Freedom of additional signals on genes: on the combination of DNA mechanics, genetics and translation speed

Zuiddam, M.

Chapter 2

Physics behind the mechanical nucleosome positioning code

This chapter is based on Zuiddam, Everaers and Schiessel, 2017, Phys. Rev. E. [31]

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. In this chapter 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.

2.1 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 [32]. 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 carriers.

Possibly the most important biological consequence of sequence dependent DNA mechanics is its impact on the positioning of DNA spools, called nucleosomes. 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 [33] 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

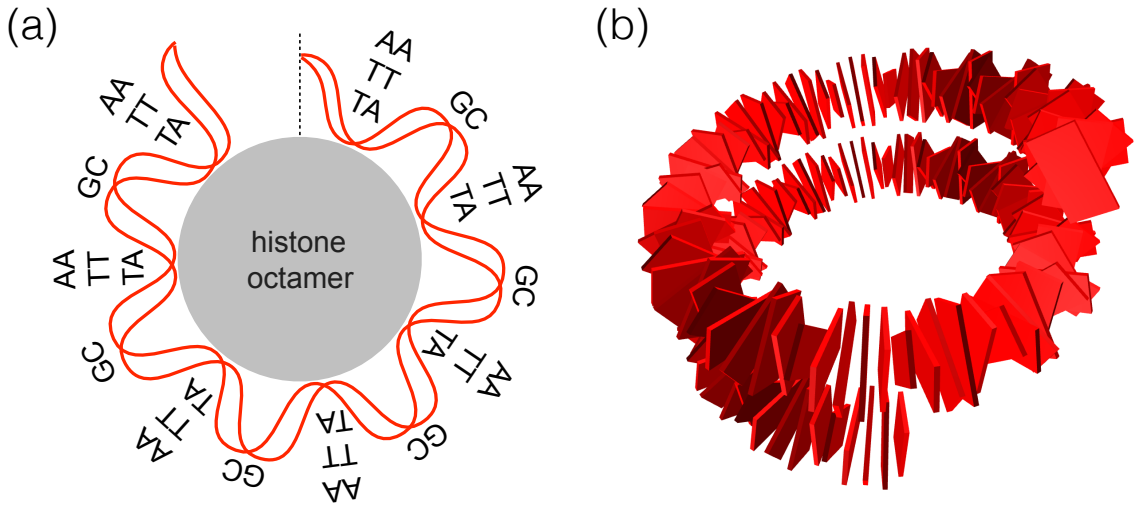

Figure 2.1: (a) The probability of finding GC steps peaks at positions where the major groove of the DNA faces the histone octamer (every 10^{th} bp) whereas TT, AA and TA are all in phase and have their peaks in between where the minor groove faces the cylinder. These are key rules from the so-called "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.

path, a left-handed superhelix of one and three quarter turns.

This structure makes nucleosomes ideal "readers" of mechanical cues. Firstly, 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 [34]) 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. Secondly, 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 10^{th} 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. 2.1(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, 9], mechanically encoded retention of a small fraction of nucleosomes in human sperm cells allowing transmission of paternal epigenetic information [9, 35] or the positioning of six million nucleosomes around nucleosome inhibiting barriers in human somatic cells [8].

However, what is still missing is a deeper understanding of the physics underlying nucleosome positioning rules. An example, mentioned in [14], 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 [12, 13] in which the conformation of a DNA molecule is described by the positions and orientations of its base pairs that are modelled as rigid objects. These nucleosome models have been used to predict nucleosome stability and positioning [15–17, 20–22, 36], forces and torques on the wrapped DNA [37], nucleosome mobility along DNA [18] and the response of nucleosomes to external forces [19]. One recent study [14] 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. 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. 2.1(b)). Assuming such an idealized shape (as done in [20–22, 36]) instead of trying to imitate details of the crystal structure (as done in [14–19]) 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) [14, 17–19, 36]. Variants of our approach are in principle applicable to any model that freezes the DNA into a fixed configuration like it is done in [15, 16, 20–22].

The goal of this chapter is not to come up with yet another tool for nucleosome positioning. Based on the more complete model [14] 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 [9]. 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 "favour" the most costly positions on the wrapped DNA. To achieve this an analytical approach as presented here is indispensable.

In the next section we introduce our model. In Section 2.3 we explain how it can be solved using transfer matrices. This is followed by two sections that develop approximations that allow to come to a detailed understanding of the nucleosome positioning rules: in Section 2.4 we take a limited number of neighbours around the given base pair step into account to derive upper and lower bounds for the probabilities of its occurrence, and in Section 2.5 we introduce the average neighbour 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 Section 2.6, and a conclusion is provided in the final section.

	q^{roll} [rad]	Q^{roll} $\left[\frac{k_B T_r}{\text{rad}^2}\right]$	q^{tilt} [rad]	Q^{tilt} $\left[\frac{k_B T_r}{\text{rad}^2}\right]$
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

Table 2.1: Parametrization used to calculate the dinucleotide energy, Eq. 2.5 to Eq. 2.7. The symbols q and Q denote the intrinsic value and the stiffness of roll or tilt.

2.2 Model

Our model is based on the rigid base pair model [12, 13], a coarse grained representation of the DNA double helix, that treats the base pairs as rigid plates. Neighbouring 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. 2.2. We force this DNA model into a superhelix to mimic the bending of the DNA inside a nucleosome, neglecting the non-uniform bending of the nucleosomal DNA observed in its crystal structure [33]. As the general nucleosome positioning rules hold all along the wrapped part [5], we expect that these simplifications do not affect the rules whose origin we aim to understand here. In addition, motivated by the observation that the basic nucleosome positioning rules can be rationalized by discussing energy costs involved in the roll and tilt degrees of freedom [14], we only account for them and neglect contributions from the other degrees of freedom. This makes the model easier to analyse. The contribution of twist and any cross terms between the rotational degrees of freedom will be discussed in Appendix A.1. There we will show that neglecting these terms does not affect the main positioning rules of our model.

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 [38], see Table 2.1 for a list of the parameters for roll and tilt.

In order 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. [36]):

$$\vec{r}(s) = [R\cos(s/R_{\text{eff}}), R\sin(s/R_{\text{eff}}), -(P/2\pi R_{\text{eff}})s], \tag{2.1}$$

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

$$[\hat{t}(s), \hat{n}(s), \hat{b}(s)] = \left[\frac{d\vec{r}}{ds}, \frac{d\vec{t}}{ds} \Big/ \left|\frac{d\vec{t}}{ds}\right|, \vec{t} \times \vec{n}\right] \tag{2.2}$$

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. 2.2(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) right-handedly around the superhelix:

$$\begin{aligned}[\hat{x}(p), \hat{y}(p), \hat{z}(p)] =& [\hat{n}(s)\cos(\theta p + \phi) - \hat{b}(s)\sin(\theta p + \phi), \\ & -\hat{n}(s)\sin(\theta p + \phi) - \hat{b}(s)\cos(\theta p + \phi), \hat{t}(s)],\end{aligned} \tag{2.3}$$

with $p = s(L-1)/(2\pi R_{\text{eff}}\alpha) + 1/2$ the positions of the dinucleotide (right in between 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/minimum roll and tilt. To reflect the approximately 10 bp helical pitch of the DNA inside the nucleosome, 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 [39] 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 $P = 25.9\,\text{Å}$, $R = 41.9\,\text{Å}$, $\alpha = 1.84$, and $L = 147$ [36], we find expressions for the angles q_p^i, $i \in \{\text{roll}, \text{tilt}, \text{twist}\}$ given by:

$$\begin{aligned}[q_p^{\text{roll}}, q_p^{\text{tilt}}, q_p^{\text{twist}}] =& [\Gamma\cos(2\pi p/10 - 147\pi/10), \\ & \Gamma\sin(2\pi p/10 - 147\pi/10), q^{\text{twist}}],\end{aligned} \tag{2.4}$$

with $\Gamma \approx 0.0796$ rad and $q^{\text{twist}} \approx 10.17/(2\pi)$ 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 [36]. This angle is given by $\arccos\{\vec{t}[s(p)] \cdot \vec{t}[s(p+1)]\} \approx 0.0788$, which is a great

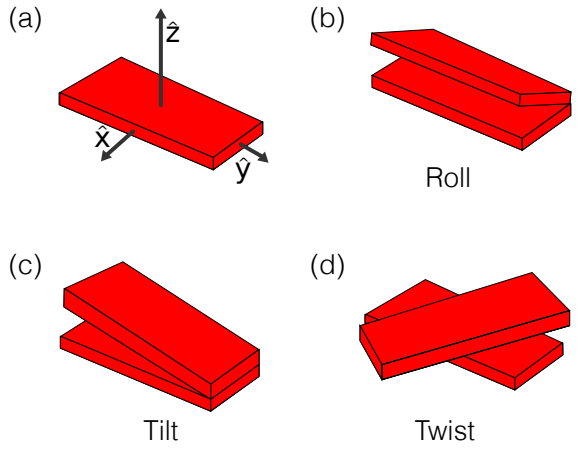

Figure 2.2: The rotational degrees of freedom between neighboring bp in the rigid base pair model. Each base pair has a coordinate system (a) which can be used to describe the relative orientation between two plates. In our model we account only for energy contributions from (b) roll and (c) tilt but neglect contributions from (d) twist. Also the translational degrees are not considered.

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 with a periodicity of 10.17. This may seem counter-intuitive. However, if the twist were equal to $2\pi/10$, the configuration of the plates would be a ring instead of a superhelix.

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:

$$E_p(a,b) = E_p^{\text{roll}}(a,b) + E_p^{\text{tilt}}(a,b) \tag{2.5}$$

with

$$E_p^{\text{roll}}(a,b) \equiv \frac{1}{2} Q^{\text{roll}}(a,b) \left[q_p^{\text{roll}} - \bar{q}^{\text{roll}}(a,b)\right]^2, \tag{2.6}$$

and

$$E_p^{\text{tilt}}(a,b) \equiv \frac{1}{2} Q^{\text{tilt}}(a,b) \left[q_p^{\text{tilt}} - \bar{q}^{\text{tilt}}(a,b)\right]^2. \tag{2.7}$$

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 $\bar{q}^{\text{tilt}}(a,b)$.

2.3 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 [5, 32]. 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 p^{th} 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. In order 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

$$P_p(a,b) = \frac{\sum\limits_{\substack{n_1,\ldots,n_L \\ n_p=a, n_{p+1}=b}} \exp\left[-\beta \sum\limits_{i=1}^{L-1} E_i(n_i, n_{i+1})\right]}{\sum\limits_{n_1,\ldots,n_L} \exp\left[-\beta \sum\limits_{i=1}^{L-1} E_i(n_i, n_{i+1})\right]} \tag{2.8}$$

where we sum over all possible states $n_i \in \{A, T, C, G\}$, with β the inverse temperature. The probability given by Eq. 2.8 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 Ref. [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, 36, 38, 40] (see Ref. [41] 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

$$\langle n|T_i|m\rangle \equiv \exp\left[-\beta E_i(n,m)\right] \tag{2.9}$$

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

$$P_p(a,b) = \frac{\sum\limits_{n_1,n_L} \langle n_1| T_1 \ldots T_{p-1} |a\rangle \langle a|T_p|b\rangle \langle b| T_{p+1} \ldots T_{L-1} |n_L\rangle}{\sum\limits_{n_1,n_L} \langle n_1| T_1 \ldots T_{p-1} T_p T_{p+1} \ldots T_{L-1} |n_L\rangle}. \tag{2.10}$$

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 neighbours, 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.

2.4 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 k^{th}-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 $k \geq 1$ is then given by

$$P_p(a,b) = \frac{\langle x_k| \prod_{i=p-k}^{p-1} T_i \, |a\rangle \, \langle a|T_p|b\rangle \, \langle b| \prod_{j=p+1}^{p+k} T_j \, |y_k\rangle}{\langle x_k| \prod_{i=p-k}^{p+k} T_i \, |y_k\rangle} \tag{2.11}$$

with

$$\langle x_k| \equiv \frac{1}{c_k} \sum_n \langle n| \, T_1 T_2 ... T_{p-k-2} T_{p-k-1} \tag{2.12}$$

and

$$|y_k\rangle \equiv \frac{1}{d_k} \sum_n T_{p+k+1} T_{p+k+2} ... T_{L-2} T_{L-1} \, |n\rangle \, , \tag{2.13}$$

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 k^{th}-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 k^{th}-order upper/lower bound by taking the maximum/minimum of Eq. 2.11 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 $0 < x_{n,k} \leq 1$, $0 < y_{n,k} \leq 1$. 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 further 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 k^{th}-order upper and lower bounds of the probability, $P^{(k)}_{\text{max},p}(a,b)$ and $P^{(k)}_{\text{min},p}(a,b)$, by

simply evaluating the probability for all 16 possible combinations of pure states. This leads to the expressions

$$P^{(k)}{}_{\max,p}(a,b) = \max_{|x_k^*\rangle,|y_k^*\rangle\in B} \frac{\langle x_k^*| \prod_{i=p-k}^{p-1} T_i \,|a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j \,|y_k^*\rangle}{\langle x_k^*| \prod_{i=p-k}^{p+k} T_i \,|y_k^*\rangle} \tag{2.14}$$

and

$$P^{(k)}{}_{\min,p}(a,b) = \min_{|x_k^*\rangle,|y_k^*\rangle\in B} \frac{\langle x_k^*| \prod_{i=p-k}^{p-1} T_i \,|a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j \,|y_k^*\rangle}{\langle x_k^*| \prod_{i=p-k}^{p+k} T_i \,|y_k^*\rangle}. \tag{2.15}$$

We prove now the expression for the k^{th}-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. 2.11. To prove Eq. 2.14, 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

$$P_p(a,b) = \frac{\sum_{n\in B}\sum_{m\in B} x_{n,k}\bar{\mathcal{T}}_{nm}y_{m,k}}{\sum_{n\in B}\sum_{m\in B} x_{n,k}\mathcal{T}_{nm}y_{m,k}}. \tag{2.16}$$

Without loss of generality, we assume that

$$\frac{\bar{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}} = \min\left(\frac{\bar{\mathcal{T}}_{AA}}{\mathcal{T}_{AA}}, \frac{\bar{\mathcal{T}}_{AT}}{\mathcal{T}_{AT}}, ..., \frac{\bar{\mathcal{T}}_{GG}}{\mathcal{T}_{GG}}\right) \tag{2.17}$$

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 $x_{i,k}y_{j,k}$:

$$\begin{aligned}\frac{\partial P_p(a,b)}{\partial(x_{i,k}y_{j,k})} =& \frac{\sum_{n\in B}\sum_{m\in B} x_{m,k}\mathcal{T}_{nm}y_{n,k}\mathcal{T}_{ij}\left(\frac{\bar{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}} - \frac{\bar{\mathcal{T}}_{nm}}{\mathcal{T}_{nm}}\right)}{\left(\sum_{n\in B}\sum_{m\in B} x_{n,k}\mathcal{T}_{nm}y_{m,k}\right)^2} \\ \leq& 0.\end{aligned} \tag{2.18}$$

The less-than or equal to sign follows from the fact that $\tilde{\mathcal{T}}_{nm}$, $\mathcal{T}_{nm}$, $x_{m,k}$ and $y_{n,k}$ are non-negative for all n, m and from Eq. 2.17. Because the derivative is non-positive, the probability is non-increasing 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 $\frac{\tilde{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}}$ from Eq. 2.16) (this can be checked by inserting $x_{i,k}y_{j,k} = 0$ in Eq. 2.16). 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:

$$P^{(k)}_{\max,p}(a,b) = \frac{x_{r,k}\bar{\mathcal{T}}_{rs}y_{s,k}}{x_{r,k}\mathcal{T}_{rs}y_{s,k}} = \frac{\bar{\mathcal{T}}_{rs}}{\mathcal{T}_{rs}}. \tag{2.19}$$

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 k^{th} 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. 2.14.

For the zeroth-order bounds, where no neighbours are taken into account, a similar result holds. This can be obtained in the same manner as Eq. 2.14 and Eq. 2.15, therefore no proof is provided. These bounds are given by

$$P^{(0)}_{\max,p}(a,b) = \max_{|x_0^*\rangle,|y_0^*\rangle \in B} \frac{\langle x_0^*|a\rangle \langle a|T_p|b\rangle \langle b|y_0^*\rangle}{\langle x_0^*| T_p |y_0^*\rangle} = 1 \tag{2.20}$$

and

$$P^{(0)}_{\min,p}(a,b) = \min_{|x_0^*\rangle,|y_0^*\rangle \in B} \frac{\langle x_k^*|a\rangle \langle a|T_p|b\rangle \langle b|y_k^*\rangle}{\langle x_k^*| T_p |y_k^*\rangle} = 0. \tag{2.21}$$

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 neighbour 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:

$$P^{(k)}_{\max,p}(a,b) \geq P^{(k+1)}_{\max,p}(a,b) \geq P_p(a,b) \tag{2.22}$$

$$P_p(a,b) \geq P^{(k+1)}_{\min,p}(a,b) \geq P^{(k)}_{\min,p}(a,b). \tag{2.23}$$

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 straighforward to prove. Consider Eq. 2.14 for the $(l+1)^{\text{th}}$ 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. 2.14 the states $\langle x_l^*| = \langle n| T_{p-l}$ and $|y_l^*\rangle = T_{p+l}|m\rangle$. We find the l^{th}-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 l^{th}-order maximum. Therefore, the $(l+1)^{\text{th}}$-order upper bound cannot be higher than the l^{th}-order upper bound. The same reasoning holds for the lower bounds.

2.5 The average neighbour 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 neighbours 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 neighbours 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 neighbours 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 neighbour 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 neighbours. 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

$$\langle f(x)\rangle_x = \frac{1}{4}\sum_{x\in\{A,T,C,G\}} f(x), \tag{2.24}$$

$$\langle g(x,y)\rangle_{x,y} = \frac{1}{16}\sum_{x,y\in\{A,T,C,G\}} g(x,y), \tag{2.25}$$

we define the average neighbour energy approximation of the probability as

$$P_p^{(e)}(a,b) \equiv \frac{\begin{array}{l}\exp\left[-\beta\langle E_{p-1}(n_{p-1},a)\rangle_{n_{p-1}}\right] \\ \quad\times\exp\left[-\beta E_p(a,b)\right] \\ \quad\quad\times\exp\left[-\beta\langle E_{p+1}(b,n_{p+2})\rangle_{n_{p+2}}\right]\end{array}}{\sum\limits_{n_p,n_{p+1}}\begin{array}{l}\exp\left[-\beta\langle E_{p-1}(n_{p-1},n_p)\rangle_{n_{p-1}}\right] \\ \quad\times\exp\left[-\beta E_p(n_p,n_{p+1})\right] \\ \quad\quad\times\exp\left[-\beta\langle E_{p+1}(n_{p+1},n_{p+2})\rangle_{n_{p+2}}\right]\end{array}}. \tag{2.26}$$

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 neighbours of ab. We have calculated the error introduced by using the average neighbour energy approximation and found it not to be larger than 3.5 percent at any position for any dinucleotide, see Appendix A.2.

Next we provide an explanation why this approximation works so well for our model. Our strategy is to bring the approximated probability, Eq. 2.26, and the full probability, Eq. 2.8, 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. 2.8):

$$P_p^{(e)}(a,b) = \frac{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}: \\ n_p=a,n_{p+1}=b}} \exp\Big[-\beta\sum_{i=1}^{L-1}\langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big]}{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}}} \exp\Big[-\beta\sum_{i=1}^{L-1}\langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big]}. \tag{2.27}$$

The hats above n_{p-1} and n_{p+2} denote that these variables are not to be summed over. The nominator factorises 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. 2.26.

We will now make the exact probability (Eq. 2.8) look more like the approximation in the form of (Eq. 2.27). By substituting the function $C_p(i,j)$, defined as

$$C_p(m,o) \equiv \frac{\frac{1}{4}\sum\limits_n \exp\left[-\beta E_{p+1}(m,n) - \beta E_{p+2}(n,o)\right]}{\exp\left[-\beta\langle E_{p+1}(m,n) + E_{p+2}(n,o)\rangle_n\right]}, \tag{2.28}$$

into Eq. 2.8 twice, we obtain

$$P_p(a,b) = \frac{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}: \\ n_p=a,n_{p+1}=b}} \exp\Big[-\beta \sum_{i=1}^{L-1} \langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big] \times C_{p-2}(n_{p-2},a) C_{p+1}(b,n_{p+3})}{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}}} \exp\Big[-\beta \sum_{i=1}^{L-1} \langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big] \times C_{p-2}(n_{p-2},n_p) C_{p+1}(n_{p+1},n_{p+3})}, \tag{2.29}$$

which is indeed very similar to Eq. 2.27, 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 A.2 for details.

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

$$C_p(m,o) \approx 1 + \frac{1}{2}\beta^2 \big\langle \big[E_{p+1}(m,n) + E_{p+2}(n,o) - \langle E_{p+1}(m,n') + E_{p+2}(n',o)\rangle_{n'}\big]^2 \big\rangle_n. \tag{2.30}$$

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.

2.6 Results

2.6.1 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 10^{th} bp) whereas AA, TA, and TT are all in phase and have their peaks in between where the minor groove faces the cylinder.

Fig. 2.3(a) shows the combined probability to encounter AA, TA, TT along the nucleosome and, separately, that of GC calculated using transfer matrices, Eq. 2.10. 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

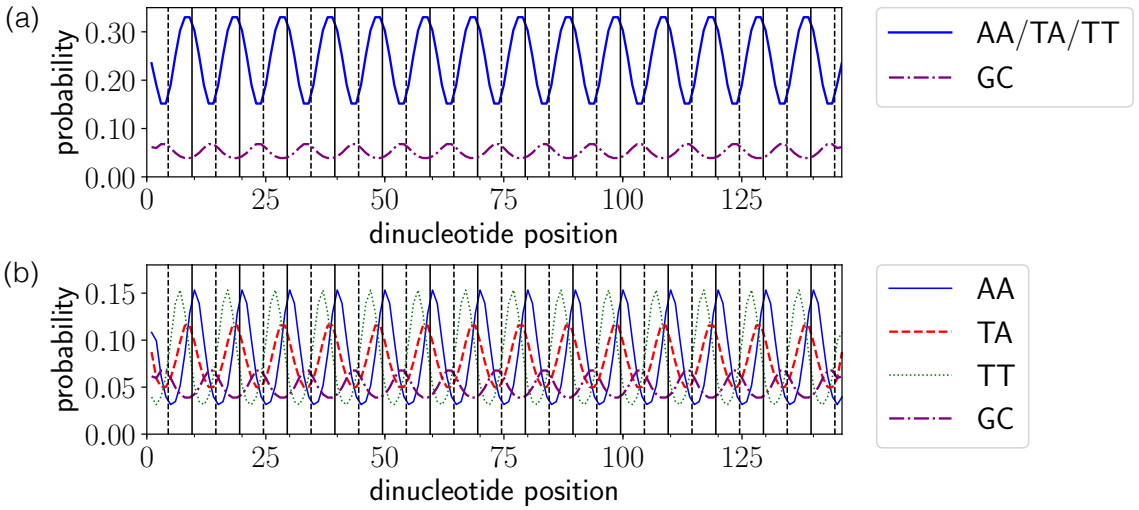

Figure 2.3: (a) The probability to find AA, TA or TT, and the probability to encounter GC at the full range of dinucleotide positions is shown. The solid and dashed vertical lines indicate minor and major bending sites (maximum negative and positive roll, respectively). The probabilities are in qualitative agreement with the well-known nucleosome positioning rules [5]. (b) Same as (a) but showing all four dinucleotide probabilities individually.

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. 2.3(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. 2.3 are, of course, simply a consequence of 10 bp periodicity in our model, see Eq. 2.6 and Eq. 2.7. 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.

2.6.2 The bounds on the probability

Fig. 2.4(a)-(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 Eq. 2.14 and Eq. 2.15). Note that the energy as defined by Eq. 2.5 to Eq. 2.7 allows also for non-integer bp positions. Even though these non-integer 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 neighbour to the left and right (first-order bounds) the bounds indicate already the qualitative behaviour of the system for some of the dinucleotides (AA, TA and TT but not GC), see Fig. 2.4(a). Accounting for two neighbours on each side (second-order bounds) provides already an excellent estimate of the dinucleotide probabilites 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. 2.4(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 neighbours involved), see Fig. 2.4(c). This allows us to define an effective order κ, similar to a correlation length:

$$P^{(k)}_{\max,p}(a,b) - P^{(k)}_{\min,p}(a,b) \approx e^{-k/\kappa}. \tag{2.31}$$

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. A.3.

2.6.3 Explaining the dinucleotide positioning rules

So far we have presented the probability distributions of a few key dinuclotides 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 neighbours matter. This is one of the reasons why the probabilities are well captured by the average neighbour 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. 2.3 shows that the probability (calculated using Eq. 2.10) 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 73-74). Moreover, TT peaks at positions of maximal positive tilt (such as position 77), while AA peaks at maximal negative tilt (e.g., at 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.

A local perspective on the dinucleotide probability fails

Table 2.1 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,

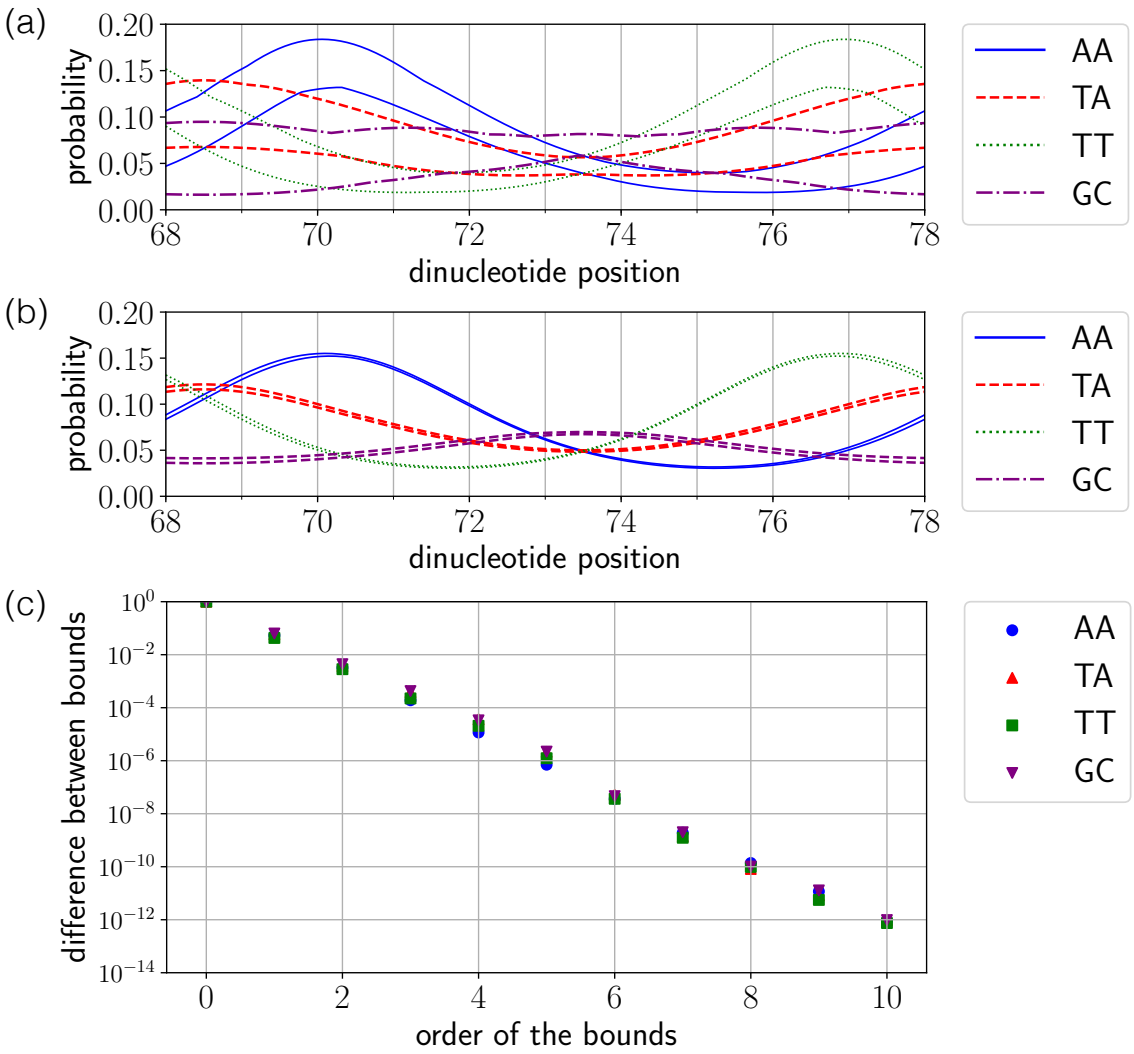

Figure 2.4: (a)-(b) Upper and lower bounds on the probabilities to have the dinucleotide AA, TA, TT or GC at several dinucleotide positions on a nucleosome. Specifically (a) depicts the first-order bounds and (b) the second-order bounds of the probability. The upper and lower bounds of the same dinucleotide have the same colour (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 neighbours k steps away from the dinucleotide of interest, decreases exponentially as the order k increases.

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

Figure 2.5: The probability of dinucleotide TA, its energy and the average of the energies of its possible neighbours are shown for 10 different positions along the nucleosome, i.e., for one full DNA helical repeat. The numbers give absolute values whereas the colours 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) colours represent relatively favourable values, red (dark gray) indicates unfavourable values. The probability follows mainly from a 'mixing' of the colours of the corresponding dinucleotide energies and average neighbour energies. The table also provides subdivisions of the TA energies into roll and tilt contributions.

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 78-79 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 2.1.

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 neighbours, suffice for some of the dinucleotides to have rather good estimates of the probability, see Fig. 2.4(a).

Neighbouring steps are equally important

The effect of the neighbours can be best understood using the average neighbour energy approximation, see Eq. 2.26. Since this is an excellent approximation, see Appendix A.2, the only terms important for the behaviour of the probability are the *energy of the dinucleotide itself*, and the *average of the energies of its possible neighbours*. 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 neighbours of *ab* with the averages of the energies of all possible neighbours of the 15 other dinucleotides.

Such information can be best presented in tabular form. Fig. 2.5 provides the relevant information for the TA dinucleotide. It presents (as numbers) the probability (obtained using Eq. 2.10) to find this dinucleotide, its energy (Eq. 2.5) and the average of the energies of its possible neighbours (see Eq. 2.26) for a 10 base pair stretch in one table (and some further information that we discuss further below). More relevant, however, are the colours assigned to each box as they indicate how these numbers compare to the values of all other possible dinucleotides. If the colour is yellow (light gray), the value is relatively favourable compared to the

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

Figure 2.6: Same as Fig. 2.5 but for GC.

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 unfavourable values, while orange (gray) indicates that this value is average.

First consider in Fig. 2.5 row "Probability": At positions 69 and 78, both associated with *negative* roll and zero tilt, TA is favourable, as we have seen in Fig 2.3. 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 (colours). Next turn to row "Average neighbour energy": the absolute values (numbers) have their lowest values at 73 and 74 but the relative values (colours) strongly prefer the opposite. Therefore, what causes the TA preference for negative roll positions is the average energy of the possible neighbours relative to the average energy of the forbidden neighbours.

Now we turn to the other rows in Fig. 2.5. These extra rows provide a subdivision of the dinucleotide energies and the average neighbour 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 neighbour 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. 2.6. This is another non-trivial dinucleotide in the sense that its behaviour is heavily affected by its possible neighbours. Positions 73 and 74, associated with large positive roll and bending towards 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 neighbours cause the probability of encountering GC to be highest at these positions and lowest at positions 69 and 78.

The complete picture

In Fig. 2.7 tables are shown that present the probabilities and relative energies for all 16 dinucleotides (again in colour code). There are seven tables corresponding to the seven rows in Fig. 2.5 and Fig. 2.6. Using these tables one can analyse preferences for each dinucleotide step individually, just as explained for TA and GC above. Moreover, for cases where the average neighbour energies dominate the positional preferences of dinucleotides (like for TA and GC), these tables allow to look up which of the possible neighbours of a given dinucleotide are favourable.

As an example, we consider again the dinucleotide TA. In Fig. 2.7(a) we see

that the probability of TA peaks at positions 69 and 78, which is not TA's intrinsic preference, Fig. 2.7(b), but that of its neighbours on average, Fig. 2.7(c). We need now to inspect the intrinsic preferences of all the possible neighbours. At position 70, three of the four possible neighbours (dinucleotides starting with an A) are favourable, namely AA, AT, and AC, see Fig. 2.7(b). Due to symmetry TT, AT, and GT are favourable at position 77, see also Fig. 2.7(b). Further details are revealed by Fig. 2.7(d), and (e) that present the roll and tilt contributions to the dinucleotide energies. It shows that AA at 70, and TT at 77 are favourable due to both their roll and tilt preferences whereas the other favourable steps, AT and AC at 70, and AT and GT at 77, prefer those positions due to roll alone. Inspecting the contributions of roll and tilt to the average neighbour energies for TA at positions 69 and 78, Fig. 2.7(f) and (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 unfavourable neighbours (AG at 70, and CT at 77) are unfavourable due to roll whereas the tilt contributions are favourable.

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 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.

Shape is more important than stiffness

The roll and tilt terms of the energy can be subdivided even further. As can been seen from equations Eq. 2.5 to Eq. 2.7 the sequence dependences 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 behaviour of our system. In Fig. 2.8 we compare two tables for dinucleotide energies: the original table on the left (identical to Fig. 2.7(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 (colour 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 [42] where molecular dynamics simulations performed on rather detailed nucleosome models revealed that nucleosome affinity is dominated by the shape of the wrapped DNA.

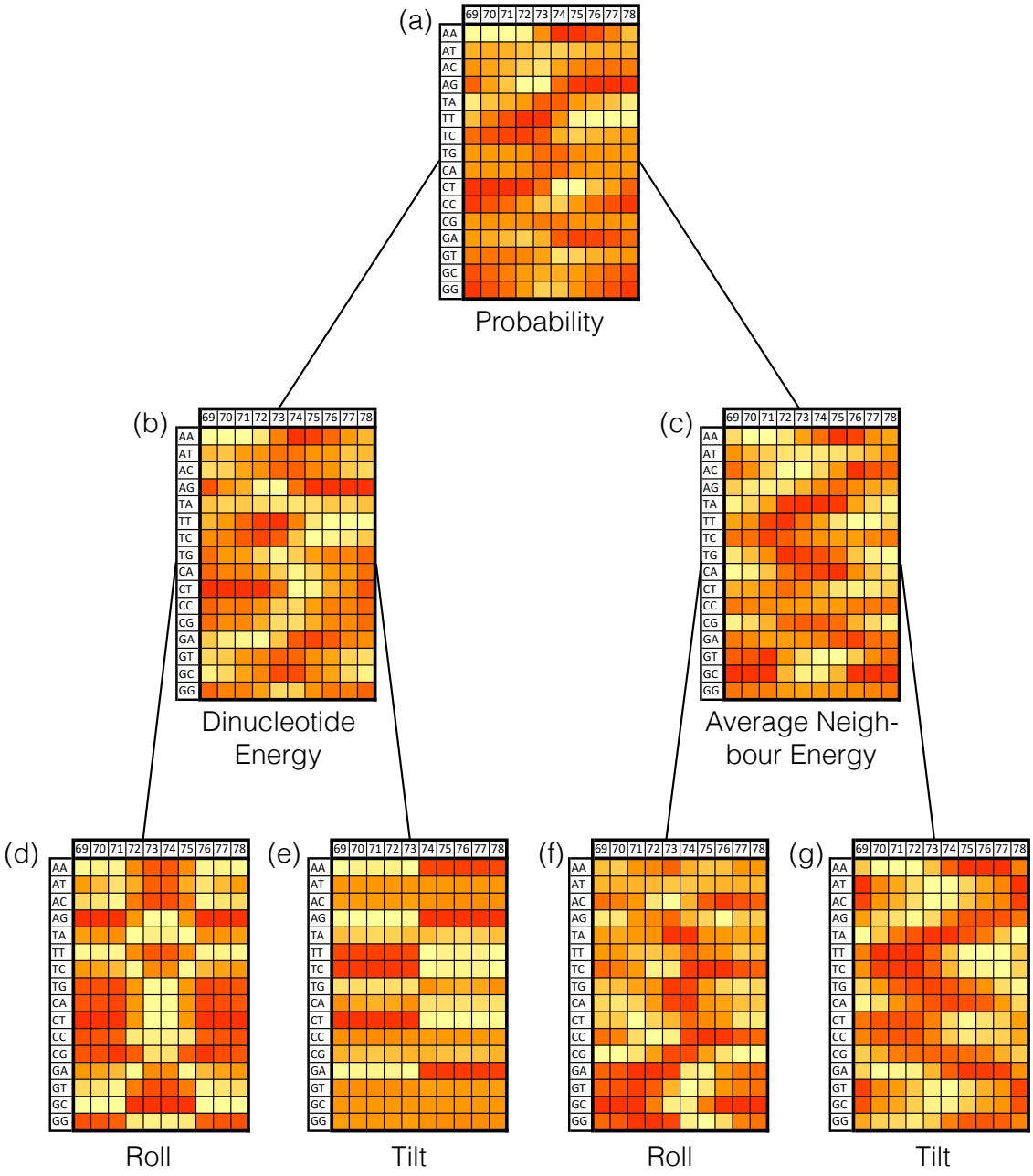

Figure 2.7: (a) Probability, (b) dinucleotide energy, and (c) average neighbour energy of all 16 dinucleotides for one DNA helical repeat. Yellow (light gray) denotes high probability/low energy, red (dark gray) low probability/high energy (relative to all other dinucleotides at the corresponding location). In addition provided are subdivisions of dinucleotide energies into (d) roll and (e) tilt, and of neighbour energies into (f) roll and (g) tilt. The colours representing the probabilities can be seen as a 'sum' of the colours of the dinucleotide energies and the average neighbour energies. The colours corresponding to the dinucleotide energies are the 'sum' of the colours for roll and tilt energies. The same holds for the neighbour energies.

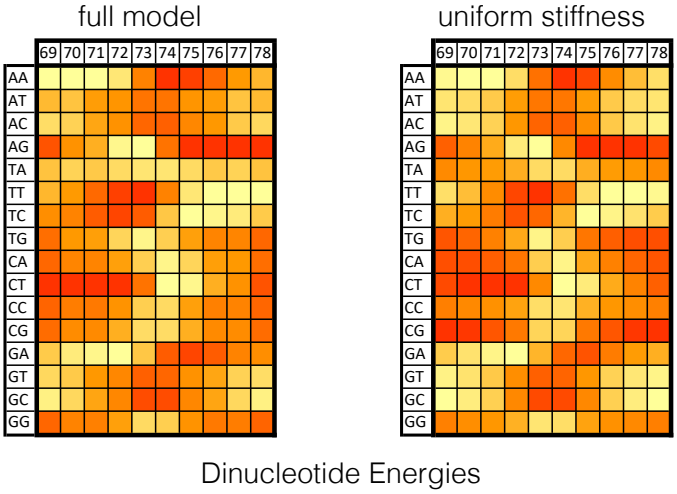

Figure 2.8: Original dinucleotide energy costs (left; same as Fig. 2.7(b)), and energy cost with all stiffnesses set to 1 (right) are shown side by side. The strong similarity between the two tables reveals that stiffnesses play only a minor role for the dinucleotide positional preferences.

Conclusion

In this chapter 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 [14]) 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 unfavourable locations. To solve this puzzle, we first introduced an approximation that, by taking a limited number of neighbours around a given dinucleotide into account, provides upper and lower bounds to its probability distribution. From this we learned that the nearest neighbours 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 neighbour 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 neighbours 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" neighbours (e.g., TA and GC).

Knowing the dinucleotide preferences of nucleosomes allows genome wide calculations of nucleosome positioning [10]. 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, 9] or the retention of nucleosomes in human sperm cells [9, 35].