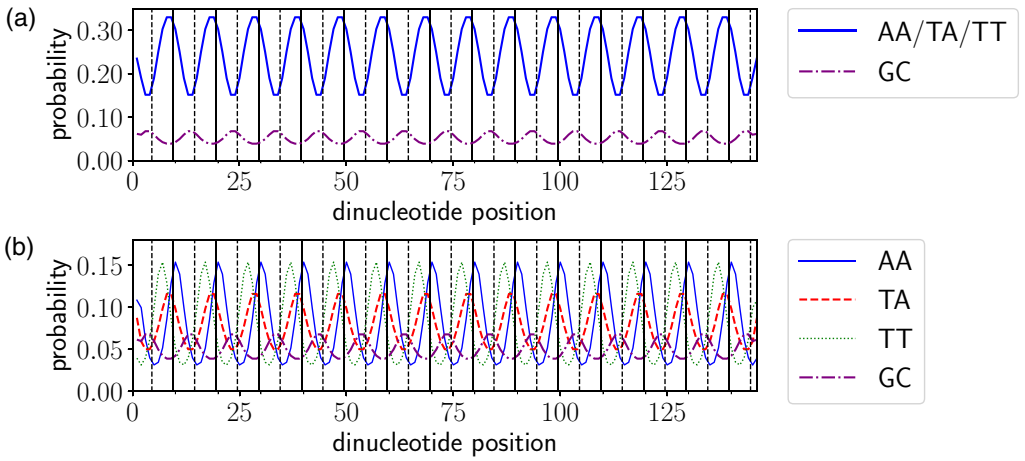

FIG. 3. (a) The probability to find AA, TA, or TT, and the probability to encounter GC at the full range of dinucleotide positions is shown. The solid and dashed vertical lines indicate minor and major groove bending sites (maximum negative and positive roll, respectively). The probabilities are in qualitative agreement with the well-known nucleosome positioning rules [5]. (b) Same as (a) but showing all four dinucleotide probabilities individually.

into Eq. (8) twice, we obtain

$$P_p(a,b) = \frac{\sum_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}: \\ n_p=a,n_{p+1}=b}} \exp\left[-\beta \sum_{i=1}^{L-1} \langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\right] C_{p-2}(n_{p-2},a) C_{p+1}(b,n_{p+3})}{\sum_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}}} \exp\left[-\beta \sum_{i=1}^{L-1} \langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\right] C_{p-2}(n_{p-2},n_p) C_{p+1}(n_{p+1},n_{p+3})}, \tag{29}$$

which is indeed very similar to Eq. (27), apart from the functions C_p. The approximation $P_p(a,b) \approx P_p^{(e)}(a,b)$ is *exact* if $C_{p-2}(n_{p-2},a)$ does not depend on a, and if $C_{p+1}(b,n_{p+3})$ does not depend on b. The approximation works well if these functions show only a weak dependence on a and b. It turns out that (for our model) the latter is true; see Appendix B for details.

The approximation gets worse with decreasing temperature. We can see this by performing a Taylor expansion in β of $C_p(m,o)$:

$$C_p(m,o) \approx 1 + \tfrac{1}{2}\beta^2 \langle [E_{p+1}(m,n) + E_{p+2}(n,o) - \langle E_{p+1}(m,n') + E_{p+2}(n',o)\rangle_{n'}]^2 \rangle_n. \tag{30}$$

Only the higher-order terms depend on m and o; these terms become increasingly important with decreasing temperature (increasing β). At room temperature the higher-order terms are not important as the various dinucleotide energies lie close to each other compared to the thermal energy. As a result, the exponential of the averages is a good approximation to the average of the exponentials and $C_p(m,o)$ shows only a weak dependence on m and o.

VI. RESULTS

A. The dinucleotide probability

Using the transfer matrix approach we calculate here the preferences of dinucleotide steps along our nucleosome model. We focus in this section on the "nucleosome positioning code" [5], which claims that high affinity sequences are characterized by the proper positioning of four dinucleotides: the probability of finding GC steps (a G followed by a C) peaks at positions where the major groove faces the protein cylinder (every 10th bp), whereas AA, TA, and TT are all in phase and have their peaks in between where the minor groove faces the cylinder.

Figure 3(a) shows the combined probability to encounter AA, TA, TT along the nucleosome and, separately, that of GC calculated using transfer matrices, Eq. (10). Both signals are 10-bp periodic in accordance with the experimental observation. Moreover, the two probabilities show the right phases: the GC signal has a peak in the center (at the nucleosomal dyad), which corresponds to a place where the major groove faces inward and the same holds for all other peaks of GC. The combined signal of AA, TA, and TT is out of phase with the GC signal and peaks at the places where the minor groove is compressed. In short, our model reproduces qualitatively the well-known nucleosome positioning rules.

More details provides Fig. 3(b), where all four dinucleotides are plotted separately. The figure shows that indeed AA, TA, and TT are all in phase with each other. Strictly speaking, however, TT peaks slightly before, and AA slightly after maxima in TA. This should be expected since TA bridges TT and AA steps. This leads to the question whether TA steps peak at the minor groove roll position because they just "happen" to bridge TT and AA steps or whether there is an intrinsic advantage for TA to peak at this position. As we explain further below, our model allows to give precise answers to such kind of questions.

Finally, we mention that the 10-bp periodicities of the signals displayed in Fig. 3 are, of course, simply a consequence of 10-bp periodicity in our model, see Eqs. (6) and (7).

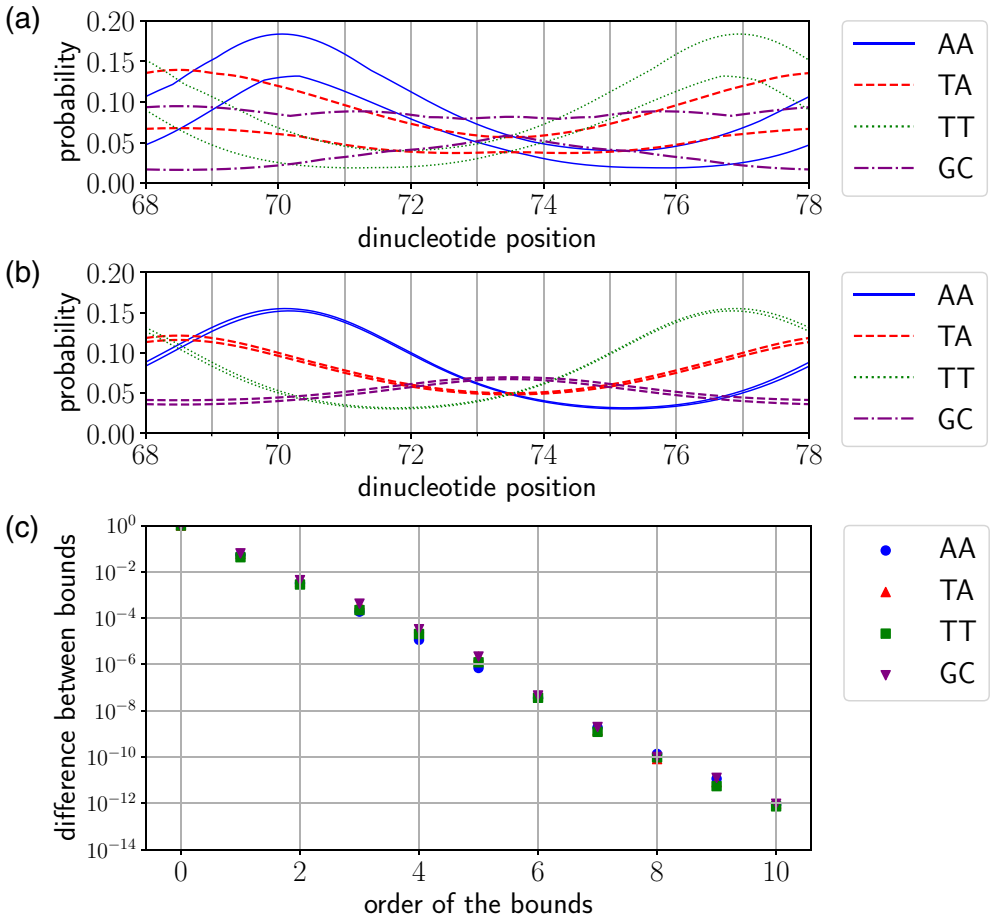

FIG. 4. (a, b) Upper and lower bounds on the probabilities to have the dinucleotide AA, TA, TT, or GC at several dinucleotide positions on a nucleosome. Specifically, (a) depicts the first-order bounds and (b) the second-order bounds of the probability. The upper and lower bounds of the same dinucleotide have the same color (line style). (c) Difference between the upper and lower bounds of the probabilities to encounter AA, TA, TT, and GC at position 79 at increasing order. The difference, and thereby the effect of the neighbors k steps away from the dinucleotide of interest, decreases exponentially as the order k increases.

However, very close to the termini of the nucleosomal DNA the probabilities deviate from this periodic signal. The short range of this boundary effect suggests that the probability of finding a dinucleotide is not affected much by far-away nucleotides. This can be demonstrated (and quantified) using the upper and lower bounds of the probability to which we now turn.

B. The bounds on the probability

Figures 4(a) and 4(b) show the first- and second-order bounds on the probability to encounter AA, TA, TT, or GC at dinucleotide position 58 through 88 using Eqs. (14) and (15). Note that the energy as defined by Eqs. (5) to (7) allows also for noninteger bp positions. Even though these noninteger positions have no physical meaning due to the discrete nature of bp sequences, we plot them here as well, as they are a useful guide for the eye. Strictly speaking, however, only the integer positions are physically meaningful.

By using only one neighbor to the left and right (first-order bounds) the bounds indicate already the qualitative behavior of the system for some of the dinucleotides (AA, TA, and TT but not GC); see Fig. 4(a). Accounting for two neighbors on each side (second-order bounds) provides already an excellent estimate of the dinucleotide probabilities as the differences between the upper and lower bounds are much smaller than the observed overall variations in the probabilities at different positions; see Fig. 4(b).

The effect of far-away bases can be characterized by one number as follows. The difference between the upper and lower bounds decays exponentially with increasing order of the bounds (i.e., increasing the number of neighbors involved); see Fig. 4(c). This allows us to define an effective order κ, similar to a correlation length:

$$P^{(k)}_{\max,p}(a,b) - P^{(k)}_{\min,p}(a,b) \approx e^{-k/\kappa}. \tag{31}$$

The value of κ is found to be approximately equal to 1.2. This shows that increasing the order of the bounds has a huge effect around $k = 1$. It also explains why only the probabilities very close to the edges of the nucleosome are not following the 10-bp periodicity. Probabilities at positions far away from the boundaries are (exponentially) less influenced by the edge and will not "feel" its presence.

While the results shown here are obtained at room temperature, the bounds remain an effective method at all possible temperatures; see Appendix. C.

TA										
Position	69	70	71	72	73	74	75	76	77	78
Probability	0.116	0.098	0.077	0.059	0.050	0.050	0.059	0.077	0.098	0.116
Dinucleotide Energy [k_BT_r]	0.703	0.676	0.535	0.273	0.050	0.050	0.273	0.535	0.676	0.703
-Roll	0.664	0.409	0.126	0.005	0.011	0.011	0.005	0.126	0.409	0.664
-Tilt	0.039	0.268	0.409	0.268	0.039	0.039	0.268	0.409	0.268	0.039
Average Neighbour Energy [k_BT_r]	1.573	1.566	1.472	1.275	1.102	1.102	1.275	1.472	1.566	1.573
-Roll	1.250	0.914	0.517	0.305	0.265	0.265	0.305	0.517	0.914	1.250
-Tilt	0.323	0.652	0.955	0.971	0.837	0.837	0.971	0.955	0.652	0.323

FIG. 5. The probability of dinucleotide TA, its energy, and the average of the energies of its possible neighbors are shown for 10 different positions along the nucleosome, i.e., for one full DNA helical repeat. The numbers give absolute values, whereas the colors indicate how the corresponding value of the TA step compares with the values of all other possible dinucleotides at the same position. Yellow (light gray) colors represent relatively favorable values, red (dark gray) indicates unfavorable values. The probability follows mainly from a "mixing" of the colors of the corresponding dinucleotide energies and average neighbor energies. The table also provides subdivisions of the TA energies into roll and tilt contributions.

C. Explaining the dinucleotide positioning rules

So far we have presented the probability distributions of a few key dinucleotides along the nucleosome model and found good agreement with the general positioning rules. We also demonstrated, by looking at upper and lower bounds of various orders, that long-range interactions are not important, but nearby neighbors matter. This is one of the reasons why the probabilities are well captured by the average neighbor approximation. Using this approximation we explain in the following how the nucleosome positioning rules in our model emerge from the elasticities and intrinsic shapes of the various dinucleotides.

Fig. 3 shows that the probability [calculated using Eq. (10)] of TA dinucleotides peaks at positions of maximal negative roll (e.g., at positions 78 and 79), whereas the one of GC dinucleotides peaks at positions of maximal positive roll (e.g., at positions 73 and 74). Moreover, TT peaks at positions of maximal positive tilt (such as position 77), while AA peaks at maximal negative tilt (e.g., at position 70). We first discuss the rules from a purely local perspective, i.e., just considering the elasticity and geometry of the dinucleotide under consideration. From this perspective only some of these findings make sense.

1. *A local perspective on the dinucleotide probability fails*

Table I presents all the parameters that were used in our model. Inspecting this table one finds that TT and AA have large positive and negative intrinsic tilt, respectively, which is consistent with their preferred positions. In contrast to that, TA has a large positive intrinsic roll, which makes positions of maximal negative roll like 78 and 79 highly unfavorable, even though this is where this step peaks. Even more surprising are the peaks for GC at positive roll positions as this is the dinucleotide step with the smallest intrinsic roll among all dinucleotide steps; see Table I.

These findings are consistent with what we have learned from the bounds on the probabilities: zeroth-order bounds, which correspond to a purely local perspective, are not useful at all to obtain estimates of the probabilities, while first-order bounds, which include the energies of the nearest neighbors, suffice for some of the dinucleotides to have rather good estimates of the probability; see Fig. 4(a).

2. *Neighboring steps are equally important*

The effect of the neighbors can be best understood using the average neighbor energy approximation; see Eq. (26). Since this is an excellent approximation, see Appendix B, the only terms important for the behavior of the probability are the *energy of the dinucleotide itself*, and the *average of the energies of its possible neighbors*. To understand the nucleosome positioning rules we need thus to compare the energy of the dinucleotide ab with the energies of the 15 other dinucleotides *and* the average of the energies of all possible neighbors of ab with the averages of the energies of all possible neighbors of the 15 other dinucleotides.

Such information can be best presented in tabular form. Figure 5 provides the relevant information for the TA dinucleotide. It presents (as numbers) the probability [obtained

GC										
Position	69	70	71	72	73	74	75	76	77	78
Probability	0.039	0.042	0.050	0.060	0.068	0.068	0.060	0.050	0.042	0.039
Dinucleotide Energy [k_BT_r]	0.546	0.644	0.681	0.571	0.428	0.428	0.571	0.681	0.644	0.546
-Roll	0.481	0.199	0.002	0.126	0.363	0.363	0.126	0.002	0.199	0.481
-Tilt	0.065	0.445	0.679	0.445	0.065	0.065	0.445	0.679	0.445	0.065
Average Neighbour Energy [k_BT_r]	2.816	2.429	1.723	0.921	0.376	0.376	0.921	1.723	2.429	2.816
-Roll	2.167	1.648	0.926	0.352	0.070	0.070	0.352	0.926	1.648	2.167
-Tilt	0.649	0.781	0.797	0.569	0.306	0.306	0.569	0.797	0.781	0.649

FIG. 6. Same as Fig. 5 but for GC.

using Eq. (10)] to find this dinucleotide, its energy [Eq. (5)] and the average of the energies of its possible neighbors [see Eq. (26)] for a 10-bp stretch in one table (and some further information that we discuss further below). More relevant, however, are the colors assigned to each box as they indicate how these numbers compare to the values of all other possible dinucleotides. If the color is yellow (light gray), the value is relatively favorable compared to the ones of other dinucleotides *at the same dinucleotide position* (i.e., the probability is relatively high, while the energy cost is relatively low). red (dark gray) denotes unfavorable values, while orange (gray) indicates that this value is average.

First consider in Fig. 5 row "Probability": At positions 69 and 78, both associated with *negative* roll and zero tilt, TA is favorable, as we have seen in Fig 3. Next consider row "Dinucleotide energy": The dinucleotide energy of TA goes against this preference having the lowest values at positions 73 and 74, and its highest at positions 69 and 78, both in absolute values (numbers) and relative values (colors). Next turn to row "Average neighbor energy": the absolute values (numbers)

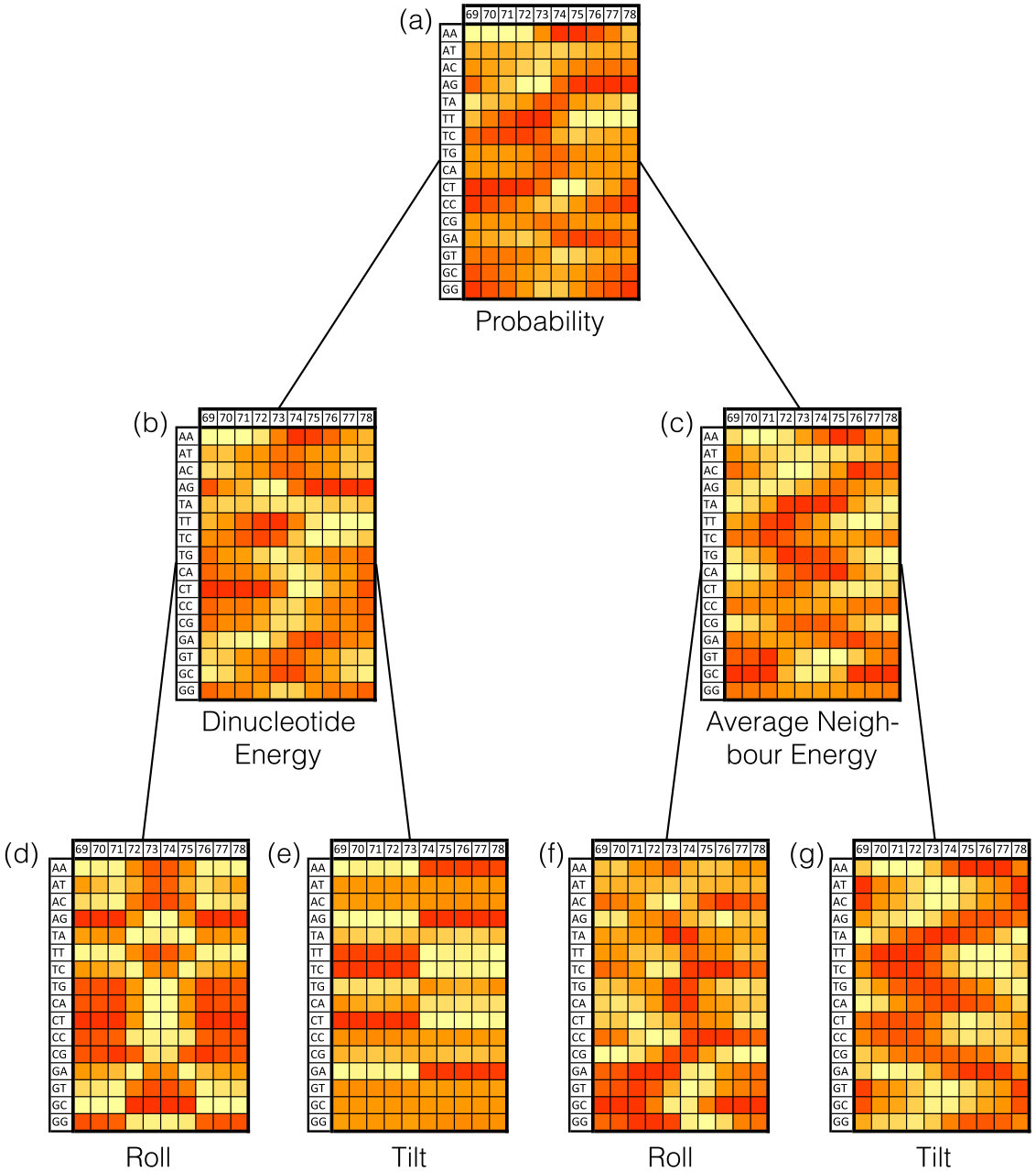

FIG. 7. (a) Probability, (b) dinucleotide energy, and (c) average neighbor energy of all 16 dinucleotides for one DNA helical repeat. Yellow (light gray) denotes high probability and low energy, red (dark gray) low probability and high energy (relative to all other dinucleotides at the corresponding location). In addition, provided are subdivisions of dinucleotide energies into (d) roll and (e) tilt, and of neighbor energies into (f) roll and (g) tilt. The colors representing the probabilities can be seen as a "sum" of the colors of the dinucleotide energies and the average neighbor energies. The colors corresponding to the dinucleotide energies are the "sum" of the colors for roll and tilt energies. The same holds for the neighbor energies.

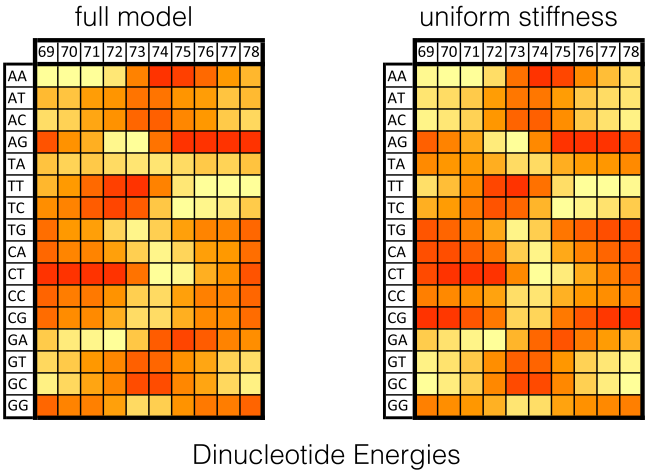

FIG. 8. Original dinucleotide energy costs [left; same as Fig. 7(b)], and energy cost with all stiffnesses set to 1 (right) are shown side by side. The strong similarity between the two tables reveals that stiffnesses play only a minor role for the dinucleotide positional preferences.

have their lowest values at 73 and 74 but the relative values (colors) strongly prefer the opposite. Therefore, what causes the TA preference for negative roll positions is the average energy of the possible neighbors relative to the average energy of the forbidden neighbors.

Now we turn to the other rows in Fig. 5. These extra rows provide a subdivision of the dinucleotide energies and the average neighbor energies into roll and tilt components. Inspecting these four extra rows reveals that the main cause for the TA preference for positions 69 and 78 lies in the average tilt contribution of the possible neighbor steps. This overrides TA's own preference (relative and absolute) for positive roll.

The same analysis as the one on TA can be performed on GC by inspecting Fig. 6. This is another nontrivial dinucleotide in the sense that its behavior is heavily affected by its possible neighbors. Positions 73 and 74, associated with large positive roll and bending toward the major groove, lead to a high dinucleotide energy of GC (compared to other steps), which has a low (positive) intrinsic roll. However, the possible neighbors cause the probability of encountering GC to be highest at these positions and lowest at positions 69 and 78.

3. *The complete picture*

In Fig. 7 tables are shown that present the probabilities and relative energies for all 16 dinucleotides (again in color code). There are seven tables corresponding to the seven rows in Figs. 5 and 6. Using these tables one can analyze preferences for each dinucleotide step individually, just as explained for TA and GC above. Moreover, for cases where the average neighbor energies dominate the positional preferences of dinucleotides (like for TA and GC), these tables allow to look up which of the possible neighbors of a given dinucleotide are favorable.

As an example, we consider again the dinucleotide TA. In Fig. 7(a) we see that the probability of TA peaks at positions 69 and 78, which is not TA's intrinsic preference, Fig. 7(b), but that of its neighbors on average, Fig. 7(c). We need now to inspect the intrinsic preferences of all the possible neighbors. At position 70, three of the four possible neighbors (dinucleotides starting with an A) are favorable, namely AA, AT, and AC; see Fig. 7(b). Due to symmetry TT, AT, and GT are favorable at position 77, see also Fig. 7(b). Further details are revealed by Figs. 7(d) and 7(e) that present the roll and tilt contributions to the dinucleotide energies. It shows that AA at 70, and TT at 77 are favorable due to both their roll and tilt preferences, whereas the other favorable steps, AT and AC at 70, AT and GT at 77, prefer those positions due to roll alone. Inspecting the contributions of roll and tilt to the average neighbor energies for TA at positions 69 and 78, Figs. 7(f) and 7(g), one learns that both degrees of freedom matter but tilt is the dominant factor. This reflects the very strong tilt preference for AA and TT but also the fact that the only unfavorable neighbors (AG at 70, CT at 77) are unfavorable due to roll whereas the tilt contributions are favorable.

Note that these considerations also explain preferred occurrences of larger motives, like, e.g., TTAA centered around negative roll positions. In addition, similar lines of arguments

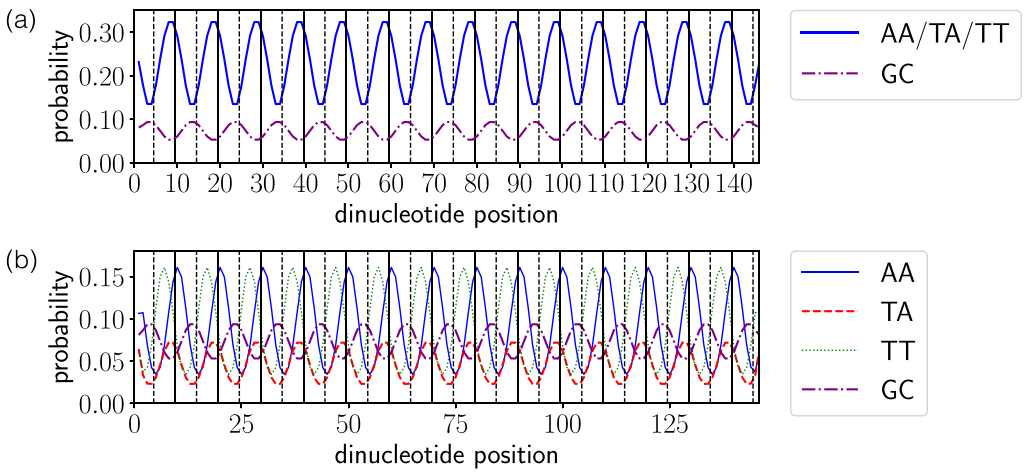

FIG. 9. Same as Fig. 3 but with including the cross-terms [see Eq. (A3)]. The positioning rules (i.e., the relative behavior of the dinucleotide probabilities at different position) has stayed the same.

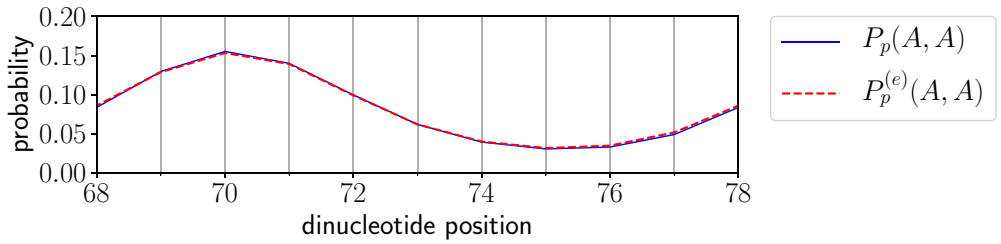

FIG. 10. The exact probability and its average neighbor energy approximation to find AA steps at all dinucleotide positions. The approximation introduces an error that is nowhere larger than 3.5%.

can be used to understand why TA is unfavored at high roll positions, like 73 and 74, or the preferences of any other dinucleotide for that matter.

4. Shape is more important than stiffness

The roll and tilt terms of the energy can be subdivided even further. As can been seen from Eqs. (5) to (7) the sequence dependencies enter the roll and tilt energies both through the intrinsic geometries and through the stiffnesses related to these two degrees of freedom. We show now that the stiffnesses are not very important to the behavior of our system. In Fig. 8 we compare two tables for dinucleotide energies: the original table on the left [identical to Fig. 7(b)] and on the right a table that is produced when we set all stiffnesses of roll and tilt to the same small value, namely to 1. Even though the specific value of the stiffness affects strongly the absolute values of the dinucleotide energies (not shown), it does hardly affect the relative values of the energies (color code). This reveals that, at least in our simplified model, the sequence preferences are largely governed by the intrinsic roll and tilt (and not the stiffnesses) of the dinucleotides. Note that this observation is consistent with the findings reported in Ref. [30], where molecular dynamics simulations performed on rather detailed nucleosome models revealed that nucleosome affinity is dominated by the shape of the wrapped DNA.

VII. CONCLUSION

In this work we obtained a detailed understanding of the physics behind nucleosome sequence preferences as they arise from the sequence-dependent geometry and elasticity of the DNA double helix. Our strategy was to build a model that is simple enough so that it can be solved analytically and complex enough to reproduce the experimentally known nucleosome positioning rules. This was achieved by forcing a coarse-grained DNA model (the rigid base pair model) along a circular path and accounting for the sequence-dependent mechanics of only the most important degrees of freedom (roll and tilt). With the help of transfer matrices, we were able to calculate the dinucleotide probabilities along our nucleosome model. These reproduce, at least qualitatively, the rules found when nucleosome position themselves freely along a long stretch of DNA (e.g., the yeast genome [6]).

However, to really understand the dinucleotide rules in detail, exactly solving the model (or simulating a more detailed version of it [10]) is not sufficient, as this system behaves rather complex. For instance, of the four "important" dinucleotides only two (AA and TT) prefer locations that correspond to their own intrinsic preferences whereas the other two (TA and GC) peak at their most unfavorable locations. To solve this puzzle, we first introduced an approximation that, by taking a limited number of neighbors around a given dinucleotide into account, provides upper and lower bounds to its probability distribution. From this we learned that the nearest neighbors influence strongly the preferences of a given dinucleotide whereas the influence of nucleotides further away is small, decreasing exponentially.

With this information at hand, we finally introduced an approximation tailored for interpreting the dinucleotide preferences. According to this average neighbor energy approximation dinucleotide preferences are dominated by two contributions: the intrinsic energy cost to place a given dinucleotide at a given position and the average energy of the possible neighbors before and after that given dinucleotide. This is an excellent approximation and allows to explain all the dinucleotide preferences found in our model. Depending on the dinucleotide at hand, a given dinucleotide is found preferentially at certain positions mainly due to its own preferences (e.g., AA and TT) or due to bringing in "good" neighbors (e.g., TA and GC).

Knowing the dinucleotide preferences of nucleosomes allows genome wide calculations of nucleosome positioning [31]. Therefore, understanding how dinucleotide preferences along nucleosomes emerge from the sequence dependent DNA mechanics, means ultimately to understand the physical underpinnings of biological processes at much larger scales as the depletion of nucleosomes at gene start sites in yeast [6,7] or the retention of nucleosomes in human sperm cells [7,8].

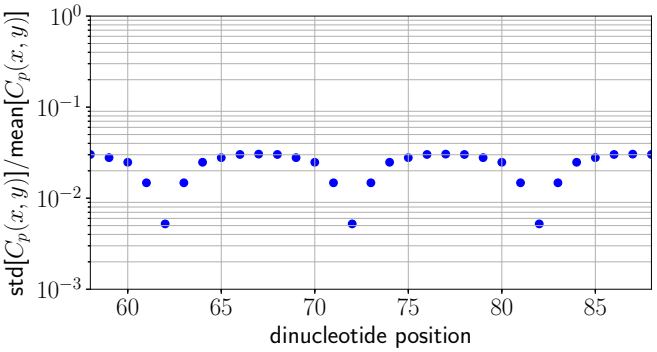

FIG. 11. The standard deviation [Eq. (B1)] divided by the mean [Eq. (B2)] of $C_p(x,y)$. As this ratio is very small at all positions p the function $C_p(x,y)$ is nearly constant, explaining the high accuracy of the average neighbor energy approximation.

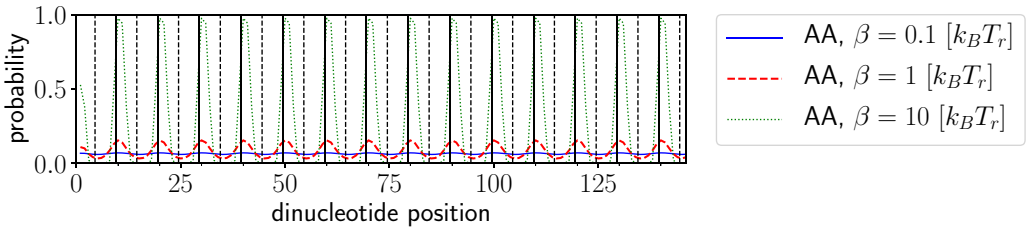

FIG. 12. The probability to obtain AA at all dinucleotide positions at several temperatures.

ACKNOWLEDGMENTS

We thank Marco Tompitak for discussions. This work is part of the Delta ITP consortium, a program of the Netherlands Organisation for Scientific Research (NWO) that is funded by the Dutch Ministry of Education, Culture, and Science (OCW).

APPENDIX A: ENERGY CONTRIBUTIONS OF TWIST AND CROSS TERMS

In Sec. II, we define the energy of a dinucleotide as the sum of the energy of roll and tilt. Keeping in mind the observation that the basic nucleosome positioning rules can be rationalized by discussing energy costs involved in the roll and tilt degrees of freedom [10], as well as our goal to reduce our model to its bare essentials, we chose to neglect the contribution of twist and of the cross terms between roll, tilt, and twist. In this appendix we will show that including the twist and cross terms does not change the qualitative agreement of our model with well-known positioning rules (see Fig. 3).

We start by defining the energy of twist and the cross-terms:

$$E^{\text{twist}}(a,b) \equiv \tfrac{1}{2} Q^{\text{twist}}(a,b)[q^{\text{twist}} - \bar{q}^{\text{twist}}(a,b)]^2 \qquad \text{(A1)}$$

and

$$E_p^{\text{cross}}(a,b) \equiv \sum_{\substack{i,j\in\{\text{roll,tilt,twist}\}\\ i\neq j}} \frac{1}{2} Q^{i,j}(a,b) \times \left[q_p^i - \bar{q}^i(a,b)\right]\left[q_p^j - \bar{q}^j(a,b)\right]. \qquad \text{(A2)}$$

The bp-step-dependent stiffnesses are now given by $Q^i(a,b)$, $i \in \{\text{roll,tilt,twist}\}$, and the corresponding intrinsic values by $\bar{q}^i(a,b)$, $i \in \{\text{roll,tilt,twist}\}$. The cross terms depend on the cross stiffnesses $Q^{i,j}(a,b)$, $i,j \in \{\text{roll,tilt,twist}\}$, $i \neq j$. (Note that, because of the constant twist, the energy associated with twist does not depend on position p but only on the dinucleotide step.) For the twist and cross terms, too, the hybrid parametrization [26] is used. We can redefine our energy as

$$E_p(a,b) = E_p^{\text{roll}}(a,b) + E_p^{\text{tilt}}(a,b) + E^{\text{twist}}(a,b) + E_p^{\text{cross}}(a,b). \qquad \text{(A3)}$$

Figure 9 was created using this redefined energy. We see that the relative behavior of the dinucleotide probabilities at different positions is the same as without the cross terms, see Fig. 3, but that the overall height of some of the probabilities have shifted. A more detailed analysis reveals that this is mainly caused by the twist contribution as well as the roll-twist coupling, whereas the other cross-terms are unimportant.

APPENDIX B: VALIDITY OF THE AVERAGE NEIGHBOR ENERGY APPROXIMATION

The average neighbor energy approximation of the probability works extremely well. We checked it for all dinucleotides and found that the largest error of this approximation occurs for the probability distribution of dinucleotide AA. Figure 10 depicts both the exact probability and its approximation for this dinucleotide. The difference between the values is always smaller than 3.5%.

To understand why this error is so small, one needs to consider the function $C_p(x,y)$, defined in Eq. (28). The average neighbor energy approximation is exact if this function is a constant (i.e., independent of x and y for each p). The approximation works well if the function is almost constant. That this is true is best seen by inspecting the standard deviation of $C_p(x,y)$, divided by its mean, and checking whether this quantity is much smaller than one. Here the

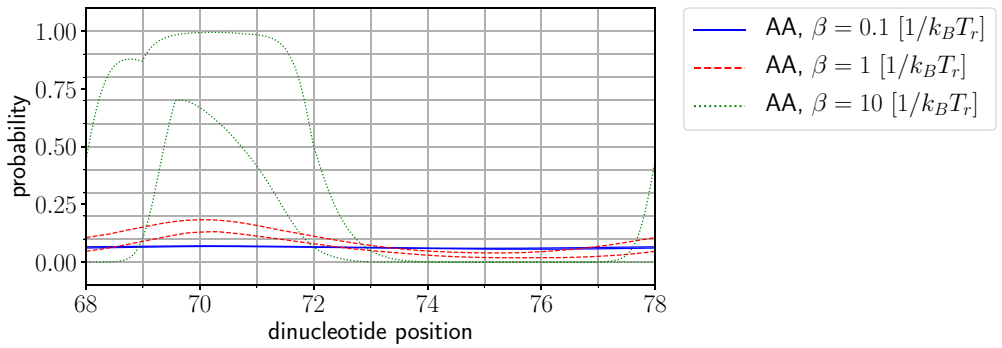

FIG. 13. The first-order bounds of the probability of encountering an AA step are shown at five different temperatures. At low temperatures, the bounds become significantly far apart from each other and only provide a qualitative description of the behavior of the probability as a function of position.

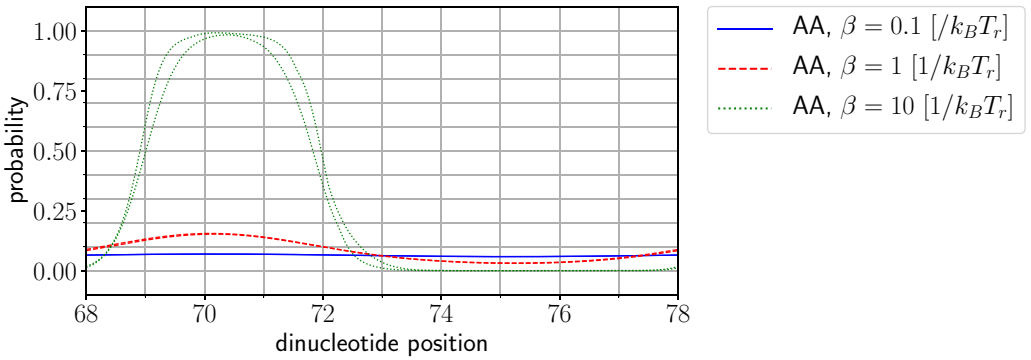

FIG. 14. The second-order bounds of the probability of encountering an AA step are shown at five different temperatures. At all temperatures the method provides a quantitative description of the probability, clearly outperforming the second-order bounds.

standard deviation and mean are defined as

$$\mathrm{std}[C_p] \equiv \sqrt{\langle \{C_p(x,y) - \mathrm{mean}[C_p]\}^2 \rangle_{x,y}}, \tag{B1}$$

with

$$\mathrm{mean}[C_p] \equiv \langle C_p(x,y) \rangle_{x,y}. \tag{B2}$$

Figure 11 shows that this ratio is indeed much smaller than one for all dinucleotide positions.

APPENDIX C: EFFECT OF TEMPERATURE ON THE PROBABILITY

The probabilities shown in the results section have all been obtained at room temperature $\beta = 1/k_BT_r$. Here we study how these probabilities change with temperature, focusing on dinucleotide AA. Its exact probability distribution for different temperatures is shown in Fig. 12. We find at temperature $\beta = 0$ a constant value $1/16$ for the probability. This is the high-temperature limit where all steps are equally probable. At low temperatures the probability varies between values close to 0 and 1, reflecting the fact that the ground state sequences becomes exceedingly important.

We also evaluated the first- and second-order bounds of the AA probability distribution at five different temperatures: $\beta = 0$, 0.1, 1, 10, and 100 (in units of $[1/k_BT_r]$); see Figs. 13 and 14. At high temperatures (low β) the bounds for both orders are very close to each other enclosing values close to $1/16$. With decreasing temperature the quality of the first-order bounds becomes poorer, giving only a rough qualitative estimate, whereas the second-order bounds continue to work well for relatively low temperatures. Note that at $\beta = 100$ the probability takes values close to 0 and 1 at most places.

We finally take the limit $\beta \to \infty$; see Fig. 15. This figure shows the second-order and third-order bounds on the probability to encounter AA at zero temperature. The only possible sequences are now ground-state sequences (due to the high level of symmetry in our model we expect many different ground states). This explains why the probability of AA can take the values 0 and 1: at several positions AA is not part of any ground-state sequence (probability is zero), while at other positions AA is part of all possible ground-state sequences (probability is 1). At some positions the method cannot determine the percentage of ground-state sequences AA is part of, resulting in bounds of 0 *and* 1.

The method of obtaining upper and lower bounds remains effective at all possible temperatures for our model, and even

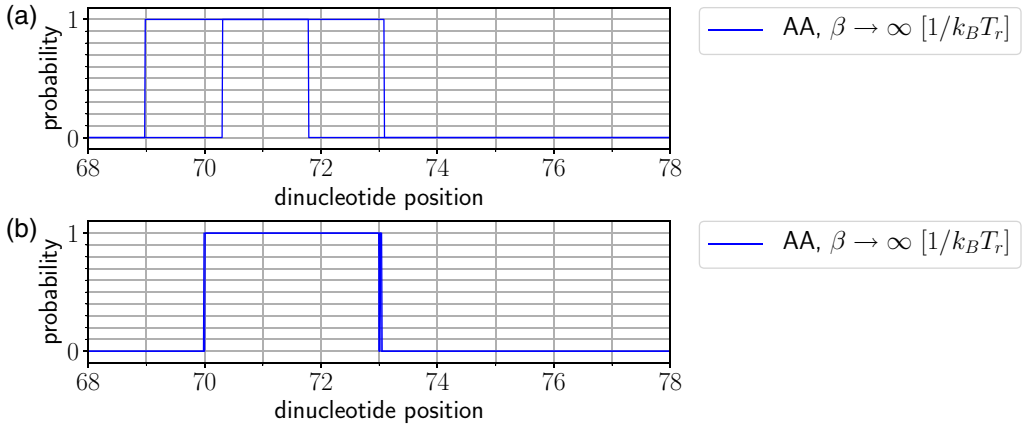

FIG. 15. (a) First-order and (b) second-order bounds on the probability to find dinucleotide AA at several dinucleotide positions on a nucleosome, in the limit of zero temperature. Higher-order bounds get increasingly sharper. At zero temperature the only possible DNA sequences are ground-state sequences, hence the bounds provide us statistics on the ground states of our system. When the lower and upper bounds are 0 and 1, AA is part of an unknown number of ground states at this position. If AA is part of all possible ground states, then the bounds are 1 and 1, and if it is not a part of the ground state, then they are 0 and 0.

provides insight into the possible ground states. Going to higher-order bounds (i.e., taking neighbors that are further away into account as well) or using the exact probability should eliminate the discrepancy between the upper and lower value. However, we are currently preparing a manuscript containing a much more efficient method of obtaining ground states of models such as the one presented in this paper.

[1] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin, Proc. Natl. Acad. Sci. USA **95**, 11163 (1998).

[2] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Nature **389**, 251 (1997).

[3] H. Schiessel, J. Phys.: Condens. Matter **15**, R699 (2003).

[4] S. C. Satchwell, H. R. Drew, and A. A. Travers, J. Mol. Biol. **191**, 659 (1986).

[5] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thaström, I. K. M. Y. Field, J. Z. Wang, and J. Widom, Nature **442**, 772 (2006).

[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] M. Tompitak, C. Vaillant, and H. Schiessel, Biophys. J. **112**, 505 (2017).

[8] T. Vavouri and B. Lehner, PLoS Genet. **7**, e1002036 (2011).

[9] G. Drillon, F. A. B. Audit, and A. Arneodo, BMC Genom. **17**, 526 (2016).

[10] B. Eslami-Mossallam, R. D. Schram, M. Tompitak, J. van Noort, and H. Schiessel, PLoS ONE **11**, e0156905 (2016).

[11] C. R. Calladine and H. R. Drew, J. Mol. Biol. **178**, 773 (1984).

[12] B. D. Coleman, W. K. Olson, and D. Swigdon, J. Chem. Phys. **118**, 7127 (2003).

[13] C. Anselmi, G. Bocchinfuso, P. D. Santis, M. Savino, and A. Scipioni, Biophys. J. **79**, 601 (2000).

[14] M. Y. Tolstorukov, A. V. Colasanti, D. M. McCandlish, W. K. Olson, and V. B. Zhurkin, J. Mol. Biol. **371**, 725 (2007).

[15] C. Vaillant, B. Audit, and A. Arneodo, Phys. Rev. Lett. **99**, 218103 (2007).

[16] S. Balasubramanian, F. Xu, and W. K. Olson, Biophys. J. **96**, 2245 (2009).

[17] A. V. Morozov, K. Fortney, D. A. Gaykalova, V. M. Studitsky, J. Widom, and E. D. Siggia, Nucl. Acids Res **37**, 4707 (2009).

[18] T. Dršata, N. Špačková, P. Jurečka, M. Zgarbová, J. Šponer, and F. Lankaš, Nucl. Acids Res. **42**, 7383 (2014).

[19] D. Norouzi and F. Mohammad-Rafiee, J. Biomol. Struct. Dyn. **32**, 104 (2014).

[20] J. Culkin, L. de Bruin, M. Tompitak, R. Phillips, and H. Schiessel (unpublished).

[21] N. B. Becker and R. Everaers, Structure **17**, 579 (2009).

[22] A. Fathizadeh, A. B. Besya, M. R. Ejtehadi, and H. Schiessel, Eur. Phys. J. E **36**, 21 (2013).

[23] L. de Bruin, M. Tompitak, B. Eslami-Mossallam, and H. Schiessel, J. Phys. Chem. B. **120**, 5855 (2016).

[24] M. Tompitak, L. de Bruin, B. Eslami-Mossallam, and H. Schiessel, Phys. Rev. E **95**, 052402 (2017).

[25] J. Wondergem, H. Schiessel, and M. Tompitak, J. Chem. Phys. **147**, 174101 (2017).

[26] N. B. Becker, L. Wolff, and R. Everaers, Nucl Acids Res. **34**, 5638 (2006).

[27] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, Nucleic Acids Res. **37**, 5917 (2009).

[28] V. B. Teif, Nucleic Acids Res. **35**, e80 (2007).

[29] G. Chevereau, A. Arneodo, and C. Vaillant, Front. Life Sci. **5**, 29 (2011).

[30] G. S. Freeman, J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. J. de Pablo, Phys. Rev. Lett. **113**, 168101 (2014).

[31] M. Tompitak, G. T. Barkema, and H. Schiessel, BMC Bioinf. **18**, 157 (2017).