The mechanical genome : inquiries into the mechanical function of genetic information

Tompitak, M.; Tompitak M.

MARCO TOMPITAK

THE MECHANICAL GENOME

THE MECHANICAL GENOME

INQUIRIES INTO THE MECHANICAL FUNCTION OF GENETIC INFORMATION

PROEFSCHRIFT

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof. mr. C. J. J. M. Stolker,

volgens besluit van het College van Promoties,

te verdedigen op woensdag 11 oktober 2017

klokke 10:00 uur

door

Marco Tompitak

geboren te Bangkok (Thailand)

in 1989

Promotores:	Prof. dr. H. Schiessel
	Prof. dr. G. T. Barkema (Universiteit Utrecht)
Promotiecommissie:	Dr. C. Vaillant (ENS de Lyon, Frankrijk)
	Prof. dr. G. J. L. Wuite (Vrije Universiteit Amsterdam)
	Prof. dr. E. R. Eliel
	Dr. D. J. Kraft
	Prof. dr. T. Schmidt
	Prof. dr. V. Vitelli

Cover design: Amar van Leeuwaarde and Marco Tompitak.

Casimir PhD Series 2017-26
ISBN 978-90-8593-310-6

An electronic version of this thesis can be found at
https://openaccess.leidenuniv.nl

Part of the work in this thesis was supported by the Netherlands Organisation for Scientific Research (NWO), as part of the Frontiers of Nanoscience program.

To my mom

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

RBP	Rigid Base Pair
WLC	Worm-Like Chain
MCMC	Markov Chain Monte Carlo
MMC	Mutation Monte Carlo
SELEX	Systematic Evolution of Ligands by EXponential enrichment
NDR	Nucleosome-Depleted Region
NAR	Nucleosome-Attracting Region
TSS	Transcription Start Site

1 INTRODUCTION

In the six decades since researchers first started to unlock the mysteries of DNA, a great deal has been learned about what could arguably be called the most important molecule to life on earth. Much also undoubtedly still remains to be discovered, and another sixty years will not likely be enough to exhaust DNA's secrets. DNA has inspired scientific inquiry in a wide range of disciplines due to its cornerstone role in the development, function, reproduction and evolution of all life as we know it. Its properties and behavior, in their myriad facets, are studied by biologists, chemists and physicists around the world.

The inquiries of any single individual, in collaboration with but a handful of others, can only hope to contribute to a small part of this panoply of scientific research. The studies presented in this thesis all probe questions on only one topic: the mechanical properties of DNA, and how they depend on the information stored in it.

The first thought likely to come to anyone's mind in association with DNA is the genetic information that it stores. It contains the blueprints for the development and functioning of all our cells, and we are the resulting product. The DNA is able to store all this information by encoding sequences of (pairs of[1]) nucleotides, four in all, usually represented by the letters A, T, C and G. A sequence of these letters encodes information by the order in which they are linked together, much like our digital systems employ sequences of ones and zeroes.

It is easy to forget that, on the physical level, a DNA sequence represents a real molecule, a physical object, with chemical and physical properties beyond the genetic information it encodes. Indeed, it must use physical entities to encode that information: DNA molecules are long and chain-like, made up of small segments (the nucleotides), which come in the four

1 DNA is a double-stranded molecule, with each strand containing essentially the same information because there is a fixed correspondence between the type of nucleotide on one strand and on the other. The pair of nucleotides is also commonly referred to as a base pair. The terminology does tend to be abused, and the word nucleotide may sometimes be encountered when actually a pair of them is meant. When the context is purely double-stranded DNA, as it is in this thesis, this is usually obvious, but it bears stressing.

varieties just mentioned. These segments differ not only in name but they are also distinct in their physical and chemical make-up.

Therefore, to say that DNA is a molecule is not entirely true. The word DNA in fact represents a class of molecules, which all differ in the nucleotide sequences they represent. A DNA molecule composed entirely of A nucleotides is not the same physical entity as one made up entirely of Cs.

Not surprisingly, then, the physical properties of a DNA molecule depend on the sequence it encodes for. One sequence may be easily bent, while another is very stiff. It may be intrinsically curved, or prefer to be very straight. These are the kinds of mechanical preferences that will be investigated in this thesis.

The elastic properties of DNA become important whenever the DNA needs to be deformed. Some sequences will resist a given deformation, while others may yield to it more easily. Examples of DNA deformation are not hard to find in nature, and occur abundantly in the interactions between DNA and various proteins. The best known and perhaps most important such interaction is that between DNA and histone octamers. In eukaryotic life, histone octamers provide little cylinders around which an organism's DNA is wrapped in order to compactify it. DNA and histones together form little spools called *nucleosomes*, which have been the object of much research. We shall look at them in more detail later in this introduction.

For the benefit of the reader, the remainder of this chapter will treat previous work[2] upon which the research in this thesis was built, in some more detail than will be provided in the chapters containing the original research. A reader familiar with the relevant literature may wish to skip this introduction.

1.1 MODELING DNA

In order to better understand DNA, we must model it; that at least is the approach of a theoretical physicist trying to contribute to biology. DNA is a complicated system that can be, and has been, modeled at various scales, ranging from all-atom models (see e.g. [4] for a review) to models that are coarse-grained and simplified at various levels (see e.g. [5]). Full atomistic models are computationally extremely costly and can only be applied to systems of limited size and simulations over limited time scales. For this

2 To some of which the author of this thesis has contributed: see [1–3].

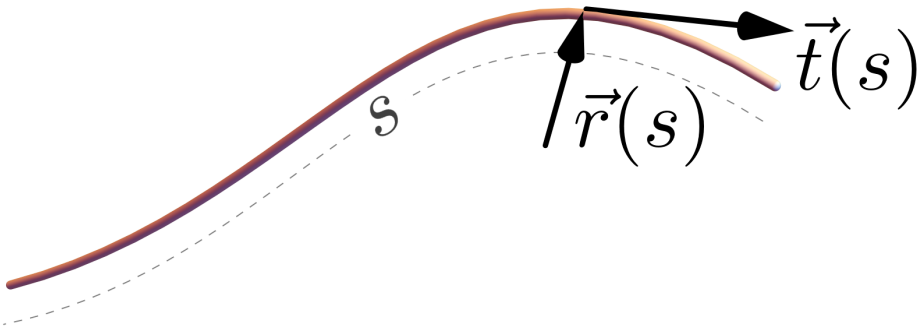

Figure 1.1: In the WLC model, DNA is represented by a smooth curve whose energy depends on the local curvature.

reason they are not suitable for the topics that the chapters of this thesis will attempt to illuminate.

On the other end of the spectrum of models we find the worm-like chain (WLC) model. This is a continuum description that models a DNA molecule as a flexible beam (see Fig. 1.1). The only degrees of freedom are bending in the x- and y-directions (if we call the direction along the DNA molecule z). Because the WLC model assumes the bending characteristics of the molecule to be isotropic, the Hamiltonian of the system depends on just one continuous parameter, the local curvature:

$$E_{WLC} = \frac{A}{2} \int \left(\frac{d\vec{t}(s)}{ds} \right)^2 ds, \tag{1.1}$$

where s parameterizes the curve followed by the DNA, $\vec{t}(s)$ is the local unit tangent vector to this curve and A is the bending modulus.

The WLC model is one of the simplest descriptions of DNA that is able to make useful predictions. Despite its simplicity, it approximates the behavior of DNA on large scales quite well. The simplifying assumptions it makes are justified if the length of DNA is long enough that local deviations average out (although we will come across a situation where this is not the case in Chapter 2).

However, such a model is really too simple for the purposes of the research presented here. Because we are interested in the effects of DNA sequence on the mechanical properties of DNA, and that sequence is encoded on the nucleotide, or base pair, level, we do not wish to coarse-grain too far beyond base pairs. In fact, we employ a model throughout

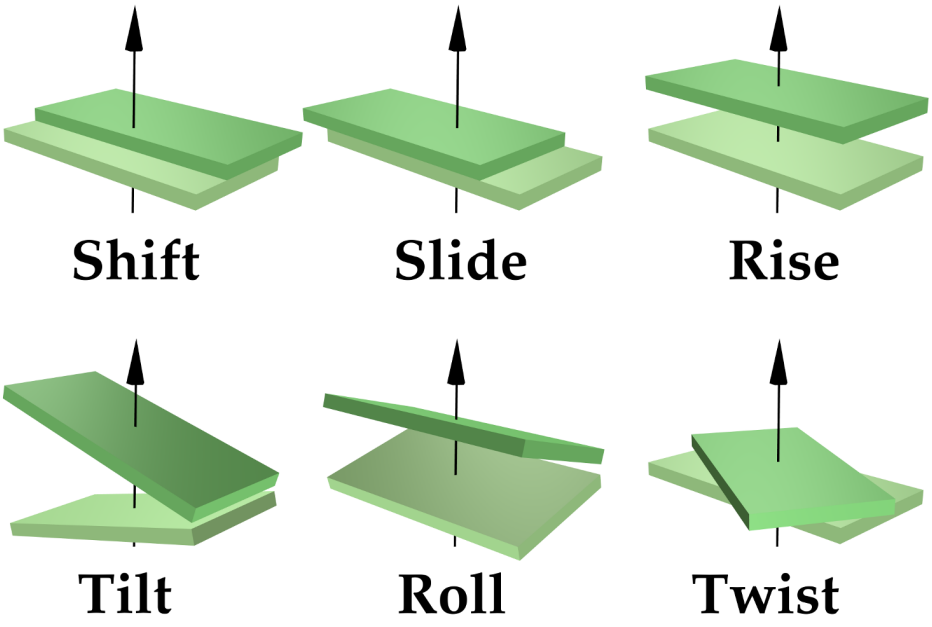

Figure 1.2: In the Rigid Base Pair model, the relation between one base pair and its neighbour is determined by six degrees of freedom.

that coarse-grains exactly at the base pair level: the Rigid Base Pair (RBP) model.

As the name implies, this model treats the base pairs of the DNA as rigid bodies [6]. Between a base pair and its nearest neighbours it assumes harmonic potentials in all degrees of freedom.

The relative position and orientation between any two rigid bodies can be described using six degrees of freedom: three translational ones and three rotational ones. In the context of DNA these degrees of freedom have been standardized relative to the base pair and given names; the translational degrees of freedom have been dubbed shift, slide and rise and the rotational ones tilt, roll and twist. Fig. 1.2 shows the meanings attached to these names.

The energetics of the RBP model are defined as a quadratic form in the differences between the six degrees of freedom just defined and their intrinsic values,

$$E_{RBP} = \frac{1}{2}\sum_n (\vec{q}_n - \vec{q}_n^0) \cdot K_n \cdot (\vec{q}_n - \vec{q}_n^0), \tag{1.2}$$

where the vectors $\vec{q}_n$ represent the degrees of freedom, $\vec{q}_n^0$ the corresponding intrinsic values and K_n are 6-by-6 stiffness matrices. The sum runs over all pairs of nearest neighbours in the DNA chain.

The RBP model lends itself well to sequence-encoded mechanics: one simply makes $\vec{q}_n^0$ and K_n dependent upon which base pairs are present. That way we have not one, but sixteen different sets of harmonic springs, one for each possible combination of two of the four base pairs.[3] Thus we obtain a fully sequence-dependent model for the mechanics of a given DNA molecule.

Various other DNA models, coarse-grained at similar levels to the RBP model, exist in the literature. The one most closely related to the RBP model, and one which may supersede it in the future, is the Rigid Base model. As one might expect, this model treats the individual bases, rather than base pairs, as rigid bodies, and puts connections between a base and its nearest neighbours along its strand, as well as the paired base [7–9]. Such a model is likely to prove more realistic than the RBP model, although its parameterization is ongoing work. Unfortunately, it is also computationally far more expensive, because the degrees of freedom in the model are not independent.

Various other approaches to coarse-grained DNA models have been proposed over the past decades, mostly at sub-basepair resolution. The interested reader may find a starting point in e.g. [10]. However, the RBP model strikes the right balance between realism – it is just able to account for sequence effects – and computational tractability, as far as the applications we have in mind here are concerned.

The RBP model therefore forms the basis of all the modeling done in this thesis. However, before we can put it into action, it needs to be parameterized. The intrinsic values and the stiffnesses needed can be found in the literature. There are two approaches: the parameters have been determined experimentally using crystallographic methods [11, 12] and they have been calculated from all-atom molecular dynamics simulations [13].

For various reasons, these two methods did not produce the same parameters; there are some significant differences in the predicted values. Previous research has concluded that the best paramaterization is actually one where the experimental values are taken for the $\vec{q}_n^0$, and the stiffnesses K_n are those found in the all-atom simulations [14]. We will have reason to reconsider this conclusion in subsequent chapters. (Our remarks on the topic can be found in Appendix A.)

3 Due to symmetry considerations (the two strands of the DNA double helix run antiparallel), there are in fact only ten independent types of spring.

We will consider only these nearest-neighbour parameterizations, in which there are 16 different springs to account for. In reality, however, the mechanical properties of a base pair step depend on their larger sequence context [15–20]. It is generally thought that one should parameterize the mechanical properties of DNA at the tetranucleotide level, but no parameterization is available that has been studied as extensively as the basic dinucleotide parameterizations of [11] and [13] (and their hybrid). We will therefore only make use of these established parameterizations. However, as we learn more about the system, it seems inevitable that the models will be updated to include longer-range effects in the future.

1.2 MODELING NUCLEOSOMES

Having settled upon the Rigid Base Pair model for our DNA, we can apply it to any system we like. There are various systems of interest in which we find deformed DNA molecules, such as overtwisted DNA that forms supercoils, DNA rings and loops, and DNA that is put under a stretching force. The system that has received the majority of scientific interest, however, has been the nucleosome.

Cells need a substantial amount of genetic information to function, which means a large quantity of DNA. The human genome, for instance, contains about 3×10^9 base pairs. A base pair being roughly 0.34 nanometers in length, and since human chromosomes come in pairs, that comes down to about two meters of DNA per human cell. All of that needs to fit inside the cell nucleus, the radius of which is only around 10^{-5} meters.

The only reason it is even possible to compress the DNA so much is that it is much thinner than it is long; the width of a double-stranded DNA molecule is only about 2 nanometers. Still, one would be hard pressed to naively try to squeeze a two-meter tangle of DNA into a cell.

Nature has found various methods to compactify DNA enough to fit inside cells. Bacteria and other prokaryotic organisms can twist up their DNA molecules so that they form compact superhelical structures [21]. Eukaryotic life has found a solution in wrapping its DNA around small protein cylinders. A complex of eight histone proteins – the histone octamer – is wrapped by 147 base pairs of DNA, in about a loop and a half.[4] Fig. 1.3 shows a representation of the resulting configuration of the DNA molecule (the histone core is not shown).

4 For a review on what is known about the structure of the nucleosome, see e.g. [22].

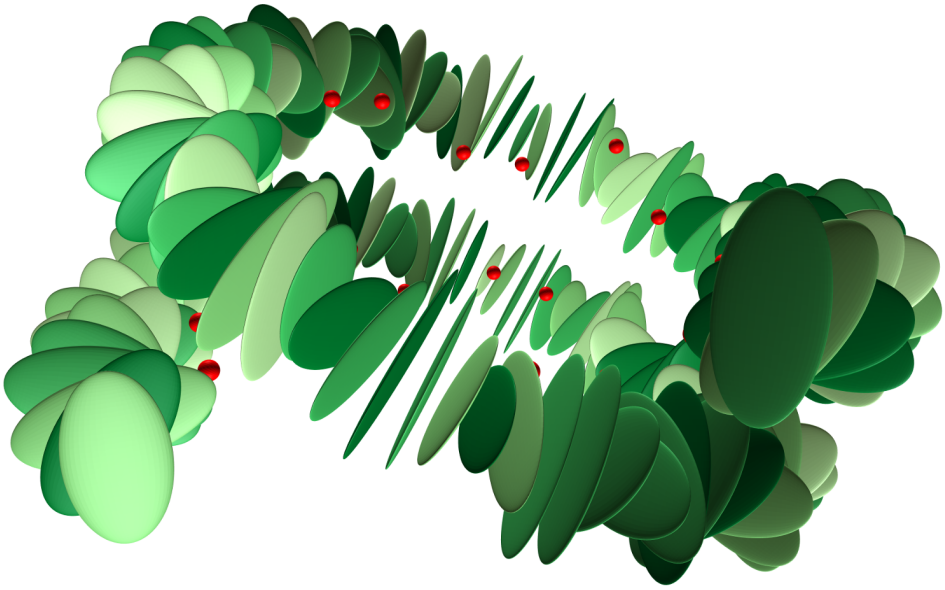

Figure 1.3: Model of a nucleosome, as described in [1]. The DNA is accounted for by the Rigid Base Pair model: the green ellipsoids represent the individual base pairs. It is wrapped around the (virtual) histone core using 28 binding sites, depicted by red circles.

With the properties of the DNA provided by the Rigid Base Pair model, a nucleosome can be modeled by forcing the DNA into the right shape. The DNA is attached to the histone core at 28 points [23], indicated in Fig. 1.3 by red spheres. As shown in that figure, these binding sites do not occur at the edges of the base pairs, but rather always halfway between one base pair and its neighbour. This is because the DNA is bound at the phosphate groups of the DNA backbones.

The backbone, and hence the phosphates, are not actually part of the Rigid Base Pair model, so it is not immediately obvious how to enforce the constraints put upon the DNA by the nucleosome. Several approaches have been put forward, such as considering the nucleosome as an ideal superhelix [24], constraining the base pairs to remain close to the places where they are found in crystallography data [25] or identifying and modeling the sites where the DNA is bound to the histone core [26].

Careful examination of the crystal structure of the nucleosome reveals a way forward in the latter case. It turns out that the phosphates of the DNA backbone lie approximately in the mid-plane between two neighbouring

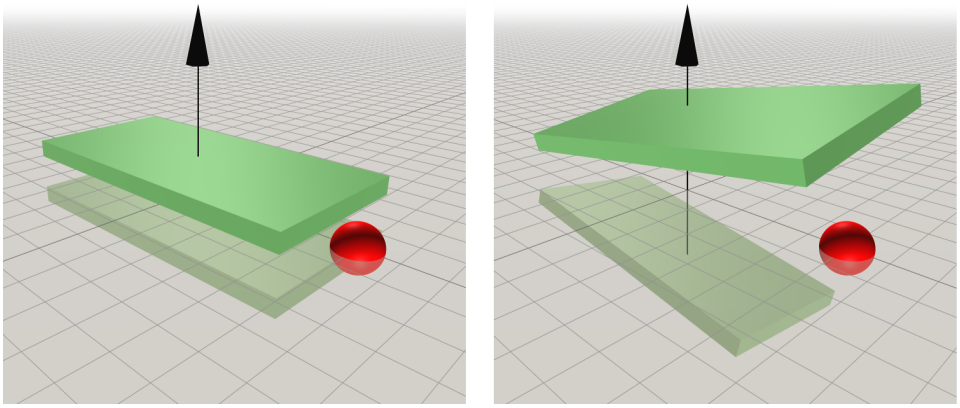

Figure 1.4: In the Eslami-Mossallam nucleosome model [1], the base pairs adjacent to a binding site are free to move as they like, as long as they do so symmetrically, in such a way that the midplane (depicted as a grey plane with a grid) and the attached phosphate (red sphere) do not move. Twenty-eight such constraints force the DNA into a nucleosomal configuration.

base pairs. The mid-plane is an imaginary plane that is exactly halfway between two base pairs, both positionally and rotationally.

The model in [26] held these mid-planes in place with harmonic potentials, with spring constants based on the thermal fluctuations of the mid-planes. However, this model contained many parameters to be fitted, and the model of [26] evolved into that of [1]. The latter model contains no free parameters, as it instead assumes rigid constraints on the mid-planes at the binding sites. The mid-plane, and the phosphate, can be kept immobile by ensuring that when one of the base pairs next to the phosphate moves or rotates, the other one moves or rotates in the opposite direction. The base pairs adjacent to a binding site are free to move, but are constrained to do so concertedly, as pictured in Fig. 1.4.

1.3 ACCESSING A SEQUENCE'S AFFINITY FOR NUCLEOSOMES

When bound to the histone core, a DNA molecule is significantly deformed. The nucleosome is therefore an archetypical example of a system in which the sequence of the DNA molecule affects the energetics [27]. As

a result, the sequence affects the positioning of nucleosomes[5] along a DNA molecule,[6] and hence it likely also influences DNA folding on the genomic scale [32]. Furthermore, since DNA wrapped into a nucleosome cannot be read out, nucleosome positioning directly influences gene expression [33, 34] (for reviews, see [35, 36]) and DNA replication [37]. Much evidence exists that evolution makes use of this sequence-dependent nucleosome affinity to organize chromatin to suit the needs of an organism [38–50].

Due to the importance to eukaryotic life of this system, one of the primary goals of the field of sequence-dependent DNA mechanics is to measure[7] and predict the affinity of a given sequence for the nucleosome. Not surprisingly, many different theoretical and computational approaches have been devised to get a grasp on sequence-dependent nucleosome affinity.

Early work looked for specific sequences and patterns that attracted or depleted nucleosomes [52–55], followed later by models that attempt to predict the nucleosome affinity for any given sequence. Some are physical models [1, 24, 26, 56–61]; many are bioinformatical, based on measured sequence preferences [62–70] and educated guesses about those preferences [71–74]. An overview of some of these models and their performance can be found in [73] and [75].

It is good to note that accurately predicting nucleosome affinity has proven difficult. As shown in [73], many models do not perform any better than a very simply predictor of nucleosome affinity: the percentage of C and G dinucleotides versus A and T (commonly referred to as the GC content). In choosing a model for the research presented in this thesis, we have therefore not looked to find or create the model that performs most accurately, but one that will allow us a deeper physical understanding of the underlying properties that determine nucleosome affinity.

---

5 Part of the literature in this field has used 'positioning' to specifically mean the placement of nucleosomes at exact locations, as opposed to more statistical measures like 'nucleosome occupancy', which has led to some confused discussion in the past [28]. We use the term 'positioning' here and in the rest of this thesis in the general sense of 'effects that bias the positions of nucleosomes', be it in a narrowly defined location or in a broad statistical sense.

6 Although it is not the only actor *in vivo*, with important roles also being played by ATP-driven chromatin remodelers. The latter are known both to override intrinsic nucleosome preferences encoded into the sequence (e.g. [29]), as well as to cooperate with them (e.g. [30]). See also [31] for a review on nucleosome positioning.

7 The history, methodology and the zoo of data on the experimental side is beyond the scope of this theoretical introduction. However, the interested reader may find a valuable starting point in [51], which reviews the results found for yeast and the methods used to obtain them. A number of key references treating other organisms can be found in the introduction to Chapter 6.

As the many references above show, there is a large zoo of models out there for predicting nucleosome affinity. The bioinformatical and phenomenological models, while they provide interesting insights into what the sequence preferences of the nucleosome are, do not teach us much about the physical origins of those preferences. Many of the physics-based approaches that have been put forward are not satisfactorily realistic, for example because they neglect degrees of freedom, or because they do not account for thermal effects. We will work with the model and methods of Eslami-Mossallam *et al.* [1], as it remedies some of these omissions. Let us therefore briefly describe how that model is used to predict nucleosome affinity.

The RBP model, when put under constraints like those of the nucleosome model described in the previous section, quickly becomes impossible to treat in closed form, and statistical methods become necessary if anything is to be learned about the system. The starting point for probing the behavior of DNA molecules under constraints has been the Markov Chain Monte Carlo (MCMC) method. This is a standard computational technique for sampling from a probability distribution that is analytically intractable. (The ideas behind MCMC methods go back to the 1930s and 1940s; the first time it was applied in its basic modern form was in 1953 [76] and this form could be considered mature in 1970 [77]. A treatment of the methods can be found in any modern textbook on computational methods for statistical physics or for statistics more generally.)

In physics, the distribution of interest is usually the Boltzmann distribution of the state space of a system. Sampling states weighted by their Boltzmann factor $e^{-E/(kT)}$ allows one to calculate thermal averages of any of the system's parameters. Naturally, MCMC methods are applied widely in biophysics, where many systems are too complex and heterogeneous for an analytic treatment.

The state space of a physical system generally consists of all the positions of its constituents – i.e. its particles or components. An MCMC simulation moves these constituents around, calculates the system's energy after the move based on the potentials in the system, compares the new energy with that of the system before the move, and decides whether or not to accept the move based on the change in energy. A basic MCMC simulation of the RBP model involves randomly moving and rotating base pairs (making sure not to violate any constraints in the model) and monitoring the elastic energies between them.

Using a standard MCMC simulation of a nucleosome with a given sequence allows us to estimate the average (meaning thermally averaged)

energy corresponding to that sequence. This is the approach used to obtain a measure of the nucleosome affinity by Eslami-Mossallam *et al.* [1]. This method is unfortunately not ideal, as it provides the average energy, while what one would like to know is the free energy. The two differ by an entropic contribution, which, however, is thought to be only a logarithmic correction. We will see how to get a measure of the actual free energy in Chapter 4.

1.4 MUTATION MONTE CARLO

We have described how to use the Eslami-Mossallam nucleosome model [1] to study a given sequence. But what if one is interested not in the properties of some specific DNA molecule, but rather of a whole class of them? As mentioned before, the term DNA does not describe a single molecule. Rather, there are many different DNA molecules, distinguishable by their sequence. So what if one wants to know something about the affinity of a DNA molecule to e.g. a nucleosome, as a function of the sequence encoded in it?

One might run an MCMC simulation for different sequences of interest and compare the average energies associated with each. However, the space of possible sequences can be large. A nucleosome wraps 147 base pairs of DNA, meaning there are 4^{147}, or roughly 10^{88} different nucleosomes. Mapping such a large sequence space becomes as untractable as integrating over all of the system's state space.

The solution is to treat the DNA sequence as just another degree of freedom, and sample it along with the spatial configuration of the DNA. This way, one can sample not just DNA configurations according to their Boltzmann weight, but also the sequences. This method has been dubbed Mutation Monte Carlo (MMC) [1].

A standard Monte Carlo simulation samples the Boltzmann distribution of a system across its state space,

$$P(\theta) = \frac{1}{Z_\theta} e^{-\beta E(\theta)}, \tag{1.3}$$

where Z_θ is the partition function, θ encodes the degrees of freedom of the system, β is the inverse temperature $1/(k_B T)$ and E is the energy of a given state.

The MMC method is a straightforward extension that includes the nucleotide sequence S of the DNA as additional degrees of freedom:

$$P(\theta, S) = \frac{1}{Z_{\theta,S}} e^{-\beta E(\theta,S)}. \tag{1.4}$$

In the case of our nucleosome model, θ represents all the inter-base-pair degrees of freedom (the q in Eq. 1.2, for all 146 pairs of successive base pairs) and S is a 147-nucleotide sequence.

Such a simulation allows us, for example, to marginalize the spatial degrees of freedom in order to calculate the probability distribution of the system in sequence space (as in [1, 78, 79]),

$$P(S) = \frac{1}{Z_{\theta,S}} \int d\theta P(S, \theta) = \frac{1}{Z_S} e^{-\beta F(S)}, \tag{1.5}$$

where $F(S)$ is the free energy of the sequence S wrapped into a nucleosome.

In such a scheme, the simulation will converge upon the high-affinity sequences for the constrained system, and one can study their statistical properties. This was the idea exploited in [1] to study the sequence preferences of the nucleosome as predicted by the RBP model. It allows one to sample the probabilities of dinucleotide steps, i.e. the combinations of pairs of neighbouring base pairs, at every point along the nucleosomal sequence. Such a sampling leads to results like those in Fig. 1.5.

The center of the x-axis denotes the dyad, or the center, of the nucleosomal sequence. We see that at this point, the nucleosome heavily favours the dinucleotides CC, CG, GC and GG. Conversely, the nucleotides AA, AT, TA and TT are strongly suppressed. Since we sampled the sequences using Boltzmann weights, this means that it costs, on average, more energy to place A/T-rich dinucleotides at the dyad than it does to place C/G-rich ones.

As we move along the nucleosome, we see oscillatory behavior in the probabilities, with the A/T and C/G dinucleotides trading places many times. The periodicity of the signals corresponds to the helical repeat length of the DNA. The preferences of the nucleosome for certain sequences stems from the fact that the DNA needs to be bent significantly to wrap around the histone core. Some dinucleotides accommodate this curvature more easily than others. Even though the shape of the DNA wrapped around the histones is close to being circular, i.e. the curvature is always in the same direction, the direction of curvature is actually con-

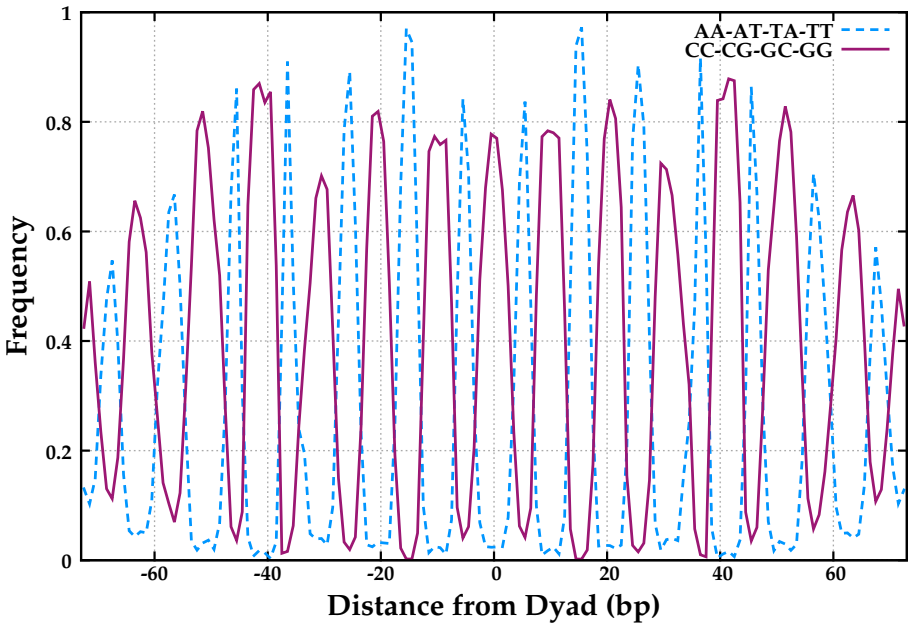

Figure 1.5: Probabilities of finding nucleotides consisting of only A and T, and nucleotides consisting only of C and G, along the nucleosone. Each curve represents the total probability of the four dinucleotides listed in the legend. The probabilities show a clear oscillatory behavior, corresponding to the helical repeat of the DNA. The distributions for this plot were produced using MMC at an artificially low temperature of 50 K to exaggerate the oscillations and speed up convergence of the simulation. More information on the temperature-dependence of the distributions can be found in [1] and Chapter 5.

tinually changing relative to the base pairs themselves. This is because the DNA is helical in nature, twisting around its own axis about once every 10 base pairs. If at one point the DNA is bent in the direction of positive roll (see Fig. 1.2), then five base pairs further the DNA is bent in the direction of negative roll, because the DNA has twisted 180° around its own axis.

This explains the most striking feature of the dinucleotide distributions found using the RBP model, as well as in experiments [64]. Much more could be said about these distributions that goes beyond the scope of this introduction; the reader can find many additional details in [1]. These kinds of distributions will play a starring role in the later chapters of this thesis.

The MMC method is the key to accessing such distributions *in silico*. The dinucleotide distributions for the nucleosome as discussed here are

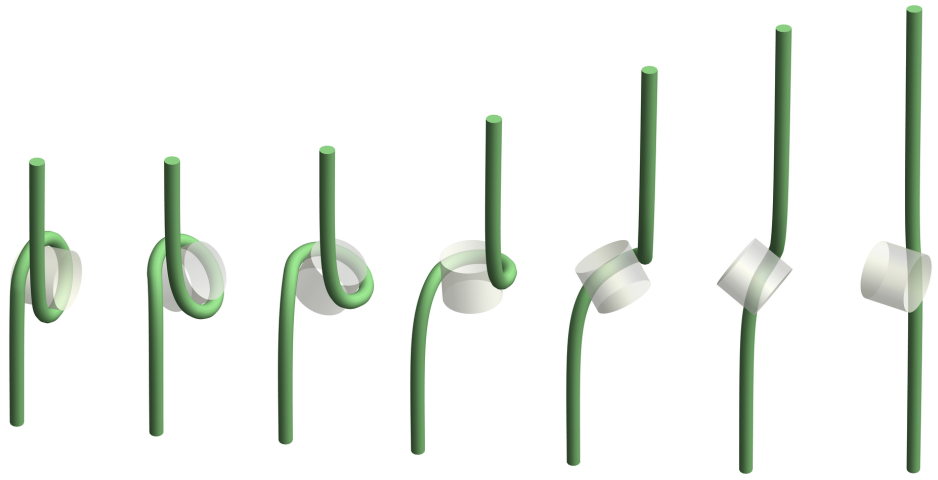

Figure 1.6: The geometry of a nucleosome unwrapping under force, with the DNA represented as a WLC.

the archetypical example considered in the literature, but the method can in principle be applied to find distributions (for dinucleotides or oligonucleotides of any other length) for DNA under arbitrary constraints. In Chapter 5 we will also look into the distributions pertaining to DNA constrained to form a ring.

1.5 MODELING NUCLEOSOME UNWRAPPING

A second application of the nucleosome model defined in Section 1.2 has been to study the unwrapping behavior of the nucleosome. When one takes the ends of the DNA wrapped into a nucleosome and pulls on them with sufficient force, the bonds between the DNA and the histone core will be broken and the DNA will unwrap from its spool.[8] Due to the geometry of the system, this is less trivial than it sounds. The DNA is looped around the histones, and when pulling on the DNA ends, rather than, say, holding the histone cylinder and pulling the DNA away from it, initially the DNA is simply tightened around the cylinder. The histone spool will need to rotate in order to let the DNA unspool.

8 In reality, the histone core is not a single, solid entity, and its constituent histones also tend to dissocciate when the DNA is unwrapped [80, 81]. However, this does not seem to greatly influence the DNA geometry, as models ignoring this subtlety describe the process well [2, 82].

During this rotation, the nucleosome under tension must pass through a state with a relatively high energetic cost, which throws up a barrier to nucleosome unwrapping. This barrier was first observed experimentally by Brower-Toland *et al.* [83], although its origin was not yet understood at that time. The barrier can be understood purely from the geometry of the process [82]; in Fig. 1.6 the states the nucleosome goes through when unwrapping are depicted and in the central picture, the nucleosome is shown in the state that exemplifies the cause of the barrier. This is the state where the histone core has made half a turn. We see that the DNA is very strongly bent where it comes away from the histones, with a curvature much stronger than in the nucleosome itself.

These ninety-degree kinks in the DNA are of course caused by the applied tension, and as such they increase in severity when the force applied is increased. This leads to the nucleosome being *kinetically protected* against unwrapping. The harder one pulls on the nucleosome, the higher the energy barrier becomes and the more it resists unwrapping.

These geometric properties of the system can be understood with a simple model like the WLC. From such a description one can obtain the average ballpark of the height of these kinetic barriers, which can serve as a basic scalar quantity to characterize the unwrapping behavior of a nucleosome. However, a much deeper look has recently proved possible. Ngo *et al.* [84] showed experimentally that a nucleosome wrapped with the Widom 601 sequence [85] unwrapped *asymmetrically*, i.e. it tended to open binding sites first on one side of the nucleosome. Such behavior cannot be captured with a WLC model, which is homogeneous and cannot capture differences between the two ends of the nucleosome. The behavior of course has everything to do with the sequence preferences of the nucleosome, and we may turn to the RBP nucleosome model from Section 1.2 to describe this new situation.

This nucleosome model consists of 147 base pairs of RBP DNA under 28 constraints. The only requirement, to allow for unwrapping, is to let some of these constraints be broken, at the cost of the binding site's adsorption energy.[9] This was done in [2] in order to understand from a theoretical perspective the results of Ngo. *et al.*

If we assume that a binding site will only open up if all of the binding sites either to its left or to its right are open (or at least that we are not interested in the situations where this does not hold), then we can characterize the unwrapping state of a nucleosome by two numbers, (L, R)

9 A description of how to arrive at the right adsorption energies can be found in [2], and the values are presented in Table 3.1.

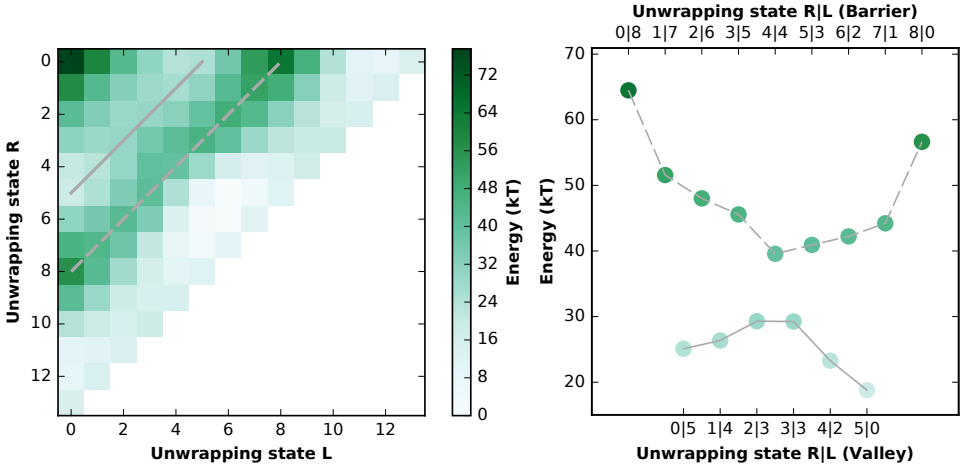

Figure 1.7: Left: Unwrapping energy landscape of the Widom 601 sequence, as a function of the number of binding sites opened from the left and from the right. The grey solid and dashed lines indicate slices through the landscape at the energy barrier and at the metastable valley, which are correspondingly plotted in the figure on the right. **Right:** The metastable valley and the unwrapping barrier both show the same asymmetry that biases the 601 nucleosome towards unwrapping from the right. (Results adapted from [2].)

denoting the number of binding sites open from the left end and from the right end. The nucleosome has 14 binding sites (each consisting of two bonds) so that we must have $L + R \leq 14$. With these two parameters, we can then map the energy landscape of the unwrapping nucleosome. This landscape is presented in Fig. 1.7 for the Widom 601 sequence, and shows that our nucleosome model also enables us to understand this asymmetric unwrapping behavior: it is encoded into the nucleosomal sequence. Specifically, this asymmetry is due to a difference in affinity between the left and right halves of the DNA sequence. A deeper explanation of these effects can be found in [2].

This incorporation of sequence preferences into the behavior of nucleosomes necessitates a new way of viewing them. Just like taking into account the sequence in the RBP model means that DNA is not one molecule, but rather a whole class with varying properties, likewise nucleosomes are not all made equal. Depending on the 147 base pairs chosen to be incorporated into the nucleosome, it prefers to unwrap from the left or from the right under tension, as we see here.

Similarly, nucleosomes may be distinguished by any number of properties, given the number of possible sequences. Beyond asymmetric unwrap-

ping, we could imagine that the sequence affects the height of the barrier, making the nucleosome more or less stably protected against forced unwrapping, as well as other processes like nucleosome breathing (spontaneous unwrapping without force) and the storage of twist defects in the DNA molecule [27]. We will take a further look into such possibilities in Chapter 3.

1.6 OUTLINE OF THIS THESIS

The introductory sections above should assist the unfamiliar reader to put the chapters that follow into their past and ongoing context. The chapters have been ordered to facilitate reading, rather than being absolutely chronological, and could be separated into two parts. Chapters 2 and 3 apply the methodologies described above to new systems. Chapter 2 applies the RBP model to a study of intrinsically curved DNA molecules. The MMC method is used in a simplified form to find DNA sequences that encode for strong intrinsic curvature, or conversely that make for a very straight DNA molecule, without any constraints yet placed upon the DNA. The chapter assesses the behavior of these molecules under a stretching force, finding that it is possible to create superhelical DNA molecules that act like small springs.

In Chapter 3 constraints are added to the RBP model. We further the study of forced nucleosome unwrapping, using the model described in Section 1.5. Using the MMC methodology we will show that we can design nucleosomal sequences with particular properties. Specifically, we will destabilize nucleosomes so that they are not strongly kinetically protected, and see that we can induce them to unwrap via specific paths.

In Chapter 4 a new model is introduced to predict the affinity of a given sequence to a nucleosome, based on the dinucleotide distributions derived from MMC simulations, as described above. This model gives an approximation scheme to the full Eslami-Mossallam nucleosome model [1] described in Section 1.2. The resulting calculations for nucleosome affinity are computationally much less expensive. After setting out the details of this model, the rest of the chapter is dedicated to determining the accuracy of the approximation.

This new model opens up many avenues of inquiry that would not be feasible to go down using the computationally expensive nucleosome model used in Chapter 3. Chapter 5 will show how the ideas behind this model can be used to go beyond the basic MMC simulation to extend to

situations where the selection pressure on DNA sequences is not linked to the physical temperature of the nucleosome system. The latter is the case in a certain class of sequence selection experiments, and presumably in real evolution.

In Chapter 6 the work firmly enters the biological realm, and genome-wide studies of nucleosome affinities are conducted. These studies would have been impossible to perform with the full RBP nucleosome model due to the amount of computing time required. The chapter presents analyses of nucleosome positioning signals in the promoter regions[10] of a wide range of organisms, and shows that the shapes of these signals are distinct signatures that separate simple, unicellular lifeforms from more complex, multicellular ones. The existence of such signals in real genomes points to the exciting possibility that mechanical evolution, the adaptation of genomic sequences to favor or disfavor certain physical structures like nucleosomes, has indeed occurred in nature.

Each of these chapters will draw its conclusions, but we will follow up with some overarching conclusions and an outlook on the future in Chapter 6.6.

As a supplement to the chapters described above, several appendices are also included. Appendix A contains some notes on the differences between the various ways of parameterizing the RBP model, as well as on how simulation temperature influences the predictions we make using MMC. Appendix B provides a numbered list of the DNA sequences that are mentioned in the chapters of this thesis. This list is referenced throughout the thesis using the format Sequence [n], so that the reader may consult the actual nucleotide sequences under discussion.

10 The part of the DNA before the start of a gene sequence, where the machinery that reads the DNA binds to it, and which is therefore in a position to 'promote' the reading of the gene.

2 FORCE RESPONSES OF DNA HELICES

This chapter is based on:
Tompitak, Schiessel and Barkema 2016 *EPL 116* 68005 [86]

In the Introduction, we started by introducing the basic model we use to describe DNA and its sequence-dependent properties. We then went on to show how we could use this model to build representations of interesting structures such as nucleosomes. Most of the chapters in this thesis indeed take an interest in such structures, but in this first chapter we will use the RBP model (see Section 1.1 in the Introduction) without yet placing it under any rigid constraints.

The sequence effects that we will study in subsequent chapters derive from a combination of two factors: the intrinsic shape of a DNA molecule with a given sequence, and the resistence of that molecule to deformation away from that intrinsic shape. For instance, a sequence will be more accommodating to form a nucleosome if its intrinsic curvature is as close as possible to the curvature required by the nucleosome, because less deformation is needed, and if its stiffness with respect to bending into the required curvature is small.

Here we specifically take a look at the properties of DNA molecules that have non-trivial intrinsic curvature. The stiffness of the molecules will still play a role, but we will choose our DNA sequences of interest by their intrinsic shape only.

2.1 INTRODUCTION

DNA with intrinsic curvature plays an important role in biological systems, influencing e.g. the positions of nucleosomes [1, 62, 87] and plectonemes [88–90] and the propensity for stretches of DNA for loop formation [91–93]. DNA molecules can intrinsically encode their preferential spatial organisation through their underlying base pair sequence. This is the first step of many layers of spatial DNA organisation on larger and larger scales, the latter steps only now beginning to be accessible to a qualitative and quantitative understanding [94].

The standard description of DNA as a worm-like chain (WLC) assumes the mechanical properties of DNA to be isotropic (see Section 1.1), and hence does not account for intrinsic curvature. Previous work has shown that, in general, this assumption is reasonable as long as the intrinsic curvature does not build up [95]. Here, we turn our attention to DNA molecules in which this assumption does not hold. Since strongly bent sequences occur in real organisms [96], it is not out of the question that long stretches of DNA with coherent curvature naturally occur, and we wish to know how this may cause predictions based on the WLC model to go wrong.

Artificial sequences with specific patterns of intrinsic curvature are also of interest. Tandem repeats of the Widom 601 sequence [85] (Sequence [1] in Appendix B) have been used as templates for reconstituting chromatin fibers [97–99]. This sequence is intrinsically curved in one specific direction, and tandem repeats of this sequence have the potential to form superhelical structures, see Fig. 2.1a.

Here, we design sequences with such superhelical intrinsic curvature *in silico* and characterize them by simulating their force response. Such structures have been studied theoretically (e.g. [100–102]) and it is known that the superhelical geometry influences their force response, if the curvature of the helix is strong enough that the persistence length of the polymer is at least comparable to the contour length of a superhelical turn [102].

The question is then whether there exist DNA sequences with an intrinsic curvature strong enough to influence their force response. It is not *a priori* clear whether we should expect this to be the case. The tandem repeats of curved sequences like 601, are not generally assumed to feature such effects, but the 601 sequence is unlikely to be the most strongly bent sequence, because it is experimentally difficult to access the entirety of the space of sequences of substantial length. To do the latter, and find the limits of how strongly curved a DNA sequence can be, we need computational methods like MMC (see [1] and Section 1.4), as described in the next section.

It turns out that strong, coherent intrinsic curvature can indeed be achieved for DNA molecules. Molecules that feature such curvature act like nanoscale helical springs, which intrinsically resist stretching. The force response predicted by the WLC model provides a poor fit in the low-force regime for these sequences. We suggest an alternative description, in which we take the superhelical structure into account.

Some study has been made of the finite-temperature force response of intrinsically curved polymers, but the problem has only been fully solved

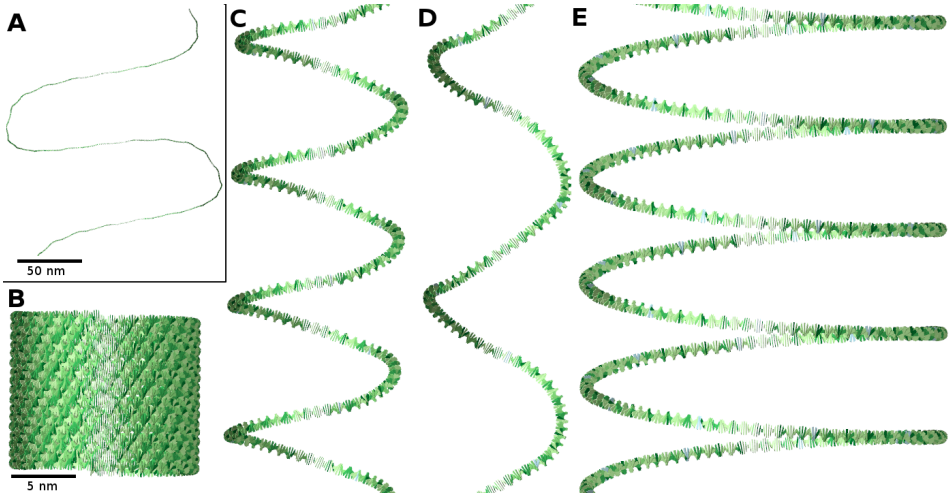

Figure 2.1: The ground state structures of tandem repeats of **A)** the Widom 601 sequence (Sequence [1] in Appendix B), **B-D)** our artificial, highly-bent DNA sequence, as predicted by the RBP model (Sequence [3]) and **E)** the natural (in its singular form) kinetoplast sequence (Sequence [2]). Helices with various shapes can be produced by varying the number of additional nucleotides between copies of the sequence. The numbers of additional base pairs in these figures are **A)**: 0; **B-D)**: 0, 1 and 2, **E)**: 2. For the geometric parameters corresponding to these structures, see Table 2.1.

in two dimensions [103]. Here we employ a discrete description [104] of a flexible helix, course-graining the DNA to the same level as does the Rigid Base Pair (RBP) model [11] (see Section 1.1), which we use to numerically assess the force response. For low forces, the force response of this discrete superhelix can be described by an extensible WLC model [105, 106] with effective values for the bending and stretching moduli. This model leads to a significantly improved prediction for the low-force regime of the force response over the naive WLC model.

2.2 DESIGNING SEQUENCES WITH INTERESTING CURVATURE

To design sequences with specific properties, we modeled the DNA using the (RBP) model (see Section 1.1) parameterized by crystallography data [11], and we employed the Mutation Monte Carlo (MMC, see Section 1.4) method to search the space of all possible sequences. The original

method [1] introduces mutation moves to a standard Monte Carlo simulation of a DNA molecule. We employed a simplified version, which performed only mutation moves and searched (through simulated annealing) for the most strongly bent sequence. The strength of instrinsic curvature of the sequence was measured by calculating the ground state configuration of the sequence and taking the inner product between the first and last tangent vectors to the DNA backbone.

We ran this algorithm on a DNA molecule consisting of 84 base pairs, and created a tandem repeat of the most strongly bent sequence thus found (Sequence [3] in Appendix B). In order to create sequences with different superhelical properties, we created repeats with additional homogeneous (sequence-averaged) DNA between the repeats, to interfere with the alignment of the direction of curvature between successive copies of the sequence. This led to DNA sequences with various values for the superhelical radius and pitch angle, see Fig. 2.1b-d and Table 2.1.

We also designed a sequence with low intrinsic curvature by taking our tandem repeat sequence and applying an MMC algorithm similar to that described above, but which maximizes the sum of the inner products between the first tangent vector, and a number of tangent vectors along the rest of the DNA molecule, at intervals of 50 base pairs (roughly 1/3 of the persistence length of DNA.) This algorithm ensures that the resulting sequence has low intrinsic curvature both locally and globally.

A major difference between this MMC simulation and the previously described one, however, is a constraint on the allowed mutations. Starting with the strongly curved sequence, we mutated it such that the distribution of dinucleotides in the sequence remained the same, and only their order was allowed to change. We achieved this by performing mutation moves that swap pieces of the sequence that both start and end with the same nucleotides. For example, ATA might be swapped with ATTA, but not with ATT. By imposing this constraint, both the nucleotide and dinucleotide contents are kept identical. This means for example that the GC content of the sequence, which is generally thought to be a good indicator of the mechanical properties of a sequence, is invariant. Even more strongly, keeping the dinucleotide content constant means that (in the RBP model) the resulting DNA molecules contain the exact same elastic components, only in a different order. Even under this constraint, we find that we can design sequences that seem similar, but in fact have vastly different intrinsic curvature properties, and therefore different force responses. Part of the resulting 4700-base-pair sequence is shown as Sequence [4] in Appendix B.

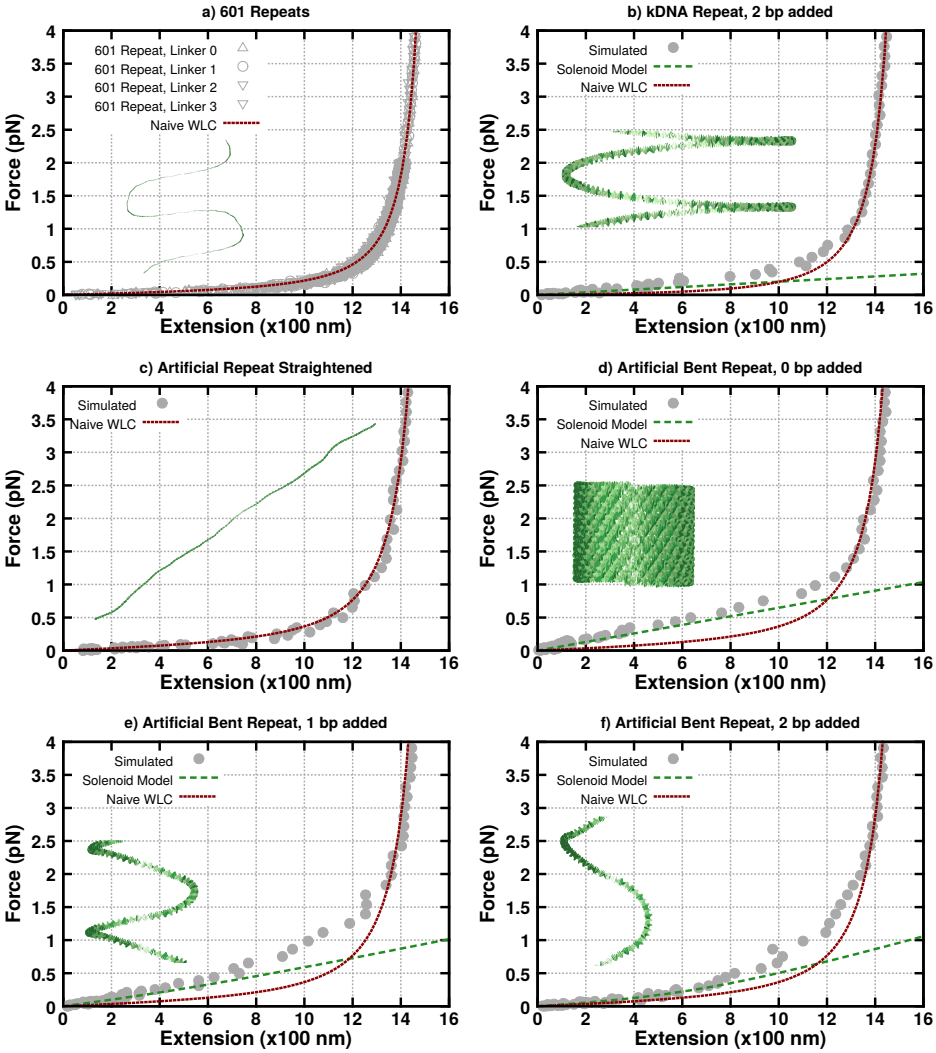

Figure 2.2: Simulated force-extension curves, WLC fits and predictions of the two-angle model for the low-force regime, for **a)** repeats of the Widom 601 sequence (Sequence [1] in Appendix B) with various numbers of additional base pairs; **b)** a repeat of the kinetoplast (kDNA) sequence (Sequence [2]); **c)** the straightened artificial sequence (Sequence [4]); **d-f)** repeats of the artificial curved sequence (Sequence [3]) with various numbers of additional base pairs. The inset pictures show the intrinsic shapes of these sequences. For the geometric parameters corresponding to these sequences, see Table 2.1.

2.3 FORCE RESPONSES

In order to measure the force response, we ran a Monte Carlo simulation of a DNA molecule with the given sequence under a number of fixed forces and in each case sampled the extension. This led to the force-extension curves presented in Fig. 2.2. In the case of the intrinsically straight sequence, the data can be fitted with the prediction from a simple WLC model [105] to good agreement, see Fig. 2.2c.

For an example of sequences with superhelicial intrinsic curvature, we first looked to the Widom 601 sequence. A repeat of this sequence indeed forms superhelices, as can be seen in Fig. 2.1a. Pulling on the helix displayed there as well as three other variants with various numbers of additional base pairs added between repeats, we obtain the force responses depicted in Fig. 2.2a. The different responses strongly overlap and do not deviate appreciably from the WLC fit. The intrinsic curvature of the 601 sequence is not strong enough to have a significant effect on the force response. This has also been observed experimentally, see Supp. Fig. 2 in Ref. [98], and can be understood from the geometry of the superhelix: the contour length of a superhelical turn is larger than the persistence length of the DNA at room temperature (see Table 2.1) meaning that the intrinsic curvature is lost to thermal fluctuations over distances at which its magnitude becomes significant [100, 102]. The result is that at this temperature, the relatively weak superhelical nature of the 601 repeat is not distinguishable from a straight molecule.

In order to get at the effects of the intrinsic curvature, we need a far more strongly bent sequence. Nature in fact provides such a sequence. A strongly curved section of DNA has been discovered in kinetoplast DNA (kDNA) [96, 107]. Taking the sequence depicted in Fig. 2 in [96] and repeating it with 2 additional base pairs in between, this DNA sequence forms a much tighter superhelix than the 601 sequence, see Fig. 2.1e. For this structure, with a single-turn contour length only 3.5 times the persistence length (see Table 2.1), the force response turns out not to fit the WLC prediction well. In Fig. 2.2b, we see a clear discrepancy for low forces. At high extension, the force response is similar, but at low extension, the tendency of the DNA to intrinsically curl up means that it acts like a spring. Hence more force is needed to stretch it, and the slope of the force-versus-extension curve is correspondingly higher.

To get a better grasp on the importance of the DNA sequence to the force response, we turn to the artificial sequences described in the previous section and compare the force responses of the helical sequences with

that of the straightened one. Since the average elastic properties of the straight and the curved versions of this sequence are identical, any difference must be due to the build-up of intrinsic curvature. In Fig. 2.1b-d and Table 2.1 we see that the artificial curved sequence forms even tighter helices, with single-helix contour lengths not much larger than the persistence length, so we expect an even larger effect.

In each of the figures 2.2d-f, the red dotted line is identical to that in Fig. 2.2c, plotted for reference. In each case there is indeed significant deviation from the WLC prediction; the more compact the helix, the stronger the deviation.

2.4 MODELING SUPERHELICAL DNA MOLECULES

The failure of the WLC model shown above is due to the assumption of no significant intrinsic curvature. The WLC force-extension curve assumes that the force response is dominated by the ironing out of thermal fluctuations around an intrinsically straight ground state. To correctly describe the situation under consideration, we must include the response due to the intrinsic resistance to stretching by the DNA molecule.

Some studies of flexible rods with intrinsic curvature exist [100–103, 108] but no analytical description at finite temperature in three dimensions is known. Since the discrepancy for DNA lies mostly in the low-force regime, we propose a partial solution that describes this regime well.

In order to account for the intrinsic curvature of the superhelical DNA molecules, we turn to a discrete model for such structures, consisting of a series of flexible, straight rods, each of which is connected to the next at fixed angles [104, 109, 110]. Two such angles are necessary to describe the orientation of one segment with respect to the next, so that this model is generally known as the two-angle model. These two angles, together with the length of the connecting rods, fix the shape of the entire superhelix.

For low forces, the backbone of the superhelix in this description behaves as an extensible worm-like chain [105, 106], with effective values for the stretching ($\tilde{\gamma}$), bending ($\tilde{A}$) and twisting ($\tilde{C}$) moduli, as well as a coupling between stretching and twisting ($\tilde{g}$) [104, 109]. With these parameters, the linear response of the superhelix is given by

$$\begin{pmatrix} F \\ M_t \\ M_b \end{pmatrix} = \begin{pmatrix} k_B T\tilde{\gamma} & k_B T\tilde{g} & 0 \\ k_B T\tilde{g} & \tilde{C} & 0 \\ 0 & 0 & \tilde{A} \end{pmatrix} \begin{pmatrix} x \\ \Omega \\ R^{-1} \end{pmatrix} \tag{2.1}$$

where F, M_t and M_b are the force, torsional torque and flexural torque applied to the system, and x, Ω and R^{-1} the extension, twist and curvature. We are interested in the dependence of F on x. If we assume no torques on the system, Eq. 2.1 reduces to

$$F = \left(k_B T \tilde{\gamma} - (k_B T)^2 \frac{\tilde{g}^2}{\tilde{C}} \right) x. \tag{2.2}$$

The effective values of the mechanical properties of the superhelical backbone depend on the elastic properties of the flexible rods that make up the superhelix, as well as the geometry of the structure, i.e. the two angles mentioned above, as described by Eqs. (96) and further in Ref. [109]. In applying this model to our DNA structures, the flexible rods represent the connections between consecutive base pairs, so their elastic properties are those of the DNA dinucleotide steps in the RBP model.

We will characterize a superhelix not by the two local angles between successive rods, but by the superhelical radius R and the superhelical rise (i.e. distance travelled along the backbone) per base pair, s_0, as these are simpler to determine numerically. The expressions for the elastic parameters in Eq. 2.1 simplify considerably if we assume that $R \gg b$ (for our helices, R is generally larger by at least a factor of 30), in which case we find

$$\tilde{\gamma} = \frac{r}{k_B T} \frac{C + (A - C) r^2}{R^2} \tag{2.3}$$

$$\tilde{g} = \frac{r^2}{k_B T} \frac{(A - C)(b^2 - s_0^2)}{bR} \tag{2.4}$$

$$\tilde{C} = r \left(A - (A - C) r^2 \right) \tag{2.5}$$

$$\tilde{A} = \frac{2rAC}{A + C - (A - C) r^2} \tag{2.6}$$

$$r = \frac{s_0}{b} \tag{2.7}$$

where A and C are the bending and twisting moduli of our DNA model, R and s_0 are as defined above, and b is the length of the flexible rods, i.e. the distance between successive base pairs, taken to be 0.34 nm. In this continuum limit, the effective spring constant of the system, i.e. the constant coefficient in Eq. 2.2, reduces to the standard classical form [102, 108].

The bending modulus can be estimated by sampling the tangent-tangent correlations in a standard Monte Carlo simulation of a homogeneous DNA

Table 2.1: Geometrical parameters (superhelical radius R and distance along the backbone between successive base pairs s_0) and number of times the persistence length of DNA fits into a single helical turn (l being the contour length of a single helical turn and l_p the persistence length) for the superhelices depicted in Fig. 2.1. The actual nucleotide sequences can be found in Appendix B.

Sequence	R (nm)	s_0 (nm)	l/l_p
Artificial + 0 bp	14.60	0.00537	1.96
Artificial + 1 bp	13.37	0.0826	1.86
Artificial + 2 bp	11.09	0.164	1.72
kDNA	26.02	0.0309	3.52
601	113.34	0.152	17.1

molecule with sequence-average properties at zero force. The twist modulus can be estimated (neglecting the cross terms in the RBP Hamiltonian) from the twist stiffnesses of the RBP model [111]. The geometrical parameters of the helices were estimated directly from the ground state structure of the sequences as follows.

We created ground-state structures for long sequences consisting of 500-1000 concatenated copies of the sequences of interest. This gave us DNA states like those depicted in Fig. 2.1 but where the superhelical backbone was much larger than the radius, reducing edge effects. We approximated the backbones of these structures by calculating the three-dimensional straight line that best fit (in the least-squares sense) the positions of all base pairs in the superhelix.

We estimated the radius of each superhelix by taking the average distance of the base pairs to their closest point on this backbone. We also used the backbone to calculate for each base pair the (signed) angle between the connecting line between the base pair and the backbone, and the same for a reference base pair. These angles ran cyclically from $-\pi$ to π, and allowed us to determine the base pairs which were an integer number of helical turns away from the first base pair. This allowed us to calculate the average distance along the backbone between two successive such base pairs, and the average number of base pairs per superhelical turn. Hence we calculated s_0, the superhelical rise (i.e. distance travelled along the backbone) per base pair. The resulting parameters are shown in Table 2.1.

Using these values to inform Eqs. 2.2–2.7, the force-extension curves for these structures are given by that of the extensible worm-like chain with our effective parameters, which can be approximated by [106]:

$$F = \left(\frac{(k_B T)^2}{\tilde{A}}\right)\left[\frac{1}{4(1 - x/L + F/K)^2} - \frac{1}{4} + \frac{x}{L} - \frac{F}{K}\right] \tag{2.8}$$

where L is the contour length of the DNA and K the effective spring constant of the backbone, written out explicitly in Eq. 2.2.

This equation gives us the force-extension curves represented by the dashed green lines in Fig. 2.2b,d-f. The agreement with the data at low forces is significantly better than that of the naive WLC prediction.

The force-extension curves of these superhelical DNA molecules have three regimes. At low forces, entropy dominates and the force response follows from the random walk statistics of the effective superhelical axis. Increasing the force, the superhelical backbone is straightened out. When the extension is approximately equal to the intrinsic contour length of the superhelical backbone, we reach a regime where the superhelix acts as a spring (the energetic regime of our effective extensible WLC) [112]. The slope of the force-extension curve becomes higher because of this additional restoring force and approaches the classical result for the linear regime of a flexible helix [102, 108]. Finally, the superhelix becomes so distorted that the effective model breaks down, and the normal high-force regime of DNA takes over.

2.5 CONCLUSIONS

We find that, while in most cases intrinsic curvature may not be an important factor to the force response of a DNA molecule, it is easy to design sequences for which the effects of intrinsic curvature are visible. Short sequences with such strong curvature also occur in nature, e.g. in kinetoplast DNA. In such cases, fitting the naive WLC force-extension curve does not give satisfactory results in the low-force regime. The intrinsic curvature needs to be taken into account, and we have provided a model to do so, at finite temperature in three dimensions, and valid for low forces.

The commonly used tandem repeats of the 601 sequence do not show appreciable effects in their force response and experimental setups using such sequences should not be affected. However, should more strongly bent sequences become desirable for such experiments in the future, care should be taken to ensure that the effects we note here are accounted for.

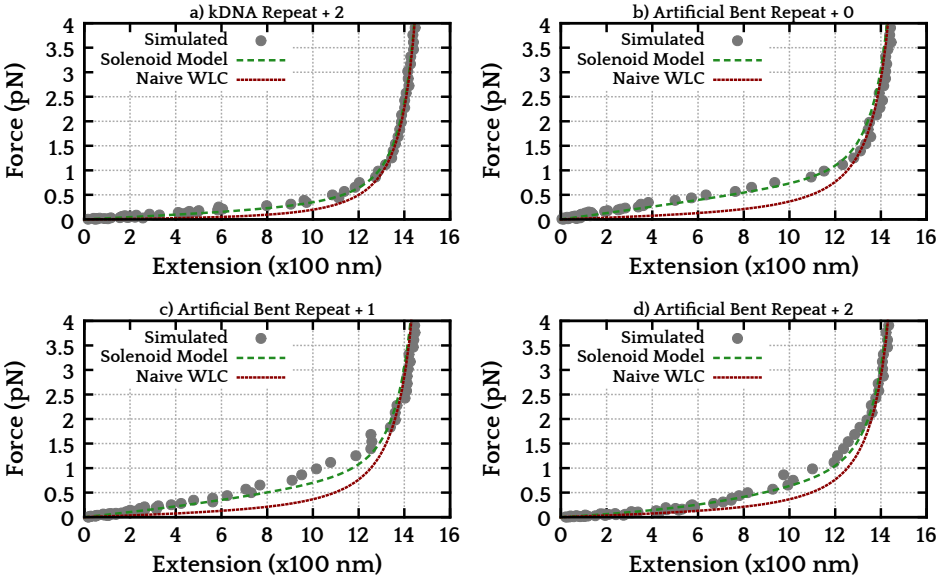

Figure 2.3: Simulated force-extension curves, WLC fits and predictions of the two-angle model combined with the extensible WLC high-force regime using an ad hoc crossover function, for **a)** a repeat of the kDNA sequence (compare Fig. 2.2b); **b-d)** repeats of the artificial curved sequence with various numbers of additional base pairs (compare Fig. 2.2d-f).

We also note that the effects to the force response of a DNA molecule can be tuned through strongly constrained mutations of the sequence. Surprisingly, both very straight and strongly bent DNA molecules, with markedly different force responses, can be obtained without altering the overall GC content, or even the overall numbers of different dinucleotides.

Finally, a practical note. The description presented here improves predictions for the low-force regime of the force responses of superhelical DNA molecules, but breaks down for high forces. One could stitch together our description for low forces with the standard high-force limit with an appropriate cross-over function. In Fig. 2.3 are shown Figs. 2.2b,d-f, but with the low-force prediction replaced by such a combined prediction for the force response. The cross-over function was chosen, rather arbitrarily, to be a multiplicative factor to the low-force function that gradually kills it off, given by $\frac{1}{2} - \frac{1}{\pi}\arctan[(z - z_c)/b]$. In all plots we used $z_c = 1300$ nm and $b = 300$ nm.

This approach gives a reasonable approximation to the simulated curves, and may be of practical use when using our low-force description in real

applications. However, the crossover function used here is arbitrary and may not work for all superhelices: there is no reason to assume that the crossover function should be independent of the superhelical parameters. Further research may lead to a more physically motivated crossover, similar to the approach of [106] for the extensible WLC.

3 DESIGNING NUCLEOSOMAL FORCE SENSORS

This chapter is based on:
Tompitak et al. 2017 *Phys. Rev. E* 95 052402 [113]

In the previous chapter we saw how the base pair sequence of a DNA molecule can significantly affect its mechanical behavior. These effects become especially important when the DNA is deformed by external constraints. The archetypical constrained DNA system is the nucleosome, as described in Section 1.2. We can learn something about the sequence preferences of the nucleosome by considering the Boltzmann probability distribution that the nucleosome imposes on the space of all possible sequences. A standard way to get a handle on this distribution is to look at dinucleotide distributions along the nucleosome, as in Section 1.4.

Section 1.5 explained how the nucleosome becomes an even richer system when we allow the bonds between the DNA and the histone core to be broken. This system was modeled in [2], based on the nucleosome model from [1], which also introduced the MMC method. In this chapter we bring these two studies together by using MMC to analyze the sequence preferences of the nucleosome unwrapping under tension, and to design nucleosomes that display the unwrapping behaviour of our choosing. Specifically, the nucleosomes can be destabilized, reducing their kinetic protection from forced unwrapping, and be made to unwrap via a specific path, much like the asymmetric unwrapping of the 601 sequence described in Section 1.5. At the same time, we keep the nucleosome stably bound when not under tension.

The fact that this is possible is in itself surprising, but also has meaning beyond the matter-of-fact conclusion that we have found a nucleosome that behaves a certain way. More importantly, it solidifies the idea we put forth at the end of Section 1.5: that the word nucleosome denotes a class of systems that may have very different properties based on the DNA sequences they contain. This chapter provides a concrete example of the differentiation that is possible among nucleosomes.

3.1 INTRODUCTION

DNA in eukaryotic cells is folded in a hierarchical series of steps into the chromatin complex. Whereas details of the higher levels are still debated, the first level of complexation is well understood: the basic repeated structure, the nucleosome, involves a short stretch of DNA, 147 base pairs (bp) in length, wrapped in 1 3/4 turns around a cylindrical aggregate of eight histone proteins. This results in a disk-shaped particle with a diameter of 11 nm and a height of 6 nm [23]. A short stretch of DNA, called the linker, connects to the next such protein spool. See also Section 1.2.

DNA is a rather stiff molecule with a persistence length of about 150 bp, or 50 nm. Therefore, wrapping the DNA into nucleosomes costs energy, which is compensated by the binding of the DNA backbones to the histone octamer at 14 binding sites [23], see Fig. 1.3. Because the deformation energy of the DNA depends on its nucleotide sequence, the affinity of a given DNA stretch to the nucleosome is dominated by the elasticity and geometry of that underlying sequence. This allows for mechanical cues to be written along DNA molecules, telling nucleosomes where to sit and where not to sit, sometimes called the "nucleosome positioning code" [62] (for earlier versions of this idea see e.g. [114] and [115]).

Remarkably, these cues can even be written on top of genes, because the degeneracy of the genetic code allows for multiplexing [1, 116–119]. Beautiful examples are nucleosome depleted regions before transcription start sites in yeast facilitating transcription initiation [64, 79] (more on this topic in Chapter 6), mechanically encoded retention of a small fraction of nucleosomes in human sperm cells allowing transmission of paternal epigenetic information [29] (also discussed in Chapter 6) or the positioning of six million nucleosomes around nucleosome inhibiting barriers in human somatic cells [50].

So far the role of the DNA sequence has mainly been seen in the positioning (or antipositioning) of nucleosomes. In other words, one scalar quantity is attributed to a 147-nucleotide stretch of DNA: its affinity to the nucleosome.[1] This, however, oversimplifies the possible roles that DNA mechanics can play for nucleosomes. Here we advocate the idea that nucleosomes form a highly diverse class of DNA-protein complexes whose diversity results from the mechanical properties of the DNA sequences involved. There are some first hints in the experimental literature that

1 Histone proteins are evolutionary well conserved, but variants exist, and they can contain posttranslational modifications. Here we neglect these effects and focus exclusively on the role of DNA elasticity.

nucleosomes can have individual properties [27], especially in the case of a nucleosome wrapped with the 601 sequence (Sequence [1] in Appendix B). Recent micromanipulation experiments on this particular nucleosome have revealed its highly asymmetric response to force [84, 120], reflecting an asymmetry in the bending energy of the wrapped DNA [2]. Such a nucleosome would act as a "polar barrier" for elongating RNA polymerases [121]. For this reason, asymmetric nucleosomes may have evolved on real genomes as well, see also [122].

The goal of this chapter is to demonstrate the possibility of designing DNA sequences that lead to special nucleosomes with non-trivial physical properties. The asymmetry of the 601 nucleosome mentioned above is still a somewhat trivial example that simply splits the affinity of the sequence in two parts (and, since it is not particularly difficult to alter the affinities of the two halves, asymmetric nucleosomes may well be the rule rather than the exception). Here we aim to construct nucleosomes that show a set of physical properties that are unlikely to emerge randomly, because they require more careful tuning of the mechanical properties of the nucleotide sequence. We decided to construct nucleosomes that show unusual responses to external tensions.

There is a wide range of experiments on nucleosomes under tension [83, 84, 123]. Most remarkably, nucleosomes can generally withstand rather high tensions without unwrapping completely. This has been explained by the combination of spool geometry and DNA stiffness [2, 82, 110, 124–126]. In order to completely unwrap, the nucleosome has to flip by 180 degrees around its symmetry axis. This leads to a high-energy transition state, the half-flipped nucleosome, between the single-wrapped and fully unwrapped nucleosome. The energy barrier arises due to two strongly bent DNA stretches in the transition state (see Fig. 1.6), which lead to a barrier with a height that increases like the square-root of the applied tension [82, 110]. Nucleosomes, through this force-induced strengthening, are kinetically protected against transient tension.

In nature, nucleosomes will be subjected to tension through the actions of various molecular motors that interact with a cell's DNA [31]. Generally, this kinetic protection is valuable in maintaining the integrity of the chromatin. There are also cases, however, where it may be beneficial to undermine this protection. One such scenario is during the anaphase of cell division, when the mother cell's DNA and its newly produced copy need to separate. This separation can partially fail, because ultrafine DNA bridges between the two copies tend to form at certain fragile sites along the genome [127, 128]. This causes tension on the DNA where the two

copies remain connected. This tension pulls apart the chromatin structure, which is thought to be a signal for repair mechanisms to target the problematic section of DNA. The main mechanism is thought to be the exposure, due to the induced force, of bare DNA, which the repair mechanism has high affinity for [129]. During this repair process, all nucleosomes are expelled from the DNA; therefore, nucleosomes that easily unwrap under tension may be helpful in promoting this repair.

Using the MMC method [1] described in Section 1.4, we will demonstrate that it is possible to construct, *in silico*, nucleosomes that behave perfectly "normal" with respect to their affinity to and their positioning along the DNA molecule, but that display a highly unusual feature in their response to force. When put under tension these nucleosomes fall apart rapidly (several orders of magnitude faster than "standard" nucleosomes) along a predefined unwrapping path. This nucleosome species serves as an example of our general idea: that nucleosomes constitute a class of DNA-protein complexes with a wide range of physical properties.

The use of the MMC method for this purpose is fundamentally no different from its application to the basic, fully wrapped nucleosome as in [1], but it does demonstrate the broad applicability of the method beyond its original purpose. One could imagine applying the same methodology to look for sequences with a range of properties: various other hypothetical nucleosome 'species' that store twist defects or are easily invaded from one side (the 'polar barriers' mentioned above); sequences that easily form DNA loops; and any other DNA system of interest.

3.2 MODELING NUCLEOSOME UNWRAPPING

We employ the same nucleosome model as in previous work [1, 2, 79] (introduced in more detail in Sections 1.2 and 1.5), in which DNA is represented by the rigid base pair model [11] (Section 1.1). This model treats the base pairs that make up a DNA molecule as rigid plates, the spatial position and orientation of which are described by six (three translational and three rotational) degrees of freedom. It assumes only nearest-neighbor interactions, placing a quadratic deformation energy between successive base pairs:

$$E = \frac{1}{2}(q - q_0) \cdot K \cdot (q - q_0), \tag{3.1}$$

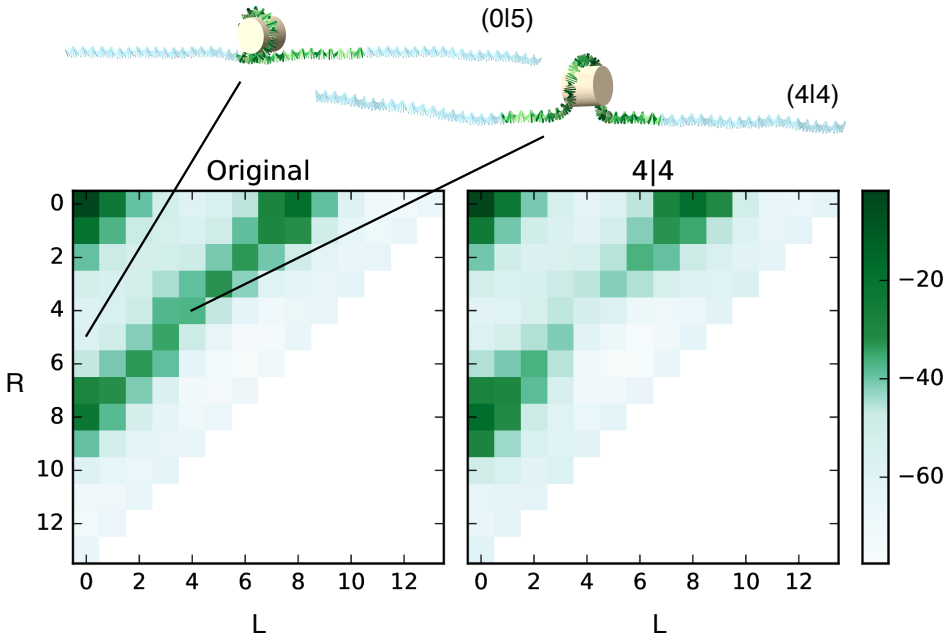

Figure 3.1: Top: two unwrapping states of the model nucleosome under tension, state $(0|5)$ (left) and state $(4|4)$ (right). Bottom left: energy landscape (in units of k_BT) of the nucleosome at position 826 of the YAL002W gene of *S. cerevisiae* under an external force of 14pN. Note that single wrapped states like $(0|5)$ are located in a metastable valley. Nucleosomes with just half a turn of wrapped DNA (e.g. $(4|4)$) form a ridge in the landscape. Bottom right: designing a special nucleosome: result of a free MMC simulation on state $(4|4)$.

where the q and q_0 are six-component vectors that encode the relative degrees of freedom between two base pairs and their equilibrium values, respectively, and K is a six-by-six stiffness matrix.

The sequence-dependence of the model comes into play because every base pair step, depending on which two nucleotides compose it, has its own stiffness and intrinsic shape. These parameters can be found in the literature [11, 13], and we use the same hybrid parameterization [14] as in [1, 2].

The DNA, modeled with the rigid base pair (RBP) model, is forced into a superhelix through a set of 28 constraints that represent the 14 binding sites to the histone octamer and which were extracted from the nucleosome crystal structure without introducing free parameters [1] (see also Section 1.2). In addition, we allow the binding sites of the nucleosome to be opened at the expense of some adsorption energy in the same way as

Table 3.1: Adsorption energies of the different binding sites along the nucleosome. Sites 1 and 14 denote the outermost binding sites, 7 and 8 the innermost. The values are taken from [26], and have been modified as described in [2].

Binding sites	Adsorption energy (kT)
1, 14	11.1
2, 13	13.1
3, 12	14.7
4, 11	11.1
5, 10	12.0
6, 9	16.3
7, 8	18.1

detailed in [2]; the adsorption energies for the different binding sites in the model are provided in Table 3.1. We added 100-base-pair tails with sequence-averaged elastic properties as handles to apply a tension. Example configurations of our model nucleosome under a tension of 14 pN are provided in Fig. 3.1.

In order to analyze the unwrapping of a nucleosome with a given sequence, we put the nucleosome in all possible unwrapping states $(L|R)$ that can be characterized by the number of binding sites opened from the left end, L, and from the right end, R. For each state $(L|R)$ we estimate the average energy from an ensemble of configurations produced by a Monte Carlo simulation. This leads to an energy landscape as a function of $(L|R)$.

3.3 DESIGNING SPECIAL NUCLEOSOMES

In Fig. 3.1 (bottom left) we depict the energy landscape for the unspooling of a particular nucleosome under an external force of 14 pN. We chose 14 pN as the force to which to attune our designer nucleotide sequences, because we wished to work at significant tension, but not such that we leave the regime of stable nucleosomes, and nucleosomes have been found to be stable under tensions of up to about this magnitude [84]. We chose a nucleotide sequence that is associated with a "normal" well-positioned nucleosome, specifically the one at position 826 of the YAL002W gene of *S. cerevisiae* (Sequence [5] in Appendix B), which has been mapped with single-nucleotide resolution *in vivo* [130] and which we have used

before to demonstrate multiplexing of mechanical cues and genetic information [1].

The unwrapping landscape shows the well-known overall features as already predicted with sequence-independent models [82, 124]: (*i*) The most expensive state is the fully wrapped state $(L, R) = (0|0)$; (*ii*) a metastable valley for nucleosomes with a single wrap, $L + R = 5$; (*iii*) a ridge for half-flipped nucleosomes with $L + R = 8$; and (*iv*) the cheapest states, nearly unwrapped nucleosomes, $L + R = 12$. Nucleosomes that are put under an external tension for a short enough time will be stuck in states with $L + R = 5$, kinetically protected by the ridge, as has been observed recently for three other sequences [84]. We expect that this feature is typical for the vast majority of nucleosomes.

However, the number of sequences into which a nucleosome can be wrapped is huge, 4^{147}, and each corresponding DNA double helix has different mechanical and geometrical properties. Could it be that among this huge sea of sequences there is a subset that leads to a very different unwrapping landscape? For example, suppose nature required a nucleosome that acted as a "force sensor", a nucleosome that is stably wrapped and positioned under normal conditions but that quickly falls off as soon as it is put under moderate tension. This might be beneficial in the detection of the ultrafine DNA bridges mentioned in Section 3.1. We are not claiming here that such nucleosome exist on real genomes but we want to check whether they could evolve in principle.

To design a nucleosome that does not get stuck in a set of metastable states we need to cut a trench through the ridge of metastable states. The ridge is caused to the largest extent by the strongly bent DNA portions of half-flipped nucleosomes, see e.g. the $(4|4)$ state shown in Fig. 3.1. What we need are nucleotide sequences that are soft or intrinsically bent in the right direction to substantially lower the cost of these bends.

Our strategy to create such sequences is to perform MMC simulations (see Section 1.4) on nucleosomes that are in an unwrapping state on top of the ridge, for example in state $(4|4)$. In order to arrive at a DNA sequence that provides a low energy cost when wrapped into this state, we use simulated annealing, i.e. gradually lowering the simulation temperature while the algorithm searches the state and sequence space of the nucleosome.

We applied this methodology to all the transition states that sit atop the energy barrier in the unwrapping landscape. Doing so gives us sequences that are favorable to these particular states, and that cut trenches through the barrier at the corresponding locations in the landscape. We performed

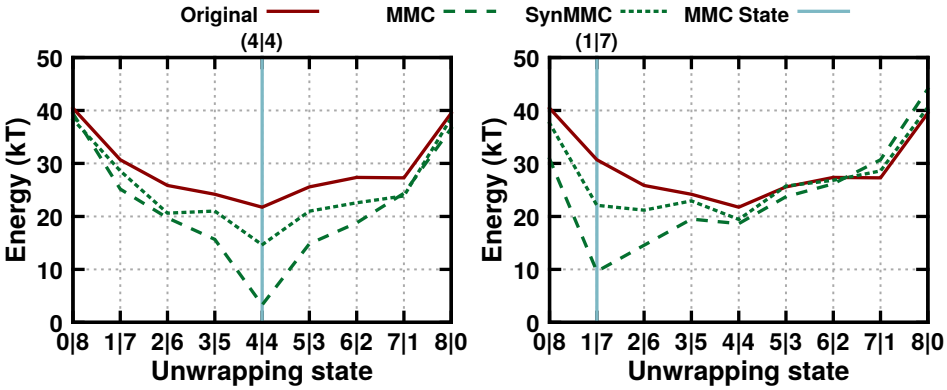

Figure 3.2: Energy along the ridge for sequences found using MMC on position 826 of the YAL002W gene of *S. cerevisiae*, held in unwrapping states (4|4) (left) and (1|7) (right). The solid line represents the original ridge. The dashed line is the ridge after free MMC and the dotted line after SynMMC.

both free MMC, where any mutation is allowed, and synonymous MMC (SynMMC), where only mutations are allowed that do not alter the protein that the DNA sequence encodes for. All the resulting sequences can be found in Appendix B, Sequences [6]–[23].

3.4 PROPERTIES OF OUR DESIGNER NUCLEOSOMES

In Fig. 3.1 (bottom right) is shown the landscape obtained from a sequence that we produced through an MMC simulation performed at state (4|4). The ridge now contains a trench at this position; see also the energy profile along the ridge, depicted in Fig. 3.2 (left). We also performed a SynMMC simulation of the same system, the result of which can also be seen in Fig. 3.2 (left), and shows that we can still dig such a trench on top of genes, albeit not as deep as in the freely mutated case. It is also possible to put a trench at an asymmetric position, see Fig. 3.2 (right), which resulted from free MMC and SynMMC on state (1|7). Taking the nucleosome on the YAL002W gene as reference, we find substantial decreases of the energy at the location of the trench, e.g. reductions of 18.4 for (4|4) and of 12.1 for (1|7) for free MMC, and of 7.1 for (4|4) and of 2.3 for (1|7) for SynMMC (here and below all energies are given in units of k_BT).

In general, changing the sequence of course affects the entire energy landscape and not just the favored state. To learn about how much the rate

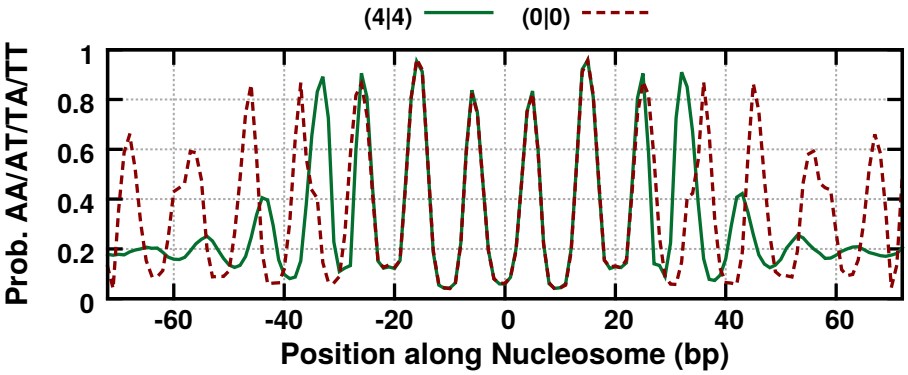

Figure 3.3: Distributions along the nucleosome of AT-rich dinucleotides (AA, AT, TA and TT, frequencies summed) from an ensemble of low-energy sequences of the fully wrapped nucleosome and of one in the $(4|4)$ unwrapping state. The central part, which is wrapped in both cases, is identical. A phase shift occurs in the perpendicularly bent unwrapped tails.

of unwrapping at the given force of 14 pN is affected, we need to calculate the total barrier height, the difference between the lowest energy state on the ridge and that in the metastable valley. Defined as such, the reference nucleosome on gene YAL002W has a barrier height of 18.5 kT. For free MMC, in all cases except $(0|8)$ and $(8|0)$, this difference was substantially reduced, e.g. to 7.4 for case $(4|4)$ and to 13.1 for case $(1|7)$. This suggests that the lifetime of the metastable state would be reduced by 2-4 orders of magnitude. For SynMMC, in five of the nine cases the lifetime is raised (e.g. twofold for case $(1|7)$ as the barrier is now 19.2), in the other cases it is lowered, specifically to 14.5 for $(4|4)$, shortening its lifetime by a factor of about 50.

What do sequences look like that feature such trenches in the landscape? To understand the typical changes in such sequences it is convenient to consider the properties of an ensemble of sequences produced by MMC (i.e. a thermal ensemble of sequences, with the probability distribution given in Eq. 1.5, as opposed to the annealing simulations we have been doing so far). Shown in Fig. 3.3 is the distribution of AA/AT/TA/TT dinucleotides found in an ensemble of 10 000 sequences for the barrier state $(4|4)$ and for the fully wrapped nucleosome. The characteristic 10-base-pair periodic signal for the fully wrapped nucleosomes are due to the well-known nucleotide preferences of high affinity sequences [1, 62, 64, 115]. For state $(4|4)$ we see that in the center of the sequence, which

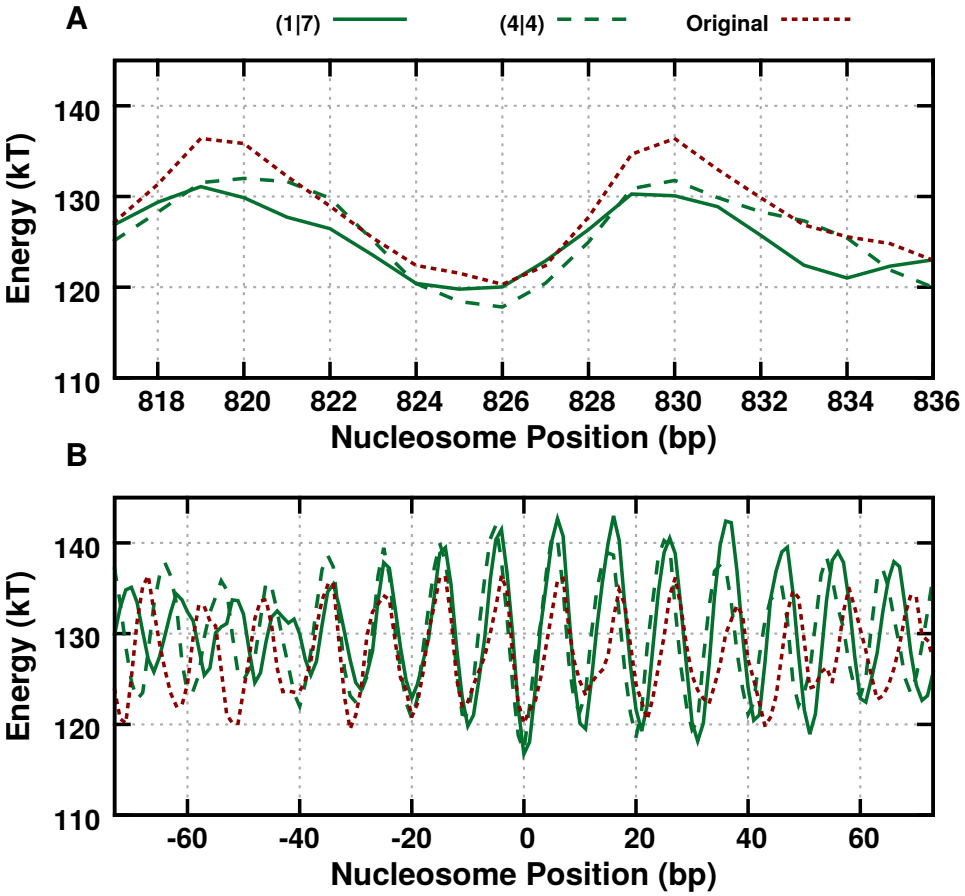

Figure 3.4: **A:** Nucleosome energy landscapes in a small neighborhood of position 826 of the YAL002W gene, with the 826-sequence replaced by the sequences found through SynMMC at states (1|7) and (4|4). In each case, the replacement sequence still provides a local minimum. **B:** Cyclical energy landscapes of sequences found through free MMC for states (1|7) and (4|4) compared to the sequence at position 826 of the YAL002W gene. There remains always a strong local minimum at position 0.

is still wrapped, the preferences are unchanged, but in the bent tails, we have a phase shift by a quarter of a period. This reflects precisely the fact that the bending direction in the DNA arms is perpendicular to the one in the wrapped portion; see the (4|4) example configuration in Fig. 3.1.

We need to check that the sequences we designed actually have good affinities for nucleosomes. In the case of SynMMC, we are modifying a genomic sequence, and we indeed find that there is still a local minimum in the energy landscape along the DNA, see Fig. 3.4A. For the sequences found using free MMC, there is no genomic context to compare to. Therefore, we shift the sequence through the nucleosome cyclically and check that the unshifted sequence is the most favourable one. In Fig. 3.4B we see that we still have strong local minima for the unshifted sequences.

Also note that in both plots in Fig. 3.4, the overall energy at the minima is similar to or reduced with respect to the original minimum. The lower energy is possible because the MMC method is not only adapting the sequence in the unwrapped part (this optimization is at odds with nucleosome affinity, as we have seen). It is also optimizing the still-wrapped part of the sequence to conform to the nucleosome, even better than the original sequence did. The result is that the sequences we designed, when fully wrapped, still give us nucleosomes which have equal or better overall affinity for the nucleosome as compared to the original sequence.

Finally we want to check that the results are not force-specific. The shape of the highly bent sections in the transition state will depend on the force: a higher force will lead to stronger, more localized curvature. Because the main feature of the sequences that facilitate crossing the barrier is likely to be the correct curvature direction, we expect our sequence optimized for 14 pN to also reduce the barrier at other forces. In Fig. 3.5 the effect is shown of the sequence modification on the barrier felt by the nucleosome at a range of forces. We see that, as expected, the barrier is significantly reduced across this entire range, and not only at the specific force at which we designed the sequence.

3.5 CONCLUSIONS

We have shown that the physical properties of nucleosomes, illustrated here through their response to an external force, depend strongly on the physical properties of the underlying nucleotide sequence. Not only can sequences position nucleosomes, but they can also equip them with special individual characteristics. Here we demonstrated this by engineering,

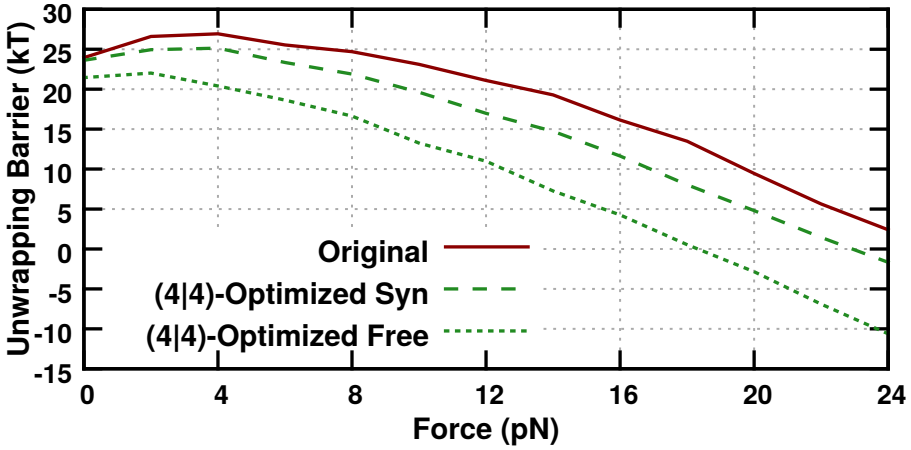

Figure 3.5: The height of the unwrapping barrier as a function of the tension applied, for the original genomic sequence, and the sequences we found through optimization for state $(4|4)$ at 14 pN, using both free and synonymous MMC. Though the sequences were optimized at a given force, they lower the barrier for the entire range of forces considered.

via our Mutation Monte Carlo algorithm, special nucleosomes that are easily unwrapped by an external force, while still being stably wrapped when no force is applied. Surprisingly, these two characteristics can be encoded into a single 147-base-pair nucleotide sequence.

One can imagine that a mechanical evolution of nucleosomes may also occur on real genomes, "speciating" nucleosomes to act as force sensors, polar barriers, twist storers and so on. What makes such an evolution special compared to ordinary evolution is that we have here a very direct mechanical connection between the 147-base-pair sequence wrapped into a nucleosome – its "genome" – and the phenotype, i.e. the set of physical properties of the nucleosome. It will be interesting to scan whole genomes for special nucleosomes and to learn in which genomic context they occur. We are currently developing the methods necessary for this endeavour.

4 A MARKOV-CHAIN MODEL FOR NUCLEOSOME AFFINITY

This chapter is based on:
Tompitak, Barkema and Schiessel 2017 *BMC Bioinformatics 18* 157 [78]

As we saw in Section 1.3, there are many models to be found in the literature that attempt to predict, for a given sequence, its affinity to nucleosomes. One approach is the biophysical one: sequence-dependent models that directly address the mechanics of DNA, such as the Rigid Base Pair Model [11] can be combined with a suitable model for the nucleosome to access the energetics of nucleosome-bound DNA [1, 2, 24–26, 56, 131]. The Eslami-Mossallam nucleosome model [1] described in Section 1.2, which forms the basis for much of the work presented in this thesis, falls into this category.

Another option is to use a bioinformatics model that defines a (Boltzmann) probability distribution on the space of all possible nucleotide sequences. The logarithm of such a probability distribution relates linearly to the free energy of a sequence when wrapped into a nucleosome. One such probability-based model has been put forward by Segal *et al.* [62], and used successfully in follow-ups to that reference [63, 64]. In this chapter we will see that this bioinformatical model can be appropriated beyond its original purpose, in that it can also be used *in silico* to provide a computationally efficient approximation to biophysical models that are themselves computationally too intensive, such as the Eslami-Mossallam nucleosome model.

For the Eslami-Mossallam nucleosome model, the resulting approximation speeds up the calculation of the affinity of a sequence for the nucleosome by a factor of around 10^5 (in an unoptimized implementation). In doing so, this approximative scheme makes it possible to use the biophysical nucleosome model of Eslami-Mossallam *et al.* [1] to analyze far larger sets of sequences. In Chapter 6 we will use it for genome-wide analyses of nucleosome positioning signals, which would not be possible with the pure biophysical model.

In this chapter we will describe the new model and perform a benchmarking analysis of the approximation to the Eslami-Mossallam nucleosome model. We will examine to what accuracy the computationally ef-

ficient model approximates the predictions of the underlying model for the first chromosome of *S. cerevisiae*, and how this accuracy depends on several factors, such as the stringency of the assumptions that go into the approximation, the size of the sequence ensemble from which the model parameters are derived and the application of smoothing filters on those parameters. In doing so, we may also indirectly draw some conclusions as to the accuracy that may be expected of models such as that of Segal *et al.* [62], which are trained on experimental sequence ensembles.

4.1 REPURPOSING THE MODEL OF SEGAL *et al.*

Since a nucleosome wraps 147 base pairs worth of DNA, the space of possible sequences contains 4^{147} or about 10^{88} possibilities. It is impossible to enumerate all of these, so a simple function is needed for the probability distribution.

Segal *et al.* do this by treating a DNA sequence as a Markov chain of order 1, where the probability of a nucleotide at a certain position depends only upon the preceding nucleotide. The probability of the sequence as a whole is the product of the probabilities of all the nucleotides it is composed of. More precisely, defining S as a sequence of length 147, consisting of nucleotides S_i with i from 1 to 147,

$$P(S) = P(\bigcap_{i=1}^{147} S_i) = P(S_{147}|\bigcap_{i=1}^{146} S_i)P(\bigcap_{i=1}^{146} S_i) \tag{4.1}$$

$$= \prod_{n=1}^{147} P(S_n|\bigcap_{i=1}^{n-1} S_i), \tag{4.2}$$

where we have applied the chain rule of probabilities. If we now introduce the assumption we mentioned earlier, that the probability of a nucleotide depends only on the preceding nucleotide, we find the expression given by Segal *et al.*, i.e.

$$P(S) = P(S_1)\prod_{n=2}^{147} P(S_n|S_{n-1}). \tag{4.3}$$

We should stress that the value of quantities like $P(S_n)$ depends not just on the value of S_n (i.e. which nucleotide is represented) but also on the position along the nucleosome, n. These probability distributions for, in the case of Segal *et al.*, dinucleotides, can be obtained by analyzing a suitable

ensemble of sequences that have high affinities for the nucleosome. Segal *et al.* generate such an ensemble from the genome they are interested in making predictions for, by mapping actual (*in vitro*) nucleosome positions along the DNA. Although the original model did not perform very well [75], this model has been applied with success – after a refinement of the model and employing a better training data set – to predicting nucleosome positions, by Field *et al.* [63] and Kaplan *et al.* [64].

These experimental probability distributions do not capture only the intrinsic mechanical preferences of the DNA. They also capture inherent biases in the sample (a genomic sequence necessarily contains only a small subset of all 10^{88} possible sequences) and biases of the experimental method. This makes it difficult to evaluate the accuracy of the model, since both the training of the model and its testing generally rely on the same experimental methods, and there is the risk that agreement between the model and reality is overestimated because the model correctly fits experimental artifacts. Therefore it becomes of interest to study the model in a theoretical framework, where we can isolate the purely mechanical effects.

Ensembles to inform this type of bioinformatics model can also be generated from a theoretical nucleosome model using the Mutation Monte Carlo (MMC) method (see Section 1.4, Fig. 1.5). This method adds mutation moves to a standard Monte Carlo simulation of a nucleosome, thereby sampling the Boltzmann probability distribution of pairs of sequences and spatial configurations (S, θ),

$$P(S, \theta) = e^{-\beta E(S,\theta)}. \tag{4.4}$$

By sampling the sequences during the MMC simulation, the spatial degrees of freedom of the nucleosome model are marginalized and one obtains the probability distribution of the sequences

$$P(S) = \int d\theta e^{-\beta E(S,\theta)} \tag{4.5}$$

and their free energy

$$F(S) = -kT \log(P(S)). \tag{4.6}$$

Note that in Eqs. 4.4–4.6 we have neglected the overall normalization of the probability distributions by the partition function Z, and hence a constant offset $-kT \log(Z)$ to the free energy. Because the probabilities we derive are simply relative frequencies with respect to our sequence ensem-

ble, they are inherently normalized (i.e. summing them over all possible sequences gives unity) and we have no information on the partition function. This is not usually an impediment as we are mostly interested in relative energy differences.

Also note that Eq. 4.6 gives us the free energy in units of kT, with T the simulation temperature. The physical model is defined in units of kT_r, with T_r being room temperature, so what we will want to calculate is

$$\frac{F(S)}{kT_r} = \frac{T}{T_r}\log(P(S)). \tag{4.7}$$

Sampling the entire sequence space is not feasible, but making the same assumption about long-range correlations in the sequence preferences as Segal *et al.*, we can assume that we may write our $P(S)$ as in Eq. 4.3. It turns out it is feasible to produce a sequence ensemble large enough that the distributions $P(S_i|S_{i-1})$ may be determined.

4.2 GENERALIZATION OF THE DINUCLEOTIDE MODEL

We used an MMC simulation of our nucleosome model at 1/6 of room temperature to generate an ensemble of 10^7 sequences, from which the oligonucleotide distributions were derived. At each position, we counted the number of instances of every mono-, di- and trinucleotide and divided these by the total number of sequences in order to obtain the probability distributions.

This gives us the joint probability distribution $P(S_n \cap S_{n-1})$ and not the conditional probability $P(S_n|S_{n-1})$ that we need for Eq. 4.3. This is easily remedied. We can rewrite Eq. 4.3 as

$$P(S) = P(S_1)\prod_{n=2}^{147}\frac{P(S_n \cap S_{n-1})}{P(S_{n-1})} = \frac{\prod_{n=2}^{147}P(S_n \cap S_{n-1})}{\prod_{n=2}^{146}P(S_n)}. \tag{4.8}$$

We see that we can write this equation in terms of the probability distributions of mono- and dinucleotides that we can find from a sequence ensemble. Analogously, if we want to expand the model to trinucleotides, we insert the assumption that the probability of a nucleotide depends only on the previous two (creating a Markov chain of order two) and we find

$$P(S) = \frac{\prod_{n=3}^{147}P(S_n \cap S_{n-1} \cap S_{n-2})}{\prod_{n=3}^{146}P(S_n \cap S_{n-1})}. \tag{4.9}$$

This model can thus be applied using probability distributions for di- and trinucleotides, both to be obtained from a suitable sequence ensemble. The result easily generalizes to tetranucleotides and beyond. For mononucleotides, the model simplifies to

$$P(S) = \prod_{i=1}^{147} P(S_i). \tag{4.10}$$

4.3 BENCHMARKING METHODOLOGY

Segal *et al.* test their model by predicting nucleosome positions along the genome they are studying and comparing with reality and they find that their model has some predictive power, even on genomes on which the method was not trained. However, their study is inevitably hampered by small statistics and their use of natural materials. The latter makes it difficult to judge the quality of their model.

The *in silico* methods allow us to test the model, as an approximation to the full underlying model, much more rigorously. Because we can explicitly calculate the energy of a given sequence, we can directly measure the correlation between the energy given by the theoretical nucleosome model and the probability calculated by the bioinformatics model. Using a standard Monte Carlo simulation of the nucleosome with a given sequence, we can measure the average energy

$$\langle E \rangle_S = \int d\theta E(S,\theta) e^{-\beta E(S,\theta)} \tag{4.11}$$

of the sequence. Unfortunately, calculating the free energy using the Eslami-Mossallam nucleosome model is not straightforward, and we will be comparing $\langle E \rangle_S$ as predicted by the biophysical model with $F(S)$ as predicted by the approximative model. At finite temperature, these quantities are not the same, differing by an entropic contribution. However, at low enough temperatures they converge, and for nucleosomes the entropic contribution is not strongly sequence-dependent, as we will see in Chapter 5. We will compare the predictions at 1/6th of room temperature, as some finite temperature is needed for the statistical simulations to function. In performing this comparison, we thus provide an upper limit for the discrepancy between the approximation and the real $\langle E \rangle_S$.

In order to generate an energy landscape with which to compare the results of the probability-based models, we take the first chromosome of

S. cerevisiae ($\sim 2 \times 10^5$ base pairs) and perform a Monte Carlo simulation of the nucleosome wrapped with each 147-base-pair subsequence of the chromosome, using the Eslami-Mossallam nucleosome model. After letting the simulation equilibrate, we sample the energy of the system and take the average. In order to be able to compare this energy landscape with a probability landscape, we calculate the (Boltzmann) probability distribution and normalize this over the set of sequences for which we calculated the energy, and then take the logarithm to regain our (shifted) energy landscape.

Analogously, we use the probability-based model to generate a probability landscape of the same sequence. This we normalize over the set of sequences analyzed and convert to an energy using Eq. 4.6. We find that this procedure is about five orders of magnitude faster than using the full biophysical model.

We only know the free energy up to some constant offset, but by making sure both the real energy landscape given by the energetic model and the approximate energy landscape provided by our probability-based model have the same normalization, we can readily compare the two.

In doing so, we may draw some conclusions about this kind of Markov-chain model not only as it relates to the nucleosome model we consider here, but about the assumptions that go into it in general, i.e. the explicit assumption of short-range correlations and the implicit assumption that the sequence ensemble on which the model is being trained is large enough. To test the first assumption, we extend the dinucleotide model used by Segal et al. to mononucleotides (which assumes no correlations at all) and trinucleotides (which relaxes the assumption of short-range correlations) and compare their accuracy. For the second, we examine the accuracy of these three models as a function of the ensemble size on which they are trained.

4.4 COMPARISON OF THE MONO-, DI- AND TRINUCLEOTIDE MODELS

We tested and compared three different probability-based models, namely the Segal *et al.* dinucleotide model, its simplification to mononucleotides and its extension to trinucleotides. Following the methodology outlined in the previous section, we arrive at correlation plots for the energy as given by the energetic model and as predicted by the probability-based models. The results are presented in Fig. 4.1A-C.

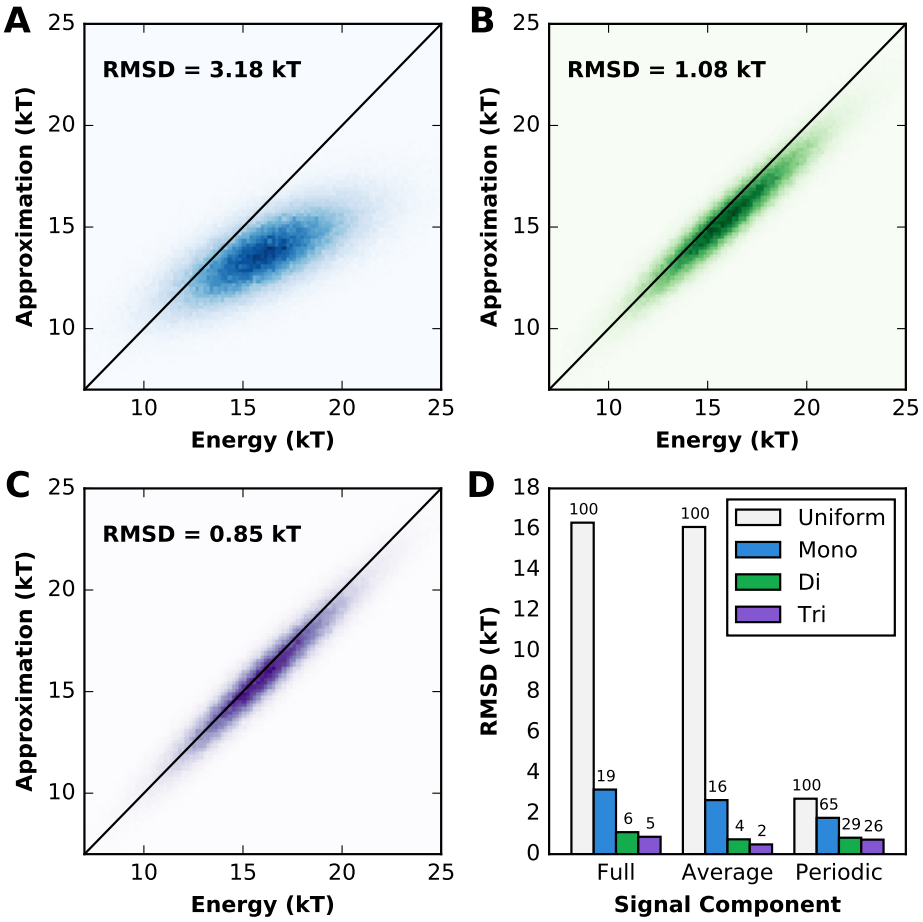

Figure 4.1: Accuracy analyses of the various models, benchmarked on the first chromosome of *S. cerevisiae*. **A**: Histogram of the energy prediction pairs of the full model and mononucleotide approximative model for the same sequences. The black diagonal indicates perfect agreement. **B, C**: As **A** for the dinucleotide and trinucleotides approximations, respectively. **D**: Comparison of the root mean square deviations of the approximative predictions from those of the full model. The grey bars indicate the RMSDs of 'bad' models, defined for the Full and Average signals as a uniform landscape, and for the periodic signal as the real landscape shifted out of phase. The other values, for the mono-, di- and trinucleotide approximations are compared with these bad models. Indicated above each bar is a percentage indicating the value relative to the corresponding bad model.

As we might expect, the longer the oligonucleotides we use, the better the agreement becomes. An important cause of the deviation from perfect agreement, apart from the spread, is a clearly visible deviation in the slope. The mononucleotide model significantly underestimates the spread in energies. This means that the mononucleotide model is not capturing effects that set sequences apart from each other. This effect is expected and should be remedied by going to longer oligonucleotides. Indeed we see this deviation greatly decreased for the dinucleotide model, and even more so for the trinucleotide model.

For a more detailed grasp on the quality of the predictions, we separate out two components of the energy landscape that are important on their own. The first is the periodicity of the energy landscape. Due to the helical nature of DNA, energy landscapes for the nucleosome show a roughly 10-base-pair periodic signal. It is important that any model for nucleosome affinity gets the frequency and phase of this periodicity right. The second property, complementary to the periodicity, is the overall energy level of the sequence. This aspect will show us how well the model captures long-range effects.

For the purposes of benchmarking, we define the local average as the 11-base-pair running average of the energy landscape, i.e. over about one period. The pure periodicity of the signal we analyze by subtracting from the signal its local average as just defined, making the signal oscillate around zero. Our benchmarking results then consist of the root-mean-square deviation (RMSD) for the full signal (already presented in Fig. 4.1A-C), for the locally averaged signal and for the pure periodicity signal.

To get a sense of what the RMSD values we find actually mean, we compare them to the RMSD value we find when we use a bad model. For the overall signal and the locally averaged signal, we define this bad model to be one that contains no sequence information at all, i.e. a perfectly uniform landscape. For the periodicity, this is not such an interesting comparison because for a periodic signal, a uniform landscape is still right twice per period. Instead we utilize as a bad model the same signal, but shifted by half a period, to push it out of phase.

RMSD values gathered from such bad models tell us about the typical size of the structures in the energy landscape that our models need to predict. We can then measure the RMSD from our benchmarked models relative to this scale. Fig. 4.1D displays the results. We see a decrease in RMSD when going to longer oligonucleotides in each of the three cases. The dinucleotide model, as used by Segal *et al.*, already performs well, with an overall RMSD of 7%. Noteworthy, it is much more accurate than the mononu-

cleotide model. However, we see that we could improve our results still by going to trinucleotides. Especially the local average is predicted much more accurately by the trinucleotide model, cutting the RMSD by about a third.

4.5 THE IMPORTANCE OF SAMPLE SIZE

Because we can produce large ensembles of sequences *in silico* with the Mutation Monte Carlo method, we are now also in a position to get a measure of how large an ensemble we need for our models to make accurate predictions.

In their 2006 study, Segal *et al.* manage to build an ensemble of $\sim 10^2$ sequences. Apart from the inherent biases that may be present in their ensemble due to their use of nonrandom yeast DNA, this is not a very large ensemble, and we should check what the effects of such limitations are.

In a later study, Kaplan *et al.* perform a similar study, where they obtain 35,000,000 sequence reads. [64] The ensemble is again trained on the yeast genome, which is some 12,000,000 base pairs long. The number 35,000,000 should therefore not be mistaken for the ensemble size. There must necessarily be many duplicate and strongly overlapping sequences in their ensemble, which arise artificially because only a small subset of sequence space is available for sampling. Giving a meaningful number for the effective sample size of such an ensemble is difficult. However, a sequence of $\sim 10^7$ base pairs can yield $10^4 - 10^5$ completely non-overlapping nucleosome sequences, which we may employ as a conservative estimate.

Later similar work using the mouse [132] and human [133] genomes has yielded larger ensembles. These genomes are two orders of magnitude larger than that of yeast, and so also provide that many more non-overlapping sequences.

In our *in silico* simulations, we built an ensemble of 10^7 independent sequences from which we derived our probability distributions. We took subsets of these sequences to see what the effects of smaller sample sizes are. The problem when statistics are small is not just that the probability distributions are less accurate. We additionally run into the issue that some rare dinucleotides simply do not appear in the ensemble at all. The estimate of their probability then becomes zero. The problem is that if any of the factors in Eq. 4.2 is zero, the entire product becomes zero, rendering the model useless.

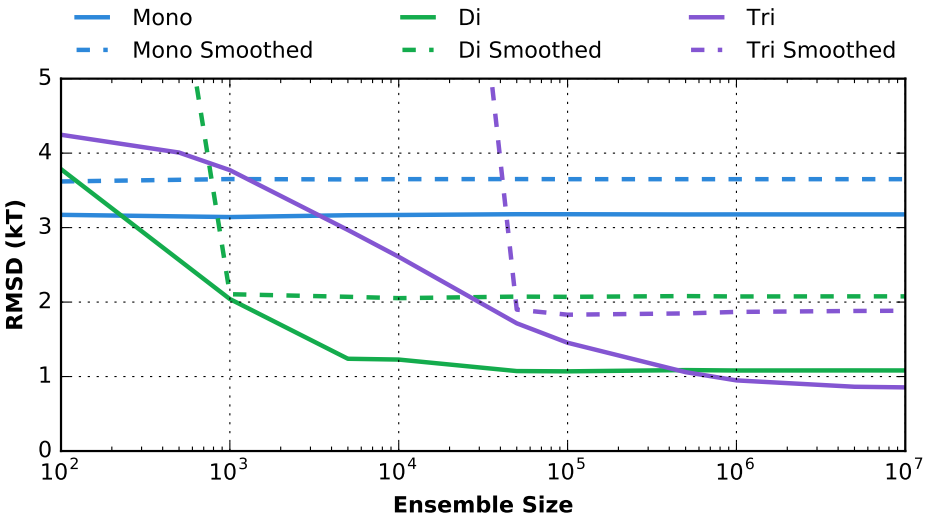

Figure 4.2: Variation of the RMSDs of the various models with the size of the sequence ensemble from which their parameters are calculated. **Solid lines:** zero-probability issues are dealt with by assuming zero information. **Dashed lines:** probability distributions are smoothed with a 3-bp running average. The performance when smoothing is strictly worse.

For Segal *et al.* and Kaplan *et al.* this problem does not arise, because they do not need to work at low temperatures, but also because they apply a smoothing to their probability distributions. They estimate the probability $P_n(S_n \cap S_n - 1)$ of a dinucleotide by averaging over not just position n, but also $n-1$ and $n+1$. This is justified by the observation that their experimental method does not provide them with a sharp resolution down to the base-pair to begin with. The effect of such smoothing is not *a priori* clear, however. In a landscape with 10-bp periodicity, taking a 3-bp running average could have averse effects. Such smoothing may not be necessary or beneficial when applied to higher-resolution data.

We therefore propose an alternative method, where instead we consider a probability of zero, for any position, a failure of the ensemble. In such a case we conclude that we simply do not have any information, i.e. we artificially insert a flat conditional probability of 0.25.

In Fig. 4.2 are presented the RMSDs of the full landscape, as predicted by our probability-based models, with probability distributions derived from various ensemble sizes. We find that smoothing the distributions

gives results that are strictly worse than simply assuming no information when an issue arises.

We can conclude from this plot that the model of Kaplan *et al.*, even with a conservative estimate for their effective ensemble size, should perform well. The dinucleotide model converges to its maximum accuracy at only 10^4 sequences. Of course, caveats surrounding the non-randomness of the DNA being sampled remain.

For larger experimental ensembles (e.g. [132] and [133]) it is advisable to move to a trinucleotide description. It requires a larger ensemble to be accurately parameterized, but starting from 5×10^5 sequences, this model becomes more accurate than the dinucleotide model.

4.6 CONCLUSIONS

With the methods available for the first time to produce sequence ensembles for nucleosome affinity based on an energetic model of the nucleosome, we investigated the capacity of a class of probability-based models to approximate real energetics. As an approximative scheme to the nucleosome model of Eslami-Mossallam et al. [1], we find errors on the order of 1 kT. This is not an insignificant disagreement, but depending on the application, this price may well be worth paying for the vast reduction in computational complexity by a factor of 10^5 (using an unoptimized implementation). Vast increases in speed can also be expected for other complex biophysical models.

Considering the assumption of short-range correlations, we find that dinucleotide models such as those used by e.g. Field *et al.* and Kaplan *et al.* already perform well, with a root mean square deviation of about 2 kT (see Fig. 4.2). However, we also find that improvement could be achieved by going to a trinucleotide model (for large enough ensemble size), and by avoiding the smoothing of the probability distributions.

We also looked into the effects of small ensemble sizes, and we find that an ensemble such as used by Field *et al.*, although caveats must be acknowledged as to likely inherent biases in their experiment, is sufficient for the dinucleotide model to reach its fundamental accuracy. For larger ensembles (10^6 or more sequences) such as provided by the mouse or human genome, however, we recommend that the trinucleotide approximation be used for higher accuracy.

We hope, however, that our work will motivate the experimental community to look into mapping nucleosomal sequence preferences experi-

mentally using more random DNA sequences than are provided by natural genomes. A starting point could be a very similar study done on DNA rings [93]. This would allow us to better examine the intrinsic sequence preferences of nucleosomes without biasing them towards a genomic context.

5 PERFORMING SELEX EXPERIMENTS *IN SILICO*

This chapter is based on:
Wondergem, Schiessel and Tompitak 2017 *submitted*

In the previous chapter, we introduced a new method that takes the idea behind Mutation Monte Carlo (see Section 1.4) and turns it into a model for nucleosome affinity. As the list of references in Section 1.3 shows, there are already a great many models out there that try to predict nucleosome affinity. In fact, the Markov chain model presented in the previous chapter is based heavily on one of them,[1] and it is reasonable to ask what we gain by adding this new method to the available zoo of models.

As already noted in the previous chapter, which introduced and benchmarked the model, we already gain something simply by effecting the wedding between biophysical and bioinformatical modeling that the MMC methodology makes possible. We obtain a computationally efficient model that can be used to analyze large numbers of sequences, but that we can still understand from a physical perspective. The underlying model we have used for the nucleosome can be defended on physical grounds to be the most realistic one that is reasonably tractable computationally, but the methodology of MMC and the Markov chain model are independent of the physical model of choice.

In Chapter 6 we will see that the model brings together not only computational efficiency and physical understanding, but also accuracy of prediction, when we apply it to its most obvious use: the analysis of real biological sequences. Before turning to biology, however, we will first examine further the relationship between MMC and the Markov chain model. The latter is not simply an approximation or a corollary to the former, as we will see. The Markov chain approach enriches the MMC methodology and extends its applicability. Through it, we will better understand and control the meaning of temperature in MMC simulations, learn about the differences between nucleosomes and DNA rings and bridge a gap between simulation and experiment.

1 That of Segal *et al.* [62], as explained in the previous chapter.

5.1 INTRODUCTION

Over the past 25 years, SELEX (Systematic Evolution of Ligands by EXponential enrichment) experiments have proven a valuable tool in identifying DNA and RNA sequences with high affinity for a large range of target molecules. This affinity can be based on any number of properties of the nucleic acids, like sequence-specific binding of the target or an RNA's ability to form stem loops. SELEX experiments have found many of their applications in clinical research: to examine the tendency of prospective therapeutic compounds to target specific genomic sequences, or designing RNA molecules that themselves interfere with the functioning of certain pathogens. (For a review, see [134].)

We will focus on the basic mechanics (elasticity and intrinsic shape) of double-stranded DNA molecules and their consequent affinity for certain complexes in which the DNA needs to be deformed. Various DNA-binding proteins are known to have DNA affinities that are dependent on the intrinsic curvature and stiffness of the underlying nucleotide sequence, such as the catabolite activator protein [135], the TATA-binding factor [136–138] and other parts of the transcriptional machinery [139–141], as well as regulatory [107, 142–144] and architectural proteins [141, 145].

However, the archetypical example is the nucleosome, a protein spool around which genomic DNA in eukaryotes is wrapped in order to compactify it [22]. The positioning of these protein spools along a genome influences the packaging of the DNA and thereby the expression of genes, as wrapped-up DNA cannot readily be read out [36]. Since DNA needs to be strongly bent in order to wrap into a nucleosome, the nucleosomal structure has a preference for sequences that facilitate this deformation. This leads to significant effects of the underlying DNA sequence on the positioning and dynamics of nucleosomes [27].

In this context, SELEX experiments have been used to look for DNA sequences with high affinity to the nucleosome [54, 55, 85] (as well as the archaeal 'nucleosome' [146]). In similar endeavors, the SELEX method has been used to look for intrinsically curved sequences [147] and to assess the sequence preferences of DNA rings [93].

In such SELEX experiments, a pool of random DNA molecules is synthesized (either completely randomly or randomly drawn from genomic sequences [148]), and these random molecules are mixed with molecules of the target type, competing to bind to them. The DNA molecules with the highest affinity will be most likely to bind to the targets. After some

time, the DNA-target complexes are extracted from the mixture, leaving behind a fraction of the DNA molecules that have a lower average affinity, and keeping a fraction with higher affinity.

By repeating this process for multiple rounds, the selective pressure on the DNA sequences increases and we end up with a smaller and smaller pool of higher and higher affinity sequences. In such a manner, the Widom 601 sequence [85] of high nucleosome affinity was discovered, and the dinucleotide probability distributions of DNA rings were mapped [93]. Although not the same on a technical level, similar experiments have been used to map the sequence preferences of nucleosomes [62–64, 115, 130, 149]. Mapping such preferences is not only an interesting goal in itself, these preferences can also be used to model sequence-dependent nucleosome affinity. Such models can in turn be employed to gain insight into the mechanical signals encoded into genomic DNA sequences [62, 64, 78, 79].

Mutation Monte Carlo (MMC, see Section 1.4 and [1]) also enables mapping of such sequence preferences. This method utilizes standard Monte Carlo simulations to sample the Boltzmann distribution associated to a modeled DNA system such as the nucleosome, and adds as a novel feature Monte Carlo moves that mutate the DNA sequence. Given a suitable model of the system of interest, this technique allows an understanding of the sequence preferences of the system from a theoretical point of view.

The MMC method is similar in many ways to the experimental SELEX method. It samples DNA sequences based on their affinity to the target. Doing so at constant finite temperature delivers probability distributions for e.g. dinucleotides (as in [1, 78, 113]), and by performing simulated annealing it searches for the sequence with the strongest affinity ([86, 113]), much as attempted in [85], leading to the 601 sequence.

However, there is also a major difference between the *in silico* method and the experimental protocols. The MMC simulation is performed at a particular temperature, which determines how stringently it selects for low-energy states and hence for high-affinity sequences. This temperature is necessarily shared by both the configurational moves, which simulate the thermal fluctuations of the system, and the mutations. In a SELEX experiment, however, the selection pressure is determined by, among other factors, the number of rounds of selection performed, and the strength of selection on the sequences is decoupled from the temperature at which the experiment is performed. This means that, despite the similarities, a MMC simulation cannot be directly taken as an *in silico* SELEX experiment.

Here we bridge this difference, such that we may apply selective pressure *in silico* at will regardless of the simulation temperature. To do so, we must examine in detail the role played by temperature in the MMC method, which we will do in Sections 5.2 and 5.3. Considering MMC simulations of both nucleosomes and DNA rings, we will find in Section 5.4 that the importance of the temperature varies from system to system.

With the tools in hand to perform simulated SELEX experiments, we first emulate the experiment performed by Rosanio *et al.* for rings [93]. In Section 5.5 we elucidate the fundamental differences between the (out-of-equilibrium) experiment of Rosanio *et al.* and our idealized equilibrium statistics, to show that a comparison is useful. After affirming this, we perform the *in silico* selection in Section 5.6 and we find both broad agreement and some striking differences between the theoretical predictions and the experimental results. Finally, in Section 5.7, we apply our SELEX simulations to tight and overwound rings, which would be difficult to treat experimentally due to the lower rate of formation of such systems.

5.2 SELEX AND MMC

In a SELEX experiment, DNA molecules compete to bind to target molecules or, in the case of DNA rings, to form closed rings in a limited amount of time [93]. The probability of a molecule with sequence S to be bound to the target instead of another, assuming equilibrium conditions, is proportional to the Boltzmann weight of that molecule's free energy when bound to the target,

$$P(S) = \frac{1}{Z} e^{-\beta F(S)}, \tag{5.1}$$

where Z is the partition function, i.e. $\sum_S e^{-F(S)/kT}$.

A single round in a SELEX experiment is then very similar to a MMC simulation. When we run a MMC simulation, we are sampling system configurations, i.e. combinations of sequences and spatial configurations (S, θ), according to their Boltzmann distribution:

$$P(S, \theta) = \frac{1}{Z} e^{-\beta E(S,\theta)} \delta(f_c(\theta)). \tag{5.2}$$

The normalization is provided by the partition function Z, obtained by integrating the numerator over all spatial degrees of freedom and summing over all sequences. In this equation we have added a delta function

to encode for the constraints on the system. In a nucleosome, there are constraints on the spatial degrees of freedom that bind the DNA to the histone core. In a DNA ring, the molecule is constrained to form a loop. The exact form of these constraints may be complex, and is captured here by a general constraint function f_c.

When we speak of the affinity of a sequence to a nucleosome or a ring, we do not make reference to any particular spatial configuration. Rather we want to take all of them into account; we need the probability of a given sequence to form a nucleosome or ring, considering the probabilities of all the possible spatial configurations the DNA may take. Then what we wish to calculate is the marginal probability distribution of the sequences:

$$P(S) = \int d\theta P(S,\theta) = \frac{1}{Z}\int d\theta e^{-\beta E(S,\theta)}\delta(f_c(\theta)). \tag{5.3}$$

This integral will not generally be tractable. In the current work, we rely on the Rigid Base Pair model [11] to provide the energy function $E(S,\theta)$. This energy function is quadratic in the degrees of freedom, making the integral above a Gaussian integral under constraints. This may be solvable for very simple constraint functions, but in general we need to resort to numerical methods like MMC.

Assuming we have a method to evaluate $P(S)$ we can consider the free energy of a given sequence:

$$P(S) = \frac{1}{Z}e^{-\beta F(S)} \rightarrow F(S) = -\frac{1}{\beta}\left(\log(P(S)) - \log(Z)\right). \tag{5.4}$$

The partition function is generally difficult to determine. In what follows we will neglect its contribution, meaning that we determine the free energy only up to a constant offset. Similarly, we will simply normalize our probability distributions as required, and drop overall factors from our equations.

However, besides this caveat, we are determining the same quantities as we would in a SELEX experiment, at least when considering only a single round. In the next section we address simulating SELEX experiments consisting of multiple rounds.

5.3 AN EFFECTIVE TEMPERATURE FOR MUTATIONS

As noted, the probability of a given sequence to survive a SELEX round depends on its free energy when bound to the target. Assuming a fraction

f is kept after a round of SELEX, the survival probability of a sequence S is

$$P_{\text{surv}}(S) = fe^{-\beta F(S)}. \tag{5.5}$$

For the sequence to survive multiple rounds, assuming selection criteria are constant from one round to the next, we multiply this probability with itself,

$$P_{\text{surv},n}(S) = f^n e^{-n\beta F(S)} = f^n e^{-\beta'_m F(S)}. \tag{5.6}$$

The fraction f can, in the case of DNA forming nucleosomes or other complexes, be constrained to be smaller than 1 by mixing together a surplus of DNA molecules with the target proteins.

Apart from the scaling with a sequence-independent prefactor, we see that applying n rounds of SELEX is equivalent to introducing an effective temperature, $T \to T' = T/n$. We call this an effective temperature, since it only applies to the selection of the sequences. In what follows we will therefore distinguish between β_m, the inverse temperature that is applied to sequence selection (i.e. the mutations in our MMC simulation), and β_s, the inverse temperature of the spatial degrees of freedom. In Eq. 5.6, the actual physical temperature of the system is not altered. We wish to replicate this effect in our MMC simulations.

The free energy in Eq. 5.4 depends on the simulation temperature and, as noted in the introduction, this temperature governs both the selection of sequences and the selection of spatial configurations during the simulation. However, there is nothing to stop us from tweaking the temperature after marginalizing out the spatial degrees of freedom. If we wish to calculate $P(S)$ at some temperature T' other than the simulation temperature T, we may simply write

$$\begin{aligned} P_{T'_m}(S) &= e^{-\beta'_m F(S)} = \left(e^{-\beta_m F(S)}\right)^{\beta'_m/\beta_m} \\ &= P_{T_m}(S)^{\beta'_m/\beta_m}, \end{aligned} \tag{5.7}$$

where the temperature subscript to $P(S)$ denotes an *effective* temperature for the mutation moves only. Note that this is distinct from changing the actual simulation temperature, in which case we must write

$$P_{T'}(S) = \int d\theta e^{-\beta' E(S,\theta)} \delta(f_c(\theta)) \tag{5.8}$$

$$= \int d\theta (e^{-\beta E(S,\theta)})^{\beta'/\beta} \delta(f_c(\theta)). \tag{5.9}$$

The question of how this expression scales with T' is not straightforward to answer and depends on the constraints placed upon the system, as we will see.

Assuming we can calculate $P(S)$, Eq. 5.7 allows us to decouple the selective pressure on the sequences from the simulation temperature, in a manner entirely analogous to how a SELEX experiment introduces an effective temperature for the sequence selection. Furthermore, we are not restricted to temperatures that are integer fractions of the physical temperature; we may choose T' as we like, even a temperature larger than the physical one.

5.4 EFFECTIVE TEMPERATURE AND SEQUENCE PREFERENCES

In order for Eq. 5.7 to be of use, we need a tractable way to calculate $P(S)$. The MMC method allows us to sample the Boltzmann distribution in sequence space of the system of interest, but sampling the full space of all possible sequences is still an impossible task for systems like the nucleosome, due to the large number of sequences.

The standard way of gaining insight into the sequence preferences of a system is by considering the probability distributions of short subsequences in the full sequence, most commonly those of dinucleotides [1, 62–64, 93, 113], which is a far more tractable problem. Those distributions capture much of the information about a system's preferences, and they can in fact be employed in calculating the affinity of sequences, as we saw in Chapter 4. We will use our trinucleotide model,

$$P(S) = P(S_1)P(S_2|S_1)\prod_{i=3}^{N} P(S_i|S_{i-1} \cap S_{i-2}), \tag{5.10}$$

where the S_i are the individual nucleotides that make up the DNA sequence. This expression for $P(S)$ assumes that the probabilities of the individual nucleotides are only strongly correlated with their nearest and next-nearest neighbours, i.e. the probability of S_i depends only on S_{i-1} and S_{i-2}. We tested this assumption extensively in Chapter 4. Using Eq. 5.10, we may sample the probability distributions of trinucleotides in our MMC simulation and from there calculate the probability or free energy of an entire sequence.

With this method for calculating $P(S)$ in hand, we can now gather an ensemble of sequences at a different mutation temperature by running a MMC simulation in sequence space only, but where we reject or accept mutations (within the Metropolis-Hastings algorithm) based on the adjusted probabilities given by Eq. 5.7.

From this new sequence ensemble we can then once again derive dinucleotide distributions to study. Comparing the distributions found using this method, with the original ones from the single-temperature MMC simulation, we may assess separately the effects of changing the mutation temperature and the spatial temperature.

We modeled DNA using the Rigid Base Pair model [11] with the standard hybrid parameterization [14]. We ran MMC simulations of nucleosomes (modeled using the Eslami-Mossallam nucleosome model [1]) and rings (modeled by connecting the first and last base pairs of the DNA using the standard sequence-dependent elasticity of the Rigid Base Pair model) at three different temperatures: room temperature, 1/2 of room temperature and 1/4 of room temperature. Then we used the method just described to independently alter the mutation temperature. The results are presented in Fig. 5.1.

The distributions for A/T-rich dinucleotides (a common set to study due to the strong preferences shown by the nucleosome for the positions of these dinucleotides) for the ring and the nucleosome show an interesting difference. In Fig. 5.1D-F, we see that the distributions we find for the ring depend strongly not just on the mutation temperature β_m, but also on the spatial temperature β_s. For the nucleosome, however, we see in Fig. 5.1A-C a strong dependence on β_m, but a far weaker dependence on β_s.

This difference can be understood in terms of the entropic contribution to the free energies of the systems. Considering a given sequence S, its free energy has a contribution from the average internal energy of a sys-

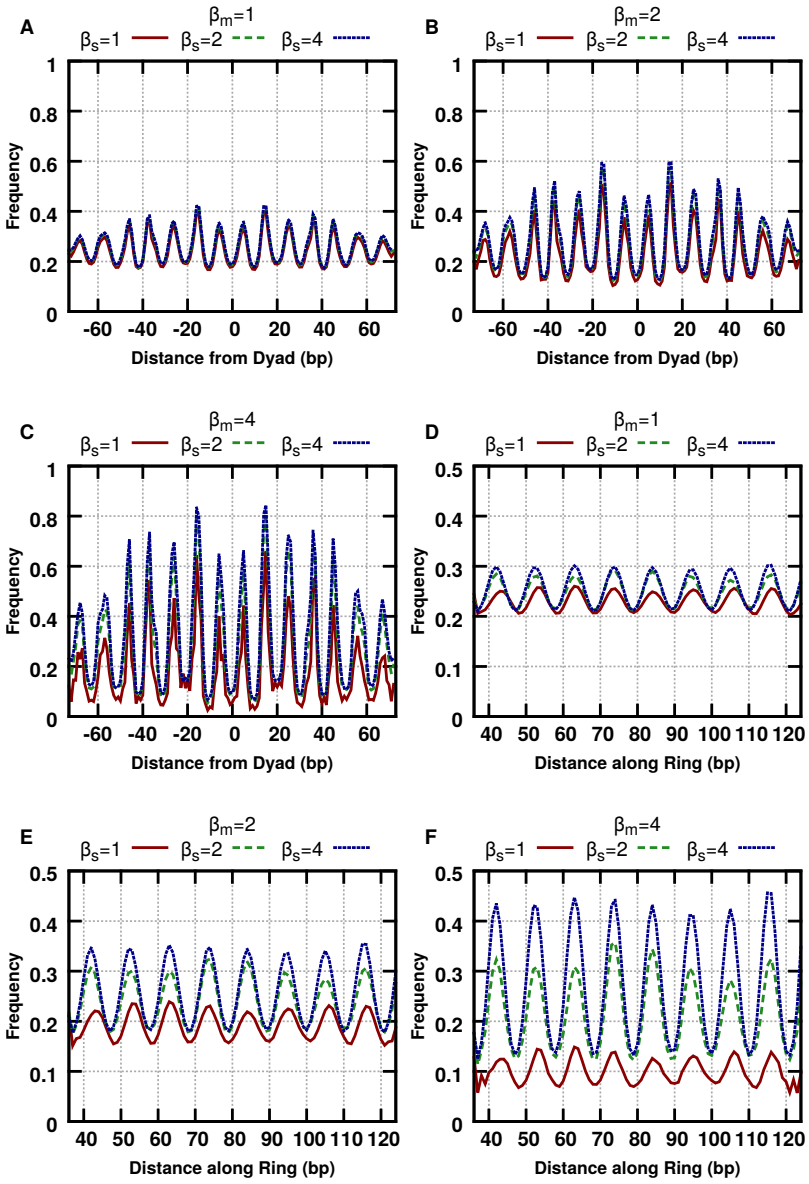

Figure 5.1: Distributions for AT-rich dinucleotides (AA, AT, TA and TT) along the nucleosome and the ring biased by the locking sequence from Rosanio *et al.* [93] (Sequence [24] in Appendix B), for different combinations of mutation temperature (β_m) and spatial temperature (β_s). The distributions are grouped by mutation temperature in order to illuminate the different effects of spatial temperature on the preferences of the nucleosome and the ring. The effect of multiple rounds of SELEX would be to raise β_m while keeping β_s constant, so one would consider the curves of the same color in successive plots.

tem, and from the entropy (denoted here by H to distinguish it from the sequence S),

$$F(S) = \langle E(S) \rangle - T_s H(S), \tag{5.11}$$

where T_s is the spatial temperature, as we are considering the system with a given sequence S.

Since the entropy is a measure of the part of configuration space that can be accessed with reasonable probability by the system, it in principle depends on the sequence. For example, for a completely free DNA molecule, a stiff sequence will limit the possible spatial configurations of the molecule more than will a sequence that bends very easily. Changing the spatial temperature will affect the accessible part of state space, and hence the contribution $T_s H(S)$, in a sequence-dependent manner.

The average energy $\langle E(S) \rangle$ also depends on temperature, but only in a sequence-independent manner. It represents the internal potential energy plus the thermal energy, simply given by the equipartition theorem:

$$\langle E(S) \rangle = E_0(S) + \frac{N}{2} k_B T_s, \tag{5.12}$$

where N is the number of degrees of freedom.

The dependence of the sequence preferences of DNA rings we find in Figs 5.1D-F are thus an entropic effect. At lower temperatures, the ring will be constrained to a smaller set of configurations, but how many depends on what the stiffness of the DNA sequence allows. Hence, lowering the spatial temperature increases the differences in affinity between sequences, leading to the larger amplitudes in Figs 5.1D-F.

For the nucleosome, the effect is much smaller. Apparently, the entropic contribution $T_s H(S)$ is not strongly sequence-dependent in this case. This was expected: because the nucleosome is a strongly constrained system, the part of configuration space that the DNA is allowed to sample is determined to a much larger degree by the constraints on the system than by the elastic properties of the DNA itself. This was already anticipated in works like [1, 2] and in Chapter 3, where the entropic contribution to the free energy of the nucleosome was neglected entirely. Using our new methodology, we are able to directly verify that this assumption is justified. However, we must conclude that the assumption does not hold for systems that are not as tightly constrained as the nucleosome, like for instance DNA rings.

5.5 AN IN SILICO SELEX EXPERIMENT FOR RINGS

Having developed the methodology to perform SELEX experiments *in silico*, we would like to compare the results of such computational treatments to experimental results. The most promising experiment to compare to is that of Rosanio *et al.* [93], the only experiment making use of completely random sequences for which the statistics we are interested in have been reported.

Rosanio *et al.* performed a SELEX experiment in which fragments consisting of 126 base pairs of DNA were made to cyclize into rings. Linear DNA fragments randomly sample bent configurations due to thermal fluctuations, and if the two ends of a fragment meet, a ligation reaction may fuse them together, creating a closed ring. The probability of a given DNA fragment cyclizing depends on its affinity to form a ring: a stiff sequence is less likely to cyclize and survive a selection round than an easily bendable one. To gain insight into the sequence preferences of rings, Rosanio *et al.* fixed 36 of the 126 base pairs to contain a predetermined sequence (Sequence [24] in Appendix B) with a known preference for bending in one direction. This biased the direction of ring formation, such that the preferences of a ring bent in a specific direction could be mapped.

We wish to mimic this experiment *in silico* by performing a MMC simulation of a DNA ring, with 36 base pairs fixed to the same sequence used by Rosanio *et al.*, and the rest free to mutate. We found that the RBP model correctly captures the fact that the 36-base-pair locking sequence biases the bending direction in the ring. Fig. 5.2 shows histograms of the rotational states of DNA rings with locking sequences, and the other 90 base pairs made into homogeneous (sequence-averaged) DNA without a coherent bending preference, sampled during a standard Monte Carlo simulation. The top panel shows the results using the Rosanio sequence, the bottom uses the artificially designed, very strongly intrinsically bent sequence from Chapter 2 (Sequence [3] in Appendix B). The fixed sequences significantly bias the ring to a subrange of rotational states; the artificially designed sequence far more strongly than does the Rosanio sequence, for which reason we will employ it later on. As a side remark, note that the energy landscape as a function of the rotational angle shows an interesting asymmetry: it is ratchet-shaped. As explained in [150, 151], a DNA ring with such a feature can be made to twirl around its backbone via a periodic change in temperature, thus acting as a molecular motor.

Before we present the results of our MMC simulation of the Rosanio *et al.* experiment, it is instructive to first discuss some significant differences.

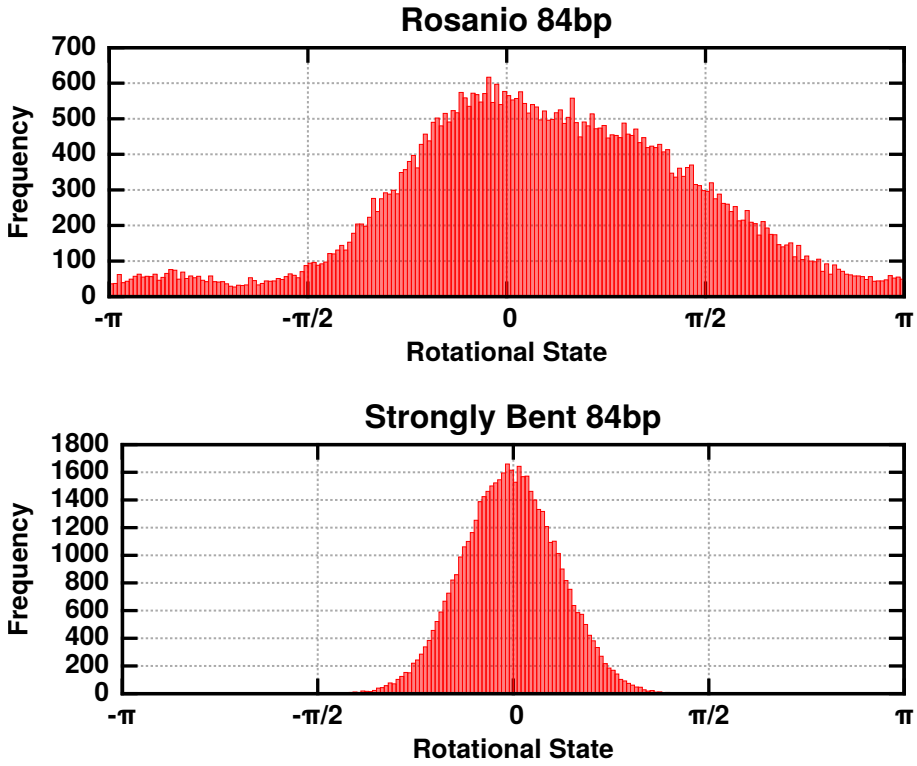

Figure 5.2: Histograms of rotational states (around its length axis) of an RBP DNA molecule forced into a ring, sampled during a standard Monte Carlo simulation, for two separate locking sequences consisting of 36 base pairs. The bias introduced using the sequence from Rosanio *et al.* [93] (Sequence [24] in Appendix B) is shown in the top panel. The bottom panel shows the bias produced using an arbitrarily selected 36-base-pair subsequence of the strongly curved 84-base-pair sequence from Chapter 2 (Sequence [3]).

The most fundamental difference is that the method employed experimentally does not actually sample Boltzmann statistics. The ligation method used by Rosanio *et al.* is irreversible, which means that their selection rounds are out-of-equilibrium processes. We should therefore first consider what probability distribution is actually sampled, and its effect on the measured sequence preferences.

The DNA fragments can be ligated into circles, for which an energy barrier exists because the DNA must be deformed. The rate of cyclization is then proportional to the Boltzmann factor of the sequence,

$$r_C = r_C(S) = \nu_C e^{-\beta F(S)}, \tag{5.13}$$

where r_C is the cyclization rate and ν_C is the attempt frequency of cyclization.

However, the DNA fragments can also be ligated to each other, causing dimerization and taking the fragments out of the pool of fragments attempting cyclization. (We are neglecting further multimerization of the dimerized fragments, which further increases the rate at which linear fragments are lost.) Assuming that the dimerization process is a second-order reaction, and defining $[L]$, $[C]$ and $[D]$ as the concentrations of linear, cyclized and dimerized fragments, respectively, and r_D as the sequence-independent rate constant for dimerization, the reaction kinetics are given by

$$\frac{d[C]_S}{dt} = r_C(S)[L]_S, \tag{5.14}$$

$$\frac{d[D]_S}{dt} = r_D[L]_S^2, \tag{5.15}$$

$$\frac{d[L]_S}{dt} = -r_C(S)[L]_S - r_D[L]_S^2. \tag{5.16}$$

In Eqs. 5.14–5.16, we have explicitly written out dependence on sequence with subscripts S. These equations hold for the concentrations of fragments with a given sequence, and we will for now only consider one sequence at a time. Therefore, in what follows we will drop the explicit subscripts.

In reality, the kinetics of fragments with different sequences are coupled, because fragments may dimerize with fragments which do not have the same sequence. This means the dimerization is much stronger than what is suggested by Eqs. 5.14–5.16. However, this additional dimerization is sequence-independent, and we will see that the dimerization component

does not alter the qualitative behavior of the system. A qualitative characterization will be sufficient for our purposes.

The probability of surviving a selection round is the probability of being cyclized at the end of the round, which is by definition

$$P(t) = \frac{[C](t)}{[C](t) + [D](t) + [L](t)} = \frac{[C](t)}{L_0}, \tag{5.17}$$

where L_0 is the concentration of free fragments at $t = 0$.

Eqs. 5.14–5.16 and Eq. 5.17 can be solved to yield

$$P(t) = -\frac{r_C}{r_D L_0}\left\{ r_C t + \log\left(\frac{r_C}{r_C + r_D L_0}\right) - \log\left(e^{r_C t} - \frac{r_D L_0}{r_C + r_D L_0}\right)\right\}. \tag{5.18}$$

The most important properties of this probability distribution can be understood in the limit of negligible dimerization (which can be physically achieved using a very low concentration of fragments). Without dimerization, the kinetics in Eqs. 5.14–5.16 simplify considerably, and Eq. 5.18 reduces to

$$P(t) = 1 - e^{-r_C t}, \tag{5.19}$$

which makes clear the saturation behavior of the probability in time. This behavior is plotted for different values of r_C in Fig. 5.3A.

In our model, we find that the free energies of the sequences vary over a multi-$k_B T$ range, and as a consequence the Boltzmann factors vary over several orders of magnitude. This means that the speed with which Eq. 5.19 saturates to 1 also varies over several orders of magnitude. This leads to a sharp division between high-affinity and low-affinity sequences: after some time, there will be a part of the sequence population that is not undergoing selection any longer. Sequences with small enough free energy (small enough being dependent upon the ligation time) all essentially have probability 1 to survive. Sequences with worse affinity are not 'guaranteed' to survive, and most will not be selected.

This behavior is clearly visible in the probability distributions imposed on the sequence space (determined by the free energies of the sequences), shown in Fig. 5.3C. The probability distribution shows a population of sequences guaranteed to survive, a population almost guaranteed not to, and a drop-off from one to the other over a span of about 4 kT.

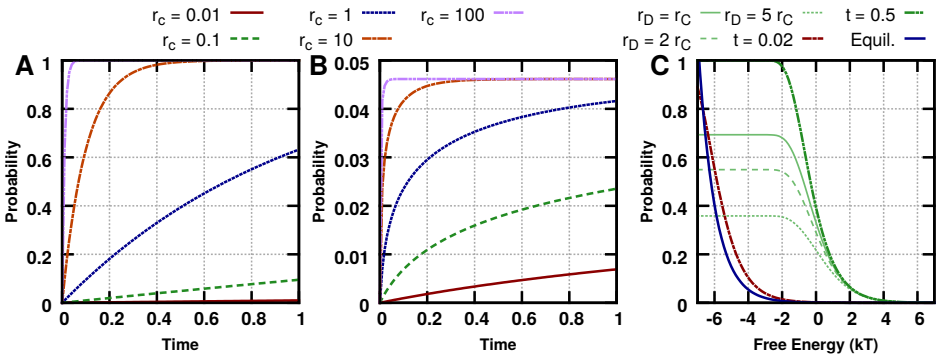

Figure 5.3: Saturation behavior in the out-of-equilibrium selection method of Rosanio *et al.* [93]. **A:** Survival probability as a function of time, without dimerization, for values of r_C spanning several orders of magnitude. **B:** As **A** with strong dimerization ($r_D/r_C = 100$). The saturation probability and how quickly it is approached change, but the overall character is similar. **C:** Probability distributions imposed on the sequence space. If the probability is not allowed to saturate, the distribution (red dot-dashed, green dot-dot-dashed curves) is similar but not identical to the Boltzmann distribution (blue solid curve). Also shown are the distributions for $t = 0.5$ with dimerization (light green curves), in which case the saturation probability is reduced but the overall shape of the distribution is maintained. The free energy range is fictive, arbitrarily chosen for the purpose of illustration, but realistic.

The shape of the drop-off resembles the Boltzmann distribution, as we see when we choose the cutoff time low enough that no sequences saturate. In fact, in the limit $t \to 0$, we find a linear regime for Eq. 5.19 where the probability becomes proportional to the Boltzmann weight; unfortunately, the constant of proportionality is linear in t and therefore the efficiency of the experiment in this limit also goes to zero. (This is exacerbated by the fact that this is only true for negligible dimerization, meaning that the concentration of fragments in the experiment must be very low as well.) We may therefore hope that, apart from the lack of selection on the saturated sequences, the selection is not qualitatively different from an equilibrium selection.

Before we show that this is the case, let us remark that the behavior of the system in the presence of dimerization is very similar to the behavior without dimerization. In Fig. 5.3B we see that, while the saturation probability and the rapidity with which the probability approaches it are both altered by the dimerization, the overall character of the plots is similar.

This is also evinced by the probability distribution in sequence space in the presence of dimerization, shown in Fig. 5.3C (light green curves). The saturation probability is different (and this is irrelevant for the competition between sequences), but the overall shape of the distribution is the same.

Let us quickly remark that the behavior we describe is actually realistic. In Fig. 5.3 we chose an arbitrary range of free energies to illustrate the behavior. However, we do see free energies in our model varying over roughly a range of this magnitude. More importantly, the experiment of Rosanio *et al.* [93] also evinces this behavior, as evidenced by Fig. 2 in that reference. This figure shows that in each round, a large percentage of fragments remains linear, meaning that in each case the selection time was chosen such that not all sequences saturate. These reaction times vary over several orders of magnitude, and the fact that at each of these selection times a meaningful selection is taking place (the probabilities do not saturate, nor go to zero), means that the Boltzmann weights of the sequences must indeed vary over several orders of magnitude.

We must also make a remark as to the behavior of the system under multiple rounds of selection. Performing one round with time t, and one with τ, we calculate the probability to survive both rounds as the product of the probabilities to survive either round, and we find

$$P(t,\tau) = 1 - e^{-\nu_C P_B t} - e^{-\nu_C P_B \tau} + e^{-\nu_C P_B (t+\tau)}. \tag{5.20}$$

If t and τ are comparable, we obtain various order terms, the lowest of which will dominate. For simplicity assume $t = \tau$, then

$$P(t,t) = 1 - 2e^{-\nu_C P_B t} + e^{-2\nu_C P_B t}. \tag{5.21}$$

In the limit of small t, we retrieve the equilibrium statistics (by expanding the expression above to leading, i.e. second, order). If we are not in this limit (which, as explained above, is likely), the effect of the second round of selection is more subtle: the closer we are to saturation, the less effect the number of rounds has, since it only affects terms that tend to zero. In general, we find a weaker effect on the strength of the selection than in the equilibrium case (Eq. 5.6).

If the ligation times of different rounds vary a lot, Eq. 5.20 will simply be dominated by the smallest ligation time. In that case, performing multiple rounds achieves little.

The question is how much the results of the experimental selection and our equilibrium simulation diverge. It turns out we can take the out-of-

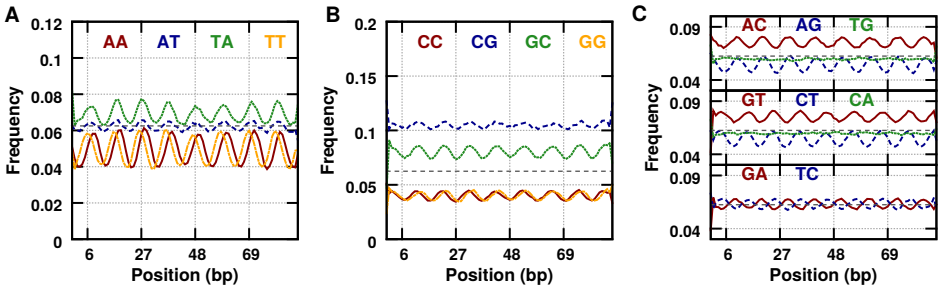

Figure 5.4: Dinucleotide distributions along a ring with the Rosanio sequence (Sequence [24] in Appendix B), obtained from a MMC simulation at room temperature, emulating a single round of SELEX. The dinucleotides have been grouped as in Fig. 3(d)-(f) in [93].

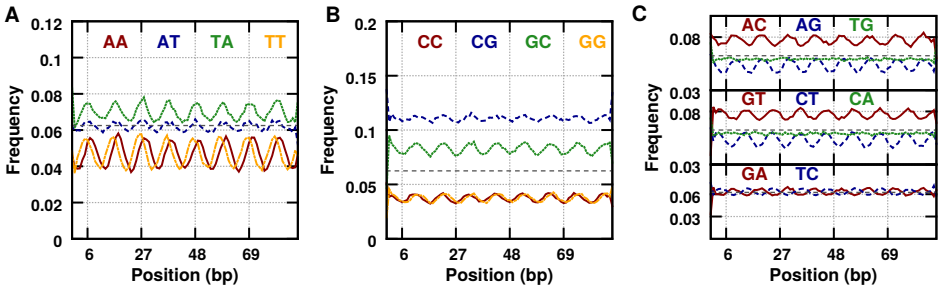

Figure 5.5: Like Fig. 5.4, but rather than sampling according to the Boltzmann distribution, sequences were selected using a hard cutoff in the free energy (as calculated using the model from Eq. 5.10 and [78]). The cut-off was placed approximately at the 99th percentile of the free energies.

equilibrium case to an extreme, modeling it as a hard cut-off on the free energies of the system, and still have a minor effect on the measured dinucleotide preferences of the system.

In order to emulate an equilibrium SELEX version of the experiment of Rosanio *et al.* [93], we performed a MMC simulation of a closed DNA ring, modeled via the RBP model with the standard hybrid parameterization [11, 13, 14]. As in the SELEX experiment, we chose a ring with 126 base pairs, of which 36 were fixed to be the locking sequence from [93]. The rest of the DNA was allowed to mutate. By sampling sequences during the simulation, we obtained a thermal sequence ensemble, from which we calculated the dinucleotide probability distributions shown in Fig. 5.4. Because we found, in Eq. 5.20 and onward, that the effect of multiple rounds of selection is small, we only simulated one round of selection.

We may now use oligonucleotide distributions calculated from the sequence ensembles we generate as input for the approximation of Eq. 5.10. Using this approximation, we performed a second simulation where we generated random sequences, and selected or discarded them using a hard cutoff on the free energy. The resulting dinucleotide distribution are shown in Fig. 5.5. We see that the calculated distributions are highly similar to each other, indicating that indeed, selecting via a Boltzmann distribution, or via a hard cut-off, both lead to very similar results. Therefore, in practice, the out-of-equilibrium nature of the experiment of Rosanio *et al.* does not make for a large difference with the equilibrium scenario.

5.6 RING SEQUENCE PREFERENCES IN VITRO AND IN SILICO

The dinucleotide distributions we find *in silico* show both similarities and differences with those found by Rosanio *et al.* [93] (compare Fig. 3(d)-(f) in that reference). First, the periodicities in the distributions, which derive from the helical nature of DNA, are very similar. The A/T-rich dinucleotides (Fig. 5.4A) are all in phase with each other, while the G/C-rich dinucleotides (Fig. 5.4B) are exactly out of phase with the former. The phasing of the other dinucleotides, shown in the three groups in Fig. 5.4C, all show phasing that resemble those found experimentally.

However, one interesting difference is the slight (1-bp) difference in phasing among the A/T-rich dinucleotides. Whereas Rosanio *et al.* find all of them peaking at exactly the same position, we find that AA generally peaks one base pair to the right of AT and TA, and TT one base pair to the left. This shift of the AA and TT dinucleotides seems to be caused by the overall preference for the TA step over the AT step. The TA step can be flanked on the left by TT but not AA, and on the right by AA but not TT. This preference of the ring is analogous to the nucleosome's preference for the TTAA tetranucleotide at the positions along the nucleosome where the minor groove faces inward [1, 152], and is therefore not unexpected.

The experimental distributions do not see this preference for the TA step over the AT step, which brings us to a more general difference between our theoretical results and the experimental distributions. Remarkably, the experimental probabilities never deviate very far from the uniform dinucleotide probability of 1/16. In our simulations, this is not the case: the probabilities take on values from around 0.04, up to around 0.12. This does not occur only locally, but several dinucleotides have an average

probability, along the entire ring, significantly different from the uniform value.

The uniformity of the experimentally obtained distributions is surprising. It is known, for instance, that the affinity of nucleosomes to sequences correlates with GC content. Therefore, e.g. the enrichment of the CG and GC dinucleotides we observe in our simulations is not unexpected. More generally, there is no reason to expect all the dinucleotides to have probabilities close to 1/16.

These unexpected features of the experimental results may be due to a property of DNA rings that the rigid base pair model does not reproduce. For example, Rosanio *et al.* find longer-range correlations in their sequences. Our model contains such interactions only indirectly, due to the thermal nature of the system and the constraints placed on the DNA, but microscopically only accounts for nearest-neighbor interactions. There is much evidence that the RBP model is an oversimplification in this regard [15–20]. However, at this time it is not clear how such nuances would give rise to the differences we observe between theory and experiment, and more research will be needed to uncover the true causes.

5.7 SELEX SIMULATION FOR SMALL AND OVERWOUND CIRCLES

A further benefit of bringing the SELEX methodology into the computational realm is that it allows for studying systems that are experimentally difficult to realize, such as very small rings, rings whose length is not an integer multiple of the helical period of DNA, or overwound rings. These all have a high energetic cost, and are therefore slow to form, as they are dependent on thermal fluctuations for ligation. In our simulations, we can simply impose the desired constraints from the beginning, and we do not need to wait for the system of interest to form spontaneously.

Fig. 5.6 presents a part of the AA/AT/TA/TT dinucleotide distributions for three different rings: the 126-base-pair ring analogous to the one used by Rosanio *et al.*, a slightly shorter, 121-base-pair ring (which leads to a slightly twisted ring because the length is not an integer multiple of the helical period), and a much shorter 84-base-pair ring, with correspondingly larger curvature. All rings are direction-biased not using the locking sequence of Rosanio *et al.*, but with the artificial, strongly bent sequence from Chapter 2 (Sequence [3] in Appendix B). We chose this sequence

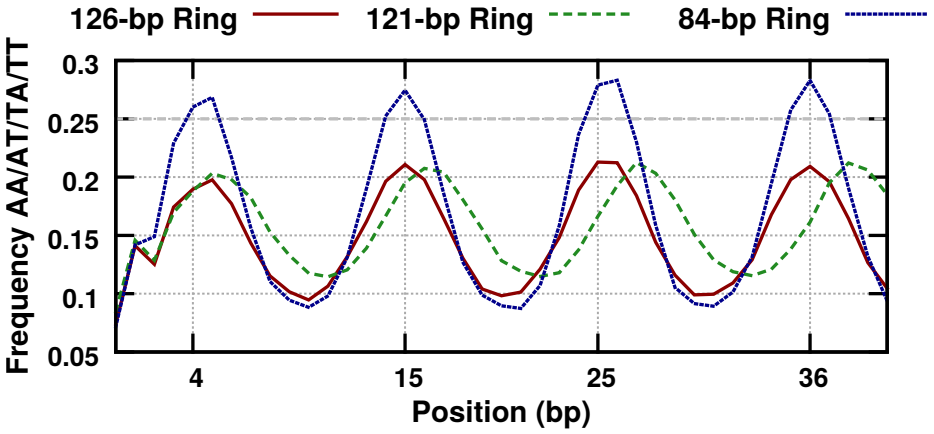

Figure 5.6: The total frequency of the A/T-rich dinucleotides (AA/AT/TA/TT) for three different rings, all directionally biased using the artificial locking sequence from [86] (see Section 5.5; Sequence [3] in Appendix B): the 126-base-pair ring considered before (red solid curve), a 121-base-pair ring, which requires over- or undertwisting of the DNA (dashed green curve), and a significantly shorter (but not overwound) 84-base-pair ring (dotted blue curve). All three curves were calculated at room temperature, with the mutation temperature reduced to 1/3 of room temperature. This was achieved as described in Section 5.3. A ring whose length is not an integer multiple of the helical repeat stretches (in this case, where the ring is underwound) the periodicity of the distributions and slightly reduces their amplitude. A tighter ring leads to larger amplitudes.

over the locking sequence from Rosanio *et al.* because of the stronger and cleaner directional bias (see Fig. 5.2).

The 84-base-pair ring is more tightly curved and therefore places a stronger selection on the sequences, leading to the higher amplitude in the frequencies. For ease of comparison, Fig. 5.6 only shows the combined frequencies of the A/T-rich dinucleotides, but the same effect applies to all individual dinucleotide frequencies.

The 121-base-pair ring is underwound by half a turn, and the periodicity in the dinucleotide frequencies is correspondingly stretched to a slightly larger period. The amplitude is not increased, as we found when shortening the ring to 84 base pairs, but is rather slightly decreased. This is in fact as expected: the locking sequence becomes less effective when the DNA is underwound, because it is designed to give coherent curvature in unconstrained DNA. The twist mismatch weakens the directional bias

imparted by the locking sequence. As for the 84-base-pair ring, these observations are conserved among all dinucleotide probabilities, not just the ones shown in Fig. 5.6.

We could underwind or overwind our rings by more than half a turn, and we would expect similar stretching and compression of the periodic nature of the frequencies. However, we will start to run into two complications. First, as already observed, the locking sequence will become less effective. (We could design locking sequences specifically for overwound or underwound DNA, but that is beyond the scope of the current work.) Second, for strongly overwound or underwound DNA it will become energetically favorable to supercoil [153, 154]. This complicates the system, because different parts of the DNA will interact and steric interactions must be taken into account.

We modeled half of a figure-eight supercoil of DNA as a simple teardrop shape as a proof-of-principle. This model consists of two constraints: we place the base pairs of our molecule along a teardrop-shaped curve, and keep the first and last base pairs fixed throughout the simulation. An example state is shown in Fig. 5.7A. Such a shape, although it is essentially two-dimensional and therefore a simplification of real three-dimensional supercoiling configurations, emulates the basic geometry of the end-loops of supercoils [155] and protein-induced DNA loops [156, 157].

Applying our methodology to such a teardrop shape, consisting of 126 base pairs, we find the dinucleotide frequencies presented in Figs. 5.7B-D. As expected, the distributions we find resemble those of a ring. However, because the curvature is not constant – it falls off towards the ends of the molecule – the amplitude of the distributions tapers off.

5.8 CONCLUSIONS

We have presented methods to emulate, *in silico*, equilibrium SELEX experiments. The MMC method [1] is akin to such experiments and can be used to select for high-affinity sequences for a given DNA system. One limitation of the MMC method was that the selection pressure on the sequences and the temperature in the simulation are linked. In an equilibrium SELEX experiment, the mutation pressure is modified in a mathematically straightforward way by the number of rounds of selection applied.

We employed the methodology of Chapter 4, which makes use of the output of a MMC simulation to build a model for sequence-dependent

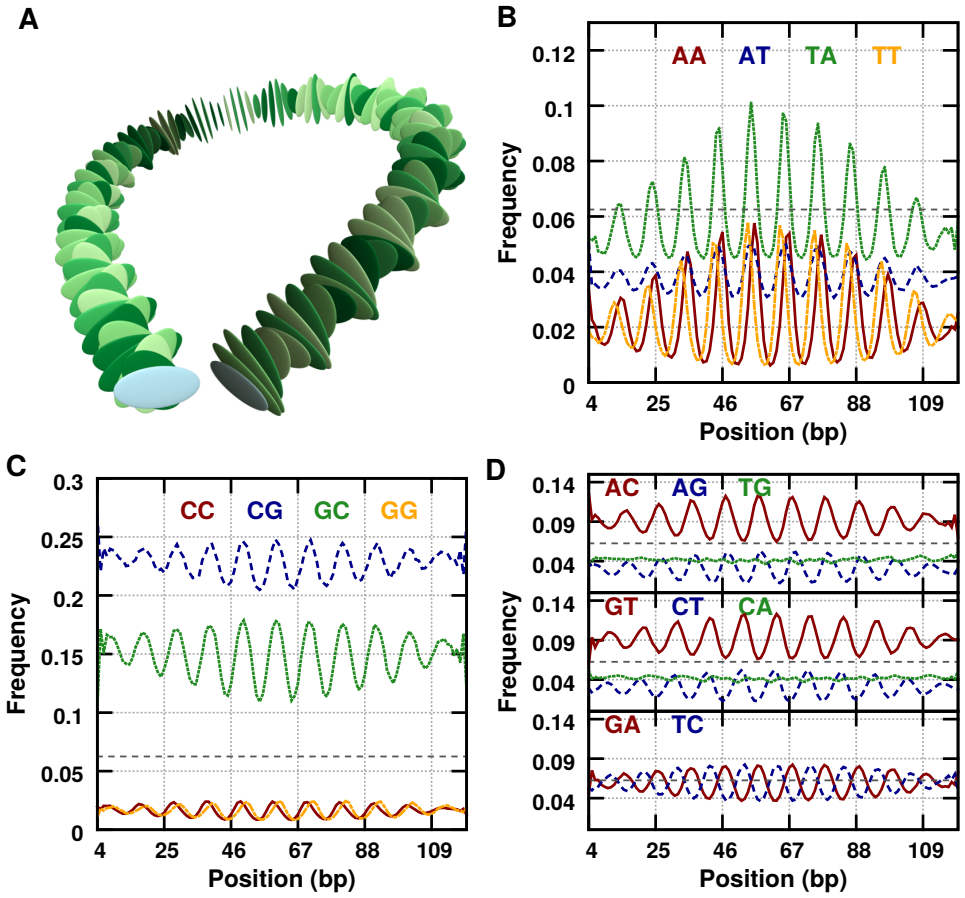

Figure 5.7: **A:** A teardrop-shaped DNA molecule, held in place at its ends. **B-D:** Dinucleotide distributions calculated for the teardrop-shaped DNA, at room temperature, with the mutation temperature reduced to 1/3 of room temperature. The teardrop is most strongly curved in the center, and more straight towards the ends of the molecule. This leads to distributions similar to those of rings, but whose amplitudes taper off towards the ends.

nucleosome affinity, to resample sequence space at a different mutation temperature, without altering the temperature employed for the spatial moves. This separation of mutation pressure and physical temperature allows us to more closely mimic the outcome of a SELEX experiment, as well as learn more about our systems in general.

We have used this new methodology to examine various systems. First, in Section 5.4, we assessed how changing the physical temperature, without changing the mutation pressure, affects the sequence preferences of nucleosomes and rings. We found that, due to the strongly constrained nature of the nucleosome, entropic contributions to the free energy do not play a large role, and consequently the sequence preference of the nucleosome are not strongly temperature-dependent (in the range between 1/4 of room temperature and room temperature.) Rings, on the other hand, are not heavily constrained systems, which means that the entropic contribution to their free energy is more important and the sequence preferences of rings depend strongly on temperature.

In Section 5.5 we considered the SELEX experiment for rings performed by Rosanio *et al.* [93]. This experiment is not an ideal equilibrium SELEX experiment because it uses irreversible reactions, and we examined what effect this has on the (non-Boltzmann) distribution the experiment imposes on sequence space. While some differences in the methodology must be noted, the effects on the measured sequence preferences turned out to be small and we were able to compare the predictions of our *in silico* SELEX experiment for rings with the experimental results. We found that the periodic nature of the dinucleotide distributions in rings is well captured by the RBP model we employed to model the DNA. However, some differences are apparent, the most striking one being that we predict significant deviation away from 1/16 in the overall frequencies of dinucleotides. For example, we find the CG dinucleotide significantly enriched, similar to what we find for nucleosomes.

The experimental distributions show very little overall variation away from 1/16, meaning that no dincleotides are significantly enriched or depleted along the entire length of the ring. While it is not unlikely that the RBP model has limitations that may prevent it from capturing the real sequence preferences of rings perfectly, we note that there is no *a priori* reason why all the dinucleotides should be equally likely, on average, and it remains an open question why this is found in the experiment.

We finally applied our methods to several systems that would be difficult to access experimentally. We considered rings that would have difficulty forming because they either consist of only a short piece of DNA,

requiring tight curvature, or because their length is not an integer multiple of the helical repeat length of DNA. We also considered a teardrop-shaped DNA molecule, which mimics a part of strongly overwound (or underwound) DNA, or protein-induced DNA loops. We showed that our methods can be used to determining the sequence preferences of such systems, opening up new possibilities of examining systems that have been inaccessible until now.

The methodology we have presented relies on a sequence-dependent description of DNA mechanics, for which role we have cast the rigid base pair model. However, the methods are general and can be used with any other underlying model, and they are straightforward to update if and when more advanced DNA models become available in the future.

6 NUCLEOSOME POSITIONING SIGNALS IN GENE PROMOTERS

Er muss sozusagen die Leiter wegwerfen,
nachdem er auf ihr hinaufgestiegen ist.

—Wittgenstein

This chapter is based on:
Tompitak, Vaillant and Schiessel 2017 *Biophys. J.* 112.3 505–511 [79]

We now leave the full Rigid Base Pair model and the Eslami-Mossallam nucleosome model [1] behind and, in this chapter, rely on the Markov-chain model of Chapter 4. At the expense of accuracy, we gain astronomically in computational cost. With this fast method in hand, we are now able to turn towards entire genomes and analyze the nucleosome affinity of billions of different sequences, and look for real nucleosome positioning signals in nature.

6.1 INTRODUCTION

Nucleosomes are the fundamental packaging units of DNA that eukaryotic organisms employ to render their genomes compact enough to fit inside a cell, consisting of about 147 base pairs worth of DNA wrapped around a histone core. This packaging also restricts access to the genome: DNA bound to histones is unavailable for coupling to many other DNA-binding complexes, such as the transcriptional machinery. Therefore, the positioning of nucleosomes along the genome interacts with gene expression, as was already realized some three decades ago [33, 34].

This interplay suggests that nucleosomes may play a role in gene regulation, and nucleosomes are in fact actively displaced in order to regulate gene expression [158, 159]. Genomic sequences may also have evolved to position nucleosomes in specific, beneficial locations. This possibility is suggested both by the fact that the degeneracy of the genetic code in principle allows for multiplexing of such positioning signals with genetic information [1, 116–119], and by the observation that the mutation patterns of DNA bound to histones differ from those of linker DNA [48].

Research into such nucleosome positioning signals, hardcoded into eukaryotic genomes, has veritably exploded over the last decade, primarily due to the development of experimental methods that allow for efficient genome-wide nucleosome mapping [160]. This research has provided insight into the importance of nucleosomal sequence preferences for chromatin organization [161], and has allowed for the creation, refinement and testing of many models for predicting nucleosome positioning along genomes [73, 75, 162]. The intrinsic nucleosome-DNA affinity of genomic sequences appears to play a significant role *in vivo* in positioning nucleosomes in certain regions of the genome, such as transcription start sites (TSSs) and origins of replication [161], alongside other effects like the presence of proteins that compete for the same DNA stretch or the action of chromatin remodellers [31, 163].

Around the TSS of *S. cerevisiae* (baker's yeast), nucleosomes have been found to be depleted on average, both *in vitro* and *in vivo* [62–64, 130, 164–167]. The persistence of this depletion *in vitro*, in the absence of active remodeling, identifies the sequence preferences of nucleosomes as the dominant cause. Those preferences have been measured and utilized in various models to explain the observed nucleosome depletion [63, 64, 71, 72, 166]. These nucleosome-depleted regions (NDR) in gene promoters are thought to be encoded into the genomic sequence to allow RNA polymerases more ready access to the TSS, thereby facilitating transcription [62]. This is not only of interest for an understanding of the workings of natural genomes, but has also recently been put forward as an interesting engineering method to modulate transcription in synthetic genomes [168].

Since the earliest studies on baker's yeast, inquiries into nucleosome positioning have been extended to the genomes of many other organisms, such as *S. pombe* [149, 169] and various other species of yeast [170], *C. elegans* [171, 172], *Plasmodium falciparum* [173], flies [174], zebrafish [175], *Arabidopsis thaliana* [176], mice [177–179] and humans [133, 178, 180–183]. Most of these studies were conducted *in vivo*, and therefore do not allow for isolation of effects encoded into the genomic sequences. This body of research shows, however, that sequence effects alone are not generally sufficient to explain *in vivo* observations [163]. An important role is also played by the active regulation of transcription. In yeast, the promoters of actively transcribed genes show much more pronounced nucleosome depletion than those of inactive genes [169].

In human cells, as in yeast, NDRs were found *in vivo* only for actively expressed genes [180]. However, *in vitro* nucleosome mapping reveals that the human genome does not share yeast's strategy of depletion-by-default.

Instead, it was found that promoter regions in the human genome showed enhanced nucleosome occupancy. One interpretation is that this is a reflection of the differentiated nature of human cells: it may be more beneficial to keep genes relatively inaccessible by default, and to actively open up the promoter region only when needed [133, 182]. This idea seems to be countered by newer results, however, which find stronger intrinsic nucleosome-attracting regions (NARs) for housekeeping genes than for tissue-specific genes, directly opposite of what one would expect [29]. Those results indicate that the function of the NARs in the human genome may be to retain nucleosomes in sperm cells (in which most nucleosomes are removed from the chromatin) and so pass on epigenetic information to the next generation.

Whichever is the case, these ideas raise the question whether the presence of an NDR in yeast versus that of an NAR in humans might be a general distinguishing feature between unicellular and multicellular life. In order to answer this question, we utilize a purely mechanics-based model for the sequence-dependent DNA-nucleosome affinity to predict *in vitro* nucleosome positioning signals, and compare the signals encoded into the promoter regions of a wide range of genomes.

6.2 METHODS

6.2.1 Data acquisition

Let us briefly summarize the origins of all the experimental data used in this chapter. All genomic sequences and gene (cDNA) data were downloaded from ensemblgenomes.org, release 31 [184]. The *in vitro* nucleosome map produced by Kaplan *et al.* [64] was retrieved from GEO accession number GSE13622. The map from Valouev *et al.* [133] was downloaded from [185]. The map from Locke *et al.* [186] was downloaded from [187]. The data from Ercan *et al.* [172] was taken directly from Fig. 1C in that reference. TSS locations in *S. cerevisiae* were derived from [188] in the manner described in [189].

6.2.2 Model

The model used for the work in this chapter is the trinucleotide approximation to the Eslami-Mossallam nucleosome model [1] described in Chapter 4, with one major alteration. For the current work, the parameterization

of the Eslami-Mossallam nucleosome model was changed from the hybrid parameterization described in [1], to a parameterization informed solely by crystallography data [11]. We found that this improves its applicability to long-range effects. See Appendix A for more information.

6.2.3 Sequence analysis

For every genome analyzed, we calculated the averaged signal as follows. For every annotated gene, we looked up the location of the TSS, and extracted the 1146 bp before and after. For each of the resulting sequences, we calculated a probability landscape for nucleosome positioning using the trinucleotide model mentioned above. We would like to calculate occupancies from these landscapes and average over all genes. Unfortunately, because the probabilities vary over several orders of magnitude, the number of genes is generally not large enough to provide a meaningful average; it tends to be dominated by the highest probabilities. Therefore, we instead consider the average energy landscape for a given organism.

From the predicted probabilities, an energy landscape can be calculated up to a constant shift, since such a probability is the normalized Boltzmann weight of a state. We took the average of the energy landscapes of all the sequences as a representative energy landscape for a given organism. For each bp (-1000 to +1000) we then calculated the nucleosome occupancy by summing the Boltzmann probabilities of all 147 nucleosome positions that lead to that bp being covered by the nucleosome. This gives us a prediction of the intrinsic nucleosome affinity encoded in the genomic sequences.

6.3 OPPOSING NUCLEOSOME OCCUPANCY SIGNALS IN YEAST AND HUMAN GENOMES

The high-coverage *S. cerevisiae* nucleosome maps provide the standard testing ground for any model designed to predict nucleosome occupancy [51]. Applying our nucleosome affinity model from Chapter 4, we find a peak in the free energy of the nucleosome in the promoter regions of *S. cerevisiae* (Fig. 6.1), which correctly predicts experimentally observed NDRs in these regions. The comparisons, for regions centered on the TSSs and on the start codons, are shown in Fig. 6.2A and B, respectively.

For the human genome, a map of *in vitro* nucleosome occupancy has been published by Valouev *et al.* [133], and, as predicted by Tillo *et al.* [182],

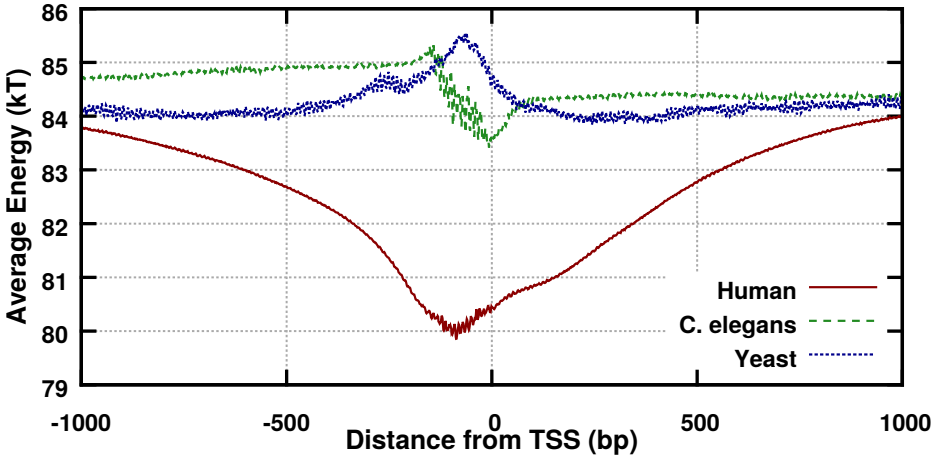

Figure 6.1: Average free energy landsapes in the promoter regions of the human, yeast and *C. elegans* genomes.

it reveals occupancy signals opposite to that of yeast: human promoters seem to encode for high, rather than low, nucleosome occupancy. Vavouri and Lehner [29] similarly find an increased retention of nucleosomes when nucleosomes are depleted in human sperm cells. Correspondingly, when applying our model to the promoter regions of the human genome, we find a very strong NAR around the TSS, due to a vast dip in the free energy (Fig. 6.1), as can be seen in Fig. 6.2C.

Initially surprisingly, the signal found by Valouev *et al.* is an order of magnitude smaller than that predicted by our model and that found by Vavouri and Lehner. This discrepancy can be explained when we consider that the nucleosome density cannot exceed 1 per 147 bp due to excluded volume. The experiment attempts to measure enrichment of nucleosomes in the promoter regions relative to the average density of nucleosomes. Unlike in experiments that look at nucleosome depletion or retention, the excluded volume between nucleosomes puts a limit on how strong the enrichment can be in practice.

This is the reason for the discrepancy between the *in vitro* results of Valouev *et al.* and ours and those of Vavouri and Lehner. In order to approximate the effects of steric interactions, we applied Percus' equation [191] to our average energy landscapes, and solved it as described in [192]. The solution depends on the chemical potential of the nucleosomes binding to the DNA (see also [190]), which we adjust to achieve a good fit with the *in vitro* data. We see that steric interactions can indeed explain the very

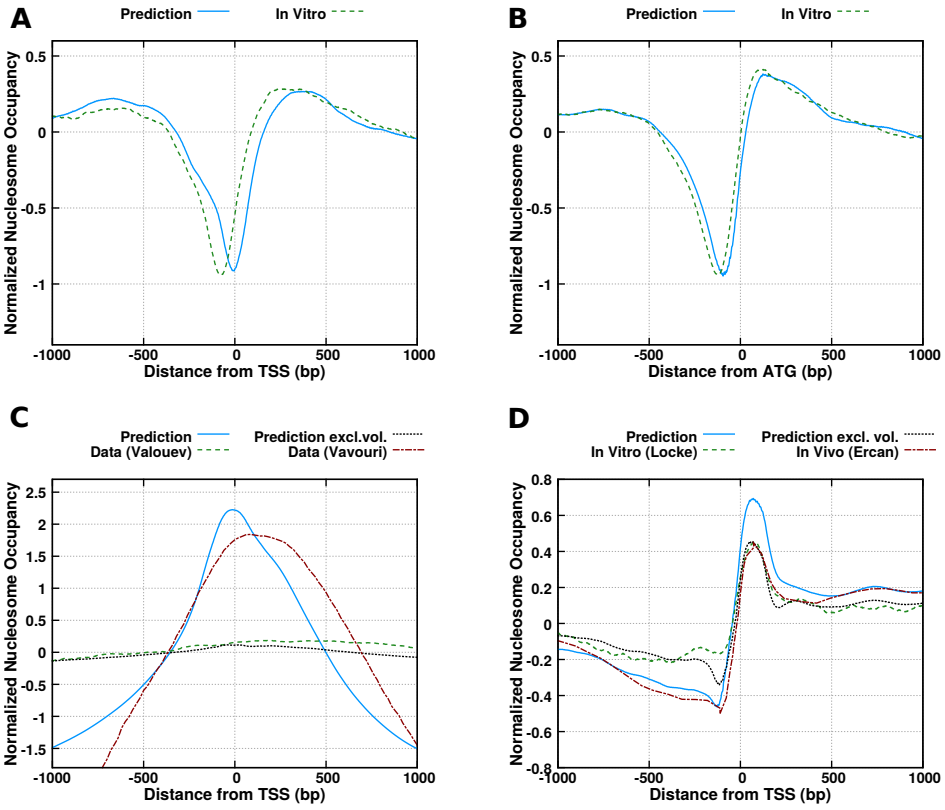

Figure 6.2: Comparison of predicted and measured intrinsic nucleosome positioning signals in promoter regions. The quantities plotted are the natural logarithms of the occupancies and the signals have been normalized such that they average to zero. In all plots, the solid blue curves are our predictions in the limit of low nucleosome density, which give an account of the strength of the signals intrinsically encoded. The dashed green curves represent *in vitro* measurements. The dotted black curves are predictions taking into account steric interactions. Using the same treatment as in [190], these curves have a free parameter $\tilde{\mu} = \mu - \langle E \rangle$, i.e. the difference between the chemical potential and the average energy of the landscape, which we determined to be -8.5 kT for yeast (curves not shown due to similarity with the low-density limit), -5.7 kT for *C. elegans* and -1.38 kT for humans. **A, B:** *S. cerevisiae*, average nucleosome occupancy centered on the TSS and start codons, respectively. Data from [64]. **C:** Like A, for *H. sapiens*. The *in vitro* data is from [133]. Additionally shown is the nucleosome retention signal from [29]. **D:** Like A, for *C. elegans*. The *in vivo* data is from [172], the *in vitro* data is from [186].

weak signal for humans (dotted black curve in Fig. 6.2C) as well as the apparent overshoot of our prediction for C. elegans (dotted black curve in Fig. 6.2D).

This means that also at physiological conditions, the nucleosome density will be saturated at much smaller values due to steric interactions. However, we stress that independent of this saturation effect, a nucleosome at the peak of the nucleosome occupancy signal will be strongly energetically bound, and so hinder transcription if it is not actively removed, as well as be more stable under a nucleosome-depleting force.

The results of Vavouri and Lehner [29] when examining where nucleosomes are retained when they are depleted from chromatin in human sperm are more in line with our predictions, as can also be seen in Fig. 6.2C. When depleting nucleosomes, excluded-volume interactions are not a constraint and our predictions can be probed. Although these authors studied a special *in vivo* situation, the nucleosome retention signals were found to correlate strongly with DNA sequence. Because the depletion of nucleosomes in sperm is an out-of-equilibrium process, and our model therefore does not make direct numerical predictions for this situation, we note the similarity between our predictions and the *in vivo* nucleosome retention signal.

We thus have interesting observations and predictions on two ends of a spectrum. A very simple, unicellular eukaryote shows nucleosome depletion as its most prominent, intrinsically encoded nucleosome positioning feature. A complex multicellular one shows high nucleosome occupancy instead. What happens in between these two extremes?

In Fig. 6.2D we present a comparison between our predicted nucleosome occupancy signal (for the underlying free energy landscape, see Fig. 6.1) for the nematode *C. elegans* and the signals found *in vitro* by Locke *et al.* [186] and *in vivo* by Ercan *et al.* [172]. We find remarkable agreement in the shape of the signal, indicating that the data is indeed indicative of intrinsically encoded nucleosome positioning. Somewhat surprisingly, the *in vitro* and *in vivo* signals are similar to each other, which is not as strongly the case for yeast, and even less so for humans (see e.g. Fig. 3 in [29]). It has been noted that an *in vivo* nucleosome occupancy map of *C. elegans* lacks many of the features that distinguish *in vivo* maps from *in vitro* maps of yeast, such as strongly phased nucleosomes. Valouev *et al.* [171] find much flexibility in nucleosome positions in *C. elegans*. Such variability may average out some of the effects of active remodeling, rendering the two maps similar.

C. elegans seems to show a nucleosome positioning signal that is a hybrid of the signals found in the yeast and human genomes. It has an NDR upstream of the TSS, like yeast, but it also shows a significant NAR just after the TSS.

6.4 INTRINSIC NUCLEOSOME POSITIONING SIGNALS ARE INDICATIVE OF MULTICELLULARITY

The hybrid behavior in *C. elegans* may be hypothetically explained. As suggested by Tillo *et al.* [182], organisms may wish to tune their genomic sequences to intrinsically deactivate genes that are active only in some cell types, while intrinsically activating those that are common to all of its cells. In unicellular life, most genes will not be permanently silenced, leading to an overall average depletion signal. In complex multicellular life, the signal may be dominated by the many genes that are intrinsically deactivated, leading to an overall attractive signal. *C. elegans* may then represent a range of organisms where the two contributions are more equal, leading to both a depleted region just before the TSS (where it is also observed in yeast) and an attractive region just after (the peak in occupancy in the human genome is also skewed towards the right).

The results of Vavouri and Lehner [29], however, suggest that, at least in the human genome, the hypothesis of Tillo *et al.* does not hold, and the function of the NARs is to retain nucleosomes in sperm cells. The hybrid signal we find in *C. elegans* may in this case similarly play a dual role of facilitating initiation of transcription but at the same time assisting in nucleosome retention.

We can extend our observation of these signals to other genomes using our model. We mapped the nucleosome positioning signals for promoters in genomes across the tree of life and discovered organisms that have intrinsically encoded NDRs and NARs, as well as many that fall into the hybrid category. The full set of signals found and described below are presented in Fig. 6.3.

Most archaea (14 genomes analyzed) show a signal similar to that of yeast, in that a nucleosome-depleted region is the most prominent feature. Archaea are unicellular organisms that do not have histone octamers, but employ only tetramers of (archaeal) histones to compactify their DNA. We expect these tetramers to obey positioning rules similar enough to nucleosomes that our model is predictive of their occupancy. We therefore analyzed the octamer affinity landscapes, for the sake of comparison to

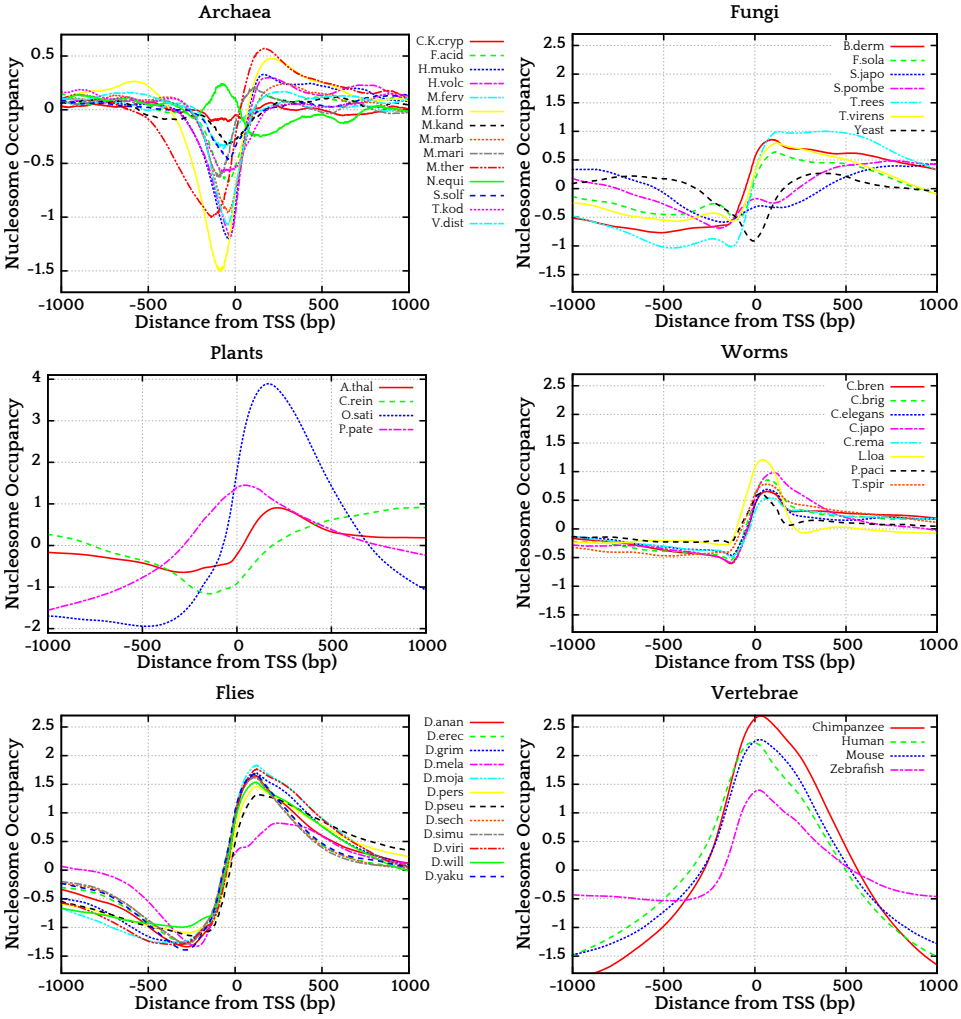

Figure 6.3: The full set of nucleosome positioning signals in the promoter regions of the organisms analyzed for this study.

eukaryotes, even though archaea do not possess them. The signals show that these simple unicellular organisms almost all fall into the depletion-by-default category.

Fungi (7 genomes analyzed) show somewhat more diverse signals than the archaea. While *S. cerevisiae* has a prominent NDR, many of the other fungi analyzed lack both a localized depleted region and a localized attractive region, but retain a step-function signal centered on the TSS. Fungal cells are not highly differentiated, but some fungi are dimorphic (they switch between unicellular and filamentous states), possibly causing these more hybrid-like signals.

Plants (4 genomes analyzed) come in many forms, from unicellular algae to complex multicellular life. As expected, we see various signals. The genome of *C. reinhardtii*, a unicellular alga, shows an NDR. Among the multicellular plants, we see two signals with a strong NAR, and one with hybrid behavior.

Among animals (24 genomes analyzed) we also find various signals. In worms, like *C. elegans*, we find both hybrid signals and more NAR-like signals. *D. melanogaster* and other members of its genus show strong hybrid signals, with a swift rise in nucleosome occupancy at the TSS. Finally, the zebrafish genome and all mammalian genomes analyzed (human, chimpanzee and mouse) have strong NARs.

We see a clear separation between unicellular and multicellular organisms. Though some signals from unicellular lifeforms show some hybrid characteristics, the dominant feature is generally an NDR. All multicellular genomes, on the other hand, either encode for high nucleosome occupancy in the promoter region, or show hybrid signals. This distinction persists across the eukaryotic phylogenetic tree and is clearly visible in Fig. 6.4, where we have plotted a representative set of signals, divided into unicellular and multicellular classes.

We finally note that, as was expected (see Section 1.3), these signals qualitatively correlate well with GC content – see Fig. 6.5 – which suggests that GC content is a prominent factor in shaping mechanical signals in promoter regions. Note however that, while GC content may be a good predictor of the nucleosome occupancy signals (the visual similarity between Figs. 6.4 and 6.5 is striking), it does not provide a numerical value for the occupancy without some sort of model.

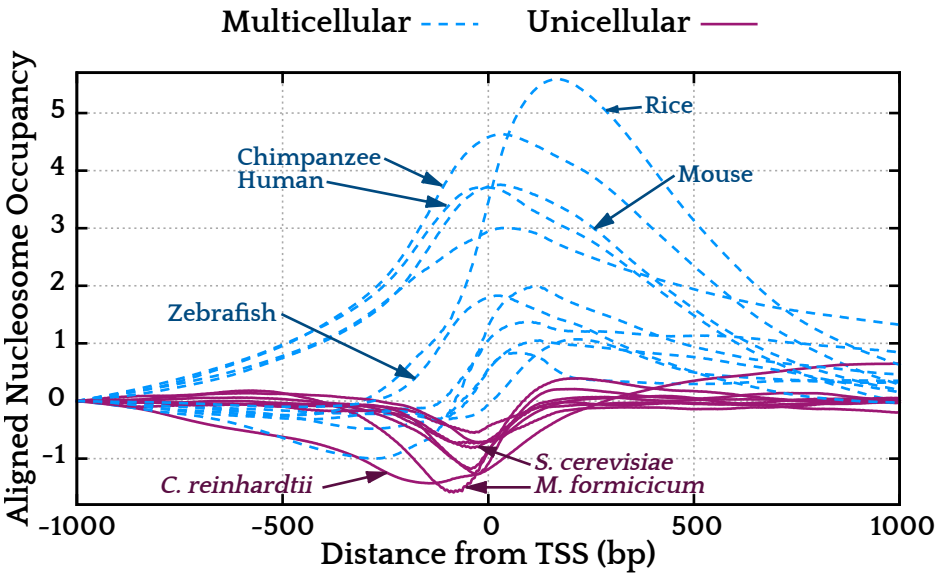

Figure 6.4: A representative selection of nucleosome positioning signals from various genomes. As a visual aid, the signals have been shifted vertically such that the logarithmic nucleosome occupancy at position -1000 is 0. The signals clearly fall into two distinct classes, based on whether the organism is unicellular or multicellular.

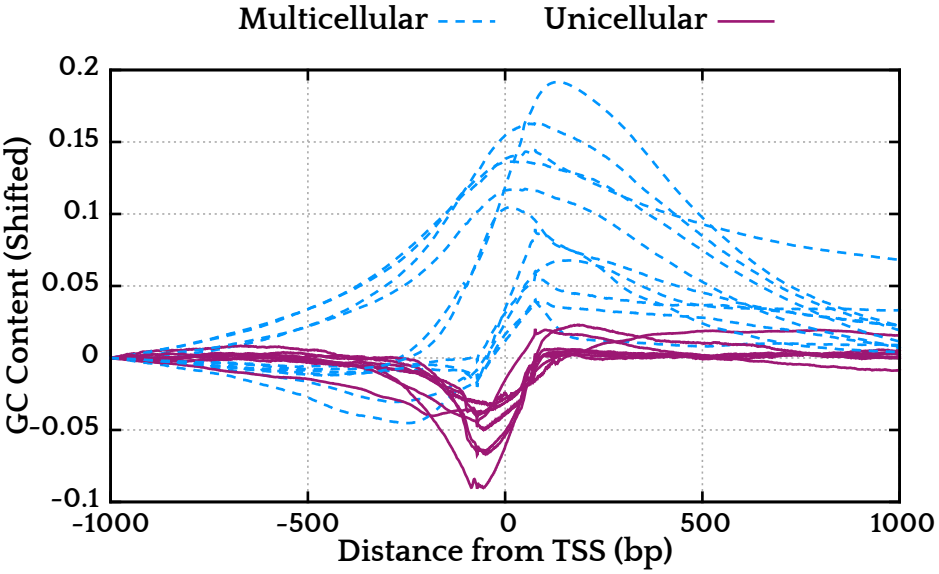

Figure 6.5: Average GC content in the promoter regions of the organisms for which the nucleosome occupancy signals were presented in Fig. 6.4.

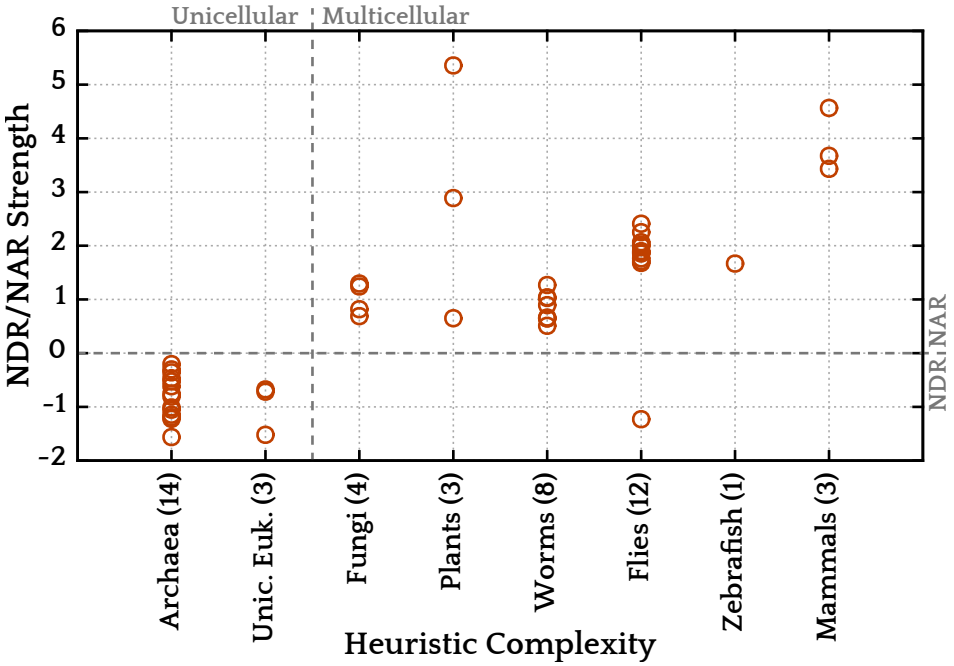

Figure 6.6: Promoter nucleosome positioning signal strength grouped by a heuristic measure of complexity of the organisms. The numbers in parentheses indicate how many genomes fall in each category.

6.5 INTRINSIC NUCLEOSOME POSITIONING SIGNALS CORRELATE WITH COMPLEXITY

One proposed measure for organism complexity is the number of different cell types an organism possesses [193], and the ideas presented here clearly have a link to this measure. Unfortunately, numerical data describing the numbers of cell types does not appear to be readily available in the literature, so we were unable to define a numerical measure of complexity. Therefore, we have restricted ourselves to ordering the organisms, by making assumptions about the cell type numbers. From simple to complex, we list: archaea, unicellular eukaryotes, filamentous and dimorphic fungi, multicellular plants, nematodes, Drosophila flies, zebrafish, and mammals.

We then considered the strength and direction of the NDR/NAR signals. To quantify this, we calculated the maximum and minimum of the signal and took the difference with the signal value at position -1000 relative to the TSS. We then took the largest of these two values (in the

absolute sense) and designated this value as the signal's strength (not in the absolute sense; a dominant NDR gives a negative signal strength).

The signal strength as thus defined clearly distinguishes unicellular and multicellular lifeforms (Welch's $t(39.051) = 10.5512$, p-value 5.4×10^{-13}) and the signals for multicellular organisms show correlation with our complexity ordering (Spearman $r_s = 0.52$, p-value 82.3×10^{-3}), as shown in Fig. 6.6. The ordering of the organisms is almost certainly imperfect, for example because all multicellular plants have been lumped together; without more accurate knowledge of the cell type numbers, there is no way to place them more realistically. However, the NDR/NAR strengths show a tentative trend. All unicellular eukaryotes have a negative signal strength, indicating an NDR, as noted in the previous section. All multicellular eukaryotes (with one exception, *D. melanogaster*) have a stronger NAR than NDR, and the strength of this NAR roughly increases with complexity. This observation concurs with the hypothesis of Tillo *et al* [182]. Our expectation based on that hypothesis would be that a more differentiated organism will have more genes that are nucleosome-occupied by default, leading to a higher NAR signal. It is not clear what purpose this correlation might serve in the context of nucleosome retention in the germline.

6.6 CONCLUSIONS

We found that the recently discovered fact that the human genome, unlike the yeast genome, encodes (on average) for an NAR rather than an NDR in the promoter region, is in fact a universal feature of multicellular life. The hypothesis put forth by Tillo *et al.* [182] is that this NAR suppresses gene transcription and that this suppression helps an organism with differentiated cell types manage its gene expression. Genes that are not needed in every cell type are suppressed by default, and only activated in those cells where they are necessary. In unicellular lifeforms, however, most genes will be in constant use, and keeping those genes easily accessible is more favorable.

On the other hand, Vavouri and Lehner [29] have found that the NARs found in humans in fact serve a different purpose, namely the retention of certain nucleosomes in sperm cells, and their study of the signals found for housekeeping genes versus tissue-specific genes directly contradicts the hypothesis of Tillo *et al.* The NARs we find in multicellular life may therefore instead be indicative of the need to retain nucleosomes in the germ cells of multicellular organisms.

NARs are common to complex multicellular lifeforms, while almost all unicellular lifeforms we analyzed have NDRs. In-between there is a range of organisms with hybrid positioning signals. In almost all of these signals, however, the NAR is a more prominent feature than the NDR. This leads to a clear distinction between uni- and multicellular life based on the type of nucleosome positioning signals found in the promoter regions.

Furthermore, the strength of the NAR appears to increase with organism complexity. This fits the hypothesis of Tillo *et al.* [182], since organisms with more cell differentiation will have more genes suppressed by an NAR (and possibly by stronger ones). If the purpose of the NARs is solely to retain nucleosomes in the germline, it seems that more complex life cares more strongly about retaining its nucleosomes and passing on epigenetic information. More research will be needed to explore this idea.

Given the presence of hybrid signals, we speculate that the encoding of NARs versus NDRs in promoter regions is not an all-or-nothing choice for organisms. Whether the NARs serve to close off genes by default, or to retain nucleosomes in the germline, they compete with an apparent need to create an NDR to facilitate the initiation of transcription. The organisms showing hybrid signals seem to strike a balance between the two.

We hope that our results will motivate the experimental community to expand the available catalog of *in vitro* nucleosome maps to a greater number and variation of organisms. This will help not only verify our findings but also be of great service to any follow-up inquiries into the deeper nature and meaning of the signals we have found. We also suggest that nucleosome maps be generated at lower nucleosome densities, because steric hindrance will hide strong enrichment signals.

We also hope to encourage further examination of housekeeping versus tissue-specific genes in other organisms to further test the hypothesis of Tillo *et al.* [182], and an expansion of the results of Vavouri and Lehner [29] to other organisms, in order to test whether or not nucleosome retention in the germline is a goal served by the mechanical signals we find in the genomes of other complex organisms. If so, our results raise an intriguing question: why do more complex organisms tend to favor stronger nucleosome retention?

CONCLUSIONS

The work presented in the past five chapters took different directions, but it has all revolved around the question of to what extent variations in nucleotide sequence can lead to DNA molecules with different physical behavior. Let us sum up our conclusions.

In Chapter 2 we found that, making use of the intrinsic curvature of nucleotide sequences, we can engineer DNA molecules with an atypical response to tension. These superhelical molecules act like nanoscale springs. In Chapter 3 we designed nucleosomal DNA sequences that led to nucleosomes with nontrivial physical behavior, namely unwrapping very easily when put under tension. This illustrates the possibility of nucleosome speciation.

In Chapter 4 we showed that, making use of the MMC method, we can create computationally far less costly approximations to complex biophysical models of DNA systems. In Chapter 5 we extended the methodology further, and we showed that we can decouple the selection pressure on sequences from the simulation temperature of the MMC simulation, and that we can use this to perform SELEX experiments *in silico*.

Finally, in Chapter 6, we applied the model of Chapter 4 in a biological setting and we found that nature has encoded nucleosome affinity signals into the nucleotide sequences of the promoter regions of organisms across the tree of life. We showed that only a small set of signal types seems to be used in nature, and that the classes into which these signal types divide the organisms coincide with the fundamental biological distinction between unicellular and multicellular life.

We have approached the question of the importance of nucleotide sequence from different angles. We have contributed novel methodology to the field, which has opened up questions not previously accessible with biophysical models of significant complexity. From the viewpoint of nanoengineering and designer DNA molecules, we conclude that sequences can be designed that lead to DNA molecules and, by extension, nucleosomes, that show nontrivial physical behavior. These results are expected to be only the first of many to come. Especially the results on nucleosomes lend credence to a wider idea: that nucleosomes are better viewed as a family of systems, with distinct attributes depending on the DNA sequence they contain. In a similar vein to our design of nucleosomal 'force

sensors', one can imagine designing nucleosomes that are good at storing twist, or that behave asymmetrically in their response to being invaded by transcriptional machinery ('polar barriers').

From the biological point of view, we find that evolution has also engineered parts of genomes to exhibit specific physical properties, not only in the few model organisms that have been experimentally studied so far, but in all the organisms we have enough genomic data for, whether they are animals, plants, fungi or simple unicellular lifeforms. The universality of the signals we find in real genomes is surprising, and more research will have to follow before we will fully understand what they mean.

APPENDICES

A | A NOTE ON MODEL PARAMETERIZATION

As discussed in Section 1.1, the Rigid Base Pair (RBP) model can be parameterized in various ways. It has long had a standard parameterization in what we call the hybrid or mixed parameterization: intrinsic values from crystallography data [11] and stiffnesses from molecular dynamics simulations [13]. This is primarily due to the conclusions of Becker *et al.* [14], who benchmarked the predictions of various parameterizations for the nucleosome affinity of a set of sequences against experimental measurements, and found that this hybrid parameterization gave the best results.

This conclusion has led to the use of the hybrid parameterization in much subsequent research including most of that presented in this thesis. However, we have found that this parameterization does not work in all cases, as we will explain below.

Furthermore, in Chapters 4 and 6 we employed oligonucleotide distributions for the Markov-chain model for nucleosome affinity that were derived using MMC simualtions at artificially low temperature. This helped the simulations converge more quickly. However, as we saw in Chapter 5 (the work for which was done after that for Chapters 4 and 6), temperature can be a subtle parameter to manipulate. We therefore also wish to briefly expand upon the effects of this artificially chosen temperature.

In light of what we learned during the research for Chapter 5, we later spent the computational time to generate sequence ensembles at room temperature, rather than at artificially low simulation temperatures. The analyses presented in this Appendix are all based on those room temperature simulations, to remove the complications in interpreting the results associated with temperature effects.

A.1 PARAMETERIZATIONS: CRYSTALLOGRAPHY, MOLECULAR DYNAMICS, AND THEIR HYBRIDS

In Chapter 6 we switched from parameterizing our model using the hybrid parameterization to using the pure parameters obtained solely from crystallography data [11]. The reason for this switch is that the hybrid parameterization turned out not to be able to correctly capture long-range

effects. This is evident in Fig. A.1, which shows the same predictions as Figs. 6.2A-B, with various parameterizations.

We saw in Chapter 6 that the pure crystallography prediction corresponds well to the experimentally measured signals. However, the predictions made using the pure MD and hybrid parameterizations do not capture the nucleosome depletion signal at all. Even though the hybrid parameterization was deemed to be the most accurate in the benchmarks of Becker *et al.* [14], we find here a situation where it fails completely.

We also considered the reverse hybrid parameterization, which uses stiffness parameters derived from crystallography data and shape parameters from MD simulations. This reverse hybrid parameterization still captures the signal, albeit that the amplitude is too small. Apparently, switching out the crystallography shape parameters for the MD ones reduces the ability of the model to map the signals we are looking for. However, the stiffness parameters are more important: using MD stiffness parameters destroys the signal, regardless of which shape parameters are used.

The reason for this failure is suggested by the black dotted curves depicted in Fig. A.1, which shows the average GC content (averaged using a 147-bp window) in the regions of interest. As we saw in Chapter 6, the nucleosome occupancy signals in the promoter regions of organisms across the tree of life correlate well with GC content. The problem with the parameterizations that fail at predicting these signals seems to be a lack of correlation of their predictions with the average GC content.

We can understand this observation when we look at the dinucleotide distributions derived from MMC simulations using each set of parameters. Fig. A.2 shows the distributions of A/T-rich and G/C-rich dinucleotides for the pure crystalography parameterization and the reverse hybrid parameterization. We see that, apart from the oscillatory behavior, there is an overall preference for G/C-rich dinucleotides over the A/T-rich ones. On the other hand, the distributions of the pure MD parameterization and the hybrid parameterization, plotted in Fig. A.3, do not show a strong preference for G/C or A/T.

The sequences for which Becker *et al.* [14] analyzed the predictions of the different parameterizations were all relatively short. On short scales, the precise local positioning is important, and the most important feature that a parameterization needs to capture is the periodicity in the probability distributions of the dinucleotides. It may well be that the hybrid parameterization in this regard delivers superior performance. However, for the application we consider in Chapter 6 and in this section, our interest lies in mapping long-range effects. It seems that to capture these

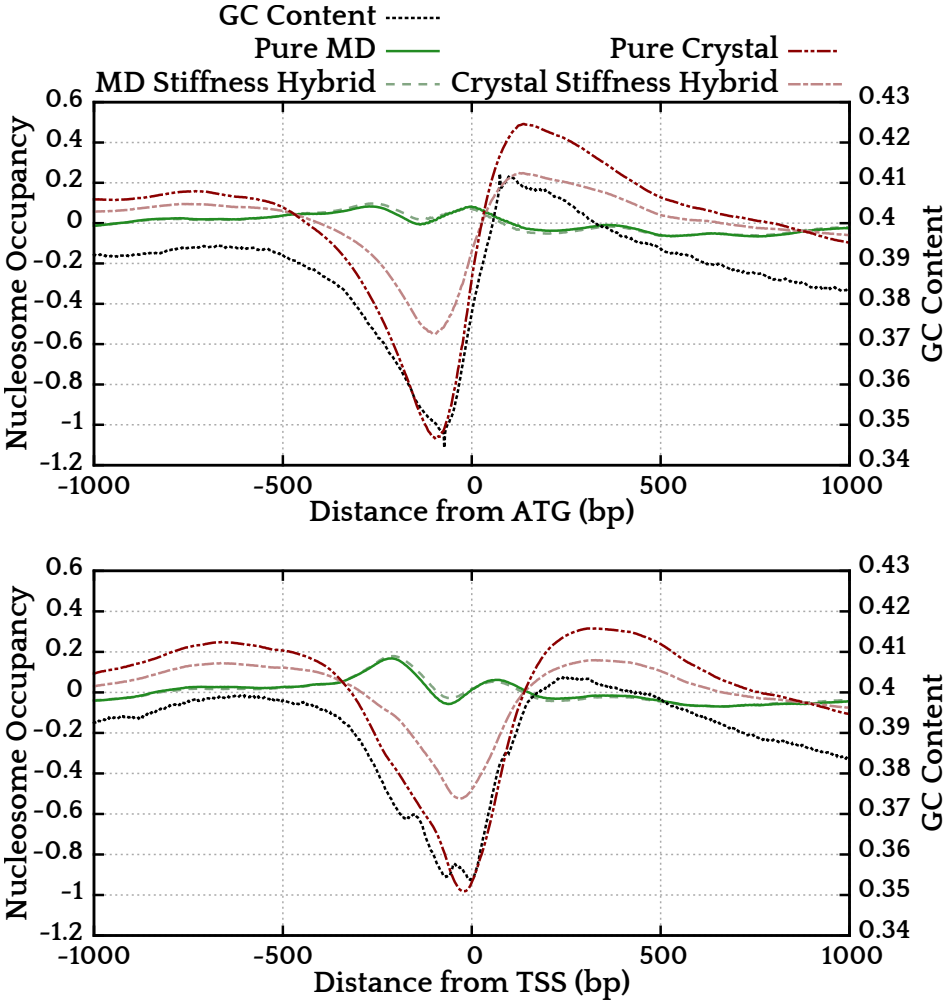

Figure A.1: Nucleosome occupancy signals in the promoter regions of *S. cerevisiae* (analogous to Figs. 6.2A-B) as predicted using the model of Chapter 4 using various parameterizations: pure MD parameters [13] *(solid green curve)*, pure crystallography parameters [11] *(dash-dot-dotted dark red curve)*, and the two possible hybrids: the MP parameterization in [14], using MD stiffness parameters with crystallography shape parameters *(dashed pale green curve)*; and the reverse of this hybrid, using crystallography stiffness parameters with MD shape parameters *(dash-dotted pale red curve)*. Also shown for reference is the average GC content in the promoter regions, smoothed over a 147-base-pair window *(black dotted curve)*.

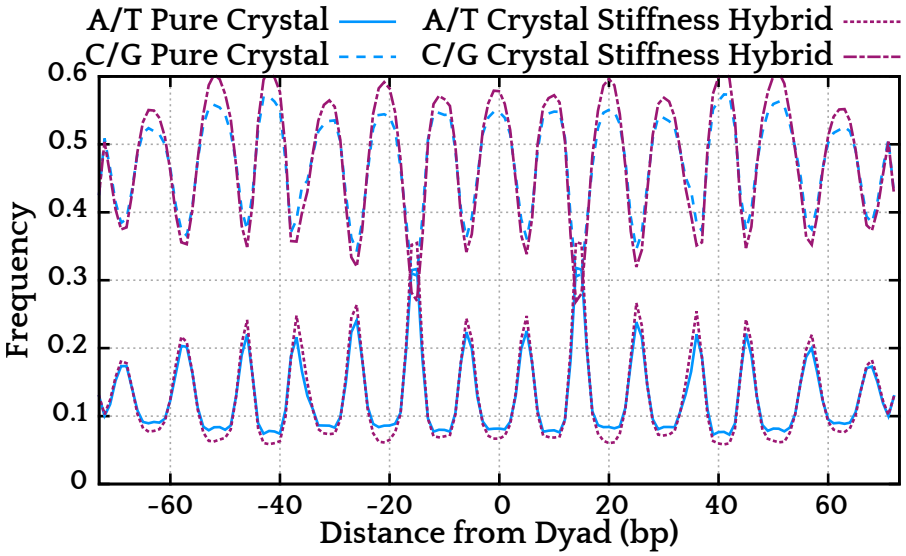

Figure A.2: Dinucleotide distributions along the nucleosome (modeled as in [1] and the rest of this thesis), with the nucleosome model parameterized by the pure crystallography parameters and the reverse hybrid parameterization (crystallography stiffness parameters, MD shape parameters). The label A/T indicates the combined probabilities of the dinucleotides AA, AT, TA and TT. The label G/C corresponds to CC, CG, GC and GG. Both parameterizations lead to a nucleosome with a significant preference for GC-rich dinucleotides.

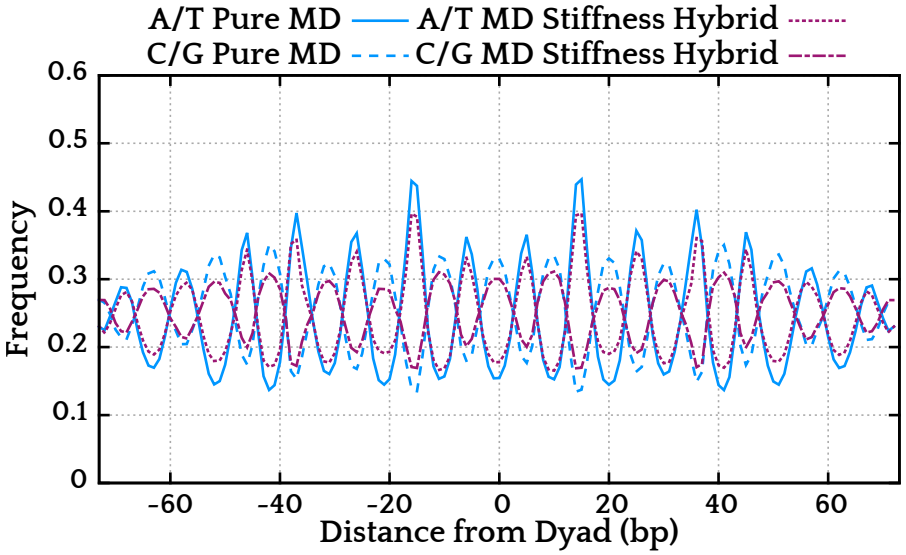

Figure A.3: As Fig. A.2 for the pure MD parameterization and the hybrid parameterization (MD stiffness parameters, crystallography shape parameters). These parameterizations show no significant preference for GC- or AT-rich dinucleotides.

long-range signals correctly, a proper correlation with GC content, and therefore a preference for high GC content, is required.

The pure crystallography parameterization does prefer high GC, but the hybrid parameterization lacks this preference. On the other hand, the pure crystallography parameterization, judging by the results of Becker *et al.* [14], does not capture the periodic signal in the distributions as well as does the hybrid parameterization. Therefore, these two parameterizations (pure crystallography and the original hybrid) may be taken to be complementary: neither performs optimally in all situations and for any given application, a careful choice should be made.

A.2 THE EFFECT OF MMC TEMPERATURE ON THE MARKOV-CHAIN MODEL

The probability distributions that inform the Markov-chain model introduced in Chapter 4 depend not only on the chosen parameterization, but also on the temperature at which the MMC simulation is run. The effect of temperature, qualitatively, is predictable: a lower simulation temperature leads to stronger preferences of the nucleosome for accommodating sequences.

Running the MMC at temperatures lower than room temperature (which is the temperature we are generally interested in) is desirable because the simulations converge more quickly. In Chapters 4 and 6 we utilized probability distributions gained from MMC simulations run at artificially low temperatures, and in which we scaled the results back up to room temperature (Eq. 4.7). The discerning reader may have noticed that this trick is in fact the same technique used in Chapter 5 to change the mutation temperature separately from the physical temperature.

Chapter 5 presents a far deeper understanding of the role of temperature in MMC simulations than was available at the time when the work for Chapters 4 and 6 was performed. Knowing now that our rescaled probabilities in fact represent the probabilities found using a low temperature for the configurational moves, and room temperature for the mutations, we should check what the effects are. We already showed in Section 5.4 that for nucleosomes, the spatial temperature has only a moderate effect on the sequence preferences (see Fig. 5.1A-C). Still, we would do well to check what effect this has on the results of Chapter 6.

To do so, we have now invested the computational time to run our MMC simulations at room temperature (all the results shown in this appendix

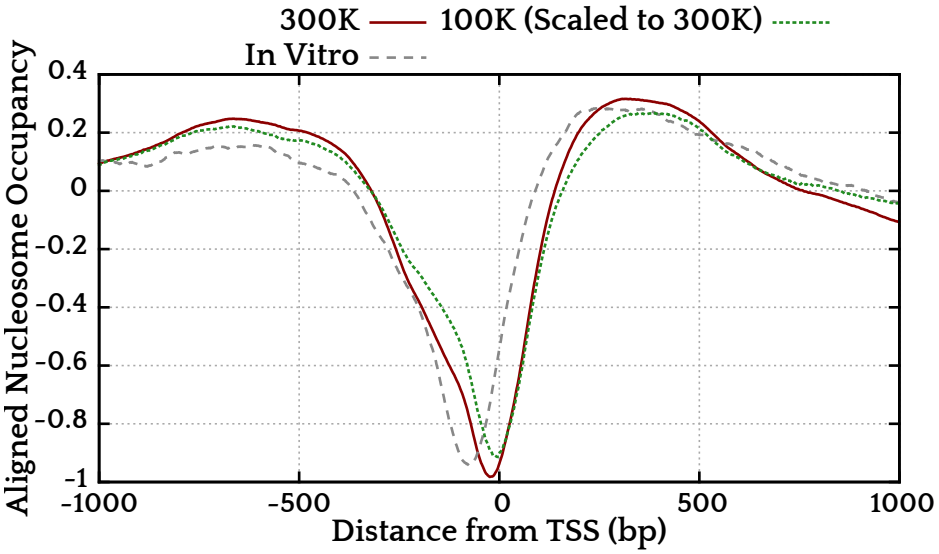

Figure A.4: As Fig. 6.2A, with the prediction using distributions derived at room temperature added. Correlation with GC content is very similar for the pure room temperature model and the low temperature model, and the predicted nucleosome occupancy signals differ little.

were derived at room temperature.) In Fig. A.4 we see that using the pure room temperature distributions leads to a slightly stronger nucleosome depletion signal. However, the differences between the two predictions are minimal, and both provide agreement with the *in vitro* data of similar quality. While this does not constitute a full do-over of the analyses performed in Chapter 6, this agreement indicates that the results of Chapter 6 are unlikely to be strongly affected by the temperature of the configurational moves in the MMC simulation, at least down to 1/3 of room temperature.

B | A LIST OF SEQUENCES OF INTEREST

For reference, this appendix lists in full detail the nucleotide sequences of particular interest mentioned in the chapters of this thesis.

[1] The Widom 601 sequence [85]:

```
CTGGAGAATC CCGGTGCCGA GGCCGCTCAA TTGGTCGTAG ACAGCTCTAG
CACCGCTTAA ACGCACGTAC GCGCTGTCCC CCGCGTTTTA ACCGCCAAGG
GGATTACTCC CTAGTCTCCA GGCACGTGTC AGATATATAC ATCCTGT
```

[2] The kinetoplast DNA sequence [96, 107]:

```
GAATTCCCAA AAATGTCAAA AAATAGGCAA AAAATGCCAA AAATCCCAAA
```

[3] The artificial, strongly curved, sequence found through MMC in Chapter 2 and Ref. [86]:

```
AACCCCCTTT AAAGAGCTTT TTAGAGCTTT TAAAGCCTCT TTAACCCTCT
TTAAACCCTC TTAAAAGCTC TTTTAAGCCC TTTT
```

[4] The straight sequence with the same dinucleotide content as (a tandem repeat of) Seq. [3], also from Chapter 2 and Ref. [86]. The entire sequence consists of 4700 nucleotides; only a fragment is given here:

```
AACCCCCTTT AAAGAGCTTT TAACCCCCTT TAAAGAGCTT TTTAGAGCTT
TTAAAGCCTC CCTTTAAGCT CTTTTACCCT CTTTAAACCC TCTAAGCTTT
TTAGAGCTTT TAAAGCCTTA AAGAGCTTTT TAGAAGCCTC TTTAAGCTCT
TTTAAAGCTC TTTTAAGCCC CCCTTTAAAA AGAGCTTTTT AAAAAGCTCT
TTTAAGCTTT AAACCCTCTT TAAAAAGAAG CCTCTCTTTT AAGCTTTTTT
AACCCTCTTA AAAGCTCTTT TAAGCTTTAA ACCCCTTAAC CCTCTTT...
```

[5] The sequence YAL002W-826, which starts at base pair position 826 of the YAL002W gene of *S. cerevisiae*. The sequences [6]-[23] are all modifications to this sequence, through either free MMC (in which case the starting sequence is actually irrelevant, but we mention it for the sake of coherence) or synonymous MMC, in various nucleosomal unwrapping states (see Chapter 3). The names follow the pattern `Y826-LR{syn}`, where L is the number of binding sites opened from the left, and R the number of sites opened from the right, and

syn is either present or not, indicating whether the sequence results from synonymous mutations or not. Nucleotides that have remained unchanged from the original sequence are printed in gray, while nucleotides that have been altered are printed in black.

CATTTTGCCC TTATTTTATT ATCGCCACAC GTTTCTTTGA TGTTTCAAGA
AACTGTTGAA CCCTCAGTAC AAAATTCTCT AGTCGTGAAT AGCTCTATTT
CATGGACTCA AAACTGTTCC AGGGTTGCTT ATTCCGTAAA TAATAAA

[6] Y826-08

AAGATAAAAG CTCTTTATAA GCCTCTTAAC CCCTATTTAA AGAGTTTTAA
GAGCCTTTAA CGGTTTAAAA GGGGTTTTGA GGGTATATTA CCCCGCGGCC
CGGCGCGCGC GCGCGGCGGC GCCGACGCGC GCGCGACCGG CGCGGTA

[7] Y826-08syn

CACTTTGCCC TCATTTTATT GAGTCCTCAC GTCTCTTTGA TGTTCCAAGA
GACCGTCGAA CCGTCTGTAC AGAACTCGTT AGTCGTTAAT TCCTCGATAT
CGTGGACGCA GAACTGTAGT CGCGTAGCGT ACAGCGTGAA TAACAAA

[8] Y826-17

ACCCCCTTTT AAGAGGAAAA GCCTCTTTAC CCCGGGTTAA AGCTCTTTAA
AGCCCTTTAA CGAGCGTTAC CTCTTTAAAG AGGGTTAAAC CGGCTACCCC
GCCCGCTCCG CGCGGTCGTC GCGCTCCGCG ACGCTCGGCG GCGCGCG

[9] Y826-17syn

CACTTCGCTT TAATATTATT GAGTCCACAC GTCTCTTTGA TGTTCCAAGA
GACCGTCGAA CCGAGCGTAC AGAATTCGCT AGTCGTAAAC AGCAGTATCA
GCTGGACGCA GAACTGCAGC CGCGTCGCGT ACTCCGTCAA TAACAAA

[10] Y826-26

GAGCGGCAAA ACCCCCTTTT AACCGGTAAA CCGGGGTTAA AGAGTTTTAA
GAGCCTTTAA TGCTCGTTTA GAGCTCTTAA CCGTTTAAAG AGGGTTAATG
CGGTCTTGGA CCGCTCGGGC TATACTCCGC GGGCGTCCGC GGCGGAG

[11] Y826-26syn

CACTTCGCTT TAATCCTTTT AAGCCCTCAC GTCTCTTTGA TGTTCCAAGA
GACCGTCGAA CCGTCGGTAC AGAACTCTTT AGTCGTAAAC AGCTCTATAA
GCTGGACGCA GAACTGTTCG CGCGTAGCGT ACAGCGTTAA TAATAAA

[12] Y826-35

GCGCGCTCGT CGCCGCTCAA AAACCCCTTT TAAAGGTTAA AGCTCTTTAA
GACCTCTTAA AAGCTCATAA AGAGCTTATA ACGAGTTTAA ACTCTTAAAG
AGGGCTTTGA GGCTACGACG CGCGCGCCGC CGGCCGGCCG CCCGCGG

[13] Y826-35syn

CACTTCGCGT TGATCCTTTT AAGCCCTCAC GTAAGTTTAA TGTTCCAAGA
GACCGTCGAA CCGAGCGTAC AGAACTCTTT AGTGGTTAAT AGCTCTATAA
GCTGGACGCA GAACTGTTCG CGGGTCGCGT ACTCCGTGAA CAATAAG

[14] Y826-44

CCGGCTCGGG CGCGGTCGTA TACGCCCTTA AACCCCCTTT AAAGAGGTAA
AGCCCTTTAT AAGCTCGTTT AAGCTCTTTA CCGCTCGTTA AAGGGCTTTT
CCGGTTTAAG AGGGGTTTAA GCCTCTATCG CCGGTCGGGT CGCGCGC

[15] Y826-44syn

CACTTCGCCC TGATCCTATT AAGCCCGCAC GTATCTCTTA TGTTCCAAGA
GACCGTCGAA CCGTCGGTAC AGAACTCTTT AGTGGTTAAT AGTAGTATTT
CCTGGACGCA GAACTGTTCG AGGGTCGCGT ACAGCGTAAA CAATAAA

[16] Y826-53

CCGCGCCCCG CCGCGTCGTC GGCGCCGCGA CGAGACTCTT TAACCCCTTT
TAAGAGTTAA CCGCGGGTAA AAGCTCTATA ACGAGCATTA AAGGCTCTTT
TAAGCGTTTA ACCTTTTAAA GGGGCTTAAA CGCGTCGGCG CGTCGCG

[17] Y826-53syn

CACTTCGCCC TGATCCTGCT AAGTCCTCAC GTATCTCTTA TGTTCCAGGA
AACCGTCGAA CCCAGCGTAC AGAACTCTTT AGTGGTTAAT AGTAGTATTT
CGTGGACTCA AAACTGTTCG AGGGTCGCGT ACTCCGTTAA TAACAAA

[18] Y826-62

CCCCGTCGGG TAACGCGTCC CGTACGCGGT ACGCCCGCGC GTCCCCCTCA
AAAACCCTCTT TAAAAGGTAA AGAGCTCTAA ACGCGCGTTA AAGGGCTTTT
ATGAGCTTTA ACCCGGGTTT ACCGGTTAAA AGGGGTTTAA CAGCTCT

[19] Y826-62syn

CACTTCGCGT TGATCCTGCT GAGCCCTCAC GTGAGCCTTA TGTTCCAGGA
AACAGTCGAA CCAAGTGTAC AGAACTCTTT AGTGGTTAAT AGTAGTATTT
CGTGGACACA AAACTGCTCG CGCGTCGCGT ACAGCGTTAA TAACAAA

[20] Y826-71

```
GCGGCTAGCG GGCCGAGAGG CGAGTCGCGG CGCCGATCGT TCGCCGACCG
ACCCCGCCTT TTACCCTCTT TAAACCGGTA ACGCGCATAA AAGGCTCTTA
AAACGCTTTA ACCCGGGGTA AAAGAGGTTA TTCGGTTTAA AGGGGGT
```

[21] Y826-71syn

```
CATTTCGCGC TGATCTTGCT AAGCCCGCAC GTCTCGCTCA TGTTCCAGGA
AACGGTCGAA CCGAGCGTTC AAAATAGTTT AGTGGTTAAT AGTAGTATTT
CGTGGACTCA AAACTGCTCG AGAGTCGCTT ACTCGGTTAA CAACAAA
```

[22] Y826-80

```
GTCCGAGGTC CGTCCGTCTA GGCCGCGCGG CGCCGCGATC GGGACGCGCG
ATACGGTCGC GCCCCCGCTT AAACCCCTTT TAAACCGGTA AGAGGCTTTT
AAAGTCTTTA ACCCGGGGTA AAGAGGGTTA TTAAAGGCTT TAATCTT
```

[23] Y826-80syn

```
CACTTCGCGC TCATACTACT ATCGCCGCAC GTGAGCCTGA TGTTCCAGGA
AACGGTGGAA CCGTCCGTTC AAAACTCTCT CGTAGTAAAT AGTAGTATTT
CGTGGACACA AAACTGCTCG CGAGTCGCTT ATAGCGTCAA TAACAAA
```

[24] The 36-base-pair locking sequence from Rosanio *et al.* [93], providing a directional bias to their DNA rings:

```
TATCTGGTGG GAAACAAGCT TCAGCGATGA GATGAG
```

BIBLIOGRAPHY

[1.] Eslami-Mossallam, B., R. D. Schram, M. Tompitak, J. van Noort, and H. Schiessel. "Multiplexing Genetic and Nucleosome Positioning Codes: A Computational Approach." *PLoS ONE* 11 (2016), e0156905.

[2.] Bruin, L. de, M. Tompitak, B. Eslami-Mossallam, and H. Schiessel. "Why Do Nucleosomes Unwrap Asymmetrically?" *J. Phys. Chem. B* 120 (2016), 5855–5863.

[3.] Culkin, J., L. de Bruin, M. Tompitak, and H. Schiessel. "Sequence dependence in nucleosome breathing." submitted (2017).

[4.] Maffeo, C., J. Yoo, J. Comer, D. B. Wells, B. Luan, and A. Aksimentiev. "Close encounters with DNA." *J. Phys. Cond. Mat.* 26 (2014), 413101.

[5.] Pablo, J. J. de. "Coarse-grained simulations of macromolecules: from DNA to nanocomposites." *Annual Review of Physical Chemistry* 62 (2011), 555–74.

[6.] Olson, W. K. et al. "A standard reference frame for the description of nucleic acid base-pair geometry." *J. Mol. Biol.* 313 (2001), 229–237.

[7.] Gonzalez, O., D. Petkevičiūtė, and J. H. Maddocks. "A sequence-dependent rigid-base model of DNA." *J. Chem. Phys.* 138 (2013), 55102.

[8.] Petkevičiūtė, D., M. Pasi, O. Gonzalez, and J. H. Maddocks. "cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA." *Nucleic Acids Res.* 42 (2014), e153.

[9.] Dršata, T. and F. Lankaš. "Theoretical models of DNA flexibility." *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3 (2013), 355–363.

[10.] Potoyan, D. A., A. Savelyev, and G. A. Papoian. "Recent successes in coarse-grained modeling of DNA." *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3 (2013), 69–83.

[11.] Olson, W. K., A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes." *Proc. Natl. Acad. Sci. USA* 95 (1998), 11163–11168.

[12.] Balasubramanian, S., F. Xu, and W. K. Olson. "DNA sequence-directed organization of chromatin: Structure-based computational analysis of nucleosome-binding sequences." *Biophys. J.* 96 (2009), 2245–2260.

[13.] Lankaš, F., J. Sponer, J. Langowski, and T. E. Cheatham III. "DNA basepair step deformability inferred from molecular dynamics simulations." *Biophys. J.* 85 (2003), 2872–2883.

[14.] Becker, N. B., L. Wolff, and R. Everaers. "Indirect readout: Detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials." *Nucleic Acids Res.* 34 (2006), 5638–5649.

[15.] Yanagi, K., G. G. Privé, and R. E. Dickerson. "Analysis of local helix geometry in three B-DNA decamers and eight dodecamers." *J. Mol. Biol.* 217 (1991), 201–214.

[16.] Packer, M. J., M. P. Dauncey, and C. A. Hunter. "Sequence-dependent DNA structure: tetranucleotide conformational maps." *J. Mol. Biol.* 295 (2000), 85–103.

[17.] Gardiner, E. J., C. A. Hunter, M. J. Packer, D. S. Palmer, and P. Willett. "Sequence-dependent DNA Structure: A database of octamer structural parameters." *J. Mol. Biol.* 332 (2003), 1025–1035.

[18.] Dixit, S. B. et al. "Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps." *Biophys. J.* 89 (2005), 3721–3740.

[19.] Fujii, S., H. Kono, S. Takenaka, N. Go, and A. Sarai. "Sequence-dependent DNA deformability studied using molecular dynamics simulations." *Nucleic Acids Res.* 35 (2007), 6063–6074.

[20.] Lavery, R. et al. "A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA." *Nucleic Acids Res.* 38 (2009), 299–313.

[21.] Kavenoff, R. and B. C. Bowen. "Electron microscopy of membrane-free folded chromosomes from Escherichia coli." *Chromosoma* 59 (1976), 89–101.

[22.] Cutter, A. R. and J. J. Hayes. "A brief review of nucleosome structure." *FEBS Letters* 589 (2015), 2914–2922.

[23.] Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. "Crystal structure of the nucleosome core particle at 2.8 A resolution." *Nature* 389 (1997), 251–60.

[24.] Morozov, A. V., K. Fortney, D. A. Gaykalova, V. M. Studitsky, J. Widom, and E. D. Siggia. "Using DNA mechanics to predict in vitro nucleosome positions and formation energies." *Nucleic Acids Res.* 37 (2009), 4707–4722.

[25.] Becker, N. B. and R. Everaers. "DNA Nanomechanics in the Nucleosome." *Structure* 17 (2009), 579–589.

[26.] Fathizadeh, A., A. B. Besya, M. R. Ejtehadi, and H. Schiessel. "Rigid-body molecular dynamics of DNA inside a nucleosome." *Eur. Phys. J. E* 36 (2013), 21.

[27.] Eslami-Mossallam, B., H. Schiessel, and J. van Noort. "Nucleosome dynamics: Sequence matters." *Advances in Colloid and Interface Science* 232 (2016), 101–113.

[28.] Pugh, B. F. "A preoccupied position on nucleosomes." *Nature Struct. Mol. Biol.* 17 (2010), 923.

[29.] Vavouri, T. and B. Lehner. "Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome." *PLoS Genetics* 7 (2011), e1002036.

[30.] Parmar, J. J., J. F. Marko, and R. Padinhateeri. "Nucleosome positioning and kinetics near transcription-start-site barriers are controlled by interplay between active remodeling and DNA sequence." *Nucleic Acids Res.* 42 (2014), 128–136.

[31.] Struhl, K. and E. Segal. "Determinants of nucleosome positioning." *Nature Struct. Mol. Biol.* 20 (2013), 267–73.

[32.] Todolli, S., P. J. Perez, N. Clauvelin, and W. K. Olson. "Contributions of Sequence to the Higher-Order Structures of DNA." *Biophys. J.* 112 (2017), 416–426.

[33.] Lorch, Y., J. W. LaPointe, and R. D. Kornberg. "Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones." *Cell* 49 (1987), 203–210.

[34.] Han, M. and M. Grunstein. "Nucleosome loss activates yeast downstream promoters in vivo." *Cell* 55 (1988), 1137–1145.

[35.] Cairns, B. R. "The logic of chromatin architecture and remodelling at promoters." *Nature* 461 (2009), 193–198.

[36.] Radman-Livaja, M. and O. J. Rando. "Nucleosome positioning: How is it established, and why does it matter?" *Dev. Biol.* 339 (2010), 258–266.

[37.] Drillon, G., B. Audit, F. Argoul, and A. Arneodo. "Ubiquitous human 'master' origins of replication are encoded in the DNA sequence via a local enrichment in nucleosome excluding energy barriers." *J. Phys. Cond. Mat.* 27 (2015), 64102.

[38.] Field, Y., Y. N. Fondufe-Mittendorf, I. K. Moore, P. Mieczkowski, N. Kaplan, Y. Lubling, J. D. Lieb, J. Widom, and E. Segal. "Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization." *Nature Genetics* 41 (2009), 438–445.

[39.] Sasaki, S. et al. "Chromatin-Associated Periodicity in Genetic Variation Downstream of Transcriptional Start Sites." *Science* 323 (2009), 401–404.

[40.] Tirosh, I., N. Sigal, and N. Barkai. "Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences." *Molecular systems biology* 6 (2010), 365.

[41.] Tsankov, A. M., Y. Yanagisawa, N. Rhind, A. Regev, and O. J. Rando. "Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization." *Genome Res.* 21 (2011), 1851–1862.

[42.] Tolstorukov, M. Y., N. Volfovsky, R. M. Stephens, and P. J. Park. "Impact of chromatin structure on sequence variability in the human genome." *Nature Struct. Mol. Biol.* 18 (2011), 510–515.

[43.] Guan, Y., V. Yao, K. Tsui, M. Gebbia, M. J. Dunham, C. Nislow, and O. G. Troyanskaya. "Nucleosome-coupled expression differences in closely-related species." *BMC Genomics* 12 (2011), 466.

[44.] Tsui, K., S. Dubuis, M. Gebbia, R. H. Morse, N. Barkai, I. Tirosh, and C. Nislow. "Evolution of nucleosome occupancy: conservation of global properties and divergence of gene-specific patterns." *Molecular and cellular biology* 31 (2011), 4348–55.

[45.] Prendergast, J. G. D. and C. A. M. Semple. "Widespread signatures of recent selection linked to nucleosome positioning in the human lineage." *Genome Res.* 21 (2011), 1777–1787.

[46.] Chen, X., Z. Chen, H. Chen, Z. Su, J. Yang, F. Lin, S. Shi, and X. He. "Nucleosomes Suppress Spontaneous Mutations Base-Specifically in Eukaryotes." *Science* 335 (2012), 1235–1238.

[47.] Langley, S. A., G. H. Karpen, and C. H. Langley. "Nucleosomes Shape DNA Polymorphism and Divergence." *PLoS Genetics* 10 (2014), e1004457.

[48.] Makova, K. D. and R. C. Hardison. "The effects of chromatin organization on variation in mutation rates in the genome." *Nature Reviews Genetics* 16 (2015), 213–223.

[49.] Zhang, T., W. Zhang, and J. Jiang. "Genome-Wide Nucleosome Occupancy and Positioning and Their Impact on Gene Expression and Evolution in Plants." *Plant Physiol.* 168 (2015), 1406–16.

[50.] Drillon, G., B. Audit, F. Argoul, and A. Arneodo. "Evidence of selection for an accessible nucleosomal array in human." *BMC Genomics* 17 (2016), 526.

[51.] Lieleg, C., N. Krietenstein, M. Walker, and P. Korber. "Nucleosome positioning in yeasts: methods, maps, and mechanisms." *Chromosoma* 124 (2015), 131–151.

[52.] Kunkel, G. R. and H. G. Martinson. "Nucleosomes will not form on double-stranded RNA or over poly(dA)-poly(dT) tracts in recombinant DNA." *Nucleic Acids Res.* 9 (1981), 6869–6888.

[53.] Prunell, A. "Nucleosome reconstitution on plasmid-inserted poly(dA) . poly(dT)." *EMBO J.* 1 (1982), 173–179.

[54.] Widlund, H. R., H. Cao, S. Simonsson, E. Magnusson, T. Simonsson, P. E. Nielsen, J. D. Kahn, D. M. Crothers, and M. Kubista. "Identification and characterization of genomic nucleosome-positioning sequences." *J. Mol. Biol.* 267 (1997), 807–817.

[55.] Cao, H., H. R. Widlund, T. Simonsson, and M. Kubista. "TGGA repeats impair nucleosome formation." *J. Mol. Biol.* 291 (1998), 253–260.

[56.] Tolstorukov, M. Y., A. V. Colasanti, D. M. McCandlish, W. K. Olson, and V. B. Zhurkin. "A Novel Roll-and-Slide Mechanism of DNA Folding in Chromatin: Implications for Nucleosome Positioning." *J. Mol. Biol.* 371 (2007), 725–738.

[57.] Tolstorukov, M. Y., V. Choudhary, W. K. Olson, V. B. Zhurkin, and P. J. Park. "nuScore: A web-interface for nucleosome positioning predictions." *Bioinformatics* 24 (2008), 1456–1458.

[58.] Miele, V., C. Vaillant, Y. D'Aubenton-carafa, C. Thermes, and T. Grange. "DNA physical properties determine nucleosome occupancy from yeast to fly." *Nucleic Acids Res.* 36 (2008), 3746–3756.

[59.] De Santis, P., S. Morosetti, and A. Scipioni. "Prediction of nucleosome positioning in genomes: limits and perspectives of physical and bioinformatic approaches." *J. Biomol. Struct. Dyn.* 27 (2010), 747–764.

[60.] Liu, H., X. Duan, S. Yu, and X. Sun. "Analysis of nucleosome positioning determined by DNA helix curvature in the human genome." *BMC Genomics* 12 (2011), 72.

[61.] Liu, G., Y. Xing, H. Zhao, J. Wang, Y. Shang, and L. Cai. "A deformation energy-based model for predicting nucleosome dyads and occupancy." *Scientific reports* 6 (2016), 24133.

[62.] Segal, E., Y. N. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. "A genomic code for nucleosome positioning." *Nature* 442 (2006), 772–8.

[63.] Field, Y., N. Kaplan, Y. N. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal. "Distinct modes of regulation by chromatin encoded through nucleosome positioning signals." *PLoS Comput. Biol.* 4 (2008), e1000216.

[64.] Kaplan, N. et al. "The DNA-encoded nucleosome organization of a eukaryotic genome." *Nature* 458 (2009), 362–366.

[65.] Peckham, H. E., R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng. "Nucleosome positioning signals in genomic DNA." *Genome Res.* 17 (2007), 1170–1177.

[66.] Gupta, S., J. Dennis, R. E. Thurman, R. Kingston, J. A. Stamatoyannopoulos, and W. S. Noble. "Predicting human nucleosome occupancy from primary sequence." *PLoS Comput. Biol.* 4 (2008), e1000134.

[67.] Reynolds, S. M., J. A. Bilmes, and W. S. Noble. "Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in Saccharomyces cerevisiae and Homo sapiens." *PLoS Comput. Biol.* 6 (2010), e1000834.

[68.] Locke, G., D. Tolkunov, Z. Moqtaderi, K. Struhl, and A. V. Morozov. "High-throughput sequencing reveals a simple model of nucleosome energetics." *Proc. Natl. Acad. Sci. USA* 107 (2010), 20998–21003.

[69.] Gabdank, I., D. Barash, and E. N. Trifonov. "FineStr: A web server for single-base-resolution nucleosome positioning." *Bioinformatics* 26 (2010), 845–846.

[70.] Guo, S. H., E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen, and K. C. Chou. "INuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition." *Bioinformatics* 30 (2014), 1522–1529.

[71.] Ioshikhes, I. P., I. Albert, S. J. Zanton, and B. F. Pugh. "Nucleosome positions predicted through comparative genomics." *Nature Genetics* 38 (2006), 1210–1215.

[72.] Yuan, G. C. and J. S. Liu. "Genomic sequence is highly predictive of local nucleosome depletion." *PLoS Comput. Biol.* 4 (2008), 0164–0174.

[73.] Tillo, D. and T. R. Hughes. "G+C content dominates intrinsic nucleosome occupancy." *BMC Bioinformatics* 10 (2009), 442.

[74.] Heijden, T. van der, J. J. F. A. van Vugt, C. Logie, and J. van Noort. "Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy." *Proc. Natl. Acad. Sci. USA* 109 (2012), E2514–22.

[75.] Liu, H., R. Zhang, W. Xiong, J. Guan, Z. Zhuang, and S. Zhou. "A comparative evaluation on prediction methods of nucleosome positioning." *Brief. Bioinform.* 15 (2013), 1014–1027.

[76.] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of state calculations by fast computing machines." *J. Chem. Phys.* 21 (1953), 1087–1092.

[77.] Hastings, W. K. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57 (1970), 97–109.

[78.] Tompitak, M., G. T. Barkema, and H. Schiessel. "Benchmarking and refining probability-based models for nucleosome-DNA interaction." *BMC Bioinformatics* 18 (2017), 157.

[79.] Tompitak, M., C. Vaillant, and H. Schiessel. "Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions." *Biophys. J.* 112 (2017), 505–511.

[80.] Böhm, V., A. R. Hieb, A. J. Andrews, A. Gansen, A. Rocker, K. Tóth, K. Luger, and J. Langowski. "Nucleosome accessibility governed by the dimer/tetramer interface." *Nucleic Acids Res.* 39 (2011), 3093–3102.

[81.] Zhang, B., W. Zheng, G. A. Papoian, and P. G. Wolynes. "Exploring the Free Energy Landscape of Nucleosomes." *Journal of the American Chemical Society* 138 (2016), 8126–8133.

[82.] Kulić, I. M. and H. Schiessel. "DNA spools under tension." *Phys. Rev. Lett.* 92 (2004), 228101–1.

[83.] Brower-Toland, B. D., C. L. Smith, R. C. Yeh, J. T. Lis, C. L. Peterson, and M. D. Wang. "Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA." *Proc. Natl. Acad. Sci. USA* 99 (2002), 1960–1965.

[84.] Ngo, T. T. M., Q. Zhang, R. Zhou, J. G. Yodh, and T. Ha. "Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility." *Cell* 160 (2015), 1135–1144.

[85.] Lowary, P. T. and J. Widom. "New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning." *J. Mol. Biol.* 276 (1998), 19–42.

[86.] Tompitak, M., H. Schiessel, and G. T. Barkema. "Force responses of strongly intrinsically curved DNA helices deviate from worm-like chain predictions." *EPL* 116 (2016), 68005.

[87.] Freeman, G. S., J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. J. de Pablo. "DNA shape dominates sequence affinity in nucleosome formation." *Phys. Rev. Lett.* 113 (2014), 1–5.

[88.] Laundon, C. H. and J. D. Griffith. "Curved helix segments can uniquely orient the topology of supertwisted DNA." *Cell* 52 (1988), 545–549.

[89.] Loenhout, M. T. J. van, M. V. de Grunt, and C. Dekker. "Dynamics of DNA Supercoils." *Science* 338 (2012), 94–97.

[90.] Bussiek, M., K. Klenin, and J. Langowski. "Kinetics of site-site interactions in supercoiled DNA with bent sequences." *J. Mol. Biol.* 322 (2002), 707–718.

[91.] Boedicker, J. Q., H. G. Garcia, S. Johnson, and R. Phillips. "DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation." *Phys. Biol.* 10 (2013), 66005.

[92.] Haeusler, A. R., K. A. Goodson, T. D. Lillian, X. Wang, S. Goyal, N. C. Perkins, and J. D. Kahn. "FRET studies of a landscape of Lac repressor-mediated DNA loops." *Nucleic Acids Res.* 40 (2012), 4432–4445.

[93.] Rosanio, G., J. Widom, and O. C. Uhlenbeck. "In vitro selection of DNAs with an increased propensity to form small circles." *Biopolymers* 103 (2015), 303–320.

[94.] Dekker, J. and L. Mirny. "The 3D Genome as Moderator of Chromosomal Communication." *Cell* 164 (2016), 1110–1121.

[95.] Vologodskaia, M. and A. Vologodskii. "Contribution of the intrinsic curvature to measured DNA persistence length." *J. Mol. Biol.* 317 (2002), 205–213.

[96.] Widom, J. "Bent DNA for gene regulation and DNA packaging." *BioEssays* 2 (1985), 11–14.

[97.] Robinson, P. J. J., L. Fairall, V. A. T. Huynh, and D. Rhodes. "EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure." *Proc. Natl. Acad. Sci. USA* 103 (2006), 6506–6511.

[98.] Kruithof, M., F.-T. Chien, A. Routh, C. Logie, D. Rhodes, and J. van Noort. "Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber." *Nature Struct. Mol. Biol.* 16 (2009), 534–40.

[99.] Schram, R. D., H. Klinker, P. B. Becker, and H. Schiessel. "Computational study of remodeling in a nucleosomal array." *Eur. Phys. J. E* 38 (2015), 85.

[100.] Panyukov, S. and Y. Rabin. "Fluctuating filaments: statistical mechanics of helices." *Phys. Rev. E* 62 (2000), 7135–46.

[101.] Kessler, D. a. and Y. Rabin. "Stretching instability of helical springs." *Phys. Rev. Lett.* 90 (2003), 024301.

[102.] Wada, H. and R. R. Netz. "Stretching helical nano-springs at finite temperature." *EPL* 77 (2007), 68001.

[103.] Benetatos, P. and E. M. Terentjev. "Stretching weakly bending filaments with spontaneous curvature in two dimensions." *Phys. Rev. E* 81 (2010), 031802.

[104.] Ben-Haïm, E., A. Lesne, and J. M. Victor. "Chromatin: a tunable spring at work inside chromosomes." *Phys. Rev. E* 64 (2001), 51921.

[105.] Marko, J. F. and E. D. Siggia. "Stretching DNA." *Macromolecules* 28 (1995), 8759–8770.

[106.] Wang, M. D., H. Yin, R. Landick, J. Gelles, and S. M. Block. "Stretching DNA with optical tweezers." *Biophys. J.* 72 (1997), 1335–1346.

[107.] Wu, H. M. and D. M. Crothers. "The locus of sequence-directed and protein-induced DNA bending." *Nature* 308 (1984), 509–513.

[108.] Love, A. E. H. *Treatise on mathematical theory of elasticity*. 2nd. Cambridge University Press, 1906.

[109.] Schiessel, H. "The physics of chromatin." *J. Phys. Cond. Mat.* 15 (2003), R699.

[110.] Schiessel, H. *Biophysics for Beginners: A Journey through the Cell Nucleus*. Pan Stanford, 2014.

[111.] Mergell, B., M. R. Ejtehadi, and R. Everaers. "Modeling DNA structure, elasticity, and deformations at the base-pair level." *Phys. Rev. E* 68 (2003), 15.

[112.] Odijk, T. "Stiff Chains and Filaments under Tension." *Macromolecules* 28 (1995), 7016–7018.

[113.] Tompitak, M., L. de Bruin, B. Eslami-Mossallam, and H. Schiessel. "Designing nucleosomal force sensors." *Phys. Rev. E* 95 (2017), 052402.

[114.] Trifonov, E. N. and J. L. Sussman. "The pitch of chromatin DNA is reflected in its nucleotide sequence." *Proc. Natl. Acad. Sci. USA* 77 (1980), 3816–3820.

[115.] Satchwell, S. C., H. R. Drew, and A. A. Travers. "Sequence periodicities in chicken nucleosome core DNA." *J. Mol. Biol.* 191 (1986), 659–675.

[116.] Itzkovitz, S. and U. Alon. "The genetic code is nearly optimal for allowing additional information within protein-coding sequences." *Genome Res.* 17 (2007), 405–412.

[117.] Cohanim, A. B. and T. E. Haran. "The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes." *Nucleic Acids Res.* 37 (2009), 6466–6476.

[118.] Itzkovitz, S., E. Hodis, and E. Segal. "Overlapping codes within protein-coding sequences." *Genome Res.* 20 (2010), 1582–1589.

[119.] González, S., A. García, E. Vázquez, R. Serrano, M. Sánchez, L. Quintales, and F. Antequera. "Nucleosomal signatures impose nucleosome positioning in coding and noncoding sequences in the genome." *Genome Res.* 26 (2016), 1532–1543.

[120.] Hall, M. A., A. Shundrovsky, L. Bai, R. M. Fulbright, J. T. Lis, and M. D. Wang. "High-resolution dynamic mapping of histone-DNA interactions in a nucleosome." *Nature Struct. Mol. Biol.* 16 (2009), 124–129.

[121.] Bondarenko, V. A., L. M. Steele, A. Újvári, D. A. Gaykalova, O. I. Kulaeva, Y. S. Polikanov, D. S. Luse, and V. M. Studitsky. "Nucleosomes Can Form a Polar Barrier to Transcript Elongation by RNA Polymerase II." *Mol. Cell* 24 (2006), 469–479.

[122.] Ramachandran, S., G. E. Zentner, and S. Henikoff. "Asymmetric nucleosomes flank promoters in the budding yeast genome." *Genome Res.* 25 (2015), 381–390.

[123.] Mihardja, S., A. J. Spakowitz, Y. Zhang, and C. Bustamante. "Effect of force on mononucleosomal dynamics." *Proc. Natl. Acad. Sci. USA* 103 (2006), 15871–15876.

[124.] Sudhanshu, B., S. Mihardja, E. F. Koslover, S. Mehraeen, C. Bustamante, and A. J. Spakowitz. "Tension-dependent structural deformation alters single-molecule transition kinetics." *Proc. Natl. Acad. Sci. USA* 108 (2011), 1885–1890.

[125.] Mochrie, S. G. J., A. H. MacK, D. J. Schlingman, R. Collins, M. Kamenetska, and L. Regan. "Unwinding and rewinding the nucleosome inner turn: Force dependence of the kinetic rate constants." *Phys. Rev. E* 87 (2013), 012710.

[126.] Lequieu, J. P., A. Córdoba, D. C. Schwartz, and J. J. de Pablo. "Tension-Dependent Free Energies of Nucleosome Unwrapping." *ACS Cent. Sci.* 2 (2016), 660–666.

[127.] Durkin, S. G. and T. W. Glover. "Chromosome Fragile Sites." *Annual Review of Genetics* 41 (2007), 169–192.

[128.] Chan, K. L., T. Palmai-Pallag, S. Ying, and I. D. Hickson. "Replication stress induces sister-chromatid bridging at fragile site loci in mitosis." *Nature Cell Biol.* 11 (2009), 753–760.

[129.] Biebricher, A. et al. "PICH: A DNA Translocase Specially Adapted for Processing Anaphase Bridge DNA." *Mol. Cell* 51 (2013), 691–701.

[130.] Brogaard, K., L. Xi, J.-P. Wang, and J. Widom. "A map of nucleosome positions in yeast at base-pair resolution." *Nature* 486 (2012), 496–501.

[131.] Vaillant, C., B. Audit, and A. Arneodo. "Experiments confirm the influence of genome long-range correlations on nucleosome positioning." *Phys. Rev. Lett.* 99 (2007).

[132.] Barozzi, I., M. Simonatto, S. Bonifacio, L. Yang, R. Rohs, S. Ghisletti, and G. Natoli. "Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers." *Mol. Cell* 54 (2014), 844–857.

[133.] Valouev, A., S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow. "Determinants of nucleosome organization in primary human cells." *Nature* 474 (2011), 516–20.

[134.] Stoltenburg, R., C. Reinemann, and B. Strehlitz. "SELEX — A (r)evolutionary method to generate high-affinity nucleic acid ligands." *Biomol. Eng.* 24 (2007), 381–403.

[135.] Kahn, J. D. and D. M. Crothers. "Protein-induced bending and DNA cyclization." *Biochem.* 89 (1992), 6343–6347.

[136.] Parvin, J. D., R. J. McCormick, P. A. Sharp, and D. E. Fisher. "Prebending of a promoter sequence enhances affinity for the TATA-binding factor." *Nature* 373 (1995), 724–727.

[137.] Schätz, T. and J. Langowski. "Curvature and sequence analysis of eukaryotic promoters." *J. Biomol. Struct. Dyn.* 15 (1997), 265–275.

[138.] Davis, N. A., S. S. Majee, and J. D. Kahn. "TATA box DNA deformation with and without the TATA box-binding protein." *J. Mol. Biol.* 291 (1999), 249–265.

[139.] Bracco, L., D. Kotlarz, A. Kolb, S. Diekmann, and H. Buc. "Synthetic curved DNA sequences can act as transcriptional activators in Escherichia coli." *EMBO J.* 8 (1989), 4289–96.

[140.] Gartenberg, M. R. and D. M. Crothers. "Synthetic DNA bending sequences increase the rate of in vitro transcription initiation at the Escherichia coli lac promoter." *J. Mol. Biol.* 219 (1991), 217–230.

[141.] Pérez-Martín, J. and V. de Lorenzo. "Clues and consequences of DNA bending in transcription." *Annual Review of Microbiology* 51 (1997), 593–628.

[142.] Yamada, H., S. Muramatsu, and T. Mizuno. "An Escherichia coli protein that preferentially binds to sharply curved DNA." *J. Biochem.* 108 (1990), 420–425.

[143.] Prosseda, G., M. Falconi, M. Giangrossi, C. O. Gualerzi, G. Micheli, and B. Colonna. "The virF promoter in Shigella: More than just a curved DNA stretch." *Molecular Microbiology* 51 (2004), 523–537.

[144.] Cloutier, T. E. and J. Widom. "DNA twisting flexibility and the formation of sharply looped protein-DNA complexes." *Proc. Natl. Acad. Sci. USA* 102 (2005), 3645–3650.

[145.] Wei, J., L. Czapla, M. A. Grosner, D. Swigon, and W. K. Olson. "DNA topology confers sequence specificity to nonspecific architectural proteins." *Proc. Natl. Acad. Sci. USA* 111 (2014), 16742–16747.

[146.] Bailey, K. A., S. L. Pereira, J. Widom, and J. N. Reeve. "Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution." *J. Mol. Biol.* 303 (2000), 25–34.

[147.] Beutel, B. A. and L. Gold. "In vitro evolution of intrinsically bent DNA." *J. Mol. Biol.* 228 (1992), 803–812.

[148.] Singer, B. S., T. Shtatland, D. Brown, and L. Gold. "Libraries for genomic SELEX." *Nucleic Acids Res.* 25 (1997), 781–786.

[149.] Moyle-Heyrman, G., T. Zaichuk, L. Xi, Q. Zhang, O. C. Uhlenbeck, R. Holmgren, J. Widom, and J.-P. Wang. "Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning." *Proc. Natl. Acad. Sci. USA* 110 (2013), 20158–63.

[150.] Kulić, I. M., R. Thaokar, and H. Schiessel. "Twirling DNA Rings - Swimming Nanomotors Ready for a Kickstart." *EPL* 72 (2005), 527–533.

[151.] Kulić, I. M., R. Thaokar, and H. Schiessel. "A DNA ring acting as a thermal ratchet." *J. Phys. Cond. Mat.* 17 (2005), S3965–78.

[152.] Collings, C. K., A. G. Fernandez, C. G. Pitschka, T. B. Hawkins, and J. N. Anderson. "Oligonucleotide sequence motifs as nucleosome positioning signals." *PLoS ONE* 5 (2010), e10933.

[153.] Vologodskii, A. V., S. D. Levene, K. V. Klenin, M. Frank-Kamenetskii, and N. R. Cozzarelli. "Conformational and thermodynamic properties of supercoiled DNA." *J. Mol. Biol.* 227 (1992), 1224–1243.

[154.] Fathizadeh, A., H. Schiessel, and M. R. Ejtehadi. "Molecular Dynamics Simulation of Supercoiled DNA Rings." *Macromolecules* 48 (2015), 164–172.

[155.] Emanuel, M., G. Lanzani, and H. Schiessel. "Multiplectoneme phase of double-stranded DNA under torsion." *Phys. Rev. E* 88 (2013), 022706.

[156.] Balaeff, A., L. Mahadevan, and K. Schulten. "Modeling DNA loops using the theory of elasticity." *Phys. Rev. E* 73 (2006), 31919.

[157.] Kulić, I. M., H. Mohrbach, R. Thaokar, and H. Schiessel. "Equation of state of looped DNA." *Phys. Rev. E* 75 (2007), 011913.

[158.] Becker, P. B. and J. L. Workman. "Nucleosome remodeling and epigenetics." *Cold Spring Harb. Perspect. Biol.* 5 (2013), a017905.

[159.] Lorch, Y. and R. D. Kornberg. "Chromatin-remodeling and the initiation of transcription." *Quarterly Reviews of Biophysics* 48 (2015), 465–470.

[160.] Tolkunov, D. and A. V. Morozov. "Genomic studies and computational predictions of nucleosome positions and formation energies." *Advances in Protein Chemistry and Structural Biology* 79 (2010), 1–57.

[161.] Iyer, V. R. "Nucleosome positioning: Bringing order to the eukaryotic genome." *Trends in Cell Biology* 22 (2012), 250–256.

[162.] Teif, V. B. "Nucleosome positioning: resources and tools online." *Brief. Bioinform.* 17 (2015), 745–757.

[163.] Zhang, Z., C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh. "A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome." *Science* 332 (2011), 977–980.

[164.] Yuan, G. C., Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. "Genome-scale identification of nucleosome positions in S. cerevisiae." *Science* 309 (2005), 626–30.

[165.] Albert, I., T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, and B. F. Pugh. "Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome." *Nature* 446 (2007), 572–576.

[166.] Lee, W., D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow. "A high-resolution atlas of nucleosome occupancy in yeast." *Nature Genetics* 39 (2007), 1235–1244.

[167.] Shivaswamy, S., A. Bhinge, Y. Zhao, S. Jones, M. Hirst, and V. R. Iyer. "Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation." *PLoS Biol.* 6 (2008), 618–630.

[168.] Curran, K. a., N. C. Crook, A. S. Karim, A. Gupta, A. M. Wagman, and H. S. Alper. "Design of synthetic yeast promoters via tuning of nucleosome architecture." *Nature Comm.* 5 (2014), 4002.

[169.] Lantermann, A. B., T. Straub, A. Strålfors, G. C. Yuan, K. Ekwall, and P. Korber. "Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae." *Nature Struct. Mol. Biol.* 17 (2010), 251–257.

[170.] Tsankov, A. M., D. A. Thompson, A. Socha, A. Regev, and O. J. Rando. "The role of nucleosome positioning in the evolution of gene regulation." *PLoS Biol.* 8 (2010).

[171.] Valouev, A. et al. "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning." *Genome Res.* 18 (2008), 1051–1063.

[172.] Ercan, S., Y. Lubling, E. Segal, and J. D. Lieb. "High nucleosome occupancy is encoded at X-linked gene promoters in C. elegans." *Genome Res.* 21 (2011), 237–244.

[173.] Bunnik, E. M., A. Polishko, J. Prudhomme, N. Ponts, S. S. Gill, S. Lonardi, and K. G. Le Roch. "DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite Plasmodium falciparum." *BMC Genomics* 15 (2014), 347.

[174.] Mavrich, T. N. et al. "Nucleosome organization in the Drosophila genome." *Nature* 453 (2008), 358–362.

[175.] Zhang, Y., N. L. Vastenhouw, J. Feng, K. Fu, C. Wang, Y. Ge, A. Pauli, P. Van Hummelen, A. F. Schier, and X. S. Liu. "Canonical nucleosome organization at promoters forms during genome activation." *Genome Res.* 24 (2014), 260–266.

[176.] Liu, M.-j., A. E. Seddon, Z. T.-y. Tsai, I. T. Major, M. Floer, G. A. Howe, and S.-h. Shiu. "Determinants of nucleosome positioning and their influence on plant gene expression." *Genome Res.* 25 (2015), 1182–1195.

[177.] Teif, V. B., Y. Vainshtein, M. Caudron-Herger, J.-P. Mallm, C. Marth, T. Höfer, and K. Rippe. "Genome-wide nucleosome positioning during embryonic stem cell development." *Nature Struct. Mol. Biol.* 19 (2012), 1185–1192.

[178.] Fenouil, R. et al. "CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters." *Genome Res.* 22 (2012), 2399–2408.

[179.] Voong, L. N., L. Xi, A. C. Sebeson, B. Xiong, J.-P. Z. Wang, and X. Wang. "Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping." *Cell* 167 (2016), 1555–1570.

[180.] Ozsolak, F., J. S. Song, X. S. Liu, and D. E. Fisher. "High-throughput mapping of the chromatin structure of human promoters." *Nature Biotechnol.* 25 (2007), 244–248.

[181.] Schones, D. E., K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. "Dynamic regulation of nucleosome positioning in the human genome." *Cell* 132 (2008), 887–98.

[182.] Tillo, D., N. Kaplan, I. K. Moore, Y. N. Fondufe-Mittendorf, A. J. Gossett, Y. Field, J. D. Lieb, J. Widom, E. Segal, and T. R. Hughes. "High nucleosome occupancy is encoded at human regulatory sequences." *PLoS ONE* 5 (2010), e9129.

[183.] Gaffney, D. J., G. McVicker, A. A. Pai, Y. N. Fondufe-Mittendorf, N. Lewellen, K. Michelini, J. Widom, Y. Gilad, and J. K. Pritchard. "Controls of Nucleosome Positioning in the Human Genome." *PLoS Genetics* 8 (2012), 1–13.

[184.] Kersey, P. J. et al. "Ensembl Genomes 2016: more genomes, more complexity." *Nucleic Acids Res.* 44 (2015), D574–80.

[185.] *Data related to Valouev et al. (2011) [133].* `http://ccg.vital-it.ch/mga/hg18/valouev11/valouev11.html`.

[186.] Locke, G., D. Haberman, S. M. Johnson, and A. V. Morozov. "Global remodeling of nucleosome positions in C. elegans." *BMC Genomics* 14 (2013), 284.

[187.] *Data related to Locke et al. (2013) [186].* `http://nucleosome.rutgers.edu/nucenergen/celegansnuc/`.

[188.] David, L., W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. "A high-resolution map of transcription in the yeast genome." *Proc. Natl. Acad. Sci. USA* 103 (2006), 5320–5325.

[189.] Vaillant, C., L. Palmeira, G. Chevereau, B. Audit, Y. D'Aubenton-Carafa, C. Thermes, and A. Arneodo. "A novel strategy of transcription regulation by intragenic nucleosome ordering." *Genome Res.* 20 (2010), 59–67.

[190.] Chevereau, G., L. Palmeira, C. Thermes, A. Arneodo, and C. Vaillant. "Thermodynamics of intragenic nucleosome ordering." *Phys. Rev. Lett.* 103 (2009).

[191.] Percus, J. K. "Equilibrium state of a classical fluid of hard rods in an external field." *J. Stat. Phys.* 15 (1976), 505–511.

[192.] Vanderlick, T. K., L. E. Scriven, and H. T. Davis. "Solution of Percuss equation for the density of hard rods in an external field." *Phys. Rev. A* 34 (1986), 5130–5131.

[193.] Valentine, J. W., A. G. Collins, and C. P. Meyer. "Morphological Complexity Increase in Metazoans." *Paleobiology* 20 (1994), 131–142.

SUMMARY

DNA, a foundational part of all life on earth that carries genetic information, is a long, chainlike molecule: a polymer. However, it is a special type of polymer, in that its constituent monomers (the nucleotides) are not all identical. They come in four varieties, generally denoted by the letters A, T, C and G. These distinct monomers are what allows DNA to encode information in the first place: the letters play a role similar to that of the ones and zeros of the binary system that our digital devices use.

However, the distinction is not purely information-theoretical. The four possible nucleotides are necessarily distinct objects, with different physical and chemical properties. Therefore, a difference in DNA sequence is not just a difference in the encoded information, it is also a difference in flexibility, intrinsic curvature, and other elastic properties of the molecule. Such differences can have far-reaching effects.

DNA regularly needs to be deformed in nature. Most importantly, it is tightly bent in order to fit into a cell. The human genome, for instance, is about two meters long. The only reason a full copy of it fits into every single tiny cell in our body is because it is ingeniously folded up. Some parts of the DNA are energetically easier to fold than others, due to the variations in elastic properties, and as a result the sequences encoded in the DNA influence this folding. We thus find ourselves with a very rich physical system, in which the information carried by the DNA is intimately connected with its physical behavior.

DNA folding is just one example of such an interplay between the information and physics associated with DNA (although it is likely the most important one). There are two main questions to ask in the broader context. First, since DNA sequence and physical behavior are so intimately linked, does nature make use of this link? It seems reasonable to expect that evolution would also explore DNA sequences based on their physical properties, if there is some benefit to be derived from it. Secondly, how can we exploit this link ourselves? Through thoughtful manipulation of sequences, what kind of properties and behavior can we bestow upon DNA?

The five projects described in this thesis are all attempts to further our grasp on these two questions. In the first project, described in Chapter 2, we take a look at how far the intrinsic curvature of DNA can be pushed.

We design DNA sequences that lead to molecules with very strong, directionally coherent curvature, such that they form superhelical structures of their own volition. Such superhelical structures look a lot like springs, and they behave similarly, being significantly more resistant to an external force than we would expect of generic DNA.

In Chapter 3, we turn to the DNA system that receives the widest interest in the literature: the nucleosome. Folding DNA into tiny cells is assisted (in eukaryotic organisms, e.g. animals, plants and fungi) by little protein cylinders around which the DNA is wrapped. The resulting DNA-protein complex is called a nucleosome. Due to the significant bending required to wrap DNA into this complex, the affinity of a piece of DNA for the nucleosome is strongly dependent on its sequence. In Chapter 3 we see how we can go beyond a simple, scalar sequence property like nucleosome affinity.

When pulling on the two ends of a piece of DNA that is wrapped around a protein cylinder, we of course expect the DNA to be peeled away. However, it turns out that, due to the geometry of that peeling process, nucleosomes are actually kinetically protected from unwrapping due to tension: there is an energetic barrier that opposes this unwrapping, and this barrier actually becomes higher, the harder you pull. The result is that unwrapping a nucleosome requires a significant amount of force, which is good because we do not in general want the nucleosomes in our cells to fall off every time they feel some tension.

It has been shown that the way in which nucleosomes unwrap, not unexpectedly, is sequence-dependent. Specifically, we know that they can be made to unwrap asymmetrically, meaning the DNA first peels away from one side only, if the DNA sequence has better nucleosome affinity in one half than in the other.

In Chapter 3 we try to push this idea further, and we design nucleosomes with a hole in their unwrapping barrier. The result is that we can make nucleosomes that are not strongly kinetically protected, and we can make them unwrap via specific pathways of our choosing. The fact that this is possible demonstrates that the nucleosome, much like the DNA polymer, should not be considered as a single complex, but rather as a class of systems, and one nucleosome can have vastly different behavior from another.

In Chapters 4 and 5 we introduce some novel methodology. Much of the work in this thesis is built upon the Mutation Monte Carlo (MMC) method invented by Behrouz Eslami-Mossallam. The idea behind this methodology is as simple as it is powerful: take a standard physical Monte Carlo

simulation of a DNA system, and add mutations to the mix. By allowing the Monte Carlo simulation to simultaneously search the conformational space of the system and its space of possible sequences, it automatically converges on sequences that have high affinity for the system, and provides us with statistics on those sequences. In Chapters 4 and 5 we expand and enrich this methodology.

Chapter 4 first shows how the MMC method can be used to generate a bioinformatical model that approximates the sequence-dependent affinity of DNA for the system being simulated with a physical model. This new model grants us significantly enhanced reach, because its approximative nature is offset by a vast gain in computational cost. It allows us to tackle problems not remotely tractable with a detailed biophysical model like the nucleosome model of Chapter 3.

The rest of Chapter 4 is dedicated to benchmarking the new model and investigating what is needed to render it accurate. In Chapter 5 we investigate in more detail the relationship between the new method of Chapter 4 and the MMC methodology. We find that it bridges the gap between MMC simulations and SELEX experiments (a type of sequence selection experiment in which sequences are selected based on their affinity to a target, such as the nucleosome, similar in many ways to the MMC method). Along the way, we gain deeper insight into how the MMC method works, and especially into the role played by temperature in an MMC simulation.

In Chapter 6 we step into the biological realm, where we put the new model from Chapter 4 to use. Thanks to the vast gain in computational efficiency, we are able to perform genome-wide analyses of real biological DNA sequences. We focus on the promoter regions of genes of a range of organisms. These are parts of the genome, in front of the genes that contain the genetic information, that influence how often the gene is utilized in a cell. An interesting role is played here by the elastic properties of the DNA sequence, and specifically by its affinity for nucleosomes. The reason is that DNA wrapped into nucleosomes cannot be read out; the nucleosomes interfere with other machinery trying to bind to the DNA. A high or low affinity in the region where the machinery for reading out genes wants to bind therefore directly influences the expression of the gene.

Real genomes are known to encode mechanical signals around such binding sites. Yeast, for instance, a simple, unicellular organism, has DNA sequences in these regions that have poor affinity for the nucleosome, in an attempt to keep the DNA accessible for reading. The human genome, on the other hand, contains the opposite signals: sequences with high

affinity, that keep nucleosomes strongly bound at these positions. This is thought to help the genome retain some of its nucleosomes in sperm cells, in which most of the nucleosomes are removed from the DNA.

Analyzing the mechanical signals in promoter regions of around 50 organisms from across the tree of life, we find a fascinating universality: unicellular organisms are all similar to yeast, and contain signals to keep nucleosomes away, while all multicellular organisms contain signals to keep nucleosomes strongly bound. Furthermore, the strength of the signals in the latter case correlates with the complexity of the organism: mammals have sequences with very good affinity, while some simpler animals like fruit flies show more moderate signals.

Whether all these signals, and this universally observed distinction between unicellular and multicellular life, serve the same purposes they are thought to do in yeast and in humans remains to be seen.

Through these inquiries and findings, the research described in this thesis has attempted to further our understanding of both of the questions posed at the beginning of this summary. We have looked into the mechanical signals that can be found in real genomes; we have pushed the limits of the properties that DNA sequences can be made to display; and we have provided new methodology that will allow us to inquire ever further into the possibilities presented by sequence-dependent DNA mechanics.

SAMENVATTING

DNA, als de drager van genetische informatie een fundamenteel onderdeel van al het leven op aarde, is een lang, kettingvormig molecuul: een polymeer. Het is echter een speciaal soort polymeer, omdat de monomeren waar DNA uit bestaat (de nucleotiden) niet allemaal identiek zijn. Er zijn vier soorten nucleotiden, meestal aangeduid met A, T, C en G. Het onderscheid tussen deze monomeren is ook wat het coderen van informatie mogelijk maakt: de letters spelen een rol analoog aan de nullen en enen van het binair systeem dat onze digitale apparaten gebruiken.

Het onderscheid is echter niet puur informatie-theoretisch. De vier nucleotiden zijn noodzakelijkerwijs niet-identieke objecten, met verschillende fysische en chemische eigenschappen. Daarom is een verschil in DNA-sequentie niet alleen een verschil in gecodeerde informatie, maar ook een verschil in flexibiliteit, intrinsieke buiging, en andere elastische eigenschappen van het molecuul. Dergelijke verschillen kunnen verstrekkende gevolgen hebben.

DNA moet regelmatig worden vervormd. Het voornaamste voorbeeld is dat DNA sterk moet worden gebogen om in een cel te passen. Het menselijk genoom, bijvoorbeeld, is zo'n twee meter lang, en past dan ook alleen maar in onze kleine cellen omdat het op ingenieuze wijze is opgevouwen. Sommige delen van het DNA zijn makkelijker op te vouwen dan andere, vanwege de variaties in de elastische eigenschappen. Het gevolg is dat de DNA-sequenties het opvouwen van het genoom beïnvloeden. Dit leidt tot een systeem met een rijke fysica, waarbij de informatie die in het DNA ligt opgeslagen nauw is verbonden met hoe het zich gedraagt.

Het compactificeren van DNA is slechts één voorbeeld van zulk samenspel tussen de informatie in en de fysica van DNA (maar waarschijnlijk het meest belangrijke voorbeeld). In bredere context zijn er twee belangrijke vragen om te stellen. Ten eerste: maakt de natuur gebruik van de nauwe verbintenis tussen DNA-sequentie en fysiek gedrag? Het ligt voor de hand dat het evolutieproces DNA-sequenties ook selecteert op gunstige fysische eigenschappen, als daarmee enig voordeel te behalen valt. Ten tweede: hoe kunnen we deze link gebruiken om DNA te manipuleren? Wat voor eigenschappen en gedrag kunnen we een DNA-molecuul meegeven door de sequentie te veranderen?

De vijf projecten beschreven in dit proefschrift zijn stuk voor stuk pogingen om onze grip op deze vraagstukken te versterken. In het eerste project, beschreven in Hoofdstuk 2, zoeken we de grenzen op van de intrinsieke buiging van DNA-moleculen. We ontwerpen DNA-sequenties die leiden tot moleculen die sterk, steeds in dezelfde richting gebogen zijn, zodat deze moleculen van zichzelf superhelische vormen aannemen. Dergelijke superhelische structuren zien eruit als kleine veren, en gedragen zich ook soortgelijk: ze bieden een sterkere weerstand tegen een externe trekkracht dan we zouden verwachten van DNA.

In Hoofdstuk 3 wenden we ons tot het DNA-systeem dat zowel in de rest van dit proefschrift, als in de literatuur, de meeste aandacht krijgt: het nucleosoom. Het opvouwen van het DNA, zodat het in cellen past, gebeurt (in eukaryoten, e.g. dieren, planten en schimmels) met behulp van kleine eiwitcilinders waar het DNA omheen gewikkeld wordt. Het resulterende DNA-eiwitcomplex noemen we een nucleosoom. Omdat DNA sterk moet worden gebogen om dit complex te vormen, hangt de affiniteit van een stuk DNA voor het nucleosoom sterk af van de sequentie. Er is en wordt veel onderzoek verricht naar de sequentie-afhankelijke affiniteit van DNA voor het nucleosoom, hoe deze affiniteit de organisatie van een genoom beïnvloedt en in hoeverre deze invloed van belang is in levende cellen. In Hoofdstuk 3 proberen we dieper te kijken dan een simpele, scalaire eigenschap zoals algehele affiniteit.

Wanneer we aan de uiteinden van een stuk DNA trekken dat om een eiwitcilinder gewonden is, verwachten we uiteraard dat we het DNA los zullen trekken. Echter blijkt dat, vanwege de geometrie van het proces, nucleosomen kinetisch beschermd zijn tegen het geforceerd afwikkelen: er bestaat een energetische barrière, en deze barrière wordt des te hoger, naarmate de kracht toeneemt. Het resultaat is dat het lostrekken van DNA een significante hoeveelheid kracht vereist, hetgeen van pas komt omdat we over het algemeen niet willen dat de nucleosomen in onze cellen uit elkaar vallen zodra er aan het DNA getrokken wordt.

Niet geheel onverwacht is de manier waarop nucleosomen afwikkelen afhankelijk van de DNA-sequentie. We weten dat nucleosomen soms asymmetrisch afwikkelen, waarmee we bedoelen dat één van de uiteinden eerder loskomt dan het andere, doordat de DNA-sequentie aan dat uiteinde minder grote affiniteit voor het nucleosoom heeft.

In Hoofdstuk 3 proberen we dit idee verder te voeren, en ontwerpen we nucleosomen met een gat in de barrière tegen het afwikkelen. Het resultaat is dat we nucleosomen kunnen maken die niet sterk kinetisch beschermd zijn, en dat we ze via specifieke paden kunnen laten afwikkelen.

Het feit dat dit mogelijk is demonstreert dat de term 'nucleosoom', net als 'DNA', niet refereert aan een enkel systeem, maar aan een hele klasse van systemen, en dat nucleosomen zeer verschillend gedrag kunnen vertonen.

In Hoofdstukken 4 and 5 introduceren we nieuwe methodologie. Veel van het werk in dit proefschrift bouwt voort op de Mutation Monte Carlo (MMC) methode van Behrouz Eslami-Mossallam. Het idee achter deze methode is even eenvoudig als krachtig: neem een Monte Carlo-simulatie van een DNA-systeem, en voeg er mutaties aan toe. Door de simulatie tegelijkertijd zowel de fysische configuraties van het systeem, als de ruimte van mogelijke DNA-sequenties te laten doorzoeken, convergeert hij automatisch naar sequenties die hoge affiniteit voor het systeem hebben, en geeft ons de statistische eigenschappen van die sequenties. In Hoofdstukken 4 and 5 breiden we deze methodologie uit.

In Hoofdstuk 4 laten we eerst zien hoe MMC kan worden gebruikt om een bioinformatisch model te genereren dat de sequentie-afhankelijke affiniteit van DNA benadert voor het systeem waarvoor we een fysisch model simuleren. Dit nieuwe bioinformatische model geeft onze methoden significant meer bereik, omdat het feit dat het een benadering is, wordt gecompenseerd door een grote besparing in computationele complexiteit. Dit stelt ons in staat om vraagstukken onder de loep te nemen die geenszins te behappen zijn met een gedetailleerd biofysisch model zoals het nucleosoommodel uit Hoofdstuk 3. De rest van Hoofdstuk 4 is gewijd aan het benchmarken van het model, en het onderzoeken van wat de voorwaarden zijn voor een zo nauwkeurig mogelijke benadering.

In Hoofdstuk 5 onderzoeken we in meer detail de relatie tussen de nieuwe methode uit Hoofdstuk 4 en de MMC-methode. We zien dat onze nieuwe methode de kloof dicht tussen MMC-simulaties en SELEX-experimenten (een klasse van experimenten waarin sequenties worden geselecteerd op hun affiniteit voor een gegeven systeem, zoals het nucleosoom; de experimentele methodologie is in veel opzichten vergelijkbaar met de computationele MMC-methode). Ook krijgen we dieper inzicht in hoe de MMC-methode werkt, bovenal in de rol van de temperatuur in een MMC-simulatie.

In Hoofdstuk 6 passen we het nieuwe model van Hoofdstuk 4 toe op biologische data. Dankzij de grote winst in computationele efficiëntie die we met dit model boeken, kunnen we gehele genomen analyseren. We richten onze blik op de promotoren van de genen van verschillende organismen. Dit zijn de delen van een genoom, die zich voor de genen bevinden, en die invloed uitoefenen op de mate waarin een gen tot expressie komt. Hierin is een interssante rol weggelegd voor de elastische eigenschappen van de

betreffende DNA-sequenties, en specifiek voor hun affiniteit voor nucleosomen. Dit omdat DNA dat in een nucleosoom is gewikkeld, niet kan worden uitgelezen; het nucleosoom zit andere systemen die aan het DNA willen binden in de weg. Een grote of kleine affiniteit in de regio waar de machinerie die de genen uitleest wil binden heeft zo direct invloed op hoe vaak een gen wordt gelezen.

Het is bekend dat echte genomen in de DNA-sequenties van promotoren voor mechanische signalen coderen. Gist, een simpel ééncellig organisme, heeft in deze regio's bijvoorbeeld DNA-sequenties met een lage affiniteit voor nucleosomen, om het DNA toegankelijk te houden. Het menselijk genoom heeft juist signalen die nucleosomen sterk aantrekken. Men denkt dat dit het genoom in staat stelt in deze regio's nucleosomen te behouden in zaadcellen, waarin de meeste nucleosomen van het genoom worden verwijderd.

Bij het analyseren van deze signalen in de promotoren van zo'n 50 verschillende organismen uit verschillende takken van de fylogenetische stamboom vinden we een opmerkelijk universele overeenkomst: ééncellige organismen lijken allemaal op gist, en hebben signalen die nucleosomen afstoten, terwijl meercellige organismen allemaal signalen bevatten die nucleosomen aantrekken. Daarnaast bestaat er in het geval van meercellige organismen een correlatie tussen hoe sterk deze signalen zijn, en hoe complex het organisme is: zoogdieren hebben DNA-sequenties met zeer hoge affiniteit voor nucleosomen, terwijl simpelere dieren, zoals fruitvliegjes, zwakkere signalen vertonen. Of al deze signalen, en de strakke scheiding die we zien tussen ééncelligen en meercelligen, in alle gevallen dezelfde functies hebben als in gist en in mensen, zal nog verder moeten worden onderzocht.

Het beschreven onderzoek en de bijbehorende bevindingen pogen antwoorden te verschaffen op de twee vragen die we aan het begin van deze samenvatting stelden. We hebben gekeken naar mechanische signalen die in echte genomen te vinden zijn; we hebben de grenzen opgezocht van de eigenschappen die we via de sequentie aan een DNA-systeem kunnen meegeven; en we hebben nieuwe methoden aangedragen die ons in staat stellen om de mogelijkheden die de sequentie-afhankelijke mechanische eigenschappen van DNA ons bieden nog verder te onderzoeken.

CURRICULUM VITAE

I was born on the 29th of October, 1989, in Bangkok, Thailand. Having spent only the first year of my life there, I grew up in Alkmaar in The Netherlands. Fascinated by physics in school, I studied Physics and Astronomy and later Theoretical Physics at the Vrije Universiteit Amsterdam, obtaining both the BSc and the MSc degree *cum laude*.

During my Master's, I performed my research training in the Gravitational Waves group at Nikhef, Amsterdam, where I worked on the data analysis pipeline of the LIGO/Virgo collaboration, and examined to what extent we might use gravitational waves to find out what neutron stars look like on the inside. I also performed historical research into the community of teachers of mathematics and physics in Dutch secondary education in the early 20th century, and how their clashes and their growing apart reflected the diverging interests of mathematicians and physicists more broadly.

After obtaining my MSc degree, Helmut Schiessel of Leiden University convinced me to make the switch to biophysics. Over the four years of my PhD project, I worked in his group on various problems pertaining to the interplay between the information that is stored in DNA and the physical properties of the resulting molecules, and how this affects the behavior of DNA in various contexts.

During my PhD, I also served as chairman of the board of the Leids Promovendi Overleg (LEO[1]), the association of PhD students in Leiden, for two years. We organised a range of events for our fellow PhDs and represented them in various ways, within the university and beyond.

1 The reason why Leids Promovendi Overleg is shortened to LEO has unfortunately been lost to the sands of time.

PUBLICATIONS

[1.] Eslami-Mossallam, B., R. D. Schram, **M. Tompitak**, J. van Noort, and H. Schiessel. "Multiplexing Genetic and Nucleosome Positioning Codes: A Computational Approach." *PLoS ONE* 11 (2016), e0156905.

[2.] Bruin, L. de, **M. Tompitak**, B. Eslami-Mossallam, and H. Schiessel. "Why Do Nucleosomes Unwrap Asymmetrically?" *J. Phys. Chem. B* 120 (2016), 5855–5863.

[3.] **Tompitak, M.**, H. Schiessel, and G. T. Barkema. "Force responses of strongly intrinsically curved DNA helices deviate from worm-like chain predictions." *EPL* 116 (2016), 68005.

[4.] **Tompitak, M.**, C. Vaillant, and H. Schiessel. "Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions." *Biophys. J.* 112 (2017), 505–511.

[5.] **Tompitak, M.**, G. T. Barkema, and H. Schiessel. "Benchmarking and refining probability-based models for nucleosome-DNA interaction." *BMC Bioinformatics* 18 (2017), 157.

[6.] **Tompitak, M.**, L. de Bruin, B. Eslami-Mossallam, and H. Schiessel. "Designing nucleosomal force sensors." *Phys. Rev. E* 95 (2017), 052402.

[7.] Wondergem, J. A. J., H. Schiessel, and **M. Tompitak**. "Performing SELEX experiments in silico." submitted (2017).

[8.] Culkin, J., L. de Bruin, **M. Tompitak**, and H. Schiessel. "Sequence dependence in nucleosome breathing." submitted (2017).

Including outreach to a wider audience:

[9.] **Tompitak, M.** and H. Schiessel. "Mechanische informatie in DNA-moleculen schrijven en lezen." *Nederlands Tijdschrift voor Natuurkunde* 83 (2017), 164.

ACKNOWLEDGEMENTS

My thanks must first and foremost be rendered to Helmut Schiessel, whose supervision has underpinned all the achievements that make up my PhD project. Without his patient assistance, his trust in my abilities, and the pleasant working environment he creates around him, I would not have gotten as far, nor enjoyed it as much. Secondly, I must thank my second promotor, Gerard Barkema, who not only lent his invaluable computational expertise to my work, but who also kept me on my toes by asking the hard questions, and who made sure I kept sight of my goals. Special thanks is due to Behrouz Eslami-Mossallam, who patiently introduced me to the field as I was getting started, and without whose Mutation Monte Carlo method most of the work presented in this thesis would not have been possible.

I must thank Cédric Vaillant, Sung Hyun Kim and Cees Dekker, with whom we collaborated on various topics, as well as the students whose projects I helped supervise: Lennart de Bruin, Jamie Culkin, Joeri Wondergem and Martijn Zuiddam. I probably learned more than I was able to teach. Artur Kaczmarczyk, Patrick van den Berg, Martín Caldarola, Bram Henneman and Remus Dame deserve to be mentioned for inspirational discussions and questions.

My personal thanks go out to my friends and colleagues at the Lorentz Institute, as well as those in the Physics of Life Processes department and other parts of the university. Fran and Marianne deserve a special thank you for keeping our institute running smoothly, and going above and beyond in their organizational activities.

I would also like to thank my friends at LEO, the Leiden PhD association, where I served two years as chairman, allowing me to assist the scientific endeavor from a different angle. Nothing we accomplished in those two years would have happened without the efforts of my fellow volunteers: Aïcha, Aleksandrina, Aquiles, Artur, Candido, Chris, Jens, Julia, Sayo, Soumya and Zohreh.

Finally, I must thank my mother, who always strongly encouraged my love of learning, without which I would not have reached this milestone.