Sequence Dependence in Nucleosome Breathing

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE
in
PHYSICS

Author :	J Culkin
Student ID :	S1574272
Supervisor :	Prof. dr. Helmut Schiessel
2^{nd} corrector :	Dr. L. Giomi

Leiden, The Netherlands, September 5, 2016

Sequence Dependence in Nucleosome Breathing

J Culkin

Huygens-Kamerlingh Onnes Laboratory, Leiden University
P.O. Box 9500, 2300 RA Leiden, The Netherlands

September 5, 2016

Abstract

The nucleosome consists of a short stretch of DNA wrapped around a protein cylinder, and is the fundamental unit of chromatin, which compacts the DNA into the cell nucleus. The nucleosome is known to transiently partially unwrap or 'breathe' *in vitro*, exposing DNA which would otherwise be sterically inaccessible to enzymes. Breathing is investigated for its potential importance *in vivo* in both essential DNA processes, and in higher-order chromatin organisation. In this thesis we present a two-parameter physical statistical model of the breathing process based on steric enzyme accessibility, the energetics of the bent DNA molecule, and the adsorption of the DNA upon the proteins. We estimate the elastic energy using Monte Carlo simulations of a coarse-grained model of the nucleosomal DNA, and we fit the model to the available experimental results. We find in agreement with experimental studies that site accessibility decays exponentially toward the centre sites, and that highly asymmetric breathing behaviour is possible due to the very sensitive dependence of breathing upon energy distribution, and in turn, sequence.

Contents

Chapter 1

Introduction

DNA molecules are many orders of magnitude longer than a cell is wide, so are highly compacted within chromatin, a complex of DNA, RNA and proteins, to fit inside the nucleus. The organisation of chromatin is an active area of research, but it is well known that at its lowest level it consists of short stretches of DNA wrapped around protein cylinders in structures called nucleosomes, as shown in figure 1.1. Nucleosomal DNA is strongly bent into a super-helical $1\frac{3}{4}$ turn, adsorbed upon the protein cylinder via hundreds of hydrogen bonds, as well as some stronger salt links between the phosphate backbone of the DNA and the proteins. These bonds are concentrated in the 14 sites where the minor groove of the DNA faces the cylinder [2]. There is an additional protein H1 positioned outside the nucleosome, thought to stabilise higher-order chromatin structure, however

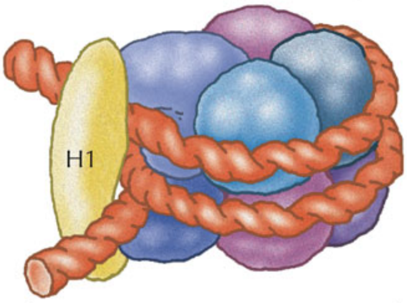


Figure 1.1: *The nucleosome: a short ($\sim$147bp) sequence of DNA wrapped around a protein cylinder in a $1\frac{3}{4}$ left-handed superhelical turn. The protein cylinder is an octamer consisting of 4 different histone components: H2A, H2B, H3 and H4. A further histone H1 is positioned outside stabilising the entry/exit. Adapted from [1]*

it is known to be transient, in constant exchange between nucleosomes [3].

A detailed structure of the nucleosome is now quite well resolved: X-ray crystallography studies have provided measurement of the interactions between base pairs, of the subsequent DNA deformation [4], and also a detailed atom-resolution mapping of the protein cylinder [5]. With this understanding, coarse-grained base pair level simulations of the nucleosome are possible, and have been used to support a growing body of work which investigates the effect of sequence upon dynamics: for example, nucleosome positioning [6], and force-induced unwrapping [7] (see review of sequence dependence [8]). This thesis continues in this theme, examining sequence dependence in nucleosomal 'breathing', a process commonly observed *in vitro*, in which thermal fluctuation drives the DNA to transiently partially unwrap from, and rewrap onto, the protein cylinder.

Nucleosomal breathing is widely studied for its potential role in essential DNA processes, gene regulation, and in dynamical chromatin structure. For example, it is known that nucleosomal DNA is inaccessible to proteins whilst it is fully wrapped on the nucleosome [8], and that breathing may be crucial in providing the required access *in vivo*. However, the *in vivo* picture is not yet clear, with multiple mechanisms proposed, including: histone modification-controlled dynamics, H1 linker histone dominated dynamics [9], and extensive ATP-dependent remodelling. Here we focus on understanding the simpler *in vitro* dynamics.

The first experiments to probe the phenomenon measured the accessibility of enzymatic restriction sites *in vitro* throughout a wrapped nucleosomal sequence ([10] and [11]). The authors found that all sites were accessible, roughly exponentially less so toward the centre. They proposed a stepwise unpeeling mechanism from the DNA ends, which later experiments have supported. Subsequent fluorescence resonance energy transfer (FRET) experiments then enabled a more direct observation of unwrapping and rewrapping, and the timescales involved. It was found in [12], using a modified 601 sequence, that the nucleosome spends $\sim$250 ms fully wrapped before spontaneously unwrapping its ends, and then rewraps within $\sim$10–50 ms. In agreement with the restriction site experiments, in [13] it was then found that this rate decreased sharply toward the centre, with the portion up to base pair 47 of the 146bp wrapped sequence only fully unwrapping once in $\sim$ 10 min.

Physical models of breathing focus on the interplay between the elastic energy of the bent DNA molecule, and the adsorption energy due to the histone-DNA interactions: unwrapping occurs when thermal fluctuations together with the elastic energy of a segment overcome its adsorption energy. The elastic energy can be simply estimated from models, for example by treating the molecule as a worm-like chain in [14]. More recently however, Monte Carlo base pair resolution computation based on the detailed crystallographic data has been possible [15], which gives sequence-dependent results.

The distribution and strength of adsorption due to histone-DNA interactions has been a common focus of investigation; whether adsorption is equally or unequally spread is unresolved, owing to difficulties with experimental approaches. Force induced unwrapping has been a common assay investigating this, in which a single wrapped nucleosome is pulled apart from both ends at once, and the dynamics are derived from sensitive measurements of the change in force. However, as analysed closely in [16], the experimental design introduced physical geometric effects that were difficult to untangle from the biochemistry. A later study [17] adopted a new design, which resolved this problem, by pulling on a single end at once. From the pauses observed in unwrapping, the authors were able to conclude that the nucleosome unpeels in 14 stages, corresponding to the 14 sites of concentrated stronger binding.

The results suggested unequal adsorption energies, which in [15] were estimated to be spread over a 7kT range, with generally the more central sites being stronger. In this thesis we will consider both equal and unequal adsorption energies - and with our breathing model will see which if either fit the restriction enzyme data better.

Existing computational models of nucleosome breathing (reviewed briefly in [8]) do not model individual base pairs, and so miss potential sequence dependence. Here we use a base pair level model of the nucleosomal DNA, and focus on understanding the early restriction enzyme experiments, and the breathing of the 5S and 601/601.2 sequences.

Chapter 2

Methods and Models

In statistical mechanics, the Boltzmann distribution for a system in equilibrium gives the probability of each of its possible states being realised, as a function of the system's temperature and the states' energies. The lower the energy of the state, the more likely it is, or the more stable it is. To determine the most likely equilibrium state of a fully or partially wrapped nucleosome, we find the state with the lowest effective adsorption energy: the sum of the adsorption energy of the DNA-protein complex, and the bending-induced elastic energy of the DNA molecule.

A free DNA molecule will adopt a certain shape in equilibrium, due to the interactions between its base pairs and other chemical groups. When the DNA is wrapped around the protein cylinder it is forced away from this shape, and like a stretched spring, we can associate an elastic energy with it, quadratically proportional to how far from its natural shape it is bent. When the DNA adsorbs upon the protein octamer, hydrogen bonds and salt links form. The strengths of these bonds can be measured by the amount of energy that would be required to break them. The stronger, more stable bond requires more energy to break, so it is measured by a negative energy. The total DNA-protein adsorption energy is the sum of all the bond energies. A stable state then must compromise: straighter DNA has lower elastic energy, but bent DNA adsorbs better, and so has lower (more negative) adsorption energy. The stable state minimises the effective adsorption energy, the sum of these energies.

We estimate the elastic energy of the DNA using Monte Carlo simulations, and fit the adsorption energy using the restriction enzyme experimental breathing data. We calculate the breathing profile of a DNA molecule from the set of energies of the partially unwrapped nucleosome, using a statistical equilibrium model with two free parameters.

2.1 Nucleosomal DNA model

We use a coarse-grained rigid base pair model of DNA, in which each base pair is represented by a rigid plate, as shown in figure 2.1. The model does not explicitly include the phosphate backbone, the protein cylinder, nor the individual bonds, but the 14 sites of concentrated direct DNA-protein interaction are implicitly represented by fixed constraints on the positions of the base pairs. Each of these sites contain 2 phosphate-histone bonds, originating from the DNA phosphate backbone, directly between pairs of nucleotides where the minor groove faces the DNA (indicated by red dots in the figure). The phosphate between successive nucleotides is known to be fixed with respect to their mid-frame [6]; by fixing the mid-frame between base pairs adjacent to the 28 phosphate-histone bonds, we force the entire molecule into the required superhelical path as shown in the figure.

In our model each base pair interacts only with its two nearest neighbours. We associate one set of degrees of freedom (d.o.f) with each free base pair step (each interaction between base pairs): three translational and three rotational, as defined in [18], and shown in figure 2.2. Note that the 28 base pair steps positioned over the phosphate-histone bonds are

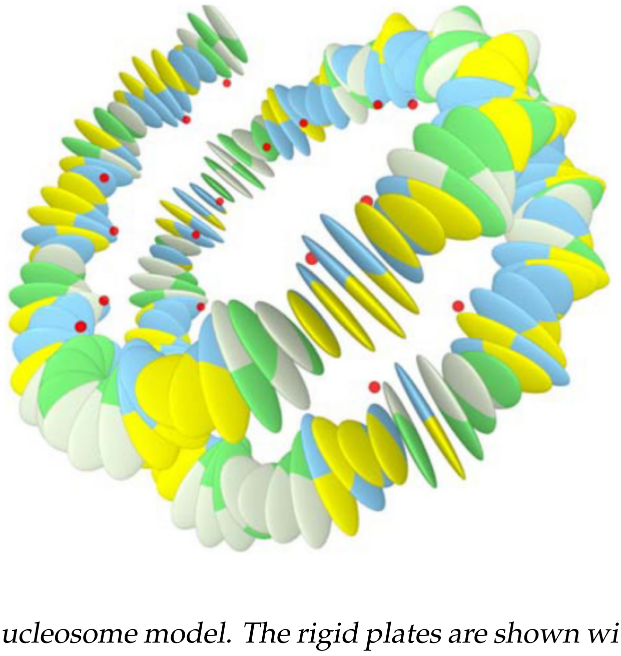


Figure 2.1: *The nucleosome model. The rigid plates are shown with colour-coded bases, and the red spots indicate the 28 fixed phosphates in the 14 concentrated binding sites. (This figure was originally published in [6])*

only free when those bonds are broken during breathing. We adopt the assumption that there are preferred intrinsic values for each d.o.f for each of the 10 possible dinucleotides* (reduced from 16 by considering symmetries), which have been experimentally determined in [4] and [19]. Further, we assume deviations from these values incur a quadratic mechanical (elastic) energy contribution from each base pair step i in the DNA molecule:

$$E^E = \sum_i \frac{1}{2} \left(q_i - q_i^0\right)^T Q \left(q_i - q_i^0\right) \tag{2.1}$$

Here q_i is a six-component vector containing the values for each d.o.f in the i^{th} base pair step, q_i^0 is the set of experimentally determined intrinsic equilibrium values for the particular base pairs in that step, and Q is the 6x6 stiffness matrix which determines the coupling strength between each d.o.f.. We follow [6] and [20] in the 'hybrid' parametrization; we use crystallographic results for the intrinsic equilibrium values, and atomistic molecular simulation results for the stiffness matrix, which [20] showed to give the most precise potentials.

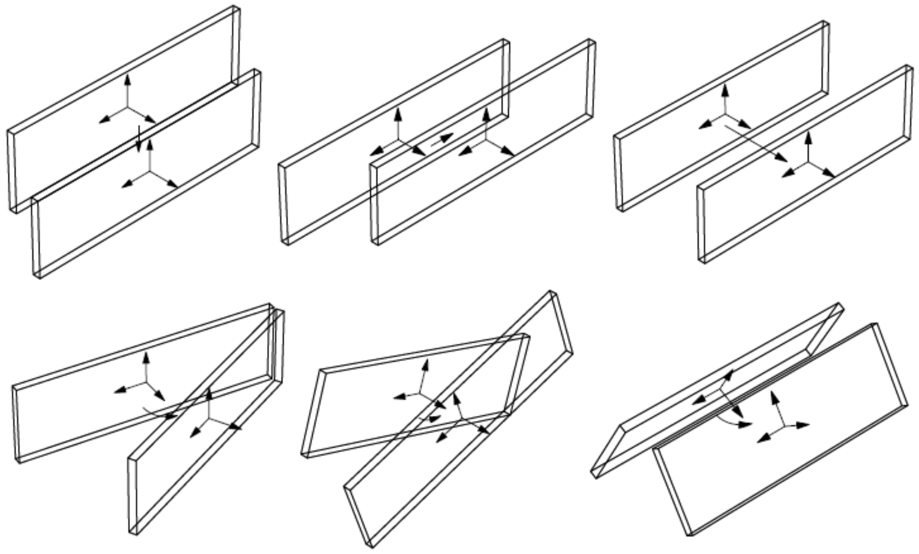


Figure 2.2: *The six degrees of freedom per base pair step: three translational and three rotational*

[*] A dinucleotide is a pair of adjacent nucleotides on one of the DNA strands, e.g. GC

2.2 Monte Carlo estimation of elastic energy

We estimate the average elastic energy of a DNA molecule fully or partially wrapped around a nucleosome core particle by the Monte Carlo method. Given an initial arbitrary configuration, rigid plate (base pairs) are randomly selected, and random movements or rotations are proposed. The standard Metropolis algorithm is used to accept or reject proposed movements, based on the resulting energy given by eq. 2.1 due to the new base pair positions. The system is first brought towards equilibrium by performing many such randomised steps, mimicking the effect of thermal fluctuation. The average energy of the conformation is then estimated by averaging over a randomised sample.

The Monte Carlo-estimated energy of the DNA molecule includes the elastic energy E^E as desired, however it also includes the kinetic energy, which we must discount. Equipartition theorem tells us that $\frac{1}{2}$kT energy is associated with each quadratic d.o.f. for a system in equilibrium. A free 147bp molecule of D.N.A. has $N_{\text{steps}} = 146$ base pair steps, each with 6 associated d.o.f, each therefore contributing $6 * \frac{1}{2} = 3kT$ of kinetic energy. However in nucleosomal DNA, for each phosphate-histone bond, we constrain one base pair step, and so remove a set of d.o.f.. The kinetic energy therefore depends on how many of the 28 possible phosphate-histone bonds are realised:

$$E^K_{\text{bonds}} = 6 \times \frac{1}{2\beta} \left(N_{\text{steps}} - n_{\text{bonds}} \right) \tag{2.2}$$

where $N_{\text{steps}} = 146$ base pair steps, and β is the inverse temperature at which the simulation is performed. Since the average elastic energy is independent of temperature, and we are removing the kinetic energy, we are free to choose the temperature. Following [8], all simulations were carried out at the low temperature of $T = \frac{1}{3}T_R \sim 100\,\text{K}$, or $\beta = 3$. We choose this as a stable average energy is reached quickly, and so simulations take less time.

As shown in [17], in nucleosomal breathing the DNA unpeels from either end in 14 steps. In each step a binding site is exposed and 2 phosphate-histone bonds are released. For a partially wrapped nucleosome, with i sites exposed from the left and j sites exposed from the right, the kinetic energy is then given by

$$E^K_{ij} = 146 - 2 \cdot (14 - i - j) \tag{2.3}$$

Subtracting the kinetic energy E^K_{ij} from the Monte Carlo estimate E^{MC}

gives the elastic energy E^E_{ij} of the bent DNA molecule

$$E^E_{ij} = E^{MC}_{ij} - E^K_{ij} \tag{2.4}$$

And finally the effective adsorption energy is the sum of the elastic energy and the adsorption energy

$$E_{ij} = E^A_{ij} + E^E_{ij} \tag{2.5}$$

We should note here that the energetics of a thermodynamic system are actually determined by the free energy, $G = U + PV - TS$, where U is internal energy, P is pressure, V is volume, T is temperature and S is entropy. The experimental assays we will examine later speak entirely in terms of the free energy. We argue that other terms than the internal energy can be neglected, for our purposes, as our breathing model will be entirely reliant only upon differences in energy between states. In the stable experimental condition, P and V are constant, so the PV term is unchanging between states; and we assume that the entropic term, though changing between unwrapping states as more DNA is free, is negligible.

2.2.1 Nucleosome positioning

The nucleosome typically wraps about 146 base pairs, meaning that a range of positions are possible for longer sequences. However, not all positions are observed. Analysis of mapped nucleosome positions both *in vitro* and *in vivo* have uncovered a number of 'sequence rules', such as avoidance of poly-A tracts, a strong preference for TT, AA and TA dinucleotides to be positioned where the minor groove faces the octamer, a preference for GC dinucleotides to be positioned where the minor groove faces away from the octamer, and consequently a roughly 10bp periodicity in likeliness of occupation as the minor groove rotates toward and away from the octamer [8]. Where the minor groove faces the octamer, the DNA is most strongly bent, hence favouring the flexible dinucleotides TT, AA and TA, and disfavouring the more rigid GC.

We are able to replicate these rules and predict nucleosome position with some success using our Monte Carlo-derived measurements of the DNA's elastic energy. We predict the preferred nucleosome positions as those which minimise the elastic energy. The accuracy of the prediction has been measured: it was found that 60% of experimentally mapped nucleosomes on yeast chromosome I fall within ±1bp of a minima in the energy [6]. The resulting energy landscapes calculated for long sequences also displays the 10bp periodicity, as seen for example in the nucleosome energy landscape produced for the full 256bp 5S sequence, shown in figure 2.3. The two positions found in vitro for the 5S sequence are indicated, each 1bp away from minima. However, in this typical case, the global minimum does not successfully predict the nucleosome position. The nucleosome position prediction does have limited success, perhaps due to the coarse-grained nature of the model.

Nucleosomes reconstituted via salt dialysis in vitro are known to form tetramer-first: H3 and H4 histones together form a tetramer, which bonds to the DNA molecule before the H1 and H2 histones complete the octamer. Consequently, for *in vitro* studies such as the restriction enzyme assays analysed in this thesis, the tetramer energy landscape also informs the nucleosome position and should therefore also be considered.

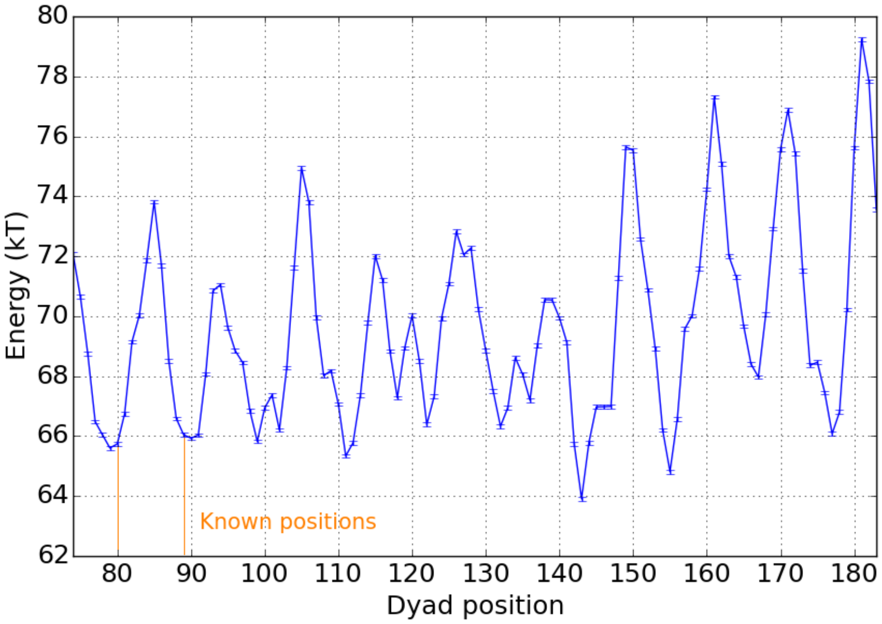


Figure 2.3: *Energy landscape calculated for full 256bp 5S sequence given in [21]. The 5S sequence is known to form nucleosomes on dyad positions 80 and 89, in each case close to a minimum, but not the global minimum.*

2.3 Breathing model

We use a simple statistical Boltzmann model for the transient unbinding and exposure of each of the 14 binding sites. We define a set of possible configurations labelled (i, j) for the nucleosome with *i* binding sites unbound from the left and *j* from the right, with at least one site still bound (Note that choosing the arguably realistic assumption that one site must remain bound is unimportant, as the states which are neglected have negligible probability.) The effective adsorption energy of each configuration determines its probability of occurrence, using the standard Boltzmann weight.

The experiments we will examine rely on enzymes binding to specific DNA sequences - their ’recognition sequence’. Due to their finite size, enzymes may need extra adjacent DNA to be unbound for there to be room for access. It has been shown [22] that this would strongly affect the exposure profiles observed in the restriction enzyme experiments. In principle the different sizes and geometry of enzymes could affect the extra exposure required, however here we make the simplifying approximation that enzymes require the same amount of extra DNA to be unbound.

In our model, the probability that a given site is accessible to an enzyme is the net probability of all configurations in which it is unbound, and there are Δ extra sites unbound on each side, to provide the enzyme the access it requires.

The probability of each configuration C_{ij} is given by the standard Boltzmann statistical weight calculated using its effective adsorption energy E_{ij}

$$C_{ij} = \frac{1}{Z} \exp\left(\frac{E_{ij}}{kT}\right) \tag{2.6}$$

where Z is the partition function over all configurations, in which at least one site remains bounds:

$$Z = \sum_{i=0,j=0,i+j<14} \exp\left(\frac{E_{ij}}{kT}\right) \tag{2.7}$$

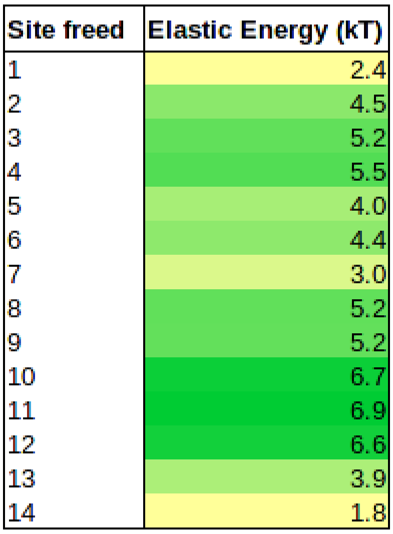
The probability that a site k is accessible to an enzyme P_k is then the sum of the probabilities of all configurations in which it is unbound (and there is still at least one site bound), and there are Δ extra sites unbound on each side. It is stated here in two terms corresponding to unwrapping from the left and right, respectively.

$$P_k = \sum_{i \geq k+\Delta, i+j<14} C_{ij} + \sum_{j>14-k-\Delta, i+j<14} C_{ij} \tag{2.8}$$

As we estimate the elastic energy, and assume the adsorption energy is equally distributed, our model has two free parameters: the adsorption energy E^A, and Δ, the number of extra unbound sites required.

2.3.1 Example breathing profiles

The widely studied artificial 601 sequence is a natural choice for calculating an example breathing profile. The 601 sequence is known to unwrap asymmetrically, preferentially from its right-hand side, and it has been suggested that this is due to that side being stiffer, i.e. that there is more elastic energy stored there [23]. We estimated the elastic energy of the sequence using Monte Carlo simulations, in all its partially wrapped states, and found indeed a significantly higher amount of elastic energy stored on the right hand side. Table 2.4 shows more elastic energy is stored in the second turn, roughly in the third quarter, as suggested in [23].



Site freed	Elastic Energy (kT)
1	2.4
2	4.5
3	5.2
4	5.5
5	4.0
6	4.4
7	3.0
8	5.2
9	5.2
10	6.7
11	6.9
12	6.6
13	3.9
14	1.8

Figure 2.4: *Elastic energy released in the 601 sequence as it unpeels from the left-hand 5′ end, deduced from Monte Carlo estimated partially wrapped energies.*

The elastic energy of nucleosomal DNA results from many interactions between the nucleotides and protein, including the nucleotide-nucleotide interactions captured in our model. As a result it is sequence dependent, and so unevenly spread along the molecule. It is useful to compare the 601 breathing profile with that of an artificial 'ideal' DNA molecule, for which the same total elastic energy is *evenly* distributed. We calculated the breathing profiles, as shown in figure 2.5, using an ansatz starting assumption for the adsorption energy, and with Δ=0 to begin with; we will address enzyme accessibility in section 2.3.3.

For the nucleosome to remain stably wrapped, it has been argued [22] that the adsorption energy should be greater in magnitude than the elastic

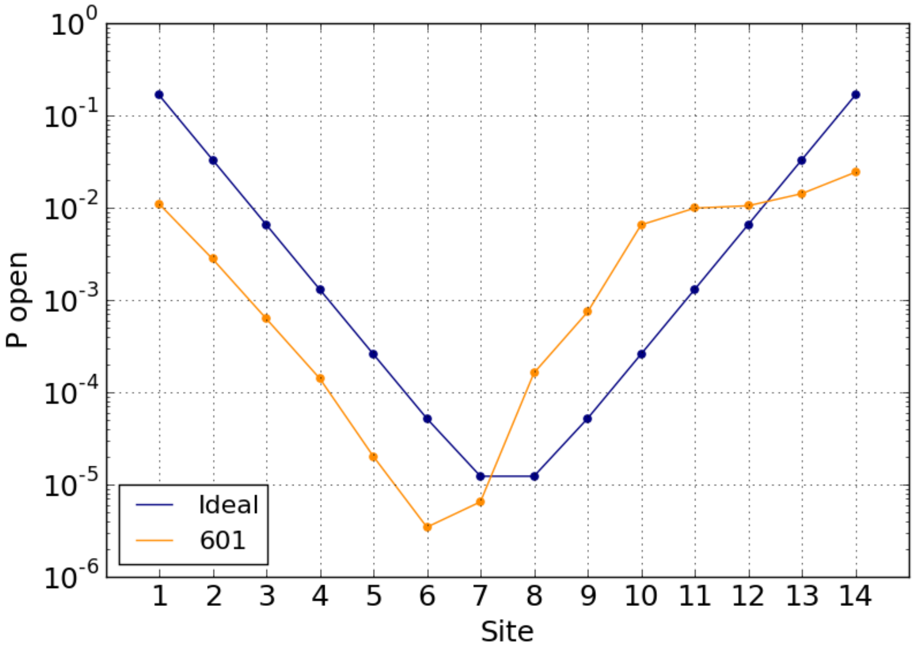


Figure 2.5: *601 and ideal breathing profiles calculated using equation 2.3. In each case, the total elastic energy is 65.5kT, the total adsorption energy is -91kT, and so the resulting total effective adsorption energy is -25.7kT. Note that an ansatz has been used for the adsorption energy used to calculate the plot: this plot merely illustrates the impact of sequence upon breathing. Note the logarithmic scale.*

energy by at least ∼1kT per binding site. This means that the total effective adsorption energy decreases in magnitude with each site opening, becoming less negative; ∼1kT energy must be 'paid' to open each site. In ideal breathing this cost is the same for each site opening, leading to a uniform exponential decrease in the Boltzmann weight probability as the nucleosome unpeels from one side. As shown in figure 2.5, the resulting site exposure probability decays exponentially toward the two central sites (7 and 8) which are equally least exposed.

The breathing profile of the 601 sequence is similar, but the uneven spread of elastic energy has some significant impacts: an asymmetry in breathing probability, with the right hand 3′ end opening much easier than the left hand 5′ end; a shift of the location of the most protected site(s); and a protection of the entire left hand side. In the next section we attempt to understand how the unevenly distributed elastic energy leads to these features.

2.3.2 Effect of energies on breathing

The Boltzmann weighted probability of a given configuration is a function of it total effective adsorption energy, the sum of the (negative) adsorption energy of the DNA-protein complex and the (positive) elastic energy of the bent DNA molecule. This means the probability of a configuration exponentially decreases as the total effective adsorption energy decreases (in magnitude), as shown in unpeeling toward the centre in figure 2.5. However, as the figure shows, breathing profiles of real sequences are not so straightforward, and are strongly affected by the magnitudes and distributions of the energies along the molecule. In this section we examine this in the simplified ideal breathing profile. It should be noted that as the Boltzmann weight dependence on the adsorption and elastic energies is equal, but opposite, an increase in elastic energy is equivalent to a decrease in adsorption energy; stronger adsorption has the same effect as a higher DNA flexibility.

A higher adsorption energy (or lower elastic energy) per site uniformly across the molecule decreases the accessibility of all sites, exampled with

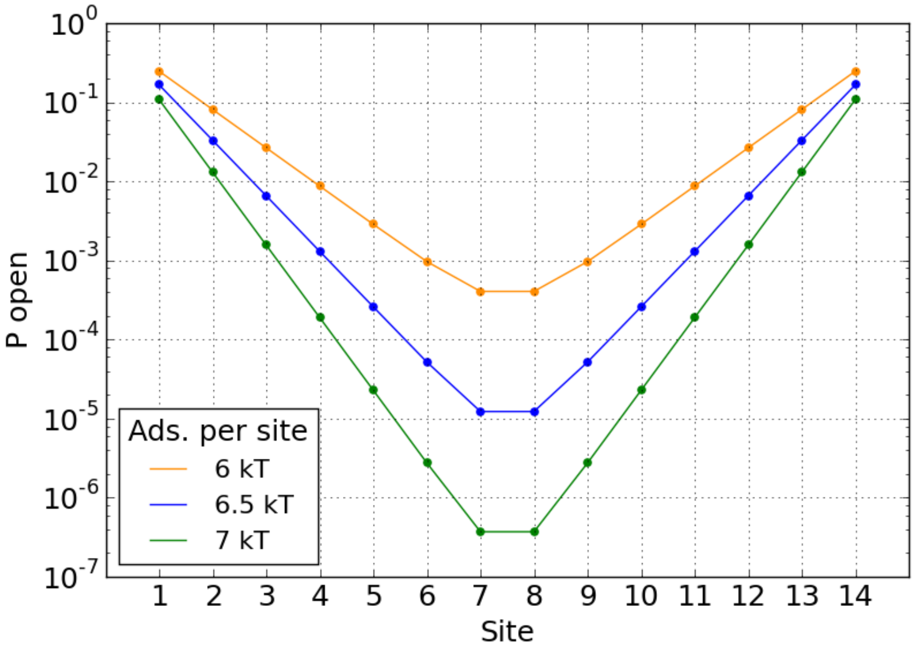


Figure 2.6: Ideal breathing profile with different values for the adsorption energy per site. In each case, there is 65.5kT elastic energy stored equally between all sites.

the 'ideal' sequence in fig. 2.6, in which the adsorption energy per site is shifted by ±0.5 kT. The probabilities of configurations with fewer bound sites, and hence smaller (less negative) total effective adsorption energies, are more strongly affected, due to the exponential dependence in the Boltzmann weight. Since sites are opened sequentially, changes in accessibility of sites cascade and affect all further site accessibilities; the probability of a site being exposed in the centre depends on the energies of all the outer configurations as well as the inner configurations. This contributes further to the central sites being relatively the most sensitive to changes in the adsorption energies, as seen in figure 2.6.

The shape of the plot is significantly altered if either of the energies are unevenly distributed. Since exposure probabilities are exponential in the energy, small imbalances can cause large changes. For example, if the two central sites (7 and 8) are bound less strongly, or equivalently if the DNA there is stiffer than elsewhere, then the total effective adsorption energy is smaller (in magnitude), and we expect the central sites to be more accessible. Figure 2.7 shows the effect of a ±1 kT change to the energy of the

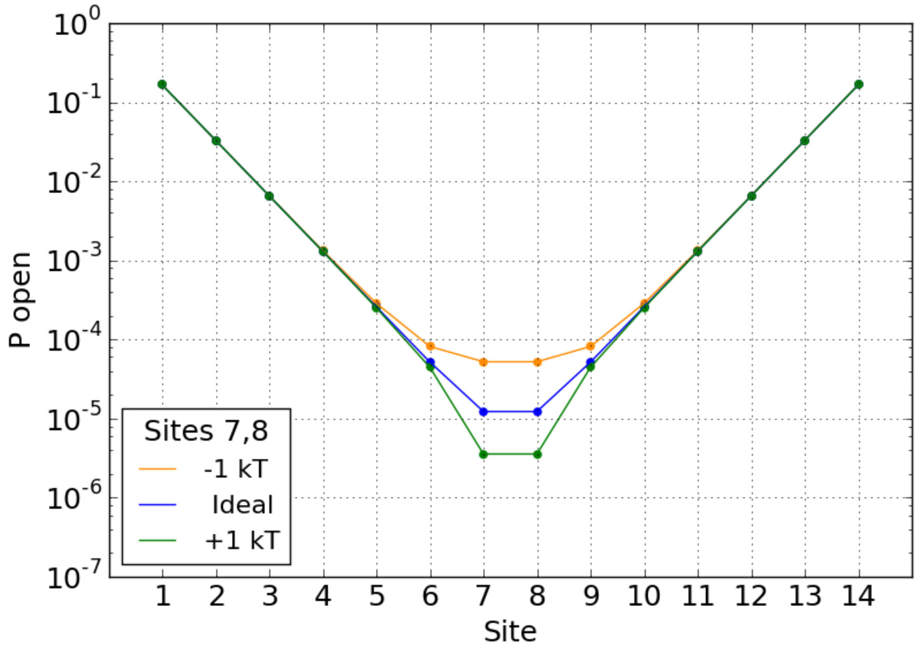


Figure 2.7: *Ideal breathing profile with increased/decreased effective adsorption energy for just the two central sites (7,8). 65.5 kT elastic energy and -91 kT adsorption energy spread evenly over the ideal sequence.*

two central sites of the 'ideal' sequence. We see that the immediately adjacent sites are also affected, since one way to access them is by the DNA unwrapping from the opposite side, through the affect central sites

Since the central sites are only exposed after one of the arms have entirely unwrapped, a change in the energies in the arm will affect the centre as well. Figure 2.8 shows that a 2 kT change to the energy of two sites (11 and 12) in the right-hand arm strongly affects the exposure probabilities of sites 6-12, and also shifts the position of most protected site(s).

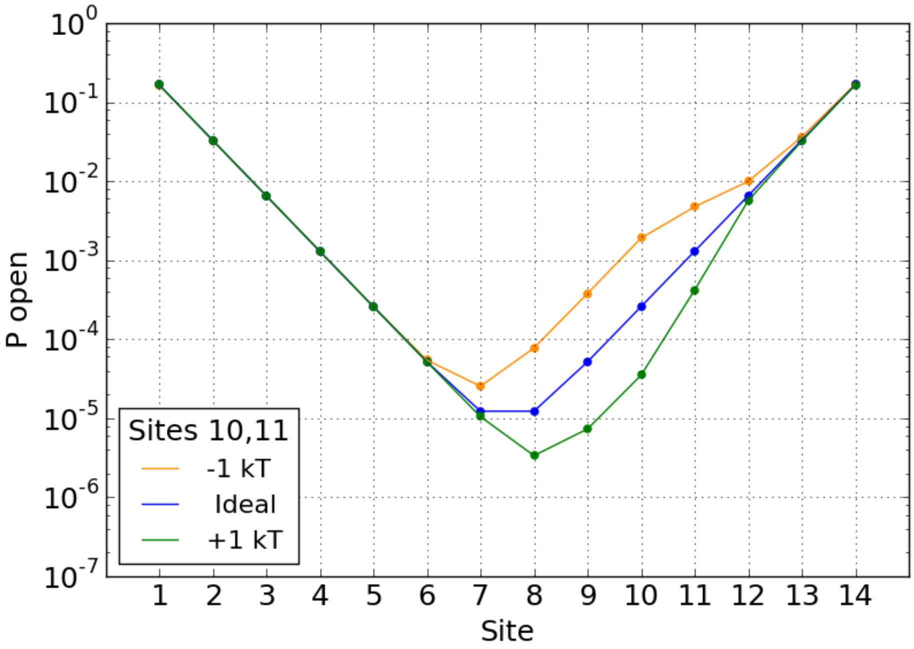


Figure 2.8: *Effect of increasing/decreasing effective adsorption energy for just two of the RH sites (11,12) of the 'ideal' sequence. 65.3 kT elastic energy and -91 kT adsorption energy spread evenly over ideal sequence.*

Finally, the effect of changing the energy of the most outer sites, for example the far right-hand sites (13,14), effects the entire breathing profile. For example, weaker outer bonds, or stiffer outer DNA, decreases the effective adsorption energy and makes those sites more accessible. This effect cascades down the entire breathing profile, as all the other sites are accessed through this one.

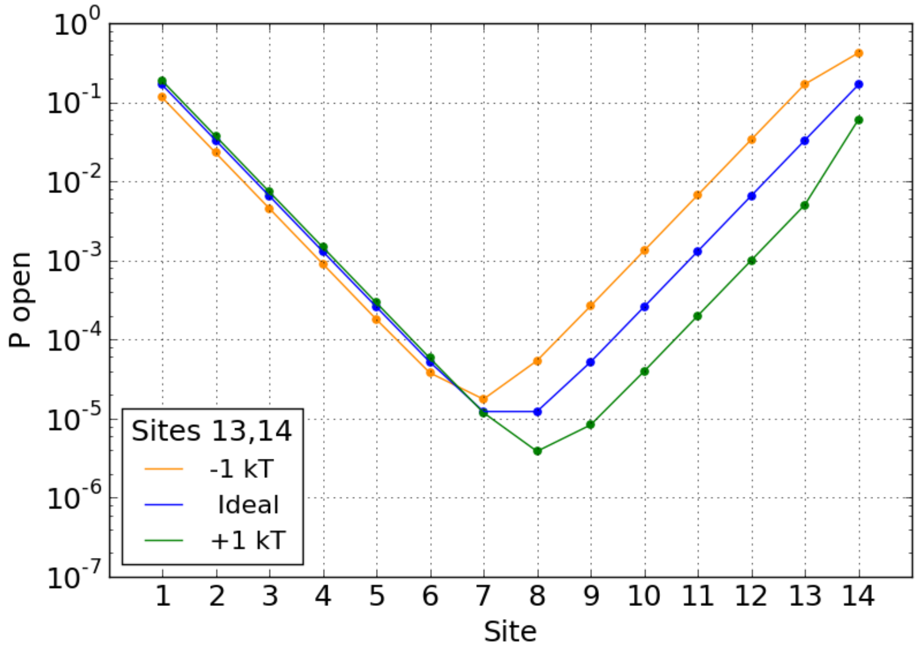


Figure 2.9: *Effect of increasing/decreasing effective adsorption energy for just two of the far RH sites (13,14) of the 'ideal' sequence. 65.3 kT elastic energy and -91 kT adsorption energy spread evenly over ideal sequence.*

In summary, it seems we can understand the 601 breathing plot from the unequal distribution of energies shown in figure 2.4. Figure 2.7 indicates that the smaller elastic energies of sites 5-7 (and so more flexible DNA) is responsible for the decrease in accessibility there, and contributes to the shift of the most protected site away from the middle. Figures 2.8 and 2.9 indicate that the higher elastic energy on the right hand side of the 601 sequence (and so smaller in magnitude effective adsorption energy) causes the increased accessibility in sites 8-12, the 'flattening' of the plot from sites 10-14, and also contributes to the shift of the most protected site to the left.

The 'flattening' of the plot on the right hand side - the almost equal accessibility of sites 10-14 - can be better understood by examining the unequal energy cost of opening each site. For a stable nucleosome, the total adsorption energy should be greater than the total elastic energy - however this need not be true at all parts of the molecule. With the ansatz assumption we made of 6.5kT for the adsorption energy per site, we find that actually the elastic energy is greater for the third quarter of the 601 sequence. This suggests that no energy cost is required to unpeel this part:

the elastic energy is enough. Figure 2.10 shows the cumulative energy cost unpeeling the 601 sequence from the right and left sides, and shows the dip in cumulative energy cost, which corresponds to a raise in the total effective adsorption energy (in magnitude). This means that, due to the stiffness of the third quarter, unpeeling it entirely leads to a more stable molecule, leading to the 'flattening' of the breathing profile. Whether this is the case for the real 601 sequence depends on the magnitude and distribution of the adsorption energies.

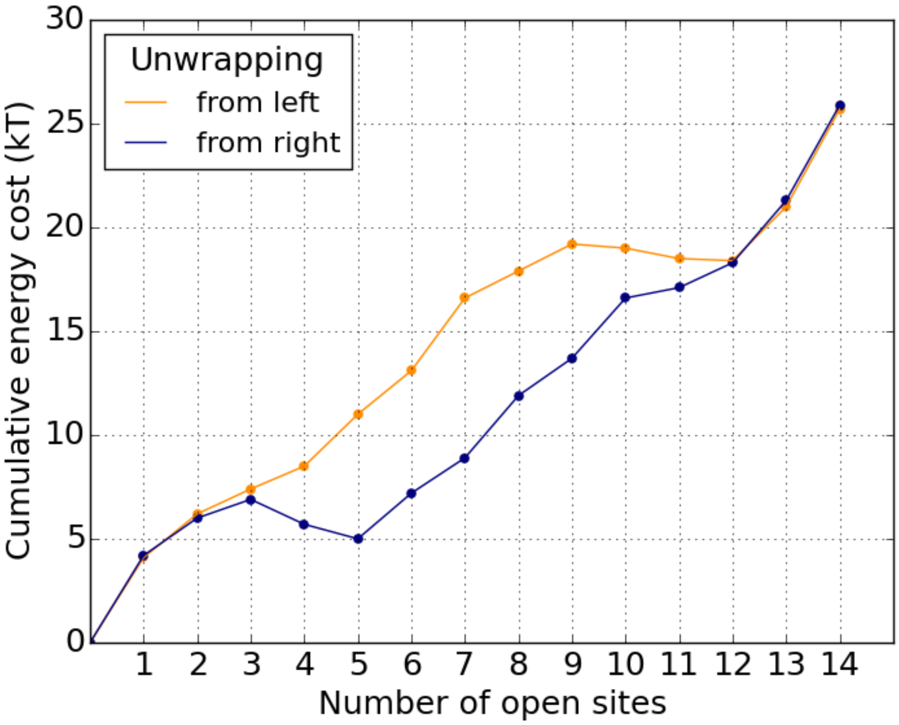


Figure 2.10: *Cumulative energy cost of unpeeling the 601 sequence from either the left or right.*

2.3.3 Enzyme accessibility

In section 2.3 we defined the free parameter Δ in the breathing model: the extra number of adjacent sites required to be open for a site for it be sterically accessible by enzymes. A theoretical study [22] has argued that 30 base pairs must be unbound either side of a site for it to be accessible, which corresponds roughly to $\Delta = 3$, since there are ~ 10 base pairs bound between each site. We will fit this parameter to the experimental data in chapter 3, but first here we examine its impact upon the example ideal and 601 breathing profiles.

Figure 2.11 shows that a higher Δ for the ’ideal’ sequence decreases all site accessibilities, exponentially more toward the centre sites (note the scale). This can be understood by considering the two terms that make up the accessibility in equation 2.3: the contributions to accessibility due to unwrapping from either of the ends. The linear decrease of the total effective adsorption energy as the nucleosome unwraps results in exponentially smaller probabilities for configurations with fewer sites. This means that one term dominates equation 2.3 for most sites - all but the central sites for which the terms are equally small. For example, the accessibility of site 2 is dominated by unwrapping from the left; the possibility of unwrapping from the right is negligible. So when delta increases, we should expect the accessibility of each site to shift to that of the site inward of it, as it is simply the exposure of that more inner site which then determines the accessibility. In the case of the ideal sequence, the evenly distributed

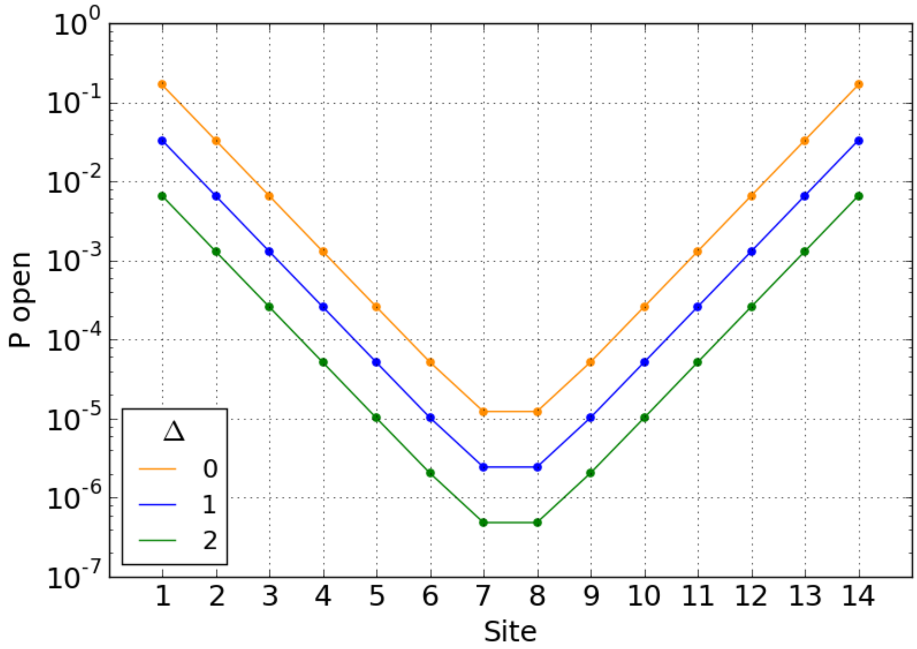


Figure 2.11: Breathing profile for the ’ideal’ DNA molecule at different values of Δ, the extra number of open sites required either side for an enzyme to bind to a site

energy, and therefore linearly decreasing total energy, leads to a uniformly exponential decrease in accessibility. However, this will not be the case with unevenly distributed energy in real sequences.

Figure 2.12 shows that the effect of increasing delta on the 601 sequence is similar to that of the ideal sequence, except for one distinctive feature: the flattening of the trough of the plot; when $\Delta \geq 3$, a small set of inner sites becomes all almost equally accessible. This is not as straightforward to understand, so again we consider the ideal profile. When $\Delta \geq 7$ (not shown here), exactly equal flattening occurs on the trough of the ideal profile. This can be understood by again considering the two terms in equation 2.3. As Δ increases, the accessibility of a site shifts to the accessibility of the site inward of it, as described already. For the ideal breathing profile, once $\Delta = 6$, the accessibilities of the two central sites have therefore been shifted to that of the outer sites. Increasing Δ any further therefore has no effect, and leads to flattening.

However, as the flattening occurs at much lower Δ for the 601 profile, there must be another explanation. Considering again the decreased energy cost to unwrap from the right hand side, shown in figure 2.10, it

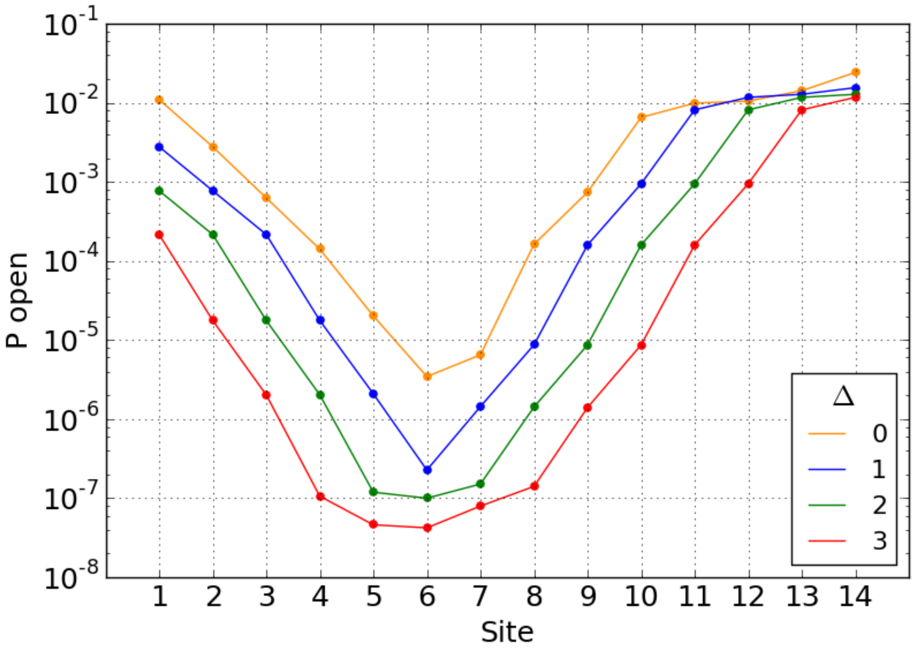


Figure 2.12: *Breathing profile for the 601 sequence at different values of Δ, the extra number of open sites required either side for an enzyme to bind to a site*

seems that in this case, unwrapping from the right makes a non-negligible contribution to the accessibility of the left hand sites. This, together with variance in energy difference between successively unwrapped configurations from either side leads to the observed flattening. This conclusion examples well the potentially strong sequence dependence of breathing profiles.

2.3.4 Approximated breathing model

There are 105 configurations (i, j) with one site still bound, leading to lengthy computation time for each sequence. An approximation to save time is to only consider the 26 configurations in which only one or the other side is unwrapped at a time, i.e. keeping $i = 0$ fixed whilst $j > 0$ and vice versa. This should make only a small impact on the probabilities, as most of the configurations neglected are quite unlikely. The simplified partition function will be:

$$Z' = \sum_{i=0}^{13} \exp\left(\frac{E_{i0}}{kT}\right) + \sum_{j=1}^{13} \exp\left(\frac{E_{0j}}{kT}\right) \tag{2.9}$$

The simplified probability of a site being open is then given, as before, by the sum of the Boltzmann weights of each state in which site k is reached opening either from the left or right, keeping the opposite side unopened.

$$P'_k = \sum_{i \geq k+\Delta} C_{i0} + \sum_{j > 14-k-\Delta} C_{0j} \tag{2.10}$$

In the approximated case, there are far fewer configurations, which means for any given site, there are fewer configurations in which it is accessible. Each configuration contributes a term to the probability of a site being open, so taking away configurations will decrease the predicted accessibility of a site - however as the normalization is also affected, its not immediately clear whether site accessibilities should go up or down. In the ideal approximated case, as shown in 2.13, we numerically find all exposure probabilities are underestimated, i.e. the entire plot is shifted down. The points are shifted $\sim$ 20%, however as the plot is logarithmic, this seemingly large error does not impact the plot qualitatively. This approximation error on real sequences is actually even less ($<$ 5%), as shown for the 601 sequence in the same figure, in which the approximate results are almost indistinguishable from the full results. We conclude that the

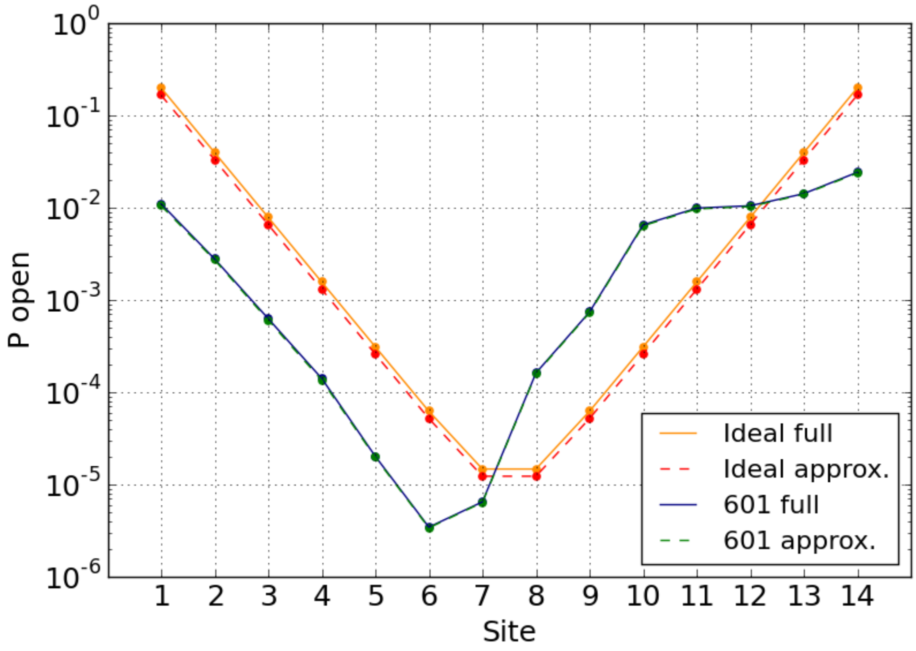


Figure 2.13: *Breathing plots for the 'ideal' and 601 sequences as calculated using the full and approximated models. The results calculated from the approximated model are all* $\sim$ 20% *lower. The average absolute error for the 601 sequence is* $\sim$ 2%

error is acceptable, and use the approximated method in the remainder of the paper.

Chapter 3

Results

Although there have been many experiments investigating nucleosomal breathing, as outlined in the introduction, so far none have been able to unveil its entire mechanics. The first experiments involving restriction enzymes are still the simplest to attempt to interpret, as they do not involve force-induced unwrapping, or FRET signals, and they also in single experiments probe large amounts of a sequence, and not just a limited section. In the following section we focus on interpreting the two restriction experiments separately - and then in conclusion together. For this section we assume that the adsorption energies are equally distributed. In section 3.2 we consider the results from other experiments, and that adsorption energy may be unequally distributed, in order to understand the restriction enzyme data better.

3.1 Restriction enzyme experiments

Restriction enzymes are proteins native to bacteria and archae which form part of a defence against foreign DNA: they bind to and cleave short (∼4bp) DNA sequences, termed restriction sites, which are specific to each enzyme. Restriction sites in the wrapped part of the DNA in a nucleosome are inaccessible, and become only transiently accessible during the dissociation in nucleosomal breathing. By incorporating restriction sites in a nucleosomal sequence, and exposing the sequence to the appropriate enzymes, the exposure of specific portions can be measured by monitoring the amount of appropriate cut pieces of DNA. To our knowledge, there are only two published reports of this type of experiment, [10] and [11]. The results of the two studies show that even the innermost sites on the DNA sequences are accessible, as shown below in fig. 3.3. Their results show a generally exponentially decay in accessibility toward the dyad, and also show that the entire 601.2 sequence has greatly reduced accessibility compared with the 5S sequence, as may be expected for the artificial high affinity sequence.

The limitations of the technique, as pointed out by the authors, must be kept in mind: the cleaving rates of the different enzymes may be dependent on their different sizes, the nature of their respective restriction sites and subsequent affinities, and the environment's temperature and ionic conditions, potentially impairing the reliability of direct comparisons. The authors use the same buffer for all assays, but vary the temperature 37 °C-65 °C dependent on the enzyme used. The authors address the concern that breathing may be temperature dependent by varying the temperature on a few of their assays. They do find a lack of correlation between temperature and accessibility (Figure 5, [10]), however their sample size is quite small. They do not address the concern that **enzymatic** accessibility may be temperature dependent; it is not clear how Δ, the number of adjacent free sites required for steric access, relates to the temperature.

A further problem is that relating an enzyme cleavage rate to a position within the nucleosome relies on there being a single stable nucleosome position, whereas for many sequences multiple positions are possible. The authors address this by selecting sequences with high-affinity for the protein octamer, which should result in a well-positioned nucleosome. However, it is unclear if they correctly identified the nucleosome position for the 601.2 sequence, as we will address later.

The authors pointed out that their *in vitro* measures of site exposure are likely affected by the absence of the linker histone H1 in their assays. As H1 is positioned outside the nucleosome interacting with and stabilising both arms of the DNA, it may strongly repress nucleosomal breathing. The authors stated they expected a quantitative effect only, however new understanding of the H1 'linker' histone challenges this. H1 histones are now known to be in constant exchange between nucleosomes, and only a fraction of nucleosomes are bound by a H1 at any time [3]. This leaves open the possibility of H1 histones only binding some nucleosomes, possibly in a sequence or post-translational modification dependent manner. More recently it has also become clearer that a large number of proteins may act similarly to the H1 histone, and that they compete and perhaps co-operate with each other, and restrict the DNA accessibility in differing ways [9]. In light of these developments, the nucleosomal accessibility observed in the assays we study here are not likely reflective even qualitatively of *in vivo* accessibilities, but they are still present the best data for understanding the underlying breathing mechanism itself.

3.1.1 Fitting the data

The results of the two restriction enzyme studies are reported per enzyme, and are only schematically mapped to the nucleosomal sequences. For each enzyme, we have identified which of the 14 direct DNA-protein binding sites that, upon opening, provides it access to its restriction site. For example, the restriction site for HindIII is found in base pairs 27-32 of the full 174bp 601.2 sequence, and so at base pairs 13-18 of the nucleosomal sequence, according to the nucleosome position reported by Widom. Table 5.1 in the appendix gives the positions of the individual site bonds, and shows that HindIII will be accessible after either the first 2 sites unbind from the left, or after 13 sites unbind from the right. By mapping each restriction site accessibility in this way, we can compare our theoretical predictions with the experimental data. However, there are two potential inadequacies of this approach that should be noted.

Firstly, some adjacent or close together restriction sites are predicted to be equally accessible as they are exposed after the same site openings. This may not be accurate, due to additional hydrogen bonds and/or for steric reasons; this is a limitation of our coarse-grained breathing model.

The second more subtle problem concerns restriction site exposure from one or the other side of the molecule unwrapping. In our breathing model, sites 1, 2, 3... 14 are accessible when either 1, 2, 3... 14 site(s) open from the left, or when 14, 13, 12... 1 site(s) open from the right. However some restriction sites can't be associated with just one site, as with HindIII above. For example, Rsa I is found in base pairs 87-90 of the full 174bp 601.2

sequence, and so at base pairs 73-76 of Widom's nucleosomal sequence. Because it's entirely between the 7th and 8th binding sites, it's accessible after either 7 sites unbind from the left, or 7 sites unbind from the right, and so cannot be associated with one site over the other. In the few cases where this problem occurs, whichever site is the closest fit is used.

To fit our model to the experimental data we use the method of least squares. We varied the two free parameters, adsorption strength per site, and extra adjacent sites Δ required for steric access. We should note that we used the square of the **log** of the accessibilities, so that the fit on the outer sites do not dominates the fit. The choice impacts only minimally upon the best fit absorption strength per site (generally lower by 0.1kT), and not at all on Δ.

3.1.2 The first study

In the first study, the 5S ribosomal RNA (rRNA) sequence of Lytechinus variegatus (the green sea urchin) was examined. The rRNA genes are some of the most highly conserved sequences across all three domains of life, and the 5S rRNA sequence is known to have a particularly high affinity for binding to the protein octamer, and for consequently having a well-positioned nucleosome. The 5S sequence does not naturally have many known restriction sites, so alterations must be made to incorporate them. Widom and Polach used a short 150bp part of the full 256bp sequence from [21], "restricted the locations of sequence changes to those regions of the sequence that are not essential for positioning" on just one half of the sequence, and spread the changes out over three constructs: construct (a) had 12 altered base pairs and contained 3 restriction sites; (b) had 15 altered base pairs and contained 2 sites; and (c) had 7 altered base pairs and contained 4 sites. They carefully ensured the nucleosome position on the three constructs was the same using gel electrophoresis and autoradiography and claimed "the histone octamers organize the DNA from positions 5 to 150 (bp)" and that due to the small length of the sequence "multiple positions of the octamer on the DNA are not anticipated".

Even if the nucleosome position is the same for the three constructs, the sequence changes could result in different elasticities stored along the molecules, or different affinities for the nucleosome, and so different breathing profiles, which would call into question the validity of combining the results of the three sequences, shown in figure 3.1. We took this into account when fitting out model to the data by calculating the breathing profiles for each of the constructs separately, and comparing to the appropriate subsets of experimental data, as shown in figure 3.2.

The Monte Carlo-estimated fully-wrapped elastic energies are slightly higher for the constructs. Whilst the total elastic energies differ by only ¡1kT, the change to the distribution of the elastic energy (± 0.1-0.7kT) does still effect the breathing dynamics, particularly on the right hand side, where most of the changes were made. Constructs (a), (b) and (c) have very similar breathing profiles, and are on average 45%, 8% and 30% more accessible than the original sequence. The differences between the breathing profiles are small but not negligible, and will be taken into account when interpreting Widom and Polach's results.

We mapped the restriction site accessibility results in figure 3.1 to the sites which expose them, and fit our model by varying the two free parameters: adsorption energy per site and Δ, the extra sites required to be

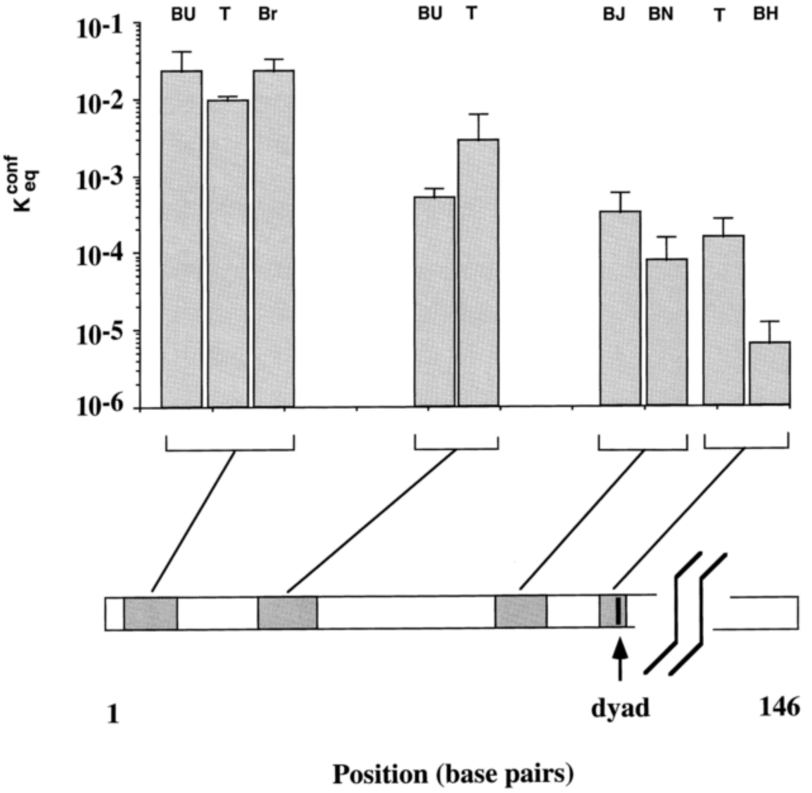


Figure 3.1: *Results of first restriction enzyme experiment in nucleosome breathing on the modified 5S sequences [10].* K_{eq}^{conf} *is the equilibrium constant for site exposure, the fraction of time a restriction site is accessible to its enzyme in equilibrium conditions. BU, T, Br etc. are the enzymes used, and their positions in the sequences are indicated. Note the logarithmic scale.*

exposed for steric access. The best fit to the limited data, by the method of least squares, results in an adsorption strength per site = 6.3kT, and surprisingly Δ=0.

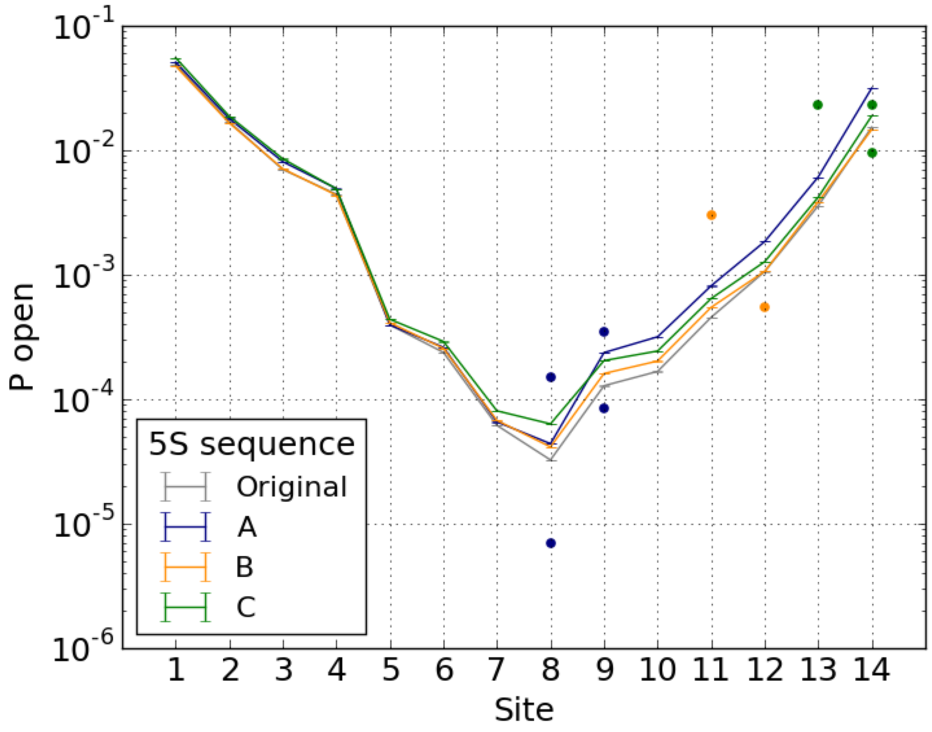


Figure 3.2: *Breathing profiles for the 5S sequence and the three constructs used by Widom and Polach, experimental (dots) and theoretical fitted predictions (lines). Fitted values of free parameters: adsorption strength per site = 6.3kT, and Δ=0. P open is the probability of a site being open at a given time. Note the logarithmic scale.*

3.1.3 The second study

In the second study, the artificial 601 sequence was used, the sequence found to have the greatest affinity for the histone octamer. For this experiment Widom and Anderson developed an altered sequence which they named 601.2, with 15 altered bases across the entire sequence, containing 12 restriction sites. In this study they probed the entire sequence, and they claimed the overall exposure pattern to be ″roughly symmetric about the mapped location of the nucleosomal center″. As figure 3.3 shows, the pattern does appear symmetric, however it should be noted that the Rsa I site at base pairs 73-76 occupies the nucleosomal centre, not the Taq I site, according to the authors reported nucleosome position.

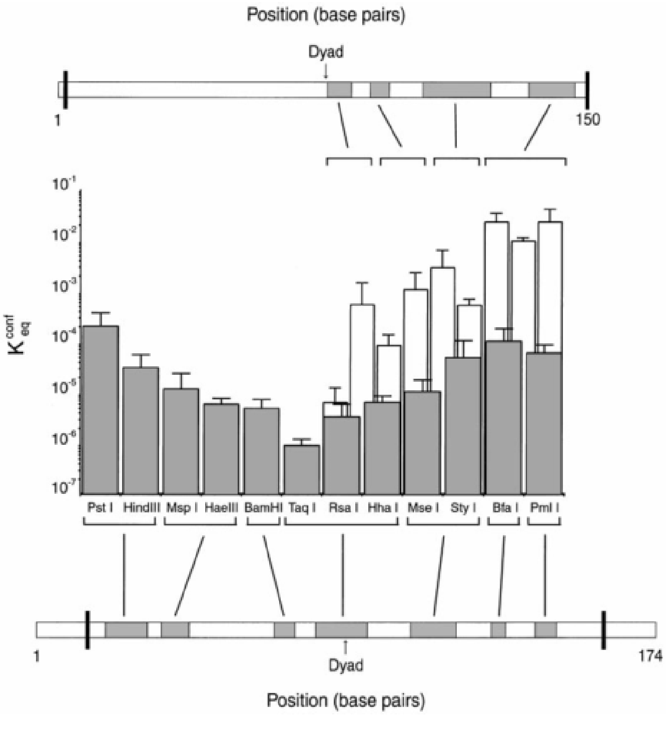


Figure 3.3: *Results of restriction enzyme experiments in nucleosome breathing, carried out by Widom et al [11]. The plot shows 601.2 sequence results (lower, grey) as well as previous 5S sequence results (upper, white). (Note that the 5S results are slightly misreported here for enzymes T and BJ, the second and fourth columns.)* K_{eq}^{conf} *is the equilibrium constant for site exposure, the fraction of time a restriction site is accessible to its enzyme in equilibrium conditions. Pst I, HindIII etc. are the enzymes used, and their positions along the two sequences are indicated. Note the logarithmic scale.*

The authors mapped the position of the 601.2 nucleosome using the enzyme exonuclease III, which digests DNA, removing nucleotides in a stepwise fashion. When the nucleosome is exposed to the enzyme, the non-nucleosomal DNA is quickly removed, but once the enzyme reaches the nucleosomal boundary there is a pause before transient breathing allows it further access (There are pauses at each of the 14 strongly bound sites). The authors radiolabelled one end of the sequence at a time and performed polyacrylamide gel electrophoresis (PAGE) on extracts to de-

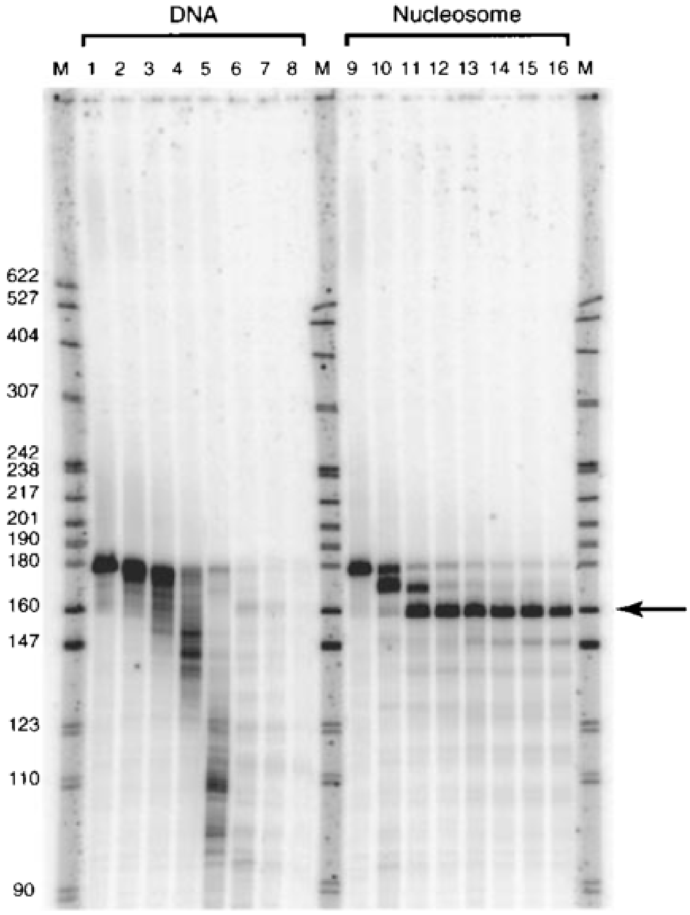


Figure 3.4: *Polyacrylamide gel electrophoresis results from [11] used to infer 601.2 nucleosome boundaries. Lanes marked M show positions of known DNA lengths, and are used as 'rulers'. Lanes 1-8 show bare DNA digestion over time, at 0,1,2,4,8,16,32 and 64 minutes. Lanes 9-16 show nucleosomal DNA digestion from the 3′ end over time at the same time points. The arrow inserted by the authors indicates where the pause in digestion they claim is at the 3′ nucleosomal boundary.*

termine the length of the intact nucleosomal DNA as a function of time. From the length of the DNA at the first pause in digestion from each end, they inferred the nucleosomal boundaries at base pairs 15-160, placing the dyad between base pairs 87 and 88 (hereafter referred to as the Widom position). However, in later studies involving the original 601 sequence, including one by one of the authors Widom, the dyad position at base pair 94 is widely reported (for example, see [24] or supplementary material in [23]). This raises the question of whether the sequence changes between 601 and 601.2 result in a $\sim$6bp shift of the nucleosome position, or rather that the 601.2 position was misreported due to the limited accuracy of the now dated technique. As we have found the breathing profile to be strongly sequence dependent, and therefore very sensitive to nucleosome position, it is crucial we have the correct position when fitting experimental data. We therefore considered whether the reported PAGE data may be consistent with a dyad positioned at base pair 94.

The authors show only the results of digestion from the 3′ end, shown in figure 3.4, and claim the pause marked by an arrow at approximately 160bp length places the 3′ nucleosome boundary at base pair 160. However, clearly visible in lanes 10 and 11, a slightly longer strand of DNA perhaps indicates another shorter pause, which the authors have not commented upon. The nucleosomal boundaries in 601 are known to be 21-168, meaning this first pause midway between length 160 and 180 base pairs is consistent with the dyad-94 position. Additionally, the 601 sequence is known to be asymmetric, and to open much easier from the 3′ end than the 5′ end [23]. This indicates the first pause in digestion from the 3′ end may be quite short, suggesting the arrow actually indicates the second pause in digestion. The accuracy of this technique must also be noted: lane 9 is supposed to show the original undigested 174bp sequence, but its length is indistinguishable from 180bp, so the accuracy is not better than $\pm$6bp. The authors did not show the result of digestion from the 5′ end, but reported an inferred boundary there at base pair 15. Although we cannot analyse those results, given that the known 601 nucleosome boundary is only $+$6bp at base pair 21, and the accuracy of the PAGE results is limited, it seems likely they may also be consistent with the dyad-94 position.

To further address this, we calculated the nucleosome and tetramer energy landscapes for both the 601 and 601.2 sequences as per section 2.2.1. These energy landscapes show the effective adsorption energy of the nucleosome as a function of dyad position, the locations of the minima being the most energetically stable positions.

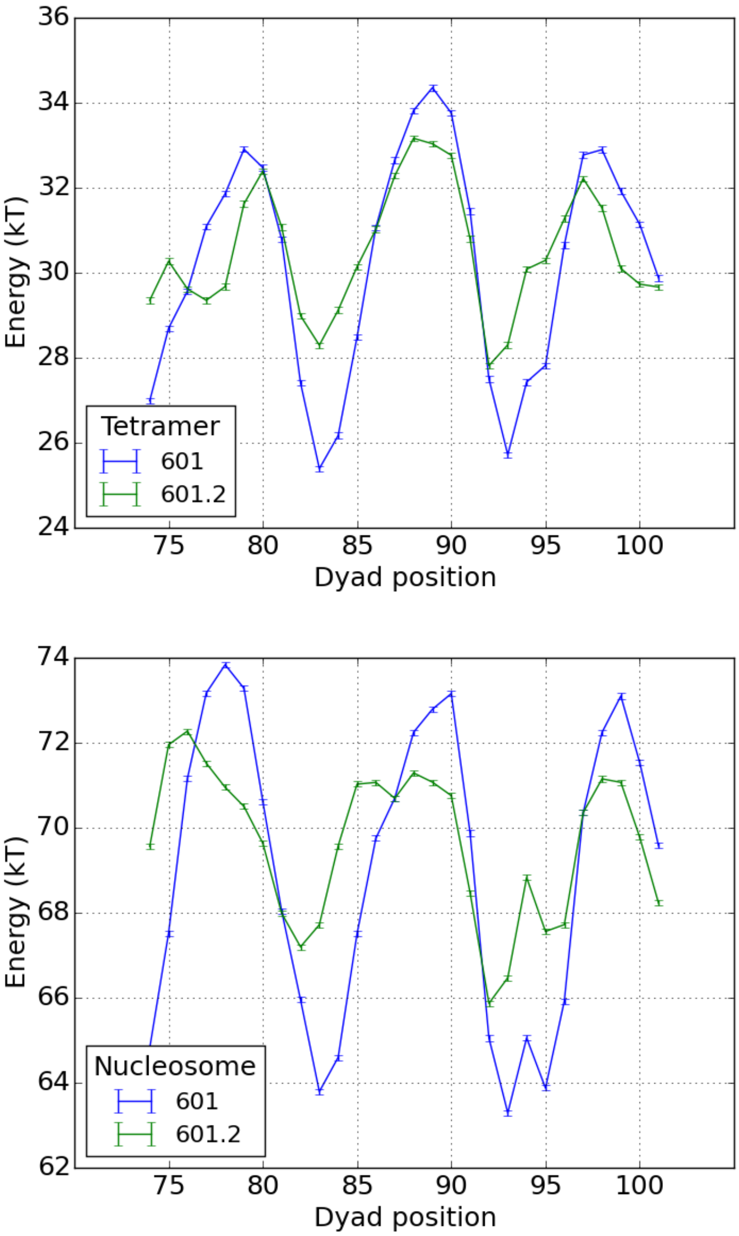


Figure 3.5: *The tetramer and nucleosome energy landscapes for the 601 and 601.2 sequences, calculated as per section 2.2.1. The position of nucleosome are predicted at the minima of the landscape, which for both the tetramer and nucleosome complexes are found at dyad positions 82/83 and 92/93, in accord with the expected 10bp periodicity. Note that Widom and Anderson report the position as between 87 and 88, here a maxima.*

We can see in figure 3.5 that the dyad-94 position is very close to minima for both the 601 and 601.2 nucleosome and tetramer energy landscapes - in most cases the global minimum. Also, all landscapes have a maximum near the Widom position, predicting it to actually be an unstable position.

We mapped the restriction site accessibility results in figure 3.3 to the sites which expose them, according to the two separate dyad positions, and in each case fit our theoretical results by varying two free parameters: adsorption strength per site and Δ, the extra sites required to be exposed for steric access. The best fit assuming the dyad-94 position is slightly better, as shown in figures 3.7 and 3.6. However, both fits result in high Δ values, the amount of extra adjacent open sites, in contrast to the fit on the 5S sequence. The original 601 nucleosome is known to breath asymmetrically, tending to open its RH side rather than its left - and this is reflected in the dyad-94 position breathing profile.

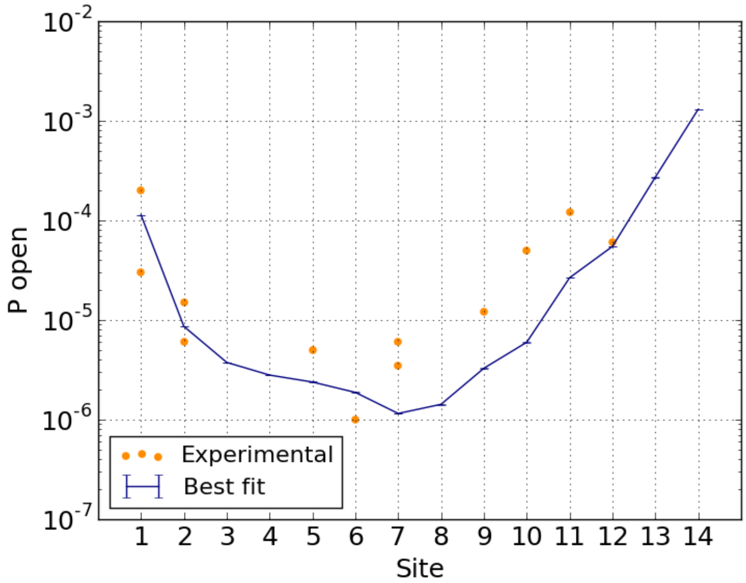


Figure 3.6: *601.2 breathing profile, experimental results against best fit model predictions, for the nucleosome position widely reported for 601 (dyad on base pair 94). with adsorption energy per site of 6.4kT and Δ=5.*

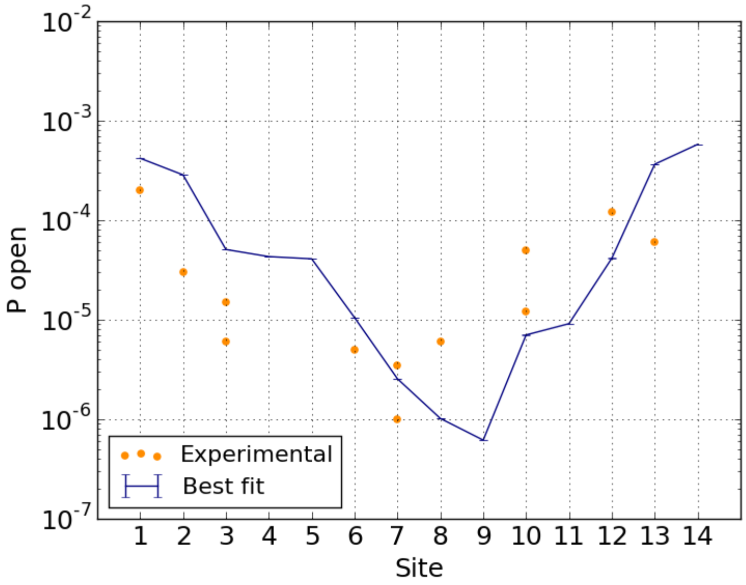


Figure 3.7: *601.2 breathing profile, experimental results against best fit model predictions, for nucleosome positioned as reported by Widom (dyad on base pair 88). The fitted parameters are: 6.7kT per site, and Δ=4.*

We conclude that it seems somewhere more likely the 601.2 position was misreported: the energy landscapes predict the dyad-94 position as stable, and Widom position as unstable; the fit assuming the dyad-94 position is slightly better; and the PAGE data seems to support either position - though perhaps they were not aware of the asymmetry of the 601/601.2 breathing profile.

Finally, we compared the 601.2 and original 601 breathing profiles. As the Monte Carlo-estimated fully wrapped elastic energy of the 601.2 sequence is $\approx 3.5kT$ lower than the 601 sequence, we expect it to be more accessible. We calculated the 601 breathing profile with the same best fit parameters of the 601.2 dyad-94 position. Figure 3.8 shows a large difference in the two profiles: the 601 sequence is almost everywhere roughly an order of magnitude less accessible than the 601.2 sequence. It is worth noting that the 15 nucleotide alterations in 601.2 do not cover the far right-hand edge, and that the accessibilities on this section are the least altered.

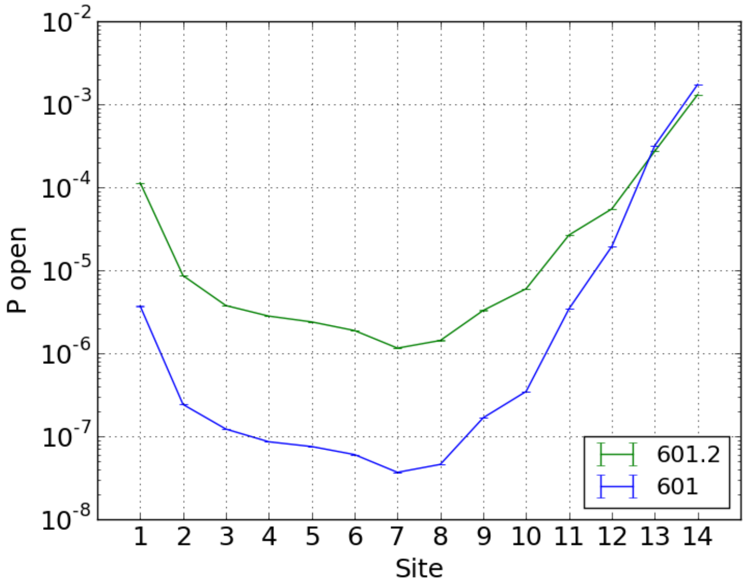


Figure 3.8: *601.2 and 601 breathing profiles compared, assuming dyad-94 nucleosome position. Calculated according to best fit parameters on 601.2 data: adsorption energy per site is 6.4kT and Δ=5.*

3.1.4 Fitting the experiments together

The artificial 601 sequence was the result of an experiment searching for the sequence with the highest affinity possible for the nucleosome, so we should expect its derivative, 601.2, to be less accessible than the 5S sequence, as indeed we find. However, fitting the two sequences 5S and 601.2 resulted in very different values for the free parameter Δ, the extra number of sites required open for enzyme access. This parameter may vary with different enzymes, or environmental conditions, but we do not expect it to be sequence-dependent. Accordingly we fit the results again, together, assuming Δ must be the same for both. We found that the fit on the 601.2 data dominated the collective fit, and consequently the same high value for Δ as before, as shown in figures 3.9 (dyad 94 position) and 3.10 (Widom position) comparing the two breathing profiles.

Previous analytic theoretical work [22] predicted the average effective adsorption energy per site to be $\gtrapprox$ 1kT, and the number of extra open adjacent base pairs required for enzyme access to be $\delta = 30 \pm 12bp$. We found the effective adsorption energy per site, assuming an equally distributed

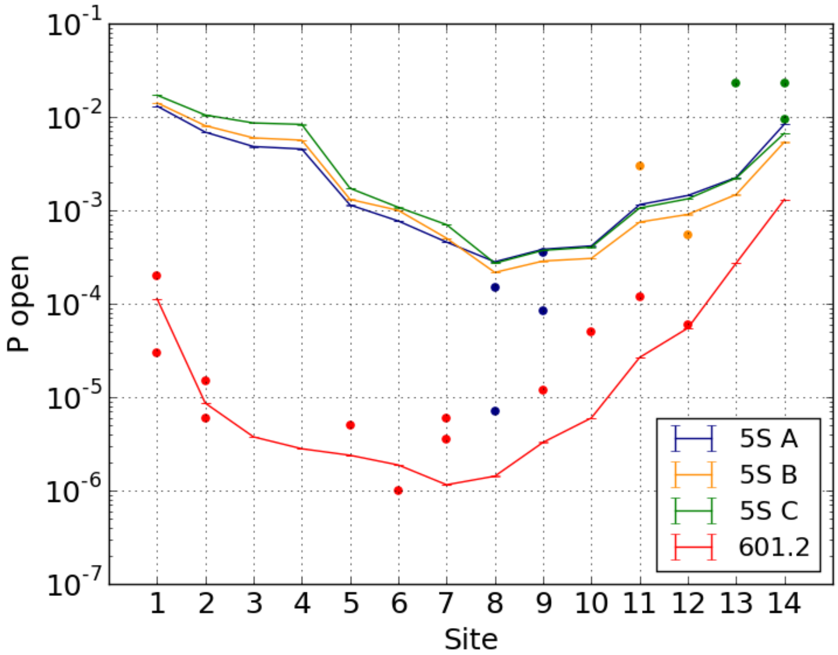


Figure 3.9: *601.2 (dyad 94 position) and 5S breathing breathing profiles resulting from collective best fit, assuming same value for Δ. Resulting free parameters, Δ=5, adsorption strengths: 601.2, 6.4kT and 5S, 5.7kT.*

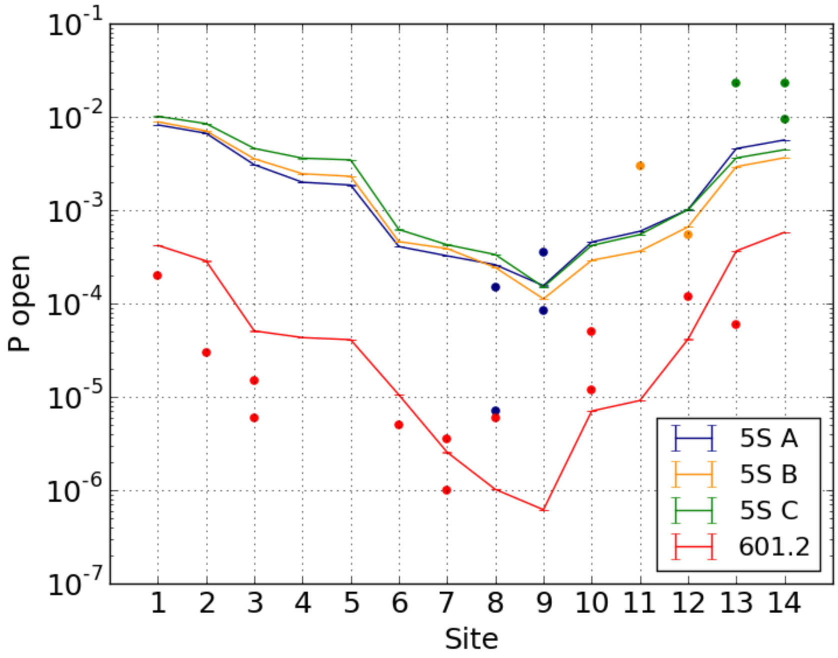


Figure 3.10: *601.2 (Widom position) and 5S breathing breathing profiles resulting from collective best fit, assuming same value for Δ. Resulting free parameters, Δ=4, adsorption strengths: 601.2, 6.7kT and 5S, 5.8kT.*

adsorption energy, to be ≈ 1.6kT for the 5S and (Widom position) 601.2 sequences, and ≈ 1.5kT for the (dyad 94) 601.2 sequence, all within the expected range. However, we found a surprisingly high values of $\Delta = 5$ (dyad 94 position) and $\Delta = 4$ (Widom position), outside and on the border of the analytic predictions, which were already surprisingly high. The implication of our fit is that a unexpectedly large amount of DNA needs to be unbound adjacent to a restriction site before an enzyme can bind to it. However, the limited success of our data fits could warrant reconsidering our assumptions.

3.2 Unequal adsorption distribution

Considering the large effect we found of the unequal distribution of energy on the breathing landscape in section 2.3.2, the assumption of equally distributed adsorption energy could be critical. If in fact adsorption energies are very different between sites, the current 2 parameter breathing model may not be even a good approximate fit to the data, leading to the potential overestimates of Δ. Here, we consider the assumption that the theoretical result of [22] that $\Delta = 3$ is correct, and that adsorption energies across sites are unequal.

In figure 3.11 we show the result of fixing Δ=3, and fitting just the (equal) adsorption energy per site to the experimental data. As in the previous fits, the 601.2 sequence adsorbs more strongly than the 5S. The total adsorption energy for the 5S is -82.6kT, compared to -91kT for the 601.2 sequence. From this figure we can infer what kind of unequal energy distribution would result in a better fit. The figure suggests the 5S sequence may actually be more strongly adsorbed toward the centre, and less so to-

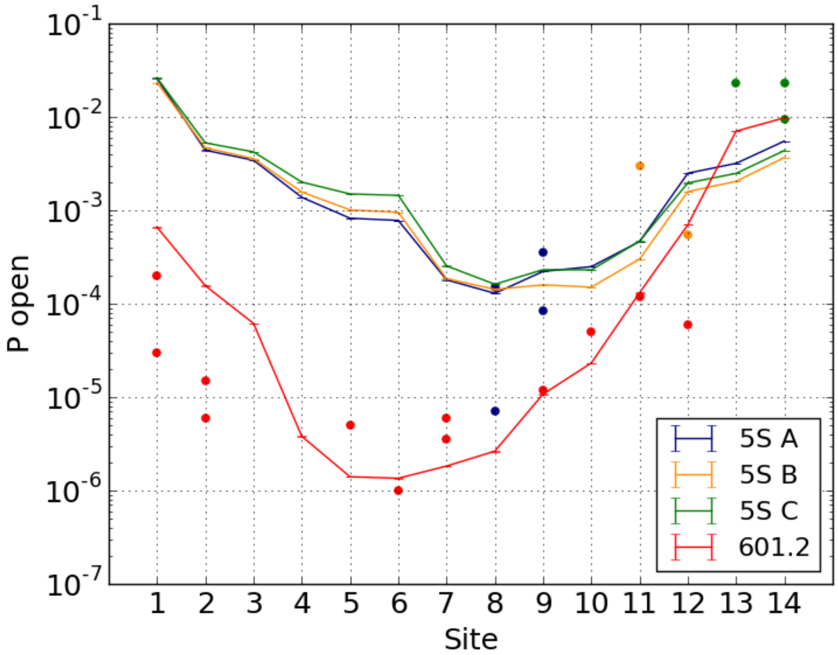


Figure 3.11: *601.2 (dyad 94 position) and 5S breathing breathing profiles, assuming $\Delta = 3$, as per [22] (so, a one parameter fit of adsorption energy per site). Resulting adsorption strengths: 601.2, 6.5kT and 5S, 5.9kT.*

ward its outer right-hand edge - so adsorption concentrated toward the centre. The figure does not suggest the same for the other sequence: it mainly suggests the left-hand side of the 601 sequence is more strongly adsorbed. It seems therefore unlikely that the distribution of adsorption energies are the same for the two sequences.

In [15] site-dependent adsorption energies were derived from the pauses in force induced unwrapping of the original 601 sequence in [17]. The resultant binding energies are higher toward the central sites, and low for the outer sites. In figures 2.7 and 2.9 we saw the effects of these binding energies on the ideal breathing profile: stronger central adsorption leads to a sharp decrease in the central sites accessibilities as well as a very small decrease on the immediately adjacent sites; and weakened outer sites actually raise the accessibilities of all sites. Oddly, this distribution may lead to a better fit to the 5S data - but not for the 601.2 data. Since the distribution was derived from 601 data, we might expect it to better explain the 601.2 data, if either. However, as it makes no sense to use the distribution for the 5S and not the 601.2 data, we have to disregard it. We conclude from the limited data that it may be the case that adsorption energy, just like elastic energy, is unequally distributed and sequence-dependent - though more robust and reliable experiments are required to settle the matter.

Chapter 4

Conclusion

Whilst the coarse-grained model presented here does seem useful in analysing the impact of energy distributions (and enzyme accessibility) upon nucleosome breathing, we've had limited success in fitting the model to the available data. It seems most likely this is due to our starting assumption of equal adsorption energy per site, however, it could also be due to flaws and inaccuracies in the experimental design of the restriction enzyme experiments we looked at: different enzymes were used to probe the different parts of the sequence; the enzymes were tested at different temperatures; different sequences with different energy distribution were used; and the indirect approach of using enzymes may also introduce unavoidable error.

Our model does concur with the enzyme restriction experiments in a few ways: the entire sequence is transiently exposed on free nucleosomes (with no linker histones); exposure decays roughly exponentially toward the centre; and the 601.2 sequence seems to adsorb much more strongly than the 5S sequence. Our nucleosome model and Monte Carlo estimates indicate that the different sequences have very different elastic energy distributions, which lead to very different breathing profiles. We have also found that it seems unlikely the adsorption energies are equal between the 14 sites, and also unlikely that they're the same for different sequences, however we have not been able to make strong conclusions on this, due to the limited reliability of the experiments.

We have found that the breathing profiles are strongly dependent on the distribution of energetics in the nucleosome, and that the energetics are (at least elastically) strongly dependent on sequence. Unfortunately, this conclusion has made it difficult to combine the results of multiple studies, as they either use different sequences, or different modifications of sequences - or sequences long enough for multiple nucleosome positions. If the underlying breathing mechanism is to be understood better, this calls for sets of experiments on the exact same sequences, and which are short enough that a single nucleosome position is guaranteed. However, we can conclude that the asymmetric breathing behaviour observed in [23] can be explained by the sequence-dependent energy distribution.

As outlined in the introduction, the breathing profiles presented in the *in vitro* restriction enzyme results may not be even qualitatively representative of the *in vivo* situation, due to the transient binding of linker histones, which bind near the entry/exit of the nucleosome, as illustrated in figure 1.1. However, it seems likely that the interplay of energy distribution and resultant asymmetric behaviour may still play important roles in breathing dynamics, and in chromatin organisation, and is worth pursuing further. Of particular note, the asymmetric breathing behaviour observed in the 601 and 601.2 sequences could play a role in mechanisms *in vivo* that selectively expose nucleosomal DNA sequences. A fruitful direction of research may lie in understanding the physical basis for how linker histones fit into this picture.

Chapter 5

Appendix

Site	Bond Locations (bp)
1	3, 7
2	15, 18
3	25, 30
4	35, 39
5	46, 50
6	56, 60
7	66, 70
8	77, 81
9	87, 91
10	97, 101
1	108, 112
12	117, 122
13	129, 132
14	140, 144

Table 5.1: *Locations of bound phosphates in the 14 sites in the nucleosome, as derived from crystallographic data in the supplemental material of [6]*

References

[1] Alberts et al. Molecular biology of the cell (4th edn.). 2006.

[2] Curt A Davey, David F Sargent, Karolin Luger, Armin W Maeder, and Timothy J Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *Journal of molecular biology*, 319(5):1097–1113, 2002.

[3] Tom Misteli, Akash Gunjan, Robert Hock, Michael Bustin, and David T Brown. Dynamic binding of histone H1 to chromatin in living cells. *Nature*, 408(6814):877–881, 2000.

[4] Wilma K Olson, Andrey A Gorin, Xiang-Jun Lu, Lynette M Hock, and Victor B Zhurkin. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19):11163–11168, 1998.

[5] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.

[6] Behrouz Eslami-Mossallam, Raoul D Schram, Marco Tompitak, John van Noort, and Helmut Schiessel. Multiplexing genetic and nucleosome positioning codes: A computational approach. *PloS one*, 11(6):e0156905, 2016.

[7] Lennart de Bruin, Marco Tompitak, Behrouz Eslami-Mossallam, and Helmut Schiessel. Why do nucleosomes unwrap asymmetrically? *The Journal of Physical Chemistry B*, 120(26):5855–5863, 2016.

[8] Behrouz Eslami-Mossallam, Helmut Schiessel, and John van Noort. Nucleosome dynamics: Sequence matters. *Advances in colloid and interface science*, 232:101–113, 2016.

[9] Jordanka Zlatanova, Corrine Seebart, and Miroslav Tomschik. The linker-protein network: control of nucleosomal DNA accessibility. *Trends in biochemical sciences*, 33(6):247–253, 2008.

[10] KJ Polach and J Widom. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *Journal of molecular biology*, 254(2):130–149, 1995.

[11] JD Anderson and J Widom. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *Journal of molecular biology*, 296(4):979–987, 2000.

[12] Gu Li, Marcia Levitus, Carlos Bustamante, and Jonathan Widom. Rapid spontaneous accessibility of nucleosomal DNA. *Nature structural & molecular biology*, 12(1):46–53, 2005.

[13] Hannah S Tims, Kaushik Gurunathan, Marcia Levitus, and Jonathan Widom. Dynamics of nucleosome invasion by DNA binding proteins. *Journal of molecular biology*, 411(2):430–448, 2011.

[14] H Schiessel. The nucleosome: A transparent, slippery, sticky and yet stable DNA-protein complex. *The European Physical Journal E*, 19(3):251–262, 2006.

[15] Arman Fathizadeh, Azim Berdy Besya, Mohammad Reza Ejtehadi, and Helmut Schiessel. Rigid-body molecular dynamics of DNA inside a nucleosome. *The European Physical Journal E*, 36(3):1–10, 2013.

[16] IM Kulić and Helmut Schiessel. Dna spools under tension. *Physical review letters*, 92(22):228101, 2004.

[17] Michael A Hall, Alla Shundrovsky, Lu Bai, Robert M Fulbright, John T Lis, and Michelle D Wang. High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nature structural & molecular biology*, 16(2):124–129, 2009.

[18] RE Dickerson. Definitions and nomenclature of nucleic acid structure components. *Nucleic acids research*, 17(5):1797–1803, 1989.

[19] Filip Lankaš, Jiří Šponer, Jörg Langowski, and Thomas E Cheatham. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophysical journal*, 85(5):2872–2883, 2003.

[20] Nils B Becker, Lars Wolff, and Ralf Everaers. Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic acids research*, 34(19):5638–5649, 2006.

[21] Robert T Simpson and Darrell W Stafford. Structural features of a phased nucleosome core particle. *Proceedings of the National Academy of Sciences*, 80(1):51–55, 1983.

[22] Peter Prinsen and Helmut Schiessel. Nucleosome stability and accessibility of its DNA to proteins. *Biochimie*, 92(12):1722–1728, 2010.

[23] Thuy TM Ngo, Qiucen Zhang, Ruobo Zhou, Jaya G Yodh, and Taekjip Ha. Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility. *Cell*, 160(6):1135–1144, 2015.

[24] Gu Li and Jonathan Widom. Nucleosomes facilitate their own invasion. *Nature structural & molecular biology*, 11(8):763–769, 2004.