STATISTICAL MECHANICS OF NUCLEOSOMES

BY RĂZVAN V. CHEREJI

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Physics and Astronomy

Written under the direction of
Prof. Alexandre V. Morozov
and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

Statistical Mechanics of Nucleosomes

by Răzvan V. Chereji

Dissertation Director: Prof. Alexandre V. Morozov

Eukaryotic cells contain long DNA molecules (about two meters for a human cell) which are tightly packed inside the micrometric nuclei. Nucleosomes are the basic packaging unit of the DNA which allows this millionfold compactification. A long-standing puzzle is to understand the principles which allow cells to both organize their genomes into chromatin fibers in the crowded space of their nuclei, and also to keep the DNA accessible to many factors and enzymes. With the nucleosomes covering about three quarters of the DNA, their positions are essential because these influence which genes can be regulated by the transcription factors and which cannot.

We study physical models which predict the genome-wide organization of the nucleosomes and also the relevant energies which dictate this organization. In the last five years, the study of chromatin knew many important advances. In particular, in the field of nucleosome positioning, new techniques of identifying nucleosomes and the competing DNA-binding factors appeared, as chemical mapping with hydroxyl radicals, ChIP-exo, among others, the resolution of the nucleosome maps increased by using paired-end sequencing, and the price of sequencing an entire genome decreased.

We present a rigorous statistical mechanics model which is able to explain the recent experimental results by taking into account nucleosome unwrapping, competition between different DNA-binding proteins, and both the interaction between histones and

DNA, and between neighboring histones. We show a series of predictions of our new model, all in agreement with the experimental observations.

Acknowledgements

I would like to express my gratitude to my advisor, Professor Alexandre Morozov, for his constant support and encouragement in my research. His insightful comments and guidance have been indispensable during the development of this dissertation. Many thanks for his patience while I was starting to learn Biophysics. Although Physics was my friend for a long time, I started to understand the beauty of Biology only after he introduced me to this field. I am grateful to the professors of my committee, Joel Lebowitz, James Broach, Anirvan Sengupta, Gyan Bhanot and Gerald Goldin. Their intuition and knowledge were a constant source of inspiration for me.

I am greatly indebted to Camelia Pop, a wonderful wife and mathematician. She has always been an example of dedication, hard work, kindness and modesty.

Many thanks to all my professors who had a contribution to my success. It has been a long journey, but they made it the most exciting one. For the last part of this journey, I would like to thank to my professors from Rutgers. Apart from the ones from my committee, there were also other professors who truly inspired me. Among them are Herbert Neuberger, Alexander Zamolodchikov, Duiliu Emanuel Diaconescu, Emil Yuzbashyan, and Kristjan Haule.

I thank my collaborators and friends, Natalia Petrenko and Nils Elfving, for their help with every Biology-related question that I had during the last years. Thanks to all my collaborators for allowing me to present parts of our joint work in my dissertation.

Thanks to my Romanian friends who made my stay at Rutgers much more pleasant: Liviu Ilinca, Vlad Vicol, Lucian Păşcuţ, Marius Beceanu, Adina Luican, Tiberiu Teşileanu, Silviu Pufu, Daniela and Tudor Prelipceanu. Even a short meeting with them, always brightened my day. I would also like to thank my fellow graduate students: George Locke, Allan Haldane, Ted Malliaris, Mohammad Ramezanali, Julie

Tsitron, Michael Manhart, Pavel Khromov, Aatish Bhatia and Manjul Apratim. They made my life at Rutgers easier.

Special thanks to my friend Diana Niţescu because she helped me when I needed it the most.

And last but not least, I thank my parents who carefully guided my first steps in school and always made all the necessary sacrifices for me to have everything that I ever needed. I love them and I dedicate my dissertation to them.

Dedication

To my family.

Table of Contents

Abstract ii

Acknowledgements iv

Dedication vi

1. Introduction 1

1.1. DNA and chromatin 1

1.2. Studies of nucleosome positioning 5

1.3. Our models of nucleosome positioning 16

2. Nucleosome positioning 20

2.1. Energetics of one-dimensional hard rods with nearest-neighbor interactions 20

2.2. Predicting two-body interactions from one-particle distribution 29

2.3. Sequence-specific energy of nucleosome formation 35

2.4. Applications 39

2.4.1. Reconstructing nucleosome energetics in a model system 39

2.4.2. Nucleosome localization by potential barriers and wells 43

2.4.3. Modeling nucleosome occupancy over transcribed regions 49

3. Nucleosome unwrapping 54

3.1. Experimental evidence of nucleosome unwrapping 55

3.2. Direct problem: Matrix solution 58

3.2.1. Single-type particles 59

3.2.2. Multiple-type particles 63

3.3. Direct problem: Recursive solution for hard-core interactions 65

3.3.1. General case 65

3.3.2. Special case: No unwrapping . . . 67
3.4. Inverse problem: Recursive solution for hard-core interactions . . . 67
3.4.1. General case . . . 68
3.4.2. Special case: No unwrapping . . . 69
3.4.3. Sequence-specific nucleosome formation energies . . . 69
3.5. Applications . . . 75
3.5.1. Nucleosome unwrapping potential . . . 75
3.5.2. Higher-order chromatin structure and linker histone energetics . 80
3.5.3. Alternative models of nucleosome unwrapping . . . 85
3.5.4. Genome-wide organization of nucleosome unwrapping states . . . 86
3.5.5. Accessibility of nucleosomal DNA to factor binding . . . 92
3.6. Nucleosome-induced cooperativity . . . 94
3.6.1. Sequence-dependent nucleosome positioning and unwrapping . . 94

4. Other joint projects . . . 100
4.1. Msn2 signaling . . . 100
4.2. Msn2-Mediator-nucleosome interplay . . . 103
4.3. Fragile nucleosomes . . . 107

5. Conclusions . . . 115

Bibliography . . . 116

References . . . 117

Appendix A. The z-transform formalism . . . 130
A.1. General method . . . 130
A.2. Applications: ideal gas, Tonks gas . . . 132
A.3. Comparison between lattice and continuous 1D fluids . . . 133

Appendix B. Alternative nucleosome unwrapping models . . . 136
B.1. Model A: Crystal structure augmented with an additional well . . . 137

B.2. Model B: Crystal structure augmented with a linear function 140
B.3. Model C: 10-bp oscillations superimposed onto a linear function 141
B.4. Model D: 11-bp oscillations superimposed onto a linear function 142
B.5. Model E: Uniform unwrapping 143
B.6. Model F: 5-bp oscillations superimposed onto a linear function 144
B.7. Model G: 5-bp stepwise unwrapping 145
B.8. Model H: 10-bp stepwise unwrapping 146
Vita 147

Chapter 1

Introduction

1.1 DNA and chromatin

Deoxyribonucleic acid (DNA) is a long polymer made of a sequence of nucleotide monomers. A nucleotide is composed of a sugar, one or more phosphate groups and a nitrogenous base. The base can be either a purine (adenine and guanine) or a pyrimidine (thymine and cytosine). DNA contains the genetic information used by all living organisms in order to function.

In eukaryotic cells, long DNA molecules must be packed within the microscopic space of the nucleus. The nucleus of every human cell has a diameter of a few microns and contains about two meters of DNA, which means a millionfold DNA compactification inside the nucleus. It is estimated that a human body has about 50 trillion cells, so that each of us has about 100 trillion meters of DNA inside our body. As a comparison, the distance between the Sun and the Earth is about 150 billion meters. Therefore, the DNA from our body is long enough to go from here to the Sun and back more than 300 times. It is also enough to circle the Earth's equator more than 2.5 million times [1]. So how is it possible to pack such long DNA molecules inside the tiny nuclei?

This packaging is realized with the help of certain proteins called histones. These are positively charged proteins, and so they bind strongly to the negatively charged DNA, which wraps around the histones and thus gets compactified. The packaging problem is very difficult. Stiff, long DNA molecules must fit inside the tiny space of the nucleus, while remaining accessible for various processes, as transcription, replication, repair, among others. The basic unit of DNA packaging is called the nucleosome. Nucleosomes are composed of 147 base-pairs (bps) of DNA wrapped around histone octamers in about two turns. The histone octamer contains two copies of the following

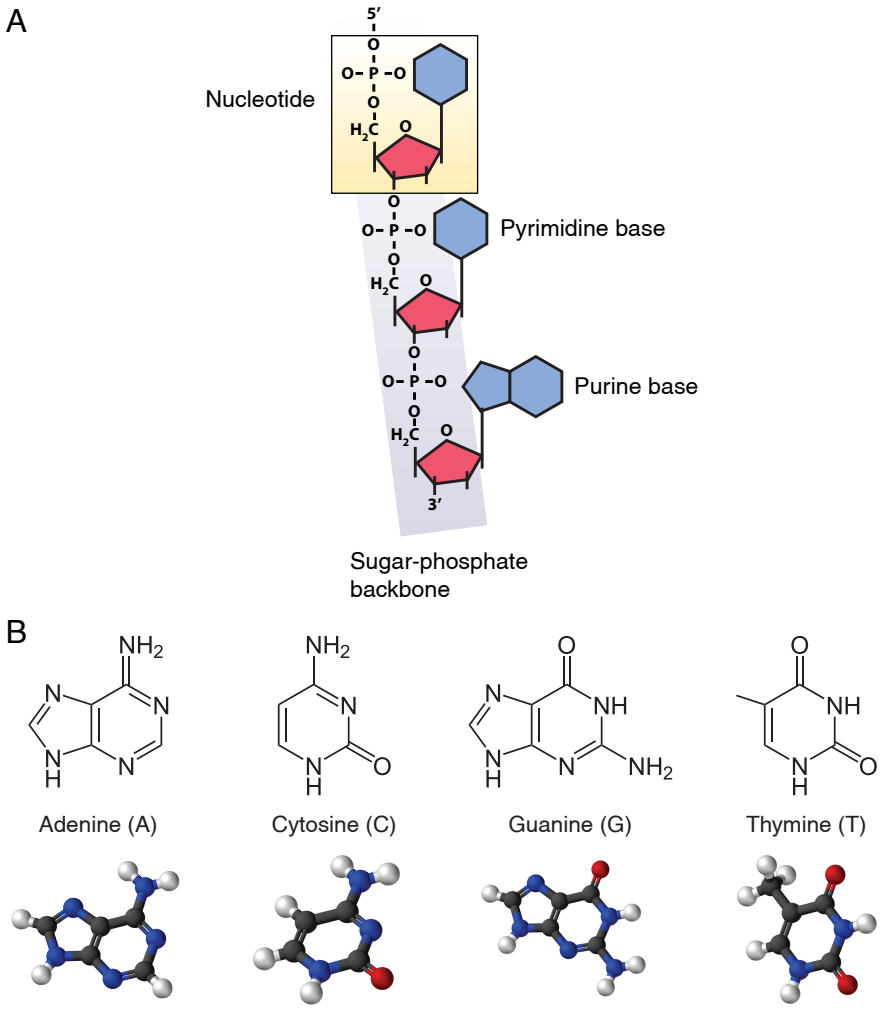


Figure 1.1: (A) DNA components: nucleotide monomer (orange area), sugar-phosphate backbone (violet area), bases (purine or pyrimidine). (B) Chemical structure of the four bases: adenine (A), cytosine (C), guanine (G) and thymine (T). For more details, see [3].

basic histones – H2A, H2B, H3 and H4 [2]. A short DNA stretch between neighboring nucleosomes is called linker DNA. There is a fifth type of histones, the linker histone H1, which binds to the edges of the nucleosomal DNA and stabilizes the nucleosome. Each histone protein has a flexible extension known as the N-terminal tail domain, which protrudes from the nucleosome surface.

Arrays of nucleosomes appear as "beads on a string" when imaged by electron microscopy (Figure 1.3), and observations of this structure were obtained four decades

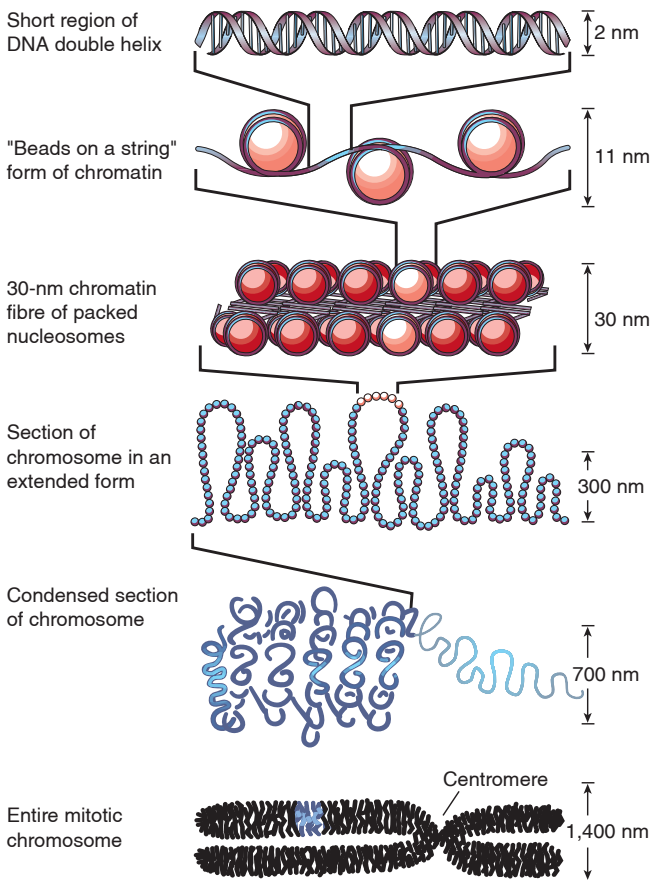


Figure 1.2: The multiple stages of DNA packaging inside the nucleus, with the corresponding length scales. Reprinted by permission from Macmillan Publishers Ltd: Nature Publishing Group, copyright (2003) [7].

ago [4, 5, 6]. Arrays of multiple nucleosomes coil together in a fiber of packed nucleosomes, known as chromatin. The chromatin fiber, which has a width of about 30 nanometers, loops and is further packaged, until this multiple folding allows the whole DNA molecule to fit inside the microscopic nucleus in each of our cells (Figure 1.2) [7]. The relationship between nucleosome positions and higher-order chromatin is not well known.

DNA packaging is not the only function of the nucleosomes. They also play a crucial role in controling DNA accessibility of many DNA-binding proteins to regulatory elements on the chromosomes. Indirectly, they influence gene expression regulation [8, 9, 10, 11, 12, 13], DNA replication [14, 15], DNA repair [14, 16], DNA recombination

[17], among others. For example, nucleosomes cause slowing down and pausing of RNA polymerase [18, 19]. Another function of the histones is that they add an epigenetic layer of information on top of the genome [20]. The histone tails can have many post-translational modifications (PTMs), and different histone variants can replace the core histones. The most studied PTMs are methylation and acetylation but there is a plethora of other possible modifications – phosphorylation, ubiquitination, sumoylation. These modifications can happen in different combinations on the residues in the N-terminal tails of histones.

Histones and the nucleosome units are not particular to humans, but are conserved among eukaryotes, from yeast to human. For this reason, it is useful and advantageous to study nucleosomes in simpler organisms, like yeast or flies. In *S. cerevisiae*, arrays of nucleosomes cover about 75% of the DNA, having a strong influence on gene regulation [8, 9, 10, 11, 12, 13]. Wrapped in nucleosomes, DNA is sterically occluded from interacting with many protein complexes, e.g. transcription factors (TFs), polymerases, recombinases and repair enzymes. Nucleosomal DNA needs to be accessible at times in order to allow different regulatory proteins to bind and conduct their biological functions. With three quarters of eukaryotic DNA being wrapped into nucleosomes, the question of how DNA-binding proteins gain access to their target sites, in living cells, is one of the puzzling problems that is still not well understood.

Access to binding sites that are buried in nucleosomes can be facilitated by ATP-dependent remodeling factors that can change the conformation of the chromatin by moving or disassembling nucleosomes [22, 23, 24, 25, 26]. How remodelers use ATP hydrolysis and convert the energy into work, necessary to move the nucleosomes, and how different remodelers pick the right nucleosomes to rearrange or disassemble them, is unknown. These ATP-dependent chromatin remodeling complexes are not always required. Some studies show that proteins can bind to their target sites even without the help from active remodeling factors [27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50].

The inter-genic regions are relatively depleted of stable nucleosomes compared to the intra-genic regions. The inter-genic regions contain unstable nucleosomes which are

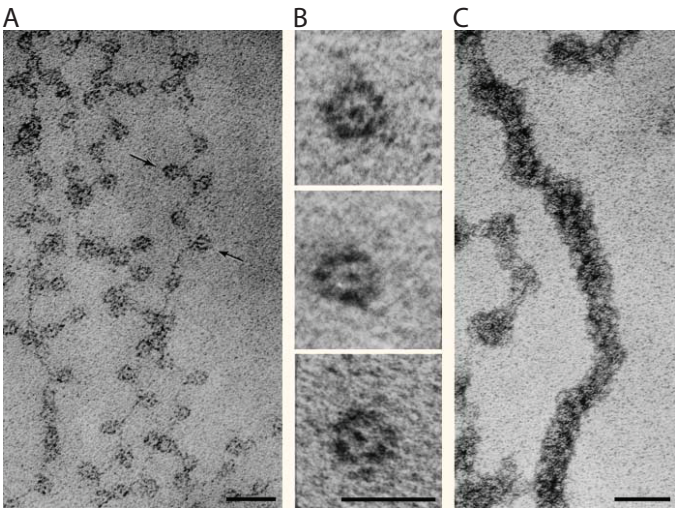


Figure 1.3: A gallery of electron micrographs of chromatin. (A) Low ionic-strength chromatin spread, the "beads on a string". Size marker: 30 nm. (B) Isolated mononucleosomes derived from nuclease-digested chromatin. Size marker: 10 nm. (C) Chromatin spread at a moderate ionic strength to maintain the 30-nm higher-order fibre. Size marker: 50 nm. Reprinted by permission from Macmillan Publishers Ltd: Nature Publishing Group, copyright (2003) [21].

easily lost in a typical nucleosome mapping experiment (MNase-seq) [51, 52]. These "fragile" nucleosomes allow easier access to the regulatory DNA-binding proteins, which need to attach to specific gene promoters in order to perform their function.

1.2 Studies of nucleosome positioning

To motivate our study, we first present a short history of experiments concerning nucleosome positioning.

The typical nucleosome mapping experiment has the following steps (Figure 1.4). Chromatin is first isolated from the cells, and then fragmented by sonication, or digested by the action of a nuclease. Chromatin digestion is usually done by micrococcal nuclease (MNase), which hydrolyses the linker DNA, while the nucleosomal DNA is protected by histones. After chromatin digestion, the nucleosomes are selected by immunoprecipitation with antibodies to histones. The selected pieces of DNA are then deproteinized and purified. Using agarose gel electrophoresis, the fragments corresponding to single

nucleosomes, with lengths of about 147 bp, are size-selected. Finally, the nucleosomal DNA pieces are identified by microarray hybridization or high-throughput sequencing. In microarray experiments, fluorescently labelled nucleosomal DNA fragments are hybridized to a microarray, in parallel with genomic DNA as a control. For each probe on the microarray, one obtains intensities corresponding both to nucleosomal DNA and to the control sample. The regions with higher-than-average ratio of these intensities correspond to the nucleosomes, while the regions with lower-than-average ratio correspond to nucleosome depleted regions (NDRs). High-throughput sequencing has an increased resolution and allows mapping of individual nucleosomes with single bp resolution.

Nucleosome mapping has been done both *in vitro* and *in vivo*. *In vitro* (Latin for "within the glass") refers to experiments performed in a controlled environment outside of a live organism, whereas *in vivo* (Latin for "within the living") indicates an experiment using a live organism, as opposed to a partial or dead organism. While *in vivo* studies reflect the nucleosome distribution in living cells, *in vitro* studies are better suited for determining the DNA sequence preferences for histone binding, because the effects of the *trans* determinants of nucleosome organization are eliminated. *In vitro*, purified histone proteins are assembled on genomic DNA either by salt dialysis [54], or by using chromatin assembly proteins [55].

The first model of nucleosome positioning was formulated by Kornberg and Stryer in 1988 [56]. This is called the barrier model of statistical positioning, and it shows that near an impenetrable barrier, the probability of finding a nucleosome is oscillatory, that is the nucleosomes are phased by the potential barrier. These potential barriers can be generated by nucleosome-excluding sequences, as Poly(dA:dT), or by histones or other DNA-binding proteins, which are strongly bound to particular loci in the genome. It was shown that, *in vivo*, several TF binding sites are covered by less nucleosomes than *in vitro* [57], which indicates that TFs can bind and create a potential barrier for histones. In yeast, there are many more well-positioned nucleosomes than in human cells, which may be explained by the fact that the human genes are longer and the potential barriers which appear at the regulatory regions are separated by much more space, such that their influence is felt by a smaller fraction of the total number of

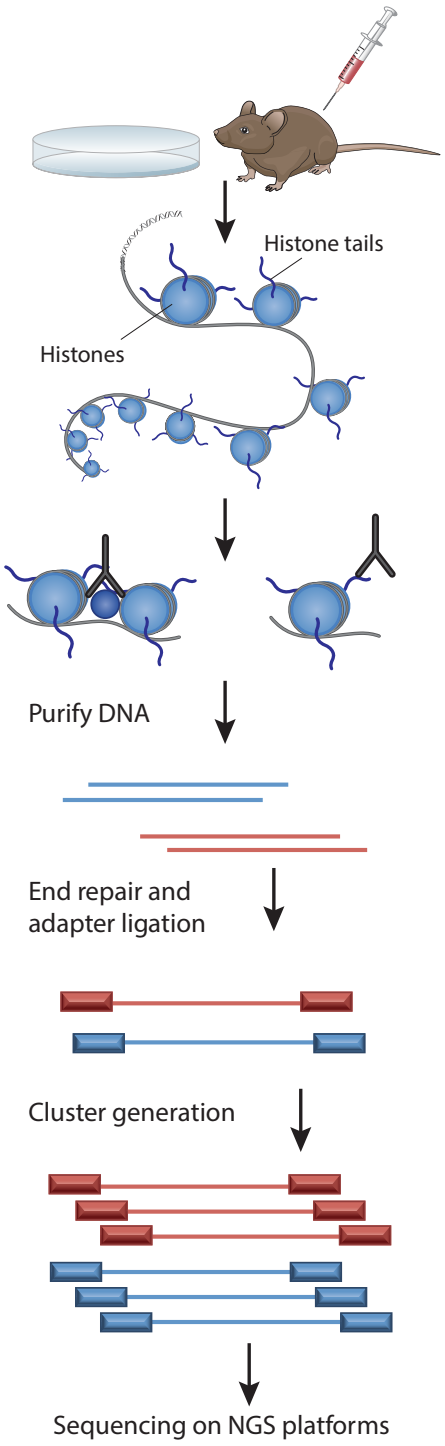


Figure 1.4: ChIP-seq experimental design. ChIP-seq is a powerful tool used to investigate protein-DNA interactions on a global scale. It is important that the appropriate controls for antibody specificity be determined before ChIP-seq is begun. After isolation of the ideal number of cells, chromatin is sheared into an ideal size range by sonication or enzymatic means, as MNase digestion. Next, high-quality antibodies are used for ChIP to enrich for factor-occupied DNA sequences. After purification of ChIP-enriched DNA, a library is constructed to allow sequencing on next-generation sequencing (NGS) platforms. Library construction typically includes end-repair, the addition of single adenosine residues, adaptor ligation and PCR with primers compatible with the sequencing platform. After cluster generation, single- or paired-end sequencing is performed on next-generation sequencing platforms. Reprinted by permission from Macmillan Publishers Ltd: Nature Publishing Group, copyright (2011) [53].

nucleosomes.

The advances of sequencing and tiling microarrays generated a lot of interest in studying nucleosome positions and the factors which affect these. The first mapping of nucleosome positions, *in vivo*, in *S. cerevisiae*, appeared in 2004. Using tiling microarrays with low resolution of about 1 kbp, Lee et al. [58] and Bernstein et al. [59] were able to observe a typical nucleosome organization at the gene promoters. They observed an NDR which is flanked by two well-positioned nucleosomes, despite the low resolution of their studies. Lee et al. [58] and Bernstein et al. [59] presented evidence that nucleosome occupancy in gene promoters is anti-correlated with the transcriptional initiation rate at the promoters. Promoters that regulate active genes were in general depleted of nucleosomes. Changes in the transcription rates were reflected also in the nucleosome occupancy of the corresponding promoters. Analyzing the change in genome-wide nucleosome organization after applying a heat shock to the cells, they found that nucleosome occupancy decreased at the promoters of induced genes and increased at the promoters of genes that were turned off.

In the following year, Yuan et al. [60] were able to map *in vivo* nucleosome positions with nearly single-nucleosome resolution. They mapped the positions of about 2000 nucleosomes over approximatively 500 kbp of *S. cerevisiae* DNA. They again found the characteristic pattern of NDRs, delimited by well-positioned nucleosomes at the gene promoters. Comparing the nucleosome-free DNA sequences, they observed that these sequences were enriched in Poly(dA:dT) tracts. These are homopolymeric stretches of deoxyadenosine nucleotides on one strand, paired with homopolymeric stretches of deoxythymidine nucleotides from the complementary DNA strand.

In 2005, Sekinger et al. [61] compared *in vivo* nucleosome organization with that obtained *in vitro*, using core histones from *HeLa* cells (the most commonly used human cell line [62]), assembled on a short piece of DNA (HIS3-PET56 region, 2.8 kbp) by gradient salt dialysis. The nucleosome pattern found *in vitro* was largely determined by the intrinsic DNA sequence preferences of histones, and it resembled the nucleosome distribution that is observed *in vivo*. Similarly for the DED1 promoter region, they also found a good agreement between *in vivo* and *in vitro* nucleosome organizations.

Both promoter regions had low histone densities and in order to test whether histone densities in promoters are lower than in the corresponding coding regions, the authors tested 4331 genes along the entire genome. They found that 78.6% of the coding regions have on average nucleosome occupancy more than two times higher than the corresponding promoters. They suggested that the yeast genome contains promoters whose DNA sequences are unfavorable for nucleosome formation, and ensure that TFs bind to these regions, rather than to the irrelevant binding motifs which are found in the non-promoter regions.

In 2007, Lee et al. [63] were able to produce the first genome-wide map of nucleosomes in *S. cerevisiae*. They were able to identify approximatively 70000 nucleosomes, which occupy about 81% of the yeast genome. They observed that the NDRs are positioned near the transcription start sites (TSSs). All the previously mentioned studies used tiling microarrays and the resolution was limited by the probe density.

ATP-dependent chromatin remodelers [24] use ATP hydrolysis in order to move or detach the histones from their preferred positions. Whitehouse et al. [64] compared the nucleosome distribution in wild-type (WT) and *isw2*Δ mutant cells. They found that in WT cells, the yeast Isw2 chromatin remodeling complex slides the +1 nucleosome (see Figure 1.5) upstream, thereby inhibiting the binding of other TFs to the promoters. The repositioning of thousands of nucleosomes which were located adjacent to regulatory sites, was found to be controlled by Isw2. A key finding was that, when ISW2 was deleted, transcription was able to initiate at cryptic start sites [65], although the mechanism that ensures the correct direction of transcription in WT cells is still not well-known.

In the same year, the first study that used high-throughput sequencing to map nucleosomes appeared [66]. Albert et al. detected the genome-wide distribution of the nucleosomes containing the H2A.Z histone variant. In the following year, Frank Pugh's lab also mapped the nucleosomes using antibodies to the H3 and H4 histones [67]. They mapped more than one million nucleosomes and confirmed the typical nucleosome pattern near TSS – a -1 nucleosome, followed by an NDR, followed by a +1 nucleosome. They also found that most of the genes have another NDR at their 3' ends. It was

found that +1 nucleosomes create barriers for the other nucleosomes, which are phased downstream.

Regular arrays of nucleosomes, as seen near TSSs, can be explained by the barrier model of nucleosome positioning, formulated by Kornberg and Stryer in 1988 [56]. Later studies of the nucleosome distribution in yeast [67, 68] support the model by Kornberg and Stryer. This canonical organization with two well-positioned nucleosomes on both sides of the TSS, separated by an NDR, is found more in the "housekeeping" genes, as those responsible for glycolysis, and less in the stress responsive genes, as those which react to unfavorable environments [69].

Yeast promoters contain A/T rich sequences, in particular Poly(dA:dT), and these DNA sequences give the DNA polymer a higher rigidity, and the formation of a nucleosome at these locations will be more difficult. In this way, the potential barrier creates an NDR and it is able to phase the nearby nucleosomes by statistical positioning (Figure 1.5).

In 2008, Shivaswamy et al. [70] also used high-throughput sequencing to map the remodeling of nucleosomes throughout the yeast genome after heat shock. They detected that in the typical promoter, after heat shock, one or two nucleosomes appeared, disappeared or were repositioned. The nucleosome positioning was found to depend on the presence of the TATA box, and to be correlated with the transcription rates. The TATA box, also called Goldberg-Hogness box, is a DNA sequence – 5'-TATAAA-3' or a variant – usually found at the binding site of RNA polymerase II in the promoter region of genes in eukaryotes. The authors of this study also observed that nucleosome remodeling causes changes in the accessibility of TFs to their binding sites.

Histones have little sequence specificity, in the sense that they do not have a binding motif, and all DNA sequences can form nucleosomes, with different affinities. The main difference between different DNA sequences which wrap around the histones is their bendability. Poly(dA:dT) sequences, and in general A/T rich sequences, are stiffer and can bend with more difficulty to form nucleosomes. Sequences that contain WW dinucleotides, with W denoting an A or T nucleotide, repeated every 10-11 bp, and interposed SS dinucleotides, with S denoting a C or G nucleotide, are more flexible and

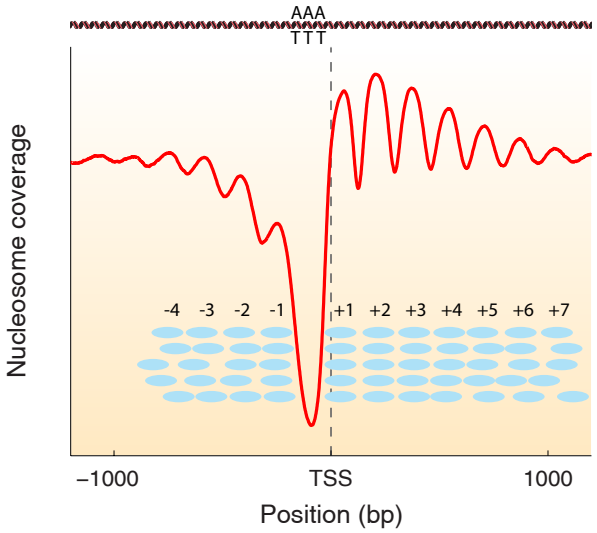


Figure 1.5: Typical *in vivo* nucleosome coverage near the TSS. A/T rich sequences which are enriched in the promoters create a stiff DNA region which disfavors nucleosome formation. The NDR and the potential barrier which is created by the A/T rich sequences may phase the nearby nucleosomes. The average nucleosome coverage (red line), or occupancy, is an average over many nucleosome configurations, corresponding to different cells. Some of these possible configurations are indicated by the blue ovals, which symbolize individual nucleosomes. Some nucleosomes are well-positioned, as the +1 nucleosome, which means that these are found in most of the cells at the same position. Other nucleosomes are fuzzier, as the +7 nucleosome, which means that in different cells these nucleosomes occupy shifted positions. The DNA and the Poly(dA:dT) sequence are not drawn to scale.

bind the histones with higher affinity [71].

Because some DNA sequences favor histone attachment and nucleosome formation, while other are stiff and disfavor nucleosome formation, the idea that *in vivo* nucleosome positions are intrinsically encoded in the DNA sequence was supported by some researchers. To test the claim that the genomic sequence is highly predictive of the *in vivo* nucleosome organization, Field et al. [72] mapped about 380000 yeast nucleosomes. They used a computational model to study the nucleosome positioning signals. They found that Poly(dA:dT) tracts are nucleosome disfavoring sequences, and important nucleosome positioning signals *in vivo*.

Using the physical properties of DNA, as the sequence-dependent DNA flexibility and the intrinsic curvature, in 2008, Miele et al. [73] computationally predicted that the

yeast promoter regions are unfavorable for nucleosome formation, and that these NDRs are bordered on both sides by regions with high nucleosome affinity. Their result shows that in addition to regulatory factors and chromatin remodeling enzymes, the physical properties of DNA play a major role in nucleosome positioning and transcription regulation.

In the following year, Morozov et al. [74] used another physical model in order to account for DNA flexibility. They computed the sequence-dependent DNA bending energies, and used these energies in order to solve the many-body problem of nucleosome positioning. They used a standard dynamic programming algorithm to position nucleosomes, taking into account the steric exclusion between neighboring histones. They found that the bending properties of DNA explain the intrinsic sequence-dependent nucleosome affinities *in vitro*.

To study the importance of DNA sequence in nucleosome positioning *in vitro*, in 2009, two different labs studied in parallel the difference between the nucleosome distributions obtained *in vivo* and *in vitro*. Using salt dialysis, Kaplan et al. [57] assembled purified chicken histone octamers on purified yeast genomic DNA. They found that *in vitro*, NDRs appear near the gene ends and near TF binding sites. Using *in vitro* data, Kaplan et al. developed a computational model which takes into account the dinucleotide and 5-mer distributions for predicting the nucleosome occupancy along the genome. Even though the model was trained on *in vitro* yeast data, the predicted nucleosome coverage correlated reasonably well with *in vivo* nucleosome distributions in *S. cerevisiae*, grown in three different media, and *C. elegans*. The authors reached the conclusion that DNA sequence preferences are responsible for most of the nucleosome organizations *in vivo*. In a second parallel study, Zhang et al. [75], performing similar *in vitro* nucleosome assembling on purified DNA from both *S. cerevisiae* and *E. coli*, which does not have histones, reached a totally different conclusion. They found that DNA preferences were not a major determinant of nucleosome organization *in vivo*, and there were other factors, more important than the DNA sequence, which made the nucleosomes to organize in a way, which could not be explained by sequence alone. They argued against a genomic code for nucleosome positioning, and suggested

that the characteristic oscillatory nucleosome distribution that appears near TSS was not generated by DNA sequences, but by phasing against a potential barrier, created during transcription initiation near TSS, also known as statistical positioning.

In 2010, Locke et al. [76] used genome-wide nucleosome maps to study the sequence specificity of the nucleosome affinity in the *S. cerevisiae*, *E. coli*, and *C. elegans* genomes. They used a statistical mechanics model of hard-rods to infer the nucleosome formation energies corresponding to every sequence of 147 nucleotides, from the nucleosome distributions which were measured in high-throughput experiments. The authors developed a series of models of increasing complexity, and concluded that the nucleosome organization could be explained by the mono- and dinucleotide distributions along the genomes, and that longer sequence motifs had a minor influence.

The idea that the DNA sequence is the main nucleosome positioning factor is still under debate. There is at least a consensus among the researchers working in the chromatin field that the DNA sequence is not the only determinant of nucleosome organization. Beyond the DNA sequence preference, there are also other factors, which influence nucleosome positioning, as for example, stacking against potential barriers [56], competition of histones with other DNA-binding proteins, as TFs Abf1 and Rap1 in *S. cerevisiae* [77], action of chromatin remodelers [24], disruption by RNA polymerase [70], among others. Transcription by RNA polymerase modifies the nucleosome organization. For example, in *S. cerevisiae*, the nucleosomes covering the highly transcribed genes are more delocalized [78]. Interestingly, the situation is reversed in Drosophila, in the sense that the genes which are active contain the regular arrays of nucleosomes, while the inactive genes contain poorly localized nucleosomes [52].

TFs compete with the histones for DNA binding, and they can also create potential barriers which oppose nucleosome formation at their binding sites. Abf1, Rap1 and Reb1 are just some examples of TFs which have binding sites near yeast promoters, which contribute to the formation of nucleosome-repulsive potential barriers [79, 80]. In human cells, CCCTC-binding factor (CTCF) and neuron-restrictive silencer transcription factor (NRSF/REST) can also generate potential barriers and phased nucleosomes [81]. Any strong binding of sequence-specific factors, in principle, can generate potential

barriers which may phase the nearby nucleosomes.

A major difference between *in vitro* and *in vivo* nucleosome distributions is that phased nucleosome arrays near TSSs appear only *in vivo*. *In vitro* we still see an NDR because of the unfavorable DNA sequences from the promoters, but this is not enough to phase the nearby nucleosomes. ATP-dependent chromatin remodelers are known to generate regular nucleosome spacing [24, 82].

In 2011, Frank Pugh's lab conducted a study [83] to determine what biochemical factors assembled the nucleosomes *in vitro* in order to resemble the organization which was observed *in vivo*. They assembled Drosophila histones on yeast DNA *in vitro*. As in the previous studies [57, 75], they obtained the NDR upstream of TSS but no regularly phased nucleosome arrays. Adding yeast whole extract did not help, which means that the simple binding of the proteins from the cell extract was not enough to phase the nearby nucleosomes. However, when ATP was added to the whole cell extract, the nucleosomes rearranged in a way which resembled the organization observed *in vivo* – stronger NDRs, phased nucleosomes near TSS, among others. They concluded that the ATP-dependent chromatin remodelers were responsible for the even-spaced nucleosome arrays. In a parallel study, Gkikopoulos et al. [84] showed that deletion of the genes encoding the chromatin remodelers Isw1, Isw2 and Chd1 resulted in a clear disruption of the regular arrays of nucleosomes downstream TSS. These two studies showed that chromatin remodelers may overcome the sequence-dependent preferences of the histones, and generate regular arrays of nucleosomes.

Because the energy of bending a DNA segment around a histone depends on its nucleotide sequence [85, 74], nucleosomes exhibit a range of *in vitro* formation energies [86, 71], although any DNA sequence can be packaged into a nucleosome. Recent work has clarified the role of sequence rules that influence nucleosome positioning. Genome-wide *in vitro* reconstitution experiments have confirmed that nucleosome architecture over promoters and genes is partially established by DNA sequence, mostly as a result of nucleosome depletion from A/T-rich, nucleosome-disfavoring sequences, on both ends of the transcript [61, 57, 75]. However, nucleosomes are not strongly localized and, on average, nucleosome occupancy is just about $20 - 30\%$ lower over NDRs compared to

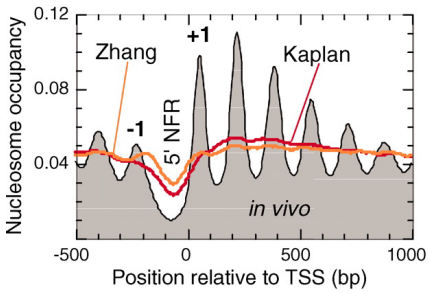


Figure 1.6: Nucleosome organization around the 5' ends of genes is not reconstituted *in vitro* with purified histones alone. Composite distribution of nucleosome midpoints, assembled *in vivo* [57] or *in vitro* [57, 75], around transcriptional start sites. Reprinted by permission from The American Association for the Advancement of Science, copyright (2011) [83].

the mean occupancy in a window which includes both the coding region and adjacent sequence (Figure 1.6). *In vivo*, the 5' and 3' NDRs flanking the transcripts are much more pronounced, with about $60-70\%$ occupancy depletion on average, with respect to the mean [57, 60, 67, 87]. They establish a striking pattern of nucleosome localization over genic regions simply due to steric exclusion, which causes nucleosomes to "phase off" potential barriers [56] (Figure 1.5). Although the exact nature of these *in vivo* barriers is unknown, and may vary between cell types and environmental conditions, they are likely established through a combined action of RNA polymerase, ATP-dependent chromatin remodeling enzymes, and DNA-binding proteins [88, 89, 90]. The reduced degree of nucleosome localization *in vitro* and shallow NDRs indicate that the sequence is only one of multiple factors which determine the nucleosome distribution *in vivo*.

Recently, the Segal and Struhl labs, which published the contradictory reports in 2009 [57, 75], wrote a joint paper [91] in which they tried to bring into agreement the two contradictory conclusions. They argued that the DNA sequence was a major determinant of the nucleosome organization, but other factors were also very important *in vivo*, which could override the sequence preference of the histones.

In 2012, a new method of mapping nucleosome centers at bp resolution was developed by Brogaard et al. [92]. They mutated the histone H4 protein to allow the positioning of a copper ion near the dyad of the nucleosome, after chromatin extraction.

Copper reacts with hydrogen peroxide creating reactive hydroxyl radicals, and these cut the DNA near the dyads. The resulting pieces of DNA are sequenced, and mapped to the genome. Using a Bayesian deconvolution algorithm, the authors computed a nucleosome positioning score for every genomic location, and generated a map of 67543 unique nucleosome positions, which covered 79.9% of the genome, and a redundant map of 351264 nucleosomes, allowing nucleosomes to overlap arbitrarily. They noticed that the center to center distances of overlapping nucleosomes have predominant values which differ by about 10 bp (DNA helical repeat). The nucleosome organization near TSS had the typical pattern – a strong depletion of nucleosomes immediately upstream TSS, flanked by two well-positioned nucleosomes on both sides. Another NDR was observed at the 3' ends of the genes. These observations confirmed the nucleosome organization that was previously obtained in MNase-seq experiments. Similar phasing of the nucleosomes was found near the autonomously replicationg sequences (ARS) which contain the origins of replication in the yeast genome, and near chromosomal centromeres.

In 2013, Locke et al. [93] studied the nucleosome organization in the worm *C. elegans*. They assembled nucleosomes, *in vitro*, on *C. elegans* genomic DNA, and compared the obtained nucleosome organization with that observed *in vivo*. The authors observed that, although the SS dinucleotides, with S denoting either a C or a G nucleotide, were the most favorable for nucleosome formation *in vitro*, the situation was different in living cells. The majority of well-positioned *in vivo* nucleosomes did not occupy thermodinamically favorable DNA sequences, as the ones observed *in vitro*. They also found that exons were in general more favorable to nucleosome formation than introns.

1.3 Our models of nucleosome positioning

Nucleosome positions and formation energies can be predicted using a thermodynamic model which takes into account steric exclusion and intrinsic histone-DNA sequence

preferences [76, 94, 95]. In this approach, sequence determinants of nucleosome energetics are inferred directly from experimentally available genome-wide nucleosome distributions. Structural regularity of the chromatin fiber imposes additional constraints on nucleosome positions [96, 97]. For example, linkers between neighboring nucleosomes become preferentially discretized with the $10-11$ bp periodicity of DNA helical twist [98]. The discretization is required to avoid steric clashes caused by the nucleosome rotating with respect to the linker DNA axis as the linker increases in length [96], and to maintain a regular pattern of protein-protein and protein-DNA contacts in the chromatin fiber [97]. Adding a short DNA segment to the linker rotates the nucleosome with respect to the rest of the fiber, causing disruption of its periodic structure. The disruption is minimized if the length of the extra segment is a multiple of $10-11$ bp, which brings the nucleosome into an equivalent rotational position.

We have recently developed a rigorous approach in which linker length discretization is described by nearest-neighbor two-body interactions in a system of non-overlapping finite-size particles [94, 95]. We have shown that it is possible to simultaneously infer one-body energies given by intrinsic histone-DNA interactions and two-body energies caused by chromatin fiber formation. We have predicted the two-body interaction from high-throughput maps of nucleosome positions on the *S. cerevisiae* genome, and demonstrated its essential role in forming nucleosome occupancy patterns over genic regions.

In Chapter 2 we present a detailed account of our theoretical framework. We develop a minimally constrained sequence-specific model of nucleosome energetics, in which the same energies are assigned to mono- and dinucleotides, regardless of their exact position within the 147 bp nucleosomal site [76]. We make a clear distinction between the two types of energies which dictate nucleosome positioning – one-body energy, given by the elastic properties of each DNA sequence and by the electrostatic interaction between the negatively charged DNA and the positively charged histones, and effective two-body interaction which appears from the geometric constraints on the chromatin fiber, that is steric clashes. We also build a minimal model in which *in vivo* nucleosomes are positioned solely by potential barriers located at each end of the transcripts.

Without invoking explicit sequence specificity, the model successfully reproduces nucleosome occupancy patterns observed *in vivo* in *S. cerevisiae*. In contrast, sequence-dependent models neglecting the additional potential barriers, can only capture the observed liquid-like, delocalized organization of *in vitro* nucleosomes [57, 75]. By combining the minimal model with sequence-specific nucleosome energies, we estimate that intrinsic histone-DNA interactions contribute to less than 30% to the height of the *in vivo* potential barriers.

To allow access to proteins, the corresponding binding sites have to be clear of nucleosomes. Two different mechanisms for site exposure have been proposed – *i*) nucleosome translocation, when nucleosomes slide along the DNA, and *ii*) nucleosome unwrapping, when outer stretches of nucleosomal DNA, from either of its ends, can transiently peel off the histone surface, unwrapping and rewrapping with a high frequency. The second scenario is supported by many authors [28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 99, 44, 45, 100, 101, 46, 18, 102, 103, 104, 105, 48, 106, 49, 107, 50, 108, 109], and will be the hypothesis of the model that we discuss in Chapter 3. The fluctuations of the DNA-histone contacts can facilitate the activity of RNA polymerases during transcription elongation [18], and they can also provide an explanation for the fast DNA repair by proteins like photolyase [43].

In Chapter 3 we develop a statistical mechanics model which allows nucleosome unwrapping, competition between different DNA-binding proteins, sequence dependent binding energies, sequence-independent potential barriers and walls, and effective two-body interactions. This model is able to explain the recent finding that many pairs of nucleosomes occupy positions that are very close to each other [92], which is impossible to explain by neglecting the nucleosome unwrapping. Our model is also able to explain the observed nucleosome-mediated cooperativity [110, 27, 29, 30, 32, 33, 40, 111, 47, 50, 49]. Using this model it is also easy to simulate the distribution of nucleosomes that collide [100], or have been partially disassembled [112, 113, 114].

A statistical mechanics model that considers the unwrapping of the nucleosomes was published recently by Teif and Rippe [105, 115, 116], where the relevant quantities are calculated using recurrent relations and dynamic programming. They solve the direct

problem, and compute the distributions of the particles, starting from the relevant energies that enter in the positioning problem. In their model, two particles are not allowed to be both unwrapped at the point of contact (as shown in Figure 2D in [116]), and for this reason, the partition function and all the final results are not exact, and this model is not appropriate for modelling colliding nucleosomes, see for example [100]. We allow both particles to be unwrapped at the point of contact, and we consider also the inverse problem, which is more interesting from an experimentalist's point of view. In any experiment which measures the binding preferences of the TFs or the arrangement of the nucleosomes in different promoters, the output of the experiment is the organization of the DNA-binding proteins. One would like to know the energies which dictate the distributions of the relevant proteins. Solving the inverse problem, one obtains these energies, and after that, one can predict what would happen in a non-natural, engineered DNA sequence, and use these sequences in different genetics experiments.

Chapter 2

Nucleosome positioning

In this chapter we present the theoretical framework which allows us to predict nucleosome organization along an entire chromosome, but also to infer relevant energetic quantities that dictate the nucleosome organization, which is observed in the experiments. Obtaining the distribution of the components of a system, when the important energies and interactions are known, is refereed to as the direct problem. Conversely, starting with the nucleosome distribution which is obtained in an experiment and solving for the energies which are able to generate this nucleosome organization, is refereed to as the inverse problem.

Although the DNA is compactified a millionfold inside the nucleus of a human cell, and the three-dimensional (3D) organization of the chromatin fiber has very important consequences, the simplest approximation in which DNA is considered as a straight linear polymer, and the long-distance interactions are neglected, is enough to explain some important features of the nucleosome organization. We present in this chapter, the detailed description of the one-dimensional (1D) lattice model of nucleosome positioning and some of its applications.

This Chapter is based on our work which was published in [94] and [95].

2.1 Energetics of one-dimensional hard rods with nearest-neighbor interactions

We model a chromosome by a 1D lattice, characterized by a finite length of L bp. This lattice is covered with histone octamers, which are approximated by 1D hard rods of length $a = 147$ bp. *S. cerevisiae*'s chromosomes have lengths ranging between about 200 kbp (chromosome I) and 1.5 Mbp (chromosome IV). In human, the chromosome

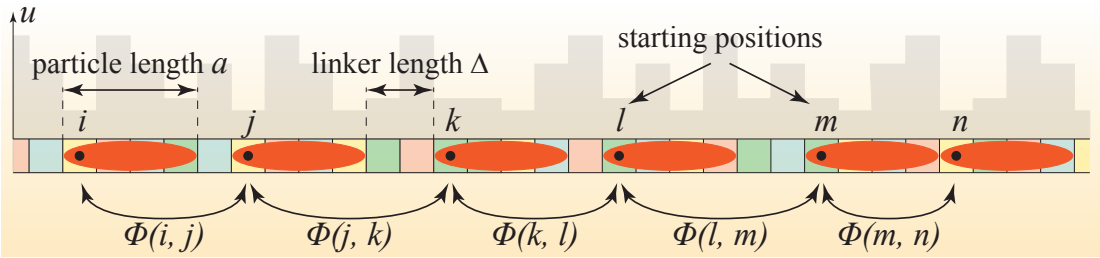


Figure 2.1: A typical configuration of six nucleosomes. Each nucleosome covers DNA sequence (represented by colored boxes) which gives the one-body energy of the nucleosome, u. The one-body energy is represented by gray bars. For simplicity in this toy model, the one-body energy is entirely determined by the base pair located at the starting position of the nucleosome. In more realistic scenarios, one-body energy is a function of the entire sequence occupied by the nucleosome. The two-body interaction, $\Phi(i, j)$, acts only between neighboring nucleosomes, where the two indices i and j represent their starting positions.

lengths are two order of magnitude larger.

We consider the problem of interacting hard rods confined to a 1D lattice (Figure 2.1). Let $u(k)$ be the binding energy of a histone that occupies bps k through $k+a-1$ on the DNA. This one-body energy has two contributions – one from the bending energy which is necessary to wrap the DNA around the histone, and another from the electrostatic interaction between the positively charged histone and the negatively charged DNA. Because of the steric clashes between neighboring histone octamers, we need to consider an interaction between the pairs of nucleosomes, which disallow the overlapping of the 1D hard rods. Let $\Phi(k, l)$ be the two-body interaction between a pair of nearest-neighbor nucleosomes with starting positions k and l, respectively.

The binding energy, $u(k)$, describes the intrinsic histone-DNA interactions, while the two-body interaction, $\Phi(k, l)$, accounts for the effects of chromatin structure. We assume that nucleosomes cannot slide off the chromosome edges, so that we impose impenetrable walls at both ends to the chromosomes,

$$u(0) = u(L - a + 2) = u(L - a + 3) = \ldots = u(L) = \infty.$$

Moreover, particle overlaps are not allowed and the two-body potential is short-range

as in the Takahashi hard-rod model [117],

$$\Phi(k,l) = \begin{cases} \infty & \text{if } l < k + a, \\ V\left(l - (k + a - 1)\right) & \text{if } k + a \leq l < k + 2a, \\ 0 & \text{if } l \geq k + 2a, \end{cases}$$

where $V(x)$ is a generic interaction which depends on the linker length, x, between the pair of neighboring nucleosomes, and $l - (k + a - 1)$ is the length of the linker DNA corresponding to the pair or nucleosomes which start at positions k and l, respectively.

For a system of N hard rods that are attached to a 1D lattice, at a fixed temperature, the canonical partition function is given by

$$Q_N = \sum_{i_1,\ldots,i_N} e^{-\beta u(i_1)} e^{-\beta \Phi(i_1,i_2)} e^{-\beta u(i_2)} \ldots e^{-\beta u(i_{N-1})} e^{-\beta \Phi(i_{N-1},i_N)} e^{-\beta u(i_N)}, \tag{2.1}$$

where k_B is Boltzmann's constant and $\beta = 1/(k_B T)$ is the inverse temperature.

Let us introduce two $l_{\max} \times l_{\max}$ matrices, where $l_{\max} = L - a + 1$ is the rightmost possible starting position of a particle of length a,

$$\langle k|e|l\rangle = e^{-\beta u(k)} \delta_{kl},$$
$$\langle k|w|l\rangle = e^{-\beta \Phi(k,l)}.$$

Here δ_{kl} is the Kronecker delta symbol, and $\langle k|M|l\rangle$ represents the element of matrix M in row k and column l, written using the Dirac notation. Of course, the element $\langle k|e|l\rangle = 0$, unless $k = l$, and so the matrix e is diagonal, and the elements $\langle k|w|l\rangle = 0$, unless $l \geq k+a$, that is the matrix w is an upper triangular matrix. By $|l\rangle$ we understand a column vector of dimension $l_{\max}$, with one at position l and zeros everywhere else, and $\langle k|$ is a row vector with one only at position k, and zeros everywhere else. Using this notation, we can rewrite Equation (2.1) as

$$\begin{aligned} Q_N &= \sum_{i_1,\ldots,i_N} \langle i_1|e|i_1\rangle \langle i_1|w|i_2\rangle \langle i_2|e|i_2\rangle \langle i_2|w|i_3\rangle \ldots \langle i_{N-1}|e|i_{N-1}\rangle \langle i_{N-1}|w|i_N\rangle \langle i_N|e|i_N\rangle \\ &= \sum_{i_1,\ldots,i_N} \langle i_1|ew|i_2\rangle \langle i_2|ew|i_3\rangle \ldots \langle i_{N-1}|ew|i_N\rangle \langle i_N|e|i_N\rangle \\ &= \langle J|(ew)^{N-1} e|J\rangle, \end{aligned}$$

where $|J\rangle = \sum_{l=1}^{l_{\max}} |l\rangle$ is a $l_{\max}$-dimensional vector with one at every position. By definition, when $N = 0$, we have that $Q_0 = 1$, i.e. there is only the empty state, and the system is found in this state with probability 1.

In our model, we consider real histones which can attach and detach from the DNA. The system can be found in different states, with various numbers of histones attached to the DNA. Therefore, we need to consider a system with a variable number of particles. For a system of hard rods at a fixed temperature, T, and fixed chemical potential, μ, the grand-canonical partition function is given by

$$\begin{aligned} Z &= \sum_{N=0}^{N_{\max}} e^{\beta N \mu} Q_N \\ &= 1 + \sum_{N=1}^{N_{\max}} \langle J|(zw)^{N-1} z|J\rangle \\ &= 1 + \sum_{N=1}^{\infty} \langle J|(zw)^{N-1} z|J\rangle \\ &= 1 + \langle J|(I - zw)^{-1} z|J\rangle, \end{aligned} \tag{2.2}$$

where $N_{\max} = \left\lfloor \frac{L}{a} \right\rfloor$ is the maximum number of particles that can fit on L bp, I is the identity matrix, and

$$\langle k|z|l\rangle = e^{\beta[\mu - u(k)]} \delta_{kl}.$$

Here we changed the upper limit of the sum to $N \to \infty$, but all the terms with $N > N_{\max}$ are vanishing because we cannot have in the lattice more than $N_{\max}$ particles without overlapping. All the configurations with overlapping particles have infinite energies due to the two-body interaction, Φ, and therefore vanishing Boltzmann weights.

The s-particle distribution functions are defined as

$$n_s(i_1, \ldots, i_s) \equiv \frac{\zeta(i_1) \ldots \zeta(i_s)}{Z} \frac{\delta^s Z}{\delta\zeta(i_1) \ldots \delta\zeta(i_s)},$$

where $\zeta(i) = e^{\beta[\mu - u(i)]}$ (see the chapter by Stell in [118]). In particular, the one-particle distribution function is

$$n_1(i) = \frac{1}{Z} \langle J|(I - zw)^{-1}|i\rangle\langle i|z|i\rangle\langle i|(I - wz)^{-1}|J\rangle, \tag{2.3}$$

and the two-particle distribution function is

$$n_2(i,j) = \frac{1}{Z} \langle J|(I - zw)^{-1}|i\rangle\langle i|z|i\rangle\langle i|w(I - zw)^{-1}|j\rangle\langle j|z|j\rangle\langle j|(I - wz)^{-1}|J\rangle. \tag{2.4}$$

The nearest-neighbor two-particle distribution function, which gives the probability of finding two nearest neighbor particles at some specified locations, is

$$\overline{n}_2(i,j) = \frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|z|i\rangle\langle i|w|j\rangle\langle j|z|j\rangle\langle j|(I-wz)^{-1}|J\rangle. \tag{2.5}$$

These relations are easy to understand. To find the probability of starting a particle at position i given by Equation (2.3), we have to add the statistical weights of all configurations that contain a particle at that position, and divide the resulting sum by the normalization factor, the partition function. Similarly, to find the probability of having a pair of nearest-neighbor particles with the starting positions i and j, given by Equation (2.5), we need to sum the statistical weights of all configurations that contain that pair of particles. Note that for short distances $j - i < 2a$, the nearest-neighbor two-particle distribution function, $\overline{n}_2(i,j)$, is identical to the unrestricted two-particle distribution $n_2(i,j)$, because there is not enough space to put an additional particle between the two particles starting at positions i and j in the lattice.

In many cases of interest the energetics of the system is unknown, but the s-particle distributions are available from experiment. Therefore, we wish to find the unknown energies, u and Φ, from the particle distributions, n_1 and $\overline{n}_2$, by inverting Equations (2.3) and (2.5). Let us define two other matrices,

$$\langle i|N|j\rangle = n_1(i)\delta_{ij},$$

and

$$\langle i|N_2|j\rangle = \overline{n}_2(i,j).$$

We now express all matrices which depend on the unknowns u and Φ, in terms of the two new matrices, N and N_2, which give the distribution of the particles, and which are typically measured in experiments. Let us first compute the following two matrix elements, $\langle i|I - N_2N^{-1}|j\rangle$ and $\langle i|I - N^{-1}N_2|j\rangle$. We obtain

$$\langle i|I - N_2N^{-1}|j\rangle = \langle i|I|j\rangle - \frac{\langle i|N_2|j\rangle}{\langle j|N|j\rangle},$$

and

$$\frac{\langle i|N_2|j\rangle}{\langle j|N|j\rangle} = \frac{\langle J|(I-zw)^{-1}|i\rangle z(i)\langle i|w|j\rangle}{\langle J|(I-zw)^{-1}|j\rangle},$$

so that,

$$\langle i|I - N_2N^{-1}|j\rangle = \frac{\langle J|(I-zw)^{-1}|j\rangle\langle i|I|j\rangle}{\langle J|(I-zw)^{-1}|j\rangle} - \frac{\langle J|(I-zw)^{-1}|i\rangle z(i)\langle i|w|j\rangle}{\langle J|(I-zw)^{-1}|j\rangle}.$$

We see that

$$\langle i|I|j\rangle = \delta_{ij},$$

which implies that

$$\langle J|(I-zw)^{-1}|j\rangle\langle i|I|j\rangle = \langle J|(I-zw)^{-1}|i\rangle\langle i|I|j\rangle,$$

and so it follows

$$\langle i|I - N_2N^{-1}|j\rangle = \frac{\langle J|(I-zw)^{-1}|i\rangle\langle i|I-zw|j\rangle}{\langle J|(I-zw)^{-1}|j\rangle}.$$

Similarly, we can compute

$$\langle i|I - N^{-1}N_2|j\rangle = \langle i|I|j\rangle - \frac{\langle i|N_2|j\rangle}{\langle i|N|i\rangle},$$

and

$$\frac{\langle i|N_2|j\rangle}{\langle i|N|i\rangle} = \frac{\langle i|w|j\rangle z(j)\langle j|(I-wz)^{-1}|J\rangle}{\langle i|(I-wz)^{-1}|J\rangle}.$$

We now obtain

$$\begin{aligned}\langle i|I - N^{-1}N_2|j\rangle =& \frac{\langle i|I|j\rangle\langle i|(I-wz)^{-1}|J\rangle}{\langle i|(I-wz)^{-1}|J\rangle} - \frac{\langle i|w|j\rangle z(j)\langle j|(I-wz)^{-1}|J\rangle}{\langle i|(I-wz)^{-1}|J\rangle} \\ =& \frac{\langle i|I-wz|j\rangle\langle j|(I-wz)^{-1}|J\rangle}{\langle i|(I-wz)^{-1}|J\rangle}. \end{aligned} \tag{2.6}$$

Equation (2.1) yields

$$\begin{aligned}\langle i|(I - N_2N^{-1})N|j\rangle =& \langle i|I - N_2N^{-1}|j\rangle\langle j|N|j\rangle \\ =& \frac{\langle J|(I-zw)^{-1}|i\rangle\langle 1|I-zw|j\rangle}{\cancel{\langle J|(I-zw)^{-1}|j\rangle}} \\ & \times \frac{1}{Z}\cancel{\langle J|(I-zw)^{-1}|j\rangle} z(j)\langle j|(I-wz)^{-1}|J\rangle \\ =& \frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|I-zw|j\rangle z(j)\langle j|(I-wz)^{-1}|J\rangle\end{aligned}$$

Summing over the second index, j, in the preceding identity, we obtain

$$\begin{aligned}\langle i|(I-N_2N^{-1})N|J\rangle =&\frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|(I-zw)z(I-wz)^{-1}|J\rangle\\ =&\frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|z\cancel{(I-wz)(I-wz)^{-1}}|J\rangle\\ =&\frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|z|J\rangle\\ =&\frac{1}{Z}\langle J|(I-zw)^{-1}z|i\rangle \end{aligned}\tag{2.7}$$

Finally, summing over the first index, i, in Equation (2.7), we obtain

$$\langle J|(I-N_2N^{-1})N|J\rangle = \frac{1}{Z}\langle J|(I-zw)^{-1}z|J\rangle = \frac{Z-1}{Z},$$

so that the partition function is given by

$$Z = \frac{1}{1-\langle J|(I-N_2N^{-1})N|J\rangle}.\tag{2.8}$$

In Equation (2.1), if we sum over the first index, we obtain

$$\begin{aligned}\langle J|I-N_2N^{-1}|j\rangle =& \frac{\cancel{\langle J|(I-zw)^{-1}(I-zw)|j\rangle}}{\langle J|(I-zw)^{-1}|j\rangle}\\ =& \frac{1}{\langle J|(I-zw)^{-1}|j\rangle}.\end{aligned}\tag{2.9}$$

From Equations (2.7) and (2.9), we have

$$\begin{aligned}z(i) =&Z\frac{\langle i|(I-N_2N^{-1})N|J\rangle}{\langle J|(I-zw)^{-1}|i\rangle}\\ =&\frac{1}{1-\langle J|(I-N_2N^{-1})N|J\rangle}\frac{\langle i|(I-N_2N^{-1})N|J\rangle}{\langle J|(I-zw)^{-1}|i\rangle}\\ =&\frac{\langle J|I-N_2N^{-1}|i\rangle\langle i|(I-N_2N^{-1})N|J\rangle}{1-\langle J|(I-N_2N^{-1})N|J\rangle}\\ =&\frac{\langle J|I-N_2N^{-1}|i\rangle\langle i|N(I-N^{-1}N_2)|J\rangle}{1-\langle J|(I-N_2N^{-1})N|J\rangle}\\ =&\frac{\langle J|I-N_2N^{-1}|i\rangle\langle i|N|i\rangle\langle i|(I-N^{-1}N_2)|J\rangle}{1-\langle J|(I-N_2N^{-1})N|J\rangle}.\end{aligned}\tag{2.10}$$

Similarly, in Equation (2.6), if we sum over the second index, j, we obtain

$$\begin{aligned}\langle i|I-N^{-1}N_2|J\rangle =& \frac{\cancel{\langle i|(I-wz)(I-wz)^{-1}}|J\rangle}{\langle i|(I-wz)^{-1}|J\rangle}\\ =& \frac{1}{\langle i|(I-wz)^{-1}|J\rangle}.\end{aligned}\tag{2.11}$$

From Equation (2.5), we can now compute

$$\begin{aligned}\langle i|w|j\rangle =&\frac{Z\langle i|N_2|j\rangle}{\langle J|(I-zw)^{-1}|i\rangle\langle i|z|i\rangle\langle j|z|j\rangle\langle j|(I-wz)^{-1}|J\rangle}\\ =&\frac{\langle i|N_2|j\rangle}{\cancel{1-\langle J|(I-N_2N^{-1})N|J\rangle}}\frac{1}{\langle J|(I-zw)^{-1}|i\rangle}\times\\ &\times\frac{\cancel{1-\langle J|(I-N_2N^{-1})N|J\rangle}}{\cancel{\langle J|I-N_2N^{-1}|i\rangle}\langle i|N|i\rangle\langle i|(I-N^{-1}N_2)|J\rangle}\times\\ &\times\frac{1-\langle J|(I-N_2N^{-1})N|J\rangle}{\langle J|I-N_2N^{-1}|j\rangle\langle j|N|j\rangle\langle j|(I-N^{-1}N_2)|J\rangle}\times\\ &\times\frac{1}{\langle j|(I-wz)^{-1}|J\rangle}.\end{aligned}$$

Using Equations (2.9) and (2.11), the preceding equation becomes

$$\begin{aligned}\langle i|w|j\rangle =&\langle i|N_2|j\rangle\cancel{\langle J|I-N_2N^{-1}|i\rangle}\frac{1}{\cancel{\langle J|I-N_2N^{-1}|i\rangle}\langle i|N|i\rangle\langle i|(I-N^{-1}N_2)|J\rangle}\\ &\times\frac{1-\langle J|(I-N_2N^{-1})N|J\rangle}{\langle J|I-N_2N^{-1}|j\rangle\langle j|N|j\rangle\cancel{\langle j|(I-N^{-1}N_2)|J\rangle}}\cancel{\langle j|I-N^{-1}N_2|J\rangle}\\ =&\frac{\langle i|N_2|j\rangle\left[1-\langle J|(I-N_2N^{-1})N|J\rangle\right]}{\langle i|N|i\rangle\langle i|(I-N^{-1}N_2)|J\rangle\langle J|I-N_2N^{-1}|j\rangle\langle j|N|j\rangle},\end{aligned}$$

and we obtain that

$$\langle i|w|j\rangle=\frac{\langle i|N^{-1}N_2N^{-1}|j\rangle\left[1-\langle J|(I-N_2N^{-1})N|J\rangle\right]}{\langle i|(I-N^{-1}N_2)|J\rangle\langle J|I-N_2N^{-1}|j\rangle}. \tag{2.12}$$

Applying logarithms to Equations (2.10) and (2.12), we obtain exact expressions for the one-body energies and two-body interactions [119, 120]

$$-\beta\left[u(i)-\mu\right]=\ln\left(\frac{\langle J|I-N_2N^{-1}|i\rangle\langle i|N|i\rangle\langle i|I-N^{-1}N_2|J\rangle}{1-\langle J|(I-N_2N^{-1})N|J\rangle}\right), \tag{2.13}$$

$$-\beta\Phi(i,j)=\ln\left(\frac{\langle i|N^{-1}N_2N^{-1}|j\rangle\left[1-\langle J|(I-N_2N^{-1})N|J\rangle\right]}{\langle i|I-N^{-1}N_2|J\rangle\langle J|I-N_2N^{-1}|j\rangle}\right). \tag{2.14}$$

If the two-body interactions different from the steric exclusion are neglected, then the remaining interaction potential is just the hard-core interaction,

$$\Phi^0(k,l)=\begin{cases}\infty & \text{if } l<k+a,\\ 0 & \text{if } l\geq k+a,\end{cases} \tag{2.15}$$

and the matrix element $\langle k|w|l\rangle$ is replaced by $\Theta(l-k-a)$, where $\Theta(l-k-a)$ is the Heaviside step function, which takes value of one if $l\geq k+a$, and it is zero otherwise.

In this case, we have that

$$\begin{aligned}\langle J|(I-zw)^{-1}|i\rangle &= \langle J|I+zw+(zw)^2+\ldots|i\rangle \\ &= Z^f_{i-1}, \\ \langle i|(I-wz)^{-1}|J\rangle &= \langle i|I+wz+(wz)^2+\ldots|J\rangle \\ &= Z^r_{i+a},\end{aligned}$$

where Z^f_i and Z^r_i are partial statistical sums which can be efficiently computed in a recursive way [74, 76]. Note that

$$Z = Z^r_1 = Z^f_{L-a+1}.$$

The partial statistical sums Z^f_i and Z^r_i account for the contributions from all possible configurations of particles confined to the boxes $[1,i]$ and $[i,L]$, respectively. It can be shown [76] that

$$\begin{aligned}Z^f_i &= \prod_{j=1}^{i} \frac{1-O(j+1)+n(j+1)}{1-O(j)}, \\ Z^r_i &= \prod_{j=i}^{L} \frac{1-O(j)+n(j)}{1-O(j)},\end{aligned}$$

where $O(i)$ is the particle occupancy, or coverage, of bp i, defined as

$$O(i) = \sum_{j=i-a+1}^{i} n(j).$$

This gives the probability of finding bp i covered by a particle, and the only particles which cover bp i are the ones that start at bps: $i-a+1, i-a+2, \ldots, i$.

Using Equation (2.3), we reproduce the previous result from [76] which can be employed to find one-body energies from one-particle distribution in the case of hard-core interactions alone. We obtain

$$-\beta\left[u^0(i)-\mu\right] = \ln\left[\frac{n(i)}{1-O(i)+n(i)}\right] + \ln\left[\prod_{j=i}^{i+a-1}\frac{1-O(j)+n(j)}{1-O(j)}\right], \tag{2.16}$$

where $u^0(i)$ denotes the one-body energy corresponding to a hard rod, which occupies the lattice sites $i,\ldots,i+146$, in the case when the particles interact only through hard-core interactions, Φ^0, given by Equation (2.15).

We showed in this Section that there is a one-to-one correspondence between the energies u and Φ on one hand, and particle distributions n_1 and n_2 on the other. Thus, if the particle distributions n_1 and n_2 are known, the energies u and Φ can be inferred exactly, from Equations (2.13) and (2.14), and vice versa, if the energies u and Φ are known, the particle distributions n_1 and n_2 can be computed from Equations (2.3),(2.4). However, in many situations the two-particle distribution is not directly available from experiments. For example, high-throughput nucleosome maps simultaneously report nucleosome positions from many cells, effectively yielding a probabilistic description of the one-particle distribution n_1. Because of this averaging over single-cell configurations, information about the pair density profile n_2 cannot be directly extracted. Nonetheless, if the two-body interactions are sufficiently strong, the one-particle distribution profile n_1 can be used to obtain information about the two-body interaction, Φ.

2.2 Predicting two-body interactions from one-particle distribution

Let us introduce the dimensionless pair distribution

$$g(i,j) = \frac{n_2(i,j)}{n_1(i)n_1(j)}. \tag{2.17}$$

Note that $g(i,j) = \overline{n}_2(i,j)/[n_1(i)n_1(j)]$ for short distances $j-i < 2a$, and that $g(i,j) = g(j-i)$ in a homogeneous system. We start with a homogeneous system of N hard-rods which interact through an arbitrary nearest-neighbor potential Φ, and then develop an approximation for the inhomogeneous case. In a translation-invariant continuous system with nearest-neighbor interactions of arbitrary strength, we have that

$$e^{-\beta\Phi(d)} = Ce^{\alpha d}g(d),$$

where C and α are positive constants [117, 121, 122, 123]. The result can also be proved for a lattice fluid of hard rods, as shown below.

Using the formalism described in Appendix A, we can compute the partition function, $Q_N(L)$, of a system of N hard rods, restricted in a box of size L bp. Using this we can find the pressure of the system, p, and obtain the following relationship between

the partition function and the pressure

$$Q_N(L) = \left(e^{\beta pb}\right)^L \left[\widetilde{F}\left(e^{\beta pb}\right)\right]^N,$$

where $\widetilde{F}(z)$ is the z transform of $f(n) = e^{-\beta\Phi(n)}$, that is

$$\widetilde{F}(z) = \sum_{n=0}^{\infty} f(n) z^{-n}.$$

We use this result to compute the conditional probability of finding an adjacent particle at a distance d from the center of a fixed particle [121, 124] by

$$\begin{aligned} P(d) &= \text{Prob}(x_N = L - d | x_{N+1} = L) \\ &= \frac{1}{Q_N} \sum_{x_{N-1}=0}^{x_N} \cdots \sum_{x_1=0}^{x_2} f(x_1) \ldots f\big((L-d) - x_{N-1}\big) f\big(L - (L-d)\big) \\ &= f(d) \frac{Q_{N-1}(L-d)}{Q_N(L)} \\ &= e^{-\beta\Phi(d)} \frac{(e^{\beta pb})^{-d}}{\widetilde{F}(e^{\beta pb})}. \end{aligned}$$

In the bulk, far from the lattice edges, we have that $n_1(i) = \rho$, and $\overline{n}_2(i, i+d) = \rho P(d)$, for all i and for $d < 2a$. The pair distribution becomes

$$\begin{aligned} g(d) &= \frac{\overline{n}_2(i, i+d)}{n_1(i) n_1(i+d)} \\ &= \frac{\rho P(d)}{\rho^2} \\ &\propto e^{-\beta pbd} e^{-\beta\Phi(d)}. \end{aligned}$$

Thus in homogeneous systems, the interaction between the particles has the form

$$e^{-\beta\Phi(d)} = C e^{\alpha d} g(d), \tag{2.18}$$

where $\alpha = \beta pb$, and C is a normalization constant.

In a more general case, the external potential breaks translational invariance, making the dimensionless pair distribution, g, dependent on the absolute position of the first particle. However, if the two-body interaction, Φ, is translationally invariant, a good approximation is provided by replacing g in Equation (2.18) with

$$P_{\text{linker}}(\Delta) = \langle g(i, i+a+\Delta) \rangle_i,$$

where by $\langle\cdot\rangle_i$ we denote the average over all initial positions i. We obtain

$$-\beta\Phi(i,j) \approx \ln\left[P_{\text{linker}}(j-(i+a))\right] + \alpha(j-i) + \ln C. \tag{2.19}$$

The constants C and α are uniquely determined by the asymptotic condition

$$\lim_{(j-i)\to\infty} \Phi(i,j) = 0.$$

Equation (2.19) provides an ansatz for reconstructing the two-body interaction Φ from

$$P_{\text{linker}}(\Delta) = \langle \overline{n}_2(i, i+a+\Delta)/[n_1(i)n_1(i+\Delta)]\rangle_i.$$

Figure 2.2 shows a numerical test of this ansatz on a 10 kbp DNA segment. We construct a random one-body energy landscape and simulate strong inhomogeneity by positioning nine potential wells with depth of $5k_BT$ at $1, 2, \ldots, 9$ kbp on the landscape. The model interaction between a pair of particles separated by a linker of size Δ is

$$\Phi(\Delta) = 5\cos\left(\frac{2\pi}{10}\Delta\right)e^{-\Delta/50},$$

measured in units of k_BT. We use the one-body energies and the two-body potential as inputs in Equations (2.3) and (2.5) to compute the dimensionless pair distribution function. The pair distribution varies significantly from bp to bp [Figure 2.2(a)], as can be expected in a system with one- and two-body energies of comparable magnitude. Following our prescription, we compute P_{linker} by averaging over all the curves in Figure 2.2(a) [Figure 2.2(b)], and employ Equation (2.19) to infer the two-body interaction, Φ [Figure 2.2(c)]. The correlation coefficient between predicted and exact two-body interactions is greater than 0.999.

If a direct measurement of the pair distribution $\overline{n}_2$ is not available, P_{linker} needs to be estimated empirically from the n_1 profile. Each nucleosome positioning data set consists of the histogram of the number of nucleosomes starting at each genomic bp i. We preprocess these data by removing all counts of height 1 from the histogram, and smoothing the remaining counts with a $\sigma = 2$ Gaussian kernel. Next, we compute the density function, $n_1(i)$, by rescaling the smoothed profile so that the maximum occupancy for each chromosome is 1. Finally, we identify all local maxima on the n_1 profile

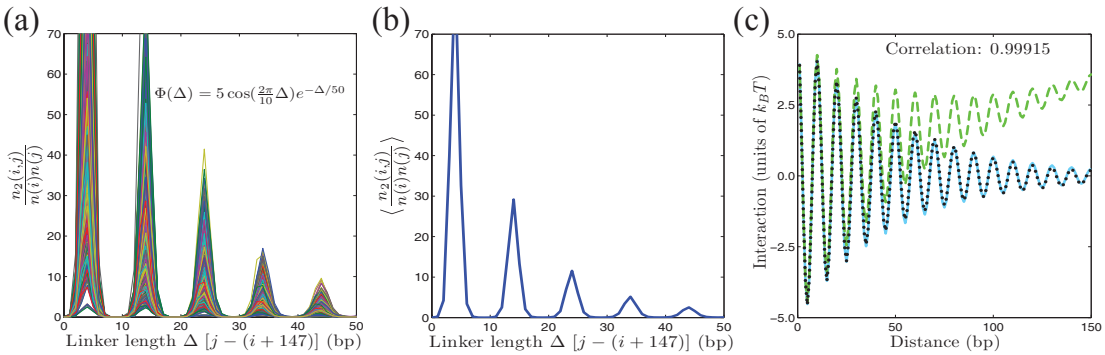


Figure 2.2: (a) $g(i,j) = n_2(i,j)/n_1(i)n_1(j)$ is plotted for a representative subset of all initial positions i in a 10^4 bp DNA segment. The one-body energies are randomly sampled from a Gaussian distribution with a mean of 2.5 k_BT and a standard deviation of 0.2 k_BT, and nine potential wells of depth 5 k_BT are added at $1, 2, \ldots, 9 \times 10^3$ bp to model a strongly inhomogeneous system. The particle distributions $n_1(i)$ and $n_2(i,j)$ are computed from one- and two-body energies, using Equations (2.3) and (2.5). (b) The function $P_{\text{linker}}(\Delta)$ is obtained by averaging $g(i,j)$ over all initial positions i. Note that $\Delta = j - (i + 147)$ represents the linker length between the two nucleosomes with starting positions i and j, respectively. (c) Exact (solid blue line) and predicted (dotted black line) two-body interactions. The predicted interaction is computed from the $-\ln(P_{\text{linker}})$ curve (dashed green line) using Equation (2.19).

and assume that they mark prevalent nucleosome positions. Specifically, for each maximum at bp i we find subsequent maxima at positions $i+146 < j_1 < j_2 < j_3 < \ldots$ in the 50 bp window. To each pair of maxima (i, j_1), $(i, j_2), \ldots$ we assign the probability that they represent neighboring nucleosomes, $n_1(i)n_1(j_1), n_1(i)[1-n_1(j_1)]n_1(j_2), \ldots$, respectively. We sum the probabilities over all initial positions i and normalize, producing an empirical estimate of P_{linker}.

Figure 2.3 demonstrates our procedure in a model system, where the preprocessing and rescaling steps were skipped since the simulated n_1 profile is noise-free and already properly normalized. Specifically, we use local maxima in the nucleosome starting probability profile [inset of Figure 2.3(a)] to obtain P_{linker} [Figure 2.3(b)]. Figure 2.3(d) shows that the two-body interaction can be reconstructed using Equation (2.19), even in the presence of one-body energies with the same periodicity. The reconstruction is facilitated by the presence of potential wells or barriers in the one-body energy profile that are strong enough to create non-uniform density of nearby nucleosomes. To find the

one-body energies, we substitute the predicted two-body energy, Φ, into Equation (2.3), which we solve numerically for z [Figure 2.3(c)]. Nucleosome occupancies inferred from predicted energies u and Φ are virtually identical to the exact profile [Figure 2.3(a)].

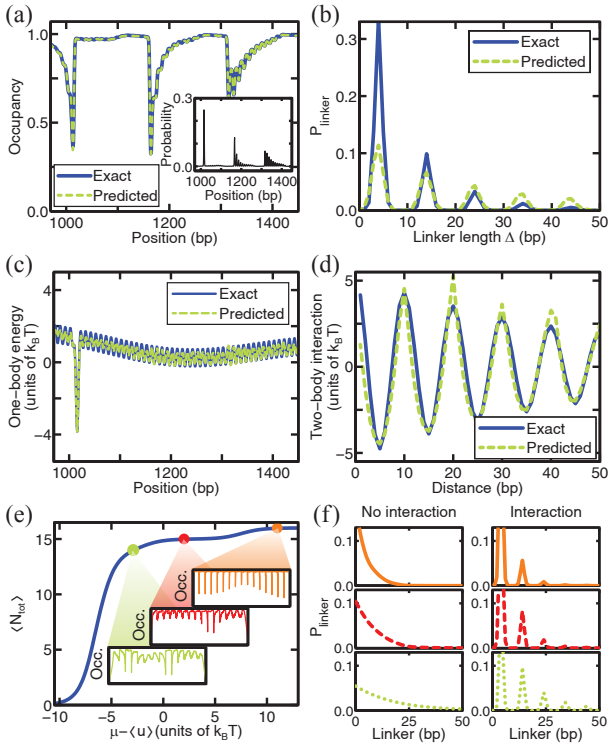


Figure 2.3: A model with 10 bp oscillations in both one-body and two-body energies. The two-body interaction is $\Phi(x) = A\cos\left(\frac{2\pi}{10}x\right)e^{-x/b}$, where $A = 5\ k_BT$ and $b = 50$ bp. For the one-body potential, 10 bp oscillations with the 0.5 k_BT amplitude are superimposed onto a smooth energy profile with two $-5\ k_BT$ potential wells separated by 1000 bp. DNA length of 2416 bp is chosen to be able to position 16 nucleosomes with 151 bp repeat length. The occupancy profile (a), the linker length distribution (b), the one-body energy (c), and the two-body interaction (d): exact (solid blue line) and predicted (dashed green line). We set $\mu - \langle u\rangle = -1\ k_BT$ in (a)–(d). The inset of (a) shows the probability of starting a nucleosome at a given bp. (e) Average number of nucleosomes $\langle N_{tot}\rangle$ vs. $\mu - \langle u\rangle$. The insets show occupancy profiles corresponding to three different chemical potentials. (f) Linker length distributions for three values of $\langle N_{\rm tot}\rangle$ shown as points in (e), with and without two-body interactions.

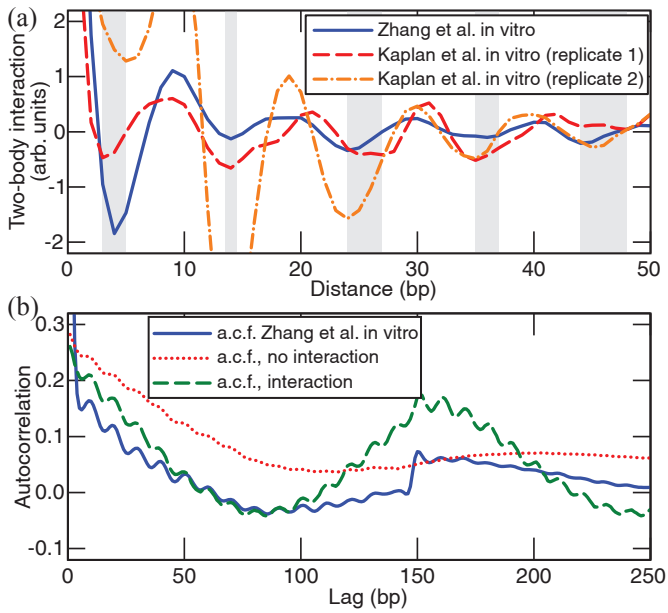


Figure 2.4: (a) Two-body interaction Φ inferred from *in vitro* maps of nucleosome positions [57, 75]. Gray bars indicate consensus positions of the minima. (b) Autocorrelation of nucleosome starting positions in one of the *in vitro* data sets [75], and of starting positions predicted using sequence-specific one-body energies from the "spatially resolved" model [76], with and without Φ. The two-body potential is from Figure 2.3, consistent with the minima of Φ observed in (a). The one-body energies have $\sigma = 0.23\ k_BT$. To account for the limited size of the *in vitro* data set, model output was degraded by randomly removing 1% of predicted nucleosome probabilities.

As the chemical potential μ is increased, nucleosomes undergo a transition in which their average number goes up in a step-like fashion [Figure 2.3(e)] [125]. In contrast to the $\Phi = 0$ case, in which linkers are distributed exponentially, the two-body interactions lead to the pronounced discretization of linker lengths [Figure 2.3(f)]. The first minimum of the two-body energy, Φ, becomes more dominant as the number of nucleosomes increases, leading to a well-positioned array with 4-bp-long linkers.

We now use Equation (2.19) to predict nearest-neighbor interactions, Φ, from genome-wide nucleosome maps [Figure 2.4(a)]. We find that despite significant experiment-to-experiment variations, all two-body potentials have minima within 1-2 bp of $5+10m$ bp, for $m = 0, 1, \ldots$ [98]. Surprisingly, there are substantial differences between two Kaplan *et al.* [57] *in vitro* replicates, with one replicate exhibiting higher values of the interaction Φ due to the pronounced depletion of nucleosomes separated by less than ten bps of DNA. Apparently, chromatin structure can undergo subtle uncontrolled changes

from experiment to experiment.

Two-body interactions are reflected in the autocorrelation of nucleosome starting positions [Figure 2.4(b)]. The oscillations in the autocorrelation function are suppressed when nucleosome positions are predicted using a sequence-specific model which neglects two-body interactions [76]. This "spatially resolved" model assigns mono- and dinucleotide energies independently at each position within the nucleosomal site, and is thus capable of capturing the 10–11 bp periodicity of one-body interactions. We find that the autocorrelation function is much closer to experiment if the two-body potential is included into the model [Figure 2.4(b)].

2.3 Sequence-specific energy of nucleosome formation

We can extract a sequence-specific component of the one-body energy by using Equations (2.13) or (2.16) to compute $u-\mu$, estimating the chemical potential, μ, and fitting the one-body energy, u, to a linear model which assigns energies to nucleotide words found within the $a = 147$ bp nucleosomal site. Assuming that the system is nearly homogeneous, we use Equations (A.3) or (A.8) from Appendix A to obtain the chemical potential of the lattice gas. After eliminating the chemical potential, μ, we fit a linear model to one-body energies, u. It was established in [76] that position-independent models, in which the energy of the nucleotide word does not depend on its exact location within the nucleosome, can be used to describe genome-wide nucleosome occupancies. Furthermore, an $N = 2$ position-independent model with just 13 fitting parameters performed as well as $N > 2$ models, where N denotes the longest word (in bp) included into the model. If both monomers and dimers contribute to the total one-body energy, the sequence-specific binding energy of a 147 bp-long nucleosomal site is given by

$$u^S = \sum_{\alpha} m_\alpha \epsilon_\alpha + \sum_{\alpha,\beta} m_{\alpha\beta} \epsilon_{\alpha\beta} + \epsilon_0, \tag{2.20}$$

where m_α is the number of nucleotides of type $\alpha \in \{A, C, G, T\}$, ϵ_α is the energy of the nucleotide α, and ϵ_0 is the overall sequence-independent offset. Similarly, $m_{\alpha\beta}$ is the number of dinucleotides of type $\alpha\beta$, and $\epsilon_{\alpha\beta}$ is the corresponding energy. In [76], word

energies were constrained by

$$\sum_{\alpha} \epsilon_{\alpha} = \sum_{\alpha} \epsilon_{\alpha\beta} = \sum_{\beta} \epsilon_{\alpha\beta} = 0,$$

yielding a 13-parameter model. Here we develop an alternative approach which does not impose any additional constraints beyond those caused by the fact that the number of mono- and dinucleotides in the 147 bp-long site is fixed.

We can express the nucleosome energies as $\mathbf{u} = \mathbf{M}\mathbf{x}$, or equivalently

$$\begin{pmatrix} u^S(1) \\ u^S(2) \\ \vdots \\ u^S(l_{max}) \end{pmatrix} = \begin{pmatrix} m_{1,1} & \cdots & m_{1,20} & 1 \\ m_{2,1} & \cdots & m_{2,20} & 1 \\ \vdots & & \vdots & \vdots \\ m_{l_{max},1} & \cdots & m_{l_{max},20} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{21} \end{pmatrix},$$

where $u^S(i)$ is the sequence-specific energy of the nucleosome that covers the DNA sequence between bps i and $i + a - 1$. We let $m_{i,1}, \ldots, m_{i,4}$ denote the number of A, C, G and T nucleotides found in that sequence, and $m_{i,5}, \ldots, m_{i,20}$ give the number of dinucleotides $AA, AC, \ldots, TG, TT$. The quantity $l_{max} = L - a + 1$ is the maximum starting position for a nucleosome, and the set of parameters $x_1, \ldots x_{21}$ represents the 21 energies from Equation (2.20), that is

$$\begin{aligned} x_1 &= \epsilon_A, \\ x_2 &= \epsilon_C, \\ &\ldots \\ x_{20} &= \epsilon_{TT}, \\ x_{21} &= \epsilon_0. \end{aligned}$$

Note that for any DNA sequence of length 147 bp,

$$\begin{aligned} m_{i,1} + m_{i,2} + m_{i,3} + m_{i,4} &= 147\ m_{i,21}, \\ m_{i,5} + m_{i,6} + \ldots + m_{i,20} &= 146\ m_{i,21}. \end{aligned}$$

This means that the rank of $\mathbf{M}$ is 19. For any linear operator, $\mathbf{M}$, the dimension of its domain (21 in our case) is equal to the sum between the dimensions of the image,

$\mathrm{im}(\mathbf{M})$, and of the kernel, $\ker(\mathbf{M})$. In our case, we have

$$\dim \ker(\mathbf{M}) = 2,$$
$$\dim \mathrm{im}(\mathbf{M}) = 19.$$

It is easy to check that the vectors

$$\mathbf{x}^* = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ -147 \end{pmatrix},$$
$$\mathbf{x}^{**} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \\ 1 \\ -146 \end{pmatrix},$$

belong to $\ker(\mathbf{M})$, so that $\ker(\mathbf{M}) = \mathrm{span}(\mathbf{x}^*, \mathbf{x}^{**})$. Any vector of parameters $\mathbf{x}$ can be uniquely decomposed as

$$\mathbf{x} = \mathbf{x}_K + \mathbf{x}^\perp,$$

with $\mathbf{x}_K$ belonging to $\ker(\mathbf{M})$, and $\mathbf{x}^\perp$ belonging to $\ker(\mathbf{M})^\perp$, the subspace orthogonal to $\ker(\mathbf{M})$. Specifically,

$$\mathbf{x}^\perp = \mathbf{x} - (\mathbf{x}^*, \mathbf{x})\mathbf{x}^* - (\mathbf{x}^{**}, \mathbf{x})\mathbf{x}^{**}.$$

The component $\mathbf{x}_K$ does not contribute to the energy of the sequence because $\mathbf{M}\mathbf{x}_K$ is

the null vector. The components of $\mathbf{x}^{\perp}$ satisfy the following relations

$$(\mathbf{x}^{\perp}, \mathbf{x}^{*}) = 0 \quad \Longrightarrow \quad \sum_{i=1}^{4} x_i^{\perp} - 147\ x_{21}^{\perp} = 0,$$
$$(\mathbf{x}^{\perp}, \mathbf{x}^{**}) = 0 \quad \Longrightarrow \quad \sum_{i=5}^{20} x_i^{\perp} - 146\ x_{21}^{\perp} = 0.$$

Thus, $\mathbf{x}^{\perp}$ containes only 19 independent parameters, and

$$x_{21}^{\perp} = \frac{1}{147} \sum_{i=1}^{4} x_i^{\perp},$$
$$x_{20}^{\perp} = 146\ x_{21}^{\perp} - \sum_{i=5}^{19} x_i^{\perp} = \frac{146}{147} \sum_{i=1}^{4} x_i^{\perp} - \sum_{i=5}^{19} x_i^{\perp},$$

which implies that

$$\mathbf{x}^{\perp} = \begin{pmatrix} x_1^{\perp} \\ \vdots \\ x_{19}^{\perp} \\ \frac{146}{147} \sum_{i=1}^{4} x_i^{\perp} - \sum_{i=5}^{19} x_i^{\perp} \\ \frac{1}{147} \sum_{i=1}^{4} x_i^{\perp} \end{pmatrix} \tag{2.21}$$

In order to compare two different sets of 21 energies (e.g. fit on different genomes), we need to eliminate the components of the two vectors included in $\ker(\mathbf{M})$. The components from $\ker(\mathbf{M})^{\perp}$ have 19 independent parameters and two redundant ones [Equation (2.21)]. The projection of the energy vector on the $\mathrm{im}(\mathbf{M})$ hyperplane is unique, and there is a one-to-one correspondence between $\mathrm{im}(\mathbf{M})$ and the parameter subspace which is orthogonal to the kernel, $\ker(\mathbf{M})^{\perp}$. In this way, every set of fitted energies uniquely determines a set of parameters, $\mathbf{x}^{\perp}$, and a sequence-specific energy, $\mathbf{u} = \mathbf{M}\mathbf{x}^{\perp}$.

For the $N = 1$ model, $\ker(\mathbf{M})$ is spanned by a single vector,

$$\mathbf{x}^{*} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -147 \end{pmatrix},$$

and $\mathbf{x}^{\perp}$ has four relevant parameters and a redundant one,

$$\mathbf{x}^{\perp} = \begin{pmatrix} x_1^{\perp} \\ x_2^{\perp} \\ x_3^{\perp} \\ x_4^{\perp} \\ \frac{1}{147}\sum_{i=1}^{4} x_i^{\perp} \end{pmatrix}.$$

Similarly, the $N = 3$ model has 85 $(4 + 16 + 64 + 1)$ fitting parameters, and there are six independent constraints on the columns of $\mathbf{M}$, so that the rank of the operator $\mathbf{M}$ is 79. The kernel of $\mathbf{M}$ is spanned by six vectors, and the parameter subspace orthogonal to the kernel, which gives the sequence energy is 79-dimensional. For the $N = 4$ and the $N = 5$ models, the total number of parameters is 341 and 1365, and the number of independent parameters is 319 and 1279, respectively.

When the $N = 2$ model described above, with 21 parameters, is trained on the energies predicted by applying Equation (2.16) to a large-scale map of nucleosomes reconstituted *in vitro* on yeast genomic DNA [75], it captures the same sequence determinants as our previously used 13-parameter model which employs additional constraints [76]. We obtain a correlation coefficient of $r = 0.9967$ between the two sequence-specific energy profiles. However, the two approaches are not equivalent, since the 21-parameter model utilizes the maximum possible number of independent fitting parameters.

2.4 Applications

In the last Section of this Chapter, we present a few applications of the theoretical framework which we described above.

2.4.1 Reconstructing nucleosome energetics in a model system

In the absence of nearest-neighbor interactions induced by chromatin structure, nucleosome formation *in vitro* is fully controlled by DNA sequence and steric exclusion. In this case, efficient procedures are available for reconstructing nucleosome positions

from formation energies [74, 126], and for inferring nucleosome energetics from experimentally available probability and occupancy profiles using Equation (2.16) [76]. However, this simple approach may lead to errors if the two-body interactions are in fact present in the system. Furthermore, many factors other than DNA sequence can affect nucleosome positioning *in vivo*, including chromatin remodeling enzymes, non-histone DNA-binding factors, and components of transcriptional machinery [88, 89, 90]. These influences are expected to create potential barriers which prevent nucleosomes from forming in certain regions, and potential wells which localize nucleosomes through favorable contacts between histones and other proteins. These effects will be lost if a purely sequence-specific model is fit to the nucleosome positioning data.

We use a simple model system to illustrate the errors caused by neglecting higher-order chromatin structure and *in vivo* potentials (Figure 2.5). We generate a random DNA fragment with length of 10^4 bp, and compute the sequence-dependent one-body energies using the 21-parameter, $N = 2$ position-independent model (see Section 2.3). The sequence-specific word energies for the model are randomly sampled from a uniform distribution, from the interval $[-0.02, +0.02]$ k_BT. Figure 2.5(a) shows sequence-dependent nucleosome energies in a representative 500 bp window (blue solid line). The window also includes one of the 3 k_BT wells placed every 2000 bp throughout the sequence to model *in vivo* effects; the total one-body energy is shown as a green dash-dot-dot line.

The total energy is used together with the two-body interaction shown in Figure 2.5(b) (blue solid line) to construct the exact one-body density profile, n_1, and the corresponding nucleosome occupancy for the DNA segment [Figure 2.5(c), green dash-dot-dot line]. If we now use Equation (2.16), which neglects the two-body interactions, to compute the one-body energies, u^0, from the nucleosome density, n_1, the predicted energy profile captures the potential wells and the sequence-specific component, but also displays spurious 10 bp oscillations caused by the "leakage" of the two-body potential, Φ, into one-body energetics [Figure 2.5(a), red dashed line]. In addition, the whole landscape is shifted downward because favorable two-body interactions are missing from the model. The "leaked" oscillations and the *in vivo* wells have no relation

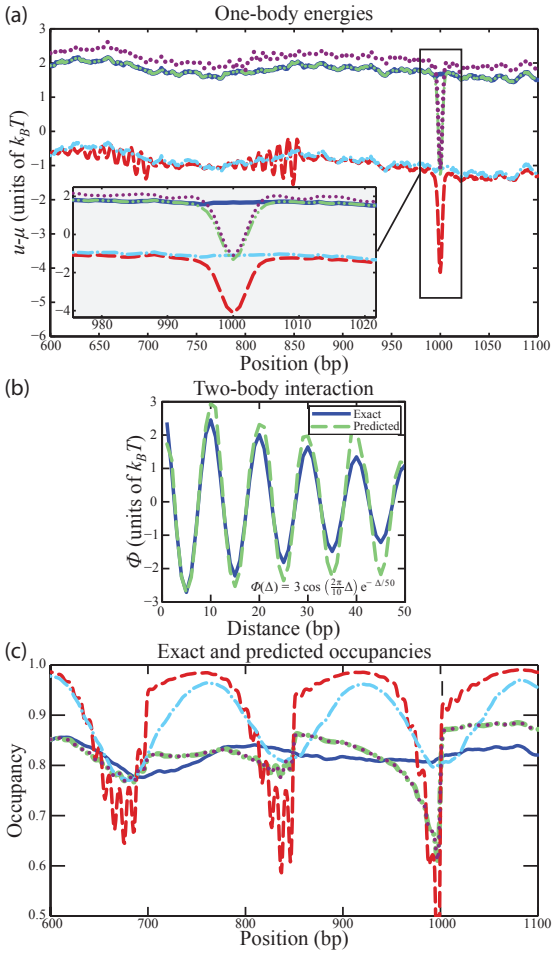


Figure 2.5: (a) One-body energies. Sequence-dependent energy given by the 21-parameter model [Equation (2.20)] (blue solid line), total energy given by the sum of the sequence-specific energies and 5 potential wells with 3 k_BT depth at $1, 3, 5, 7$, and 9 $\times 10^3$ bp designed to mimic the *in vivo* effects (green dash-dot-dot line). Energy predicted with a model that neglects two-body interactions [Equation (2.16)] (red dashed line), energy predicted by fitting the 21-parameter model to the energies from Equation (2.16) (light blue dash-dot line), a numerical solution of the full model which takes Φ into account (maroon dotted line). Inset: zoom-in on the region with the potential well. (b) Exact two-body interaction Φ (blue solid line) and predicted interaction [Equation (2.19)] (green dashed line). (c) Nucleosome occupancies. Occupancy generated by the exact sequence-specific one-body energy and the exact interaction (blue solid line), occupancy corresponding to the combined exact one-body energy (sequence-specific component and potential wells) and the exact interaction (green dash-dot-dot line). Predicted occupancy generated by the one-body energy from Equation (2.16) and predicted Φ (red dashed line), occupancy generated using predicted sequence-dependent one-body energy [Equation (2.20)] and predicted Φ (light blue dash-dot line), occupancy predicted using numerically computed one-body energies from the full model and predicted Φ (maroon dotted line).

to sequence and therefore can be removed by fitting either the 13- or the 21-parameter model to the predicted binding energy, u^0 [Figure 2.5(a), light blue dash-dot line]. The two predicted energy profiles are highly correlated with each other ($r = 0.9993$) and with the exact profile, u ($r = 0.9913$ for the 13-parameter model, and $r = 0.9915$ for the 21-parameter model), indicating that the sequence-specific component can be extracted even if the two-body interactions are not handled correctly.

Predicting occupancies from the energy profiles constructed under the $\Phi = \Phi^0$ assumption [Equation (2.15)], causes discrepancies with the exact result [Figure 2.5(c), green dash-dot-dot line]. For example, using the one-body energies, u^0, predicted with Equation (2.16) [Figure 2.5(a), red dashed line], and the two-body potential predicted with Equation (2.19) [Figure 2.5(b), green dashed line], gives an occupancy profile with higher average occupancy, sharp peaks, and enhanced 10 bp oscillations compared to the exact landscape [Figure 2.5(c), red dashed line]. This is not unexpected because the two-body potential is both imprinted in the one-body profile and included explicitly. In contrast, if $\Phi = \Phi^0$ [Equation (2.15)] is assumed at this stage as well, the exact occupancy can be restored from the one-body energy, u^0, but the origin of various contributions remains unclear as they are all lumped into the one-body landscape.

When the 21-parameter model is fit to the u^0 profile [Figure 2.5(a), light blue dash-dot line] and combined with the predicted Φ [Figure 2.5(b), green dashed line], the occupancy is off since *in vivo* potential wells cannot be captured by this model [Figure 2.5(c), light blue dash-dot line]. Nucleosomes are not strongly localized if the *in vivo* wells and barriers are absent [Figure 2.5(c), blue solid line], consistent with the relatively smooth *in vitro* occupancy profiles [57, 75]. Note that the two occupancy profiles will coincide if the mean of the predicted one-body energies is set to the correct value, eliminating the spurious offset caused by the two-body interaction Φ [Figure 2.5(a), compare blue solid and light blue dash-dot lines].

In order to reconstruct the occupancy correctly and avoid mixing one-body and two-body contributions, we need to turn to the full theory developed in Section 2.1. Inserting the predicted two-body interaction, Φ, [Figure 2.5(b), green dashed line] into Equation (2.3), and using the exact one-particle distribution profile, n_1, we obtain a

system of nonlinear equations which can be solved numerically, yielding energies that are very close to the exact result [Figure 2.5(a), maroon dotted line]. These energies and the predicted interaction, Φ, can be used to reconstruct the occupancy profile which is nearly exact [Figure 2.5(c), compare green dash-dot-dot and maroon dotted lines]. Thus we have succeeded in separating the one- and two-body energies, and in splitting off the sequence-dependent part in the former. However, the full procedure is computationally intensive and becomes inefficient if the DNA is much longer than 10^4 bp. However, longer segments may be split into manageable pieces and handled separately.

2.4.2 Nucleosome localization by potential barriers and wells

Nucleosomes in the vicinity of potential barriers and wells can be localized by steric exclusion alone [56]. This mechanism is thought to contribute to prominent nucleosome occupancy peaks in genic regions observed *in vivo*, but not *in vitro* [57, 75, 60, 67, 87, 127]. In order to understand the nature and the extent of *in vivo* nucleosome localization, we need to study nucleosome occupancy patterns created by placing a single potential barrier or potential well onto an otherwise flat one-body energy landscape.

In Figure 2.6 we show the nucleosome occupancy induced by a symmetric Gaussian barrier, with and without two-body interactions. As the chemical potential is changed to increase the average nucleosome occupancy, the oscillations become more prominent. Without two-body interactions, the peak situated closest to the barrier is always the highest and the occupancy pattern is a decaying oscillation [Figure 2.6(a)]. Strikingly, including the two-body interaction, Φ, results in a markedly different occupancy profile – oscillations are more persistent and the peak situated closest to the barrier is not always the highest one [Figure 2.6(b)].

The degree of nucleosome localization is also controlled by the width of the Gaussian barrier, in the sense that wider barriers induce less prominent oscillations, but produce stronger occupancy depletion over the potential barrier itself [Figures 2.6(c) and 2.6(d)]. The barrier height also controls the degree of depletion [Figures 2.6(e) and 2.6(f)]. Interestingly, increasing the strength of two-body interactions results in a

higher average occupancy and produces shorter peak-to-peak distances [Figure 2.6(g)]. In fact, the peak-to-peak distances, which can be interpreted as the sum of the 147 bp nucleosomal site and a linker, can be varied in a wide range, by changing either the chemical potential, μ, or the strength of the interaction, Φ, [Figure 2.6(h)].

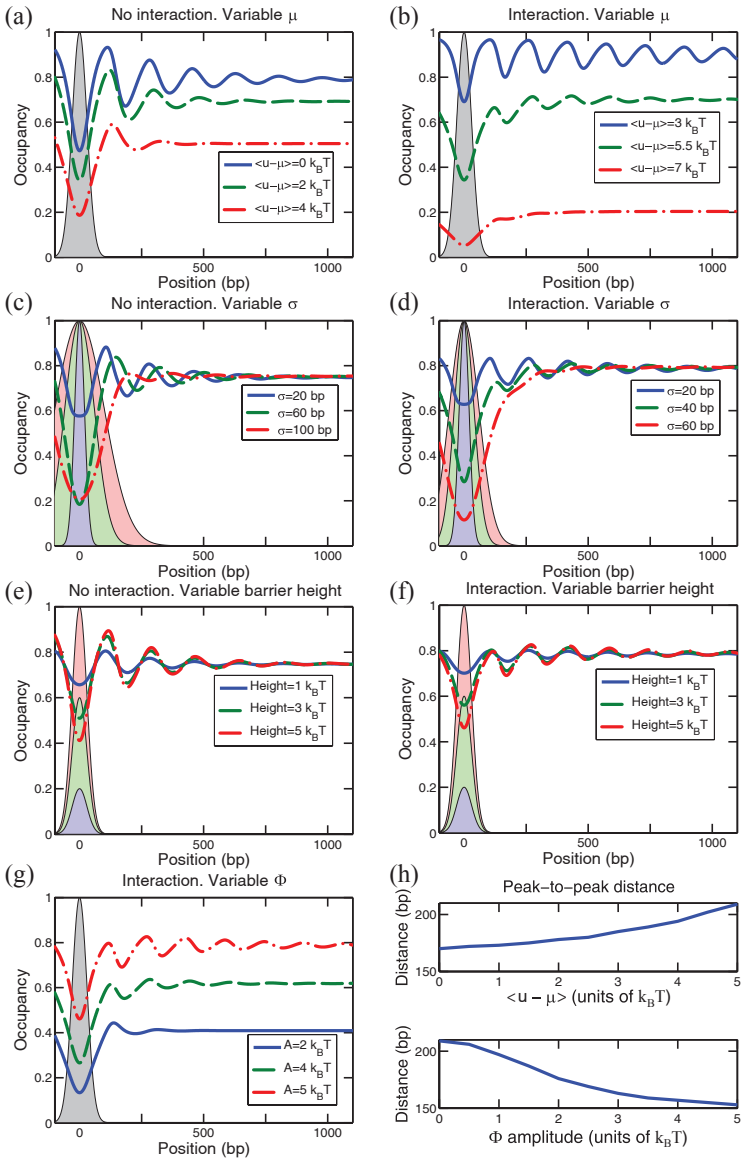


Figure 2.6: Symmetric Gaussian barrier. Occupancy profiles for the following scenarios: variable chemical potential, μ, (a) and (b), variable barrier width (c) and (d), and variable barrier height (e) and (f). Unless otherwise specified in the legend, the barrier heights are 5 k_BT, $\sigma = 30$ bp, and $\langle u-\mu\rangle = 5\ k_BT$ [in panel (c) $\langle u-\mu\rangle = 1\ k_BT$]. Panels (b), (d) and (f) have a two-body interaction $\Phi(\Delta) = A\cos(2\pi\Delta/10)\exp(-\Delta/50)$, with $A = 5\ k_BT$. (g) Occupancy profiles for variable interaction strength A. (h) Variation of the typical distance between neighboring nucleosomes as μ or A is varied. In the upper panel, we use $\Phi = \Phi^0$ [Equation (2.15)], and in the lower panel, we use $\langle u\rangle - \mu = 5\ k_BT$.

Similar conclusions can be reached if the Gaussian barrier is replaced by a symmetric Gaussian potential well (Figure 2.7). Oscillations decay less rapidly with two-body interactions, and the extent of oscillations is controlled by the chemical potential and by the depth and the width of the well. However, in this case the nucleosome situated closest to the well is always the most localized. Nucleosome occupancy in the vicinity of 5' NDRs is prominently asymmetric [75, 67]. This asymmetry can be modeled by a combination of a symmetric barrier with an adjacent potential well [94], or by a single asymmetric barrier. In Figure 2.8, we show how nucleosome localization and the degree of asymmetry in the occupancy profile vary with the chemical potential, μ, the strength of the two-body interaction, Φ, the height of the barrier, and the degree of its asymmetry.

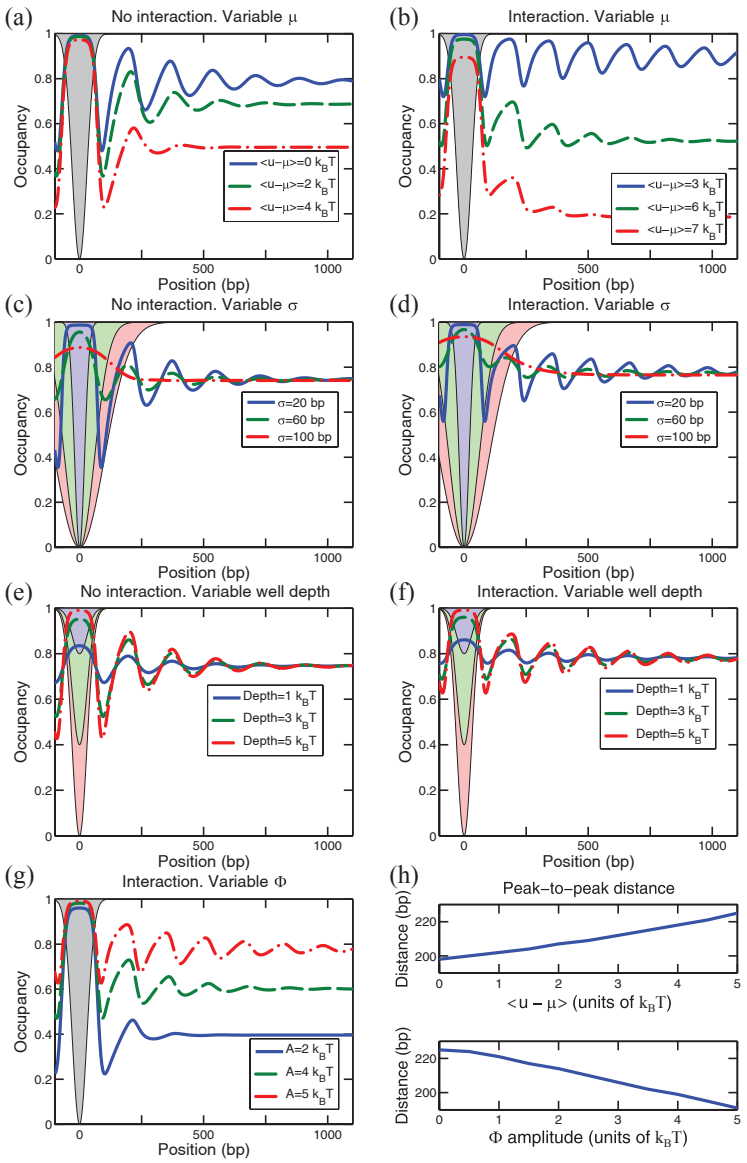


Figure 2.7: Symmetric Gaussian well. Occupancy profiles for the scenarios described in Figure 2.6. All the parameters not explicitly given in the legends are from Figure 2.6. In particular, well depths have the same magnitude as the heights of the corresponding barriers.

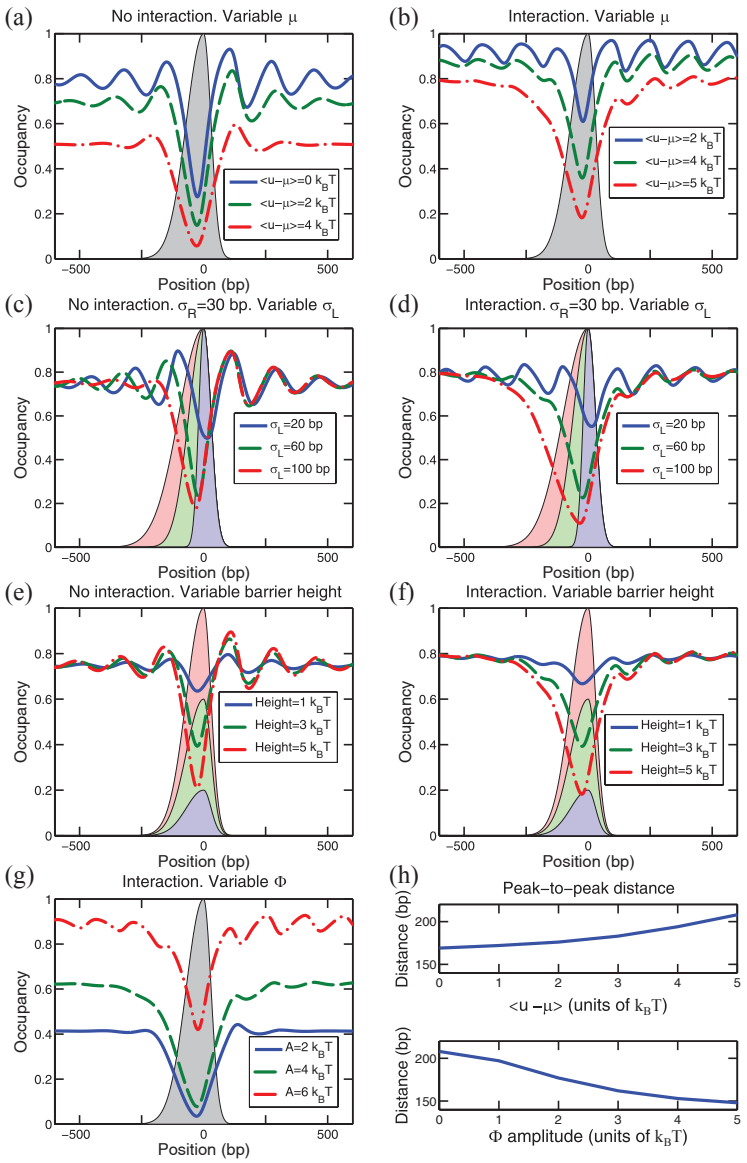


Figure 2.8: Asymmetric Gaussian barrier. Occupancy profiles for the scenarios described in Figure 2.6. Unless otherwise specified in the legend, the barrier heights are 5 k_BT, $\sigma_L = 70$ bp, $\sigma_R = 30$ bp, and $\langle u - \mu \rangle = 5\ k_BT$. In panel (c), we use $\langle u - \mu \rangle = 1\ k_BT$.

In summary, the two-body interactions, Φ, significantly modify nucleosome occupancy profiles, affecting the heights and spacings of the observed nucleosome localization peaks. However, as the interaction itself only couples neighboring nucleosomes but does not determine their absolute positions, a potential barrier or well is required to

achieve localization in the first place. Increasing the width of this feature diminishes its localization capacity. The interaction favors configurations with linker lengths corresponding to the minima of the interaction Φ, leading to linker length discretization [97, 98, 94]. By changing the interaction strength or the chemical potential, one can create occupancy patterns with different average linker lengths.

2.4.3 Modeling nucleosome occupancy over transcribed regions

The characteristic patterns of nucleosome occupancy in the region between 5' and 3' NDRs are shown in Figure 2.9. There is a pronounced lack of nucleosome localization *in vitro* [57, 75] [Figures 2.9(a) and 2.9(b)]. The 21-parameter $N = 2$ position-independent model captures this liquid-like behavior correctly, but is unable to account for the *in vivo* peaks. Since DNA sequence alone clearly cannot produce the observed degree of *in vivo* localization, we sought to construct a minimal model in which potential barriers of non-sequence origin flank each gene, and the one-body energy landscape is flat otherwise [68, 128] (Figure 2.10).

In the Kaplan et al. dataset [57], the first nucleosome is in fact the most localized and the average profile is consistent with the absence of two-body interactions [Figure 2.9(a)]. In contrast, Zawadzki et al. [127] and Mavrich et al. [87] profiles appear to be shaped by the higher-order chromatin structure [Figure 2.9(c)]. This experimental discrepancy may have resulted from under-digesting chromatin with MNase [129]. In addition, the number of active genes that presumably reside in more open, active chromatin characterized by weaker two-body interactions could vary between experiments. However, in all three cases, the *in vivo* barriers are necessary to reproduce observed localization patterns. The 5' NDR is strongly asymmetric [Figures 2.9(a) and 2.9(c)], and thus needs to be modeled either with a combination of a symmetric barrier and a potential well for the +1 nucleosome, or with a single asymmetric barrier (Figure 2.8). The height of each barrier in Figure 2.10 is adjusted to reproduce the extent of observed nucleosome depletion.

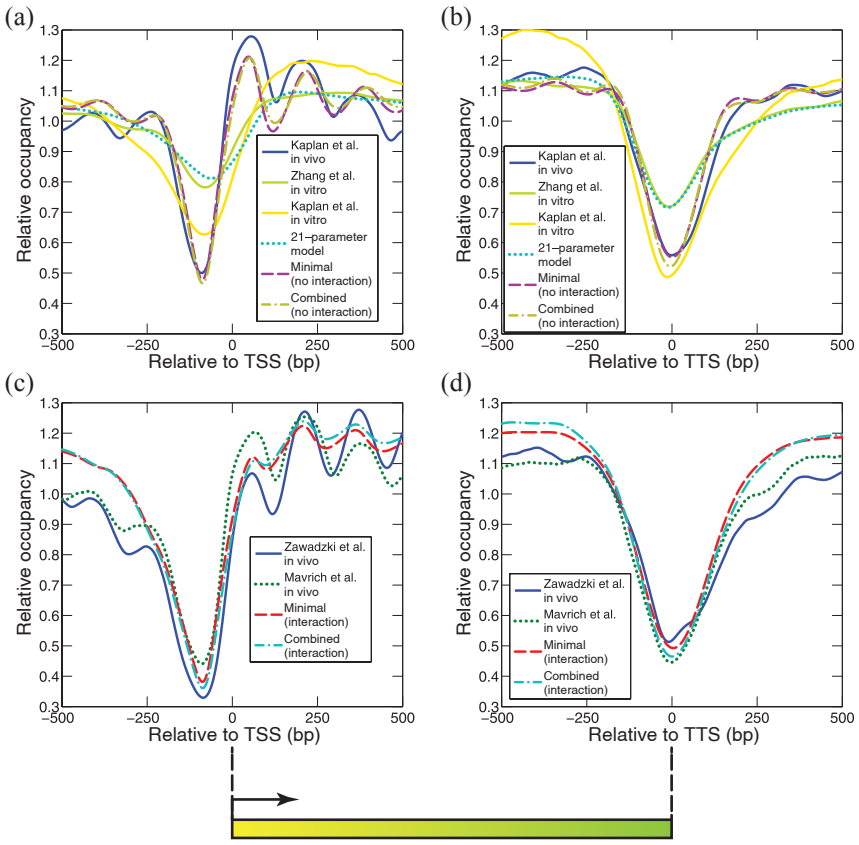


Figure 2.9: Average nucleosome occupancy in the vicinity of transcription start and termination sites (TSS and TTS, respectively). Each occupancy profile is normalized by its average in the $[-500, 500]$ bp window. (a), (b): Nucleosome occupancy observed *in vivo* (YPD medium) and *in vitro* by Kaplan et al. [57], and *in vitro* by Zhang et al. [75], and predicted using a 21-parameter $N = 2$ position-independent model, a minimal model in which nucleosomes are localized purely by means of sequence-independent potential barriers (Figure 2.10), and a combined model in which sequence-specific energies from the 21-parameter $N = 2$ model are added to the barriers from Figure 2.10. The two-body potential is turned off. Note that in [75], DNA was mixed with histones in a 1:1 mass ratio which is close to the *in vivo* value, while in [57], the ratio was 0.4:1, resulting in deeper NDRs. (c), (d): Nucleosome occupancy observed *in vivo* by Zawadzki et al. [127] and Mavrich et al. [67], and predicted using the 21-parameter $N = 2$ position-independent model, the minimal model, and the combined model. The two-body potential is given by $\Phi(\Delta) = A\cos(2\pi\Delta/10)\exp(-\Delta/50)$, with $A = 5\ k_BT$.

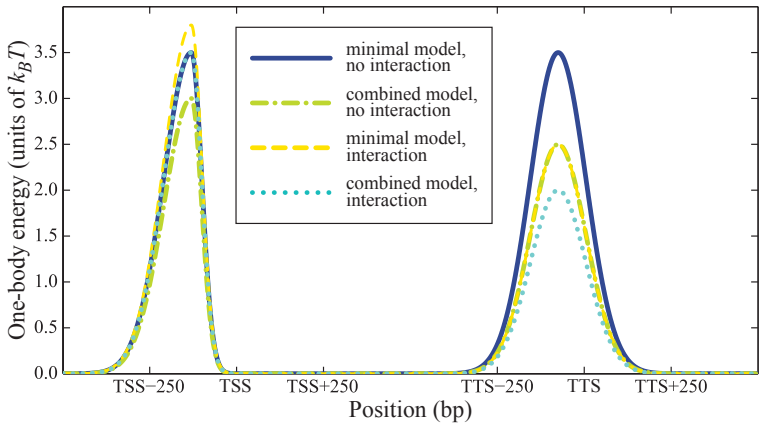


Figure 2.10: The one-body energy profiles used in Figures 2.9 and 2.11. The 5' asymmetric barrier has $\sigma_{\text{left}} = 80$ bp and $\sigma_{\text{right}} = 30$ bp. The 3' symmetric barrier has $\sigma = 80$ bp. Solid blue line: barriers used in the *in vivo* minimal model without two-body interactions [Figures 2.9(a) and 2.9(b), Figure 2.11]. Dash-dot green line: barriers used in the *in vivo* combined model without two-body interactions [Figures 2.9(a) and 2.9(b), Figure 2.11]. Dashed yellow line: barriers used in the *in vivo* minimal model with two-body interactions [Figures 2.9(c) and 2.9(d)]. Dotted light blue line: barriers used in the *in vivo* combined model with two-body interactions [Figures 2.9(c) and 2.9(d)]. The landscapes shown in the Figure are shifted vertically so that $\langle u - \mu \rangle = 0.56\ k_BT$ in the minimal model without two-body interactions, $0.62\ k_BT$ in the combined model without two-body interactions, $4.49\ k_BT$ in the minimal model with two-body interactions, and $4.62\ k_BT$ in the combined model with two-body interactions.

The average occupancy profiles are not significantly altered if sequence-specific energies from the 21-parameter $N = 2$ position-independent model are added to the barriers from Figure 2.10 (Figure 2.9, compare the combined and minimal models). The $N = 2$ model yields a standard deviation of $0.61\ k_BT$ for the energies genome-wide, consistent with the assumption that the sequence-dependent energies should be less than $1\ k_BT$, and thus the one-body landscape is still dominated by the barriers. Note that the barrier heights are reduced in the combined model because sequence-dependent nucleosome depletion over NDRs is now included explicitly.

The difference between the minimal and the combined models is more pronounced if individual occupancy profiles are displayed as a heat map (Figure 2.11). Minimal model barriers adjacent to each other on the genomic sequence, that is the 5' barriers of two divergent genes sharing a single promoter, sometimes create anomalous NDRs

with the extent of nucleosome depletion, not observed in the data [Figures 2.11(e) and 2.11(f)]. Interestingly, these effects are reduced when sequence specificity is combined with the minimal model [Figures 2.11(g) and 2.11(h)]. Comparing barrier heights in the minimal and combined models (Figure 2.10), we conclude that intrinsic histone-DNA interactions are responsible for less than 30% of the barriers.

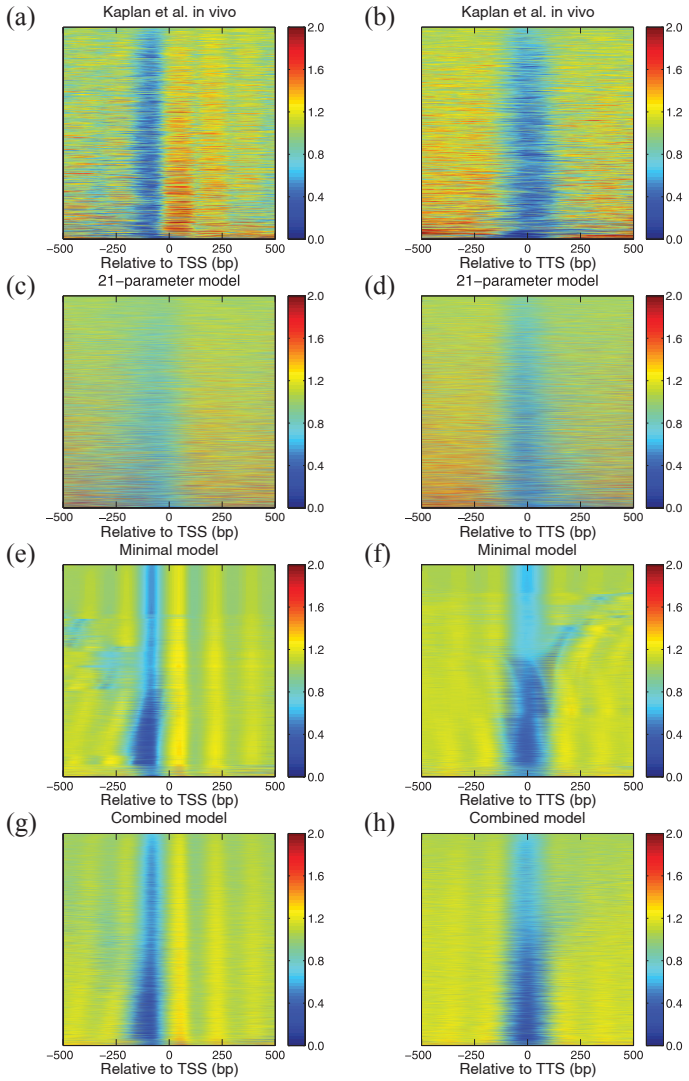


Figure 2.11: Heat maps of nucleosome occupancy around TSS and TTS for 5747 *S. cerevisiae* genes. *In vivo* nucleosomes (YPD medium) [57] (a) and (b), $N = 2$ position-independent model (c) and (d), minimal model (e) and (f), combined model (g) and (h). The minimal model is constructed by placing potential barriers from Figure 2.10 at the end of each gene onto an otherwise flat one-body energy landscape without two-body interactions. The combined model is constructed by adding sequence-specific energies from the 21-parameter $N = 2$ position-independent model (which have standard deviation of 0.61 k_BT genome-wide) to the minimal model. The occupancy for each gene is normalized by the average occupancy in the $[-500, 500]$ bp window. The experimental data [(a) and (b)] are smoothed with a $2D$ Gaussian kernel ($\sigma_X = 1$ bp and $\sigma_Y = 2$ genes). The genes are sorted in each panel in the order of increasing variance of the occupancy. The genome-wide average occupancies are 0.1508 [(a) and (b)], 0.2024 [(c) and (d)], 0.7516 [(e) and (f)], and 0.7232 [(g) and (h)].

Chapter 3

Nucleosome unwrapping

In this Chapter we present a rigorous treatment of nucleosome unwrapping. We first present some experimental observations of partially unwrapped nucleosomes, and then we construct a statistical mechanics model which explains these observations. This Chapter is based on our work submitted for publication [130].

Nucleosomes regulate many biological functions such as transcription, DNA replication, DNA repair, among others. Wrapped in nucleosomes, DNA is sterically occluded from interacting with many protein complexes, as for example transcription factors, polymerases, repair enzymes. At specific times, all nucleosomes need to have their DNA accessible to proteins that perform repair and replication tasks.

One mechanism through which DNA becomes accessible is nucleosome unwrapping. Both outer stretches of nucleosomal DNA can transiently peel off the histone surface due to thermal fluctuations, unwrapping and rewrapping with high frequency [28, 31, 34, 39, 41, 43, 99, 44, 45, 101, 46, 18, 48, 49, 108, 109]. Partial DNA unwrapping enables easier access of DNA-binding factors to their target sites which are packaged into chromatin.

Proteins can utilize spontaneous DNA unwrapping to bind to their target sites, which would favor further destabilization of the histone-DNA complex and binding of additional proteins (Figure 3.1). Since partial unwrapping of nucleosomal DNA is energetically less costly than nucleosome translocation, it is likely to play a major role in numerous DNA-mediated processes. For example, nucleosome "breathing" governs transcription dynamics of RNA polymerase [18], and may provide an explanation for fast DNA repair by photolyases [43].

Partial unwrapping of nucleosomal DNA and subsequent differential accessibility of

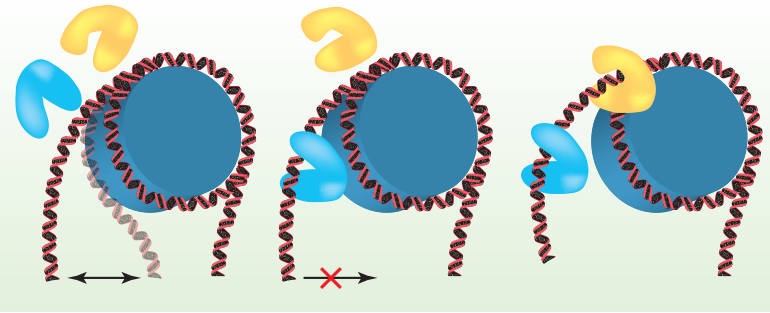


Figure 3.1: A nucleosome is a dynamic structure. Transient nucleosome unwrapping followed by factor binding prevents subsequent rewrapping, mediating further binding events.

nucleosome-covered protein-binding sites were observed in single nucleosomes [27, 28, 35, 39, 40, 41, 131, 49, 50], di-nucleosomes [100], and multi-nucleosome arrays [45, 46]. The unwrapping process was also modeled computationally [132, 47, 105, 115, 133, 48].

3.1 Experimental evidence of nucleosome unwrapping

Experimental observations of partial unwrapping of the nucleosomes appeared a long time ago [134, 135], and the first theoretical atempts to explain the dissociation of DNA from the histones followed shortly [136, 28].

Polach and Widom showed evidence for "site exposure" mechanism by which proteins may gain access to their target sites in nucleosomal DNA. They showed that nucleosomes are dynamic structures, transiently exposing DNA termini. This mechanism allows DNA-binding proteins to attach even to buried sites, in a cooperative way [29]. The same nucleosome-mediated cooperative binding was observed also by Adams and Workman [27]. They analyzed the binding of unrelated proteins to nucleosome cores *in vitro*. Even though these proteins do not bind cooperatively on naked DNA, they do so when their target sites are both under the same nucleosome.

In the following years Widom's group performed a series of studies in order to determine the rates of nucleosomal site exposure [31, 41, 49]. They also studied the sequence and position dependence of the site exposure rates [35], the effect of histone tails [36], acetylation [37], Poly(dA:dT) elements [38].

The nucleosome-induced cooperativity was observed not only *in vitro* [29], but also

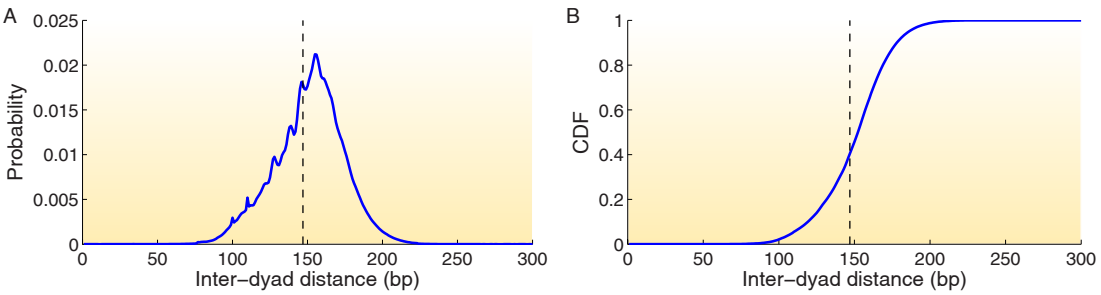


Figure 3.2: Probability distribution (A) and cumulative distribution function (B) of the inter-dyad distances reported in [92]. About 40% of the inter-dyad distances are less than 147 bp which means that the nucleosomes from the corresponding pairs must have been partially unwrapped. 147 bp distance is marked by the dashed black line.

in vivo [32, 33, 40]. More recently Tims and Widom [137] tested the nucleosome-induced cooperativity between proteins which bind on opposite sides of nucleosomes. They found no cooperativity in this case, and concluded that cooperative binding happens only when both factors bind at the same side of the nucleosome.

Using optical tweezers, different groups studied the nucleosome unwrapping when DNA was stretched [138, 139, 140, 18, 141]. Fluorescence resonance energy transfer (FRET) techniques were also used in studying nucleosome unwrapping [44, 46, 101, 113, 104]. Recently, atomic force microscopy (AFM) imaging was used to directly visualize the dynamics of nucleosomes [103, 107].

Other groups also studied processes which either require or induce nucleosome unwrapping, e.g. rapid nucleosomal DNA repair [43], accessibility of DNA packed in homogeneous nucleosome arrays, resembling the chromatin fiber [45], extensive overlapping of dinucleosome pairs [100], DNA unzipping of single molecules of nucleosomal DNA [142, 143], transcription dynamics of RNA polymerase II [18], nucleosome unwrapping induced by UV damage [104], effects of histone post-translational modifications (PTMs) to nucleosome unwrapping [106], nucleosome disassembly by the DNA mismatch repair complex hMSH2-hMSH6 [144], partially unwrapped CENP-A nucleosomes [145].

In 2012, Brogaard et al. [92] developed a new chemical approach of mapping nucleosome with base-pair resolution. They mutated the histone H4 protein and attached copper ions near the dyads of nucleosomes. Copper reacts with hydrogen peroxide creating reactive hydroxyl radicals, which cut the DNA near the dyads. The resulting

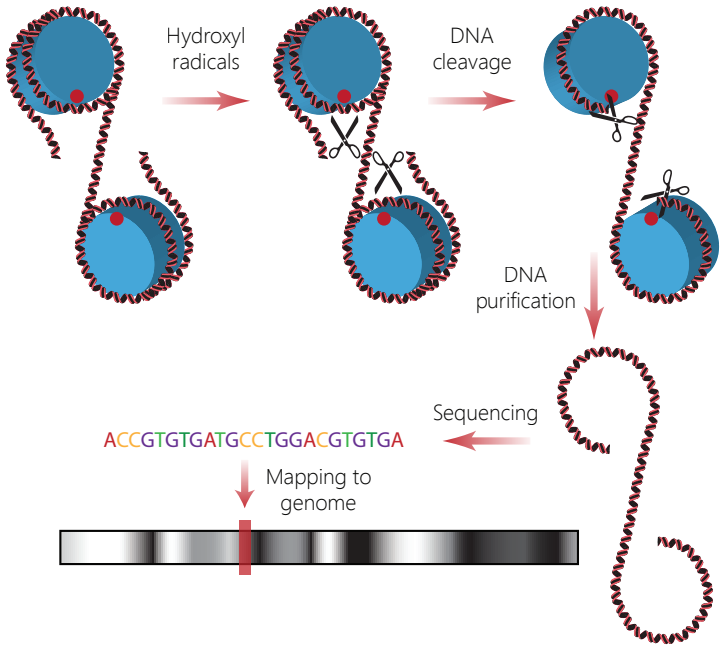


Figure 3.3: The new method of chemical mapping of nucleosome dyads. Mutant H4 histones (S47C) were modified by covalent attachment of a sulfhydryl-reactive, copper-chelating label to the cysteines. With the addition of copper and hydroxen peroxide, a localized cloud of hydroxyl radicals was produced which specifically cleaved the DNA backbone at sites symmetrically flanking nucleosome dyads. The cleavage products were isolated on an agarose gel, purified, sequenced using paired-end reads, and mapped to the *S.cerevisae* genome. Each mapped pair of reads yields a measurement of the distance between dyads of neighboring nucleosomes positioned on the same genome.

pieces of DNA are then sequenced and mapped to the yeast genome.

Using the short DNA fragments, with both ends indicating dyads of nucleosomes coming from the same cell, we compute the probability distribution of the inter-dyad distances. We observe that about 40% of these distances are less than 147 bp (Figure 3.2). This means that yeast has massive nucleosome unwrapping, genome-wide, *in vivo*. If nucleosomes were always fully wrapped then it would be impossible to obtain inter-dyad distances which are smaller than 147 bp. The distance between their centers would be in this case equal to 147 bp plus the length of the linker DNA between them (Figure 3.4).

Short fragments of MNase-protected DNA are always obtained in the MNase-seq experiments (Figure 3.5), but these could also result because of MNase over-digestion, and so they are not irrefutable proofs of spontaneous nucleosome unwrapping.

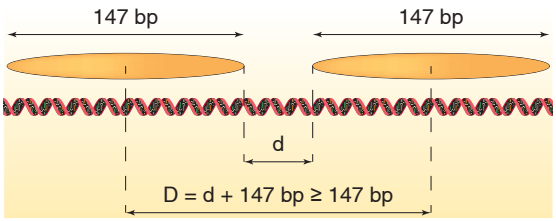


Figure 3.4: Illustration of a pair of nucleosomes, both covering 147 bp of DNA. The distance between their centers must be greater than 147 bp because of the linker between the nucleosomes.

In order to explain the previous experimental observation of nucleosome unwrapping, we developed a rigorous statistical mechanics approach which allows partial unwrapping of DNA from the histones, competition between multiple species of DNA-binding proteins, study of sequence-dependent binding energies, sequence-independent potential barriers and potential wells, and also effective two-body interactions between DNA-binding proteins. Starting from the relevant energies, we can compute the distribution of nucleosomes, and we can also infer the energies from the observed genome-wide organization of the nucleosomes. Our model explains the observed genome-wide distribution of inter-dyad distances [92], as well as the findings of earlier experiments which probed differential accessibility of nucleosome-covered binding sites [28, 39] and studied nucleosome-induced cooperativity between DNA-binding factors [27, 50]. We reproduce both the nucleosome density and the degree of unwrapping in the vicinity of TSS.

We next present the theoretical framework which allows us to predict the distribution of nucleosomes, and to infer the relevant energies which contribute to this organization genome-wide.

3.2 Direct problem: Matrix solution

Let us start with the direct problem which consists in computing the nucleosome distribution when we know the relevant energies, that is the binding energies of the histones, and the effective interaction between neighboring nucleosomes.

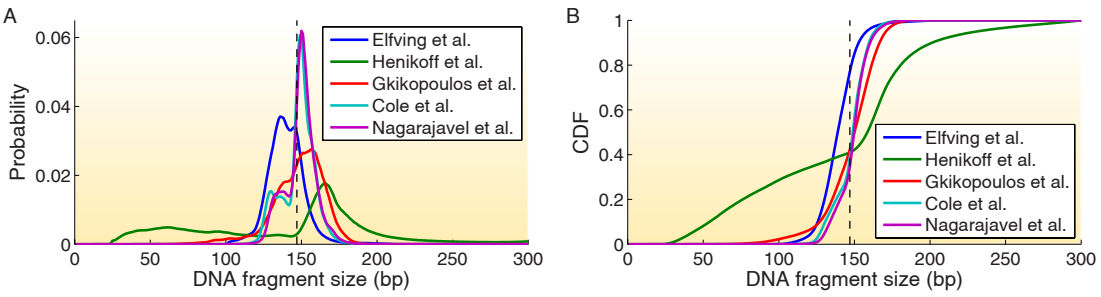


Figure 3.5: Probability density function (PDF, panel A) and cumulative distribution function (CDF, panel B) of the DNA fragment sizes which were measured in a series of nucleosome mapping experiments using paired-end sequencing. Many fragments have a length smaller than 147 bp which may indicate partially unwrapped nucleosomes. 147 bp length is marked by the dashed black line. This Figure is based on data from [78, 146, 84, 147, 148].

3.2.1 Single-type particles

We consider a system of one-dimensional rods (histones) which can attach and detach from a one-dimensional lattice (one chromosome) of L sites which represent the DNA bps. In order to model partial unwrapping of nucleosomal DNA off the surface of histone octamers, we allow the particles to cover a variable number of bps, ranging between two limits, $a_{\min}$ and $a_{\max}$. We assume that the particles cannot overlap while they are attached to the lattice. This is implemented using hard-core interactions between adjacent particles. To prevent particles from sliding off the chromosome, we fix infinite potential walls at both ends of the lattice. In addition, we allow two-body interactions between nearest-neighbor particles.

The attachment of a particle to the DNA modifies the total energy of the system in a sequence-specific manner. Physically, the binding energy may have contributions from DNA bending, electrostatic interactions, hydrogen bond formation, van der Waals contacts, etc. We denote the total one-body energy of a particle which covers bps $k, k+1, \ldots, l$ by $u(k,l)$. Note that for pairs of coordinates (k,l) such that $l-k+1 > a_{\max}$ or $l-k+1 < a_{\min}$, the binding energy $u(k,l) = \infty$, because all particles must have their lengths between $a_{\min}$ and $a_{\max}$ bps. The theory presented below is valid for arbitrary binding energies, $u(k,l)$.

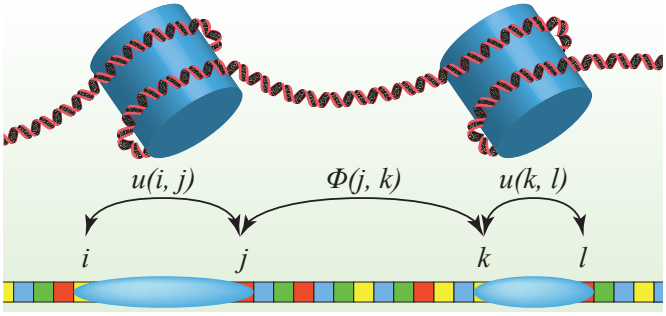


Figure 3.6: Schematic illustration of one-body and two-body potentials in a multi-nucleosome system. Nucleosomes may be partially unwrapped, resulting in variable DNA footprints.

Let $\Phi(j,k)$ be the two-body interaction between a pair of nearest-neighbor particles which cover base pairs $\ldots, j-1, j$ and $k, k+1, \ldots$ (Figure 3.6). In the case of nucleosomes, such interactions may be used to account for the effects of higher-order chromatin structure. Although we do not focus on two-body interactions in this chapter, they are included below for the sake of completion. We impose

$$\Phi(j,k) = \begin{cases} \infty & \text{if } k \leq j, \\ V(k-j-1) & \text{if } k > j, \end{cases}$$

where $V(d)$ is an arbitrary interaction potential which depends only on the linear distance, d, between two neighboring particles.

For a fixed number of particles, N, attached to the DNA, the canonical partition function is

$$Q_N = \sum_{\{i_n=1,\ldots,L\}_{n\in\{1,\ldots,2N\}}} e^{-\beta u(i_1,i_2)} e^{-\beta\Phi(i_2,i_3)} e^{-\beta u(i_3,i_4)} \ldots \times e^{-\beta u(i_{2N-3},i_{2N-2})} e^{-\beta\Phi(i_{2N-2},i_{2N-1})} e^{-\beta u(i_{2N-1},i_{2N})}, \quad (3.1)$$

where k_B is the Boltzmann constant, and $\beta = 1/(k_B T)$ is the inverse temperature. Note that with our definitions of the one-body energies, two-body interactions, and hard-wall boundary conditions, only the configurations of non-overlapping particles contribute to the partition function [Equation (3.1)].

In order to simplify the notation, we introduce two $L \times L$ matrices:

$$\langle k|e|l\rangle = e^{-\beta u(k,l)},$$
$$\langle k|w|l\rangle = e^{-\beta \Phi(k,l)}.$$

Here $\langle k|M|l\rangle$ represents the element of matrix M in row k and column l; $|l\rangle$ is a column vector of dimension L with one at position l and zero everywhere else, and $\langle k|$ is a row vector with one at position k and zero otherwise. Let $|J\rangle$ be the vector of dimension L with ones at every entry. Equation (3.1) can then be rewritten in the form

$$Q_N = \begin{cases} 1 & \text{if } N = 0, \\ \langle J|(ew)^{N-1}e|J\rangle & \text{if } N \geq 1. \end{cases}$$

Since the particles are allowed to attach and detach from the lattice, the system has a variable number of particles, and so the grand-canonical partition function for this system is

$$\begin{aligned} Z &= \sum_{N=0}^{N_{\max}} e^{\beta N \mu} Q_N \\ &= 1 + \sum_{N=1}^{N_{\max}} \langle J|(zw)^{N-1}z|J\rangle \\ &= 1 + \sum_{M=0}^{\infty} \langle J|(zw)^{M}z|J\rangle \\ &= 1 + \langle J|(I - zw)^{-1}z|J\rangle, \end{aligned} \tag{3.2}$$

where μ is the chemical potential, $N_{\max}$ is the maximum number of particles that can fit on L bp, and I is the identity matrix. Let

$$\begin{aligned} \zeta(k,l) &= \langle k|z|l\rangle \\ &= e^{\beta[\mu - u(k,l)]}. \end{aligned}$$

Note that all particle configurations with $N > N_{\max}$ do not contribute to the partition function Z, allowing us to extend the upper limit in Equation (3.2), from $N_{\max}$ to infinity.

From the partition function, we can compute the s-particle distribution functions,

exactly as we did in the previous chapter. We have

$$n_1(k,l) = \frac{\zeta(k,l)}{Z}\frac{\delta Z}{\delta\zeta(k,l)},$$
$$n_2(i,j;k,l) = \frac{\zeta(i,j)\zeta(k,l)}{Z}\frac{\delta^2 Z}{\delta\zeta(i,j)\delta\zeta(k,l)},$$

and, in general, for any positive integer s, we have

$$n_s(i_{1L},i_{1R};\ldots;i_{sL},i_{sR}) = \frac{\zeta(i_{1L},i_{1R})\ldots\zeta(i_{sL},i_{sR})}{Z}\frac{\delta^s Z}{\delta\zeta(i_{1L},i_{1R})\ldots\delta\zeta(i_{sL},i_{sR})}.$$

Using these relations, we obtain the one-particle distribution

$$n_1(k,l) = \frac{1}{Z}\langle J|(I-zw)^{-1}|k\rangle\langle k|z|l\rangle\langle l|(I-wz)^{-1}|J\rangle, \tag{3.3}$$

and the two-particle distribution

$$n_2(i,j;k,l) = \frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|z|j\rangle\langle j|w(I-zw)^{-1}|k\rangle\langle k|z|l\rangle\langle l|(I-wz)^{-1}|J\rangle. \tag{3.4}$$

In particular, the nearest-neighbor two-particle distribution is given by

$$\overline{n}_2(i,j;k,l) = \frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|z|j\rangle\langle j|w|k\rangle\langle k|z|l\rangle\langle l|(I-wz)^{-1}|J\rangle.$$

Equations (3.3) and (3.4) have an obvious interpretation. To find the probability that a particle covers positions k to l, we need to add statistical weights of all the configurations that contain such a particle, and divide the resulting sum by the partition function [Equation (3.3)]. Similarly, in order to find the probability of a pair of particles, one covering positions i to j, and the other covering positions k to l, we need to sum statistical weights of all the configurations containing such a pair of particles, and divide by the partition function [Equation (3.4)].

With the aid of one-particle distribution, $n_1(k,l)$, we define the occupancy at a bp i as the probability of finding that bp covered by any particle,

$$\mathrm{Occ}(i) = \sum_{k=i-a_{\max}+1}^{i}\ \sum_{l=\max(i,k+a_{\min})}^{k+a_{\max}-1} n_1(k,l).$$

Note that $1-\mathrm{Occ}(i)$ is the probability that bp i is not covered by any particles.

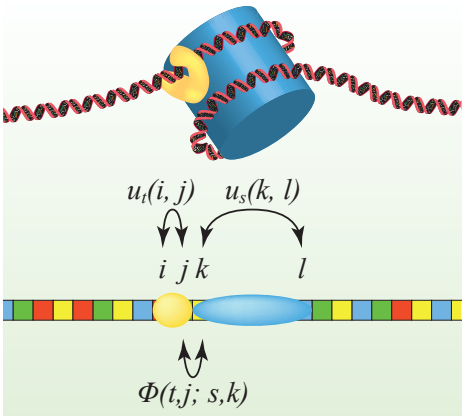


Figure 3.7: Schematic illustration of one-body and two-body potentials in a system with multiple-type particles. The model allows all particles to be in multiple stages of unwrapping. In practice, we allow nucleosomes to be partially unwrapped but the transcription factors (TFs) always have fixed DNA footprints.

3.2.2 Multiple-type particles

The above formalism can easily be extended to the case in which T types of particles are allowed to attach to the one-dimensional lattice. Let the binding energy of a particle of type $t \in \{1, \ldots, T\}$, that covers bps i to j on the lattice, be denoted by $u_t(i, j)$. The interaction between a particle of type t ending at position k, and the next particle of type s starting at position l, is denoted by $\Phi(t, k; s, l)$ (Figure 3.7). Each particle of type t, when attached to the DNA, is in contact with a number of bps ranging between $a^t_{\min}$ and $a^t_{\max}$. Thus, the binding energy $u_t(i, j) = 0$, if i and j do not satisfy the constraints $a^t_{\min} \leq j - i + 1 \leq a^t_{\max}$. Also, the two-body interaction $\Phi(t, i; s, j) = \infty$, for $j \leq i$, since the particles cannot overlap. With this notation, the grand-canonical partition function becomes

$$Z = \sum_{\text{all states}} e^{-\beta[u_{t1}(i_{1L}, i_{1R}) - \mu_{t_1}]} e^{-\beta\Phi(t_1, i_{1R}; t_2, i_{2L})} e^{-\beta[u_{t_2}(i_{2L}, i_{2R}) - \mu_{t_2}]} \ldots,$$

where μ_t is the chemical potential of the particles of type t. The sum is over all configurations of the system, which can have variable numbers of particles of any type.

Defining

$$\zeta_t(k, l) = e^{-\beta[u_t(k,l) - \mu_t]},$$

and using the matrix notation,

$$\langle t,k|z|s,l\rangle = \zeta_t(k,l)\delta_{ts},$$
$$\langle t,k|w|s,l\rangle = e^{-\beta\Phi(t,k;s,l)},$$

where δ_{ts} is the Kronecker delta symbol, the partition function becomes

$$Z = \sum_{\text{all states}} \langle t_1,i_{1L}|z|t_1,i_{1R}\rangle\langle t_1,i_{1R}|w|t_2,i_{2L}\rangle\langle t_2,i_{2L}|z|t_2,i_{2R}\rangle\dots\ .$$

Each vector $|t,i\rangle$ has dimension TL, and all of its entries are zero except for the entry in position $((t-1)L+i)$, which is one. For example, the $|1,i\rangle$ vectors have a one at position i, the $|2,i\rangle$ vectors have a one at position $L+i$, and so forth. Recall that $|J\rangle$ is the vector with TL elements all equal to one. Similarly to Equation (3.2), the partition function is

$$Z = 1 + \langle J|(I-zw)^{-1}z|J\rangle.$$

As in the case of the single-type particles, we compute the one-particle density

$$\begin{aligned} n_1^t(k,l) &= \frac{\zeta_t(k,l)}{Z}\frac{\delta Z}{\delta\zeta_t(k,l)} \\ &= \frac{1}{Z}\langle J|(I-zw)^{-1}|t,k\rangle\langle t,k|z|t,l\rangle\langle t,l|(I-wz)^{-1}|J\rangle. \end{aligned} \tag{3.5}$$

We also obtain the two-particle density

$$\begin{aligned} n_2^{t,s}(i,j;k,l) = \frac{1}{Z}&\langle J|(I-zw)^{-1}|t,i\rangle\langle t,i|z|t,j\rangle \\ &\times \langle t,j|w(I-zw)^{-1}|s,k\rangle\langle s,k|z|s,l\rangle\langle s,l|(I-wz)^{-1}|J\rangle, \end{aligned} \tag{3.6}$$

and the nearest-neighbor two-particle density

$$\begin{aligned} \overline{n}_2^{t,s}(i,j;k,l) = \frac{1}{Z}&\langle J|(I-zw)^{-1}|t,i\rangle\langle t,i|z|t,j\rangle \\ &\times \langle t,j|w|s,k\rangle\langle s,k|z|s,l\rangle\langle s,l|(I-wz)^{-1}|J\rangle. \end{aligned} \tag{3.7}$$

Equations (3.6) and (3.7) give the joint probability that a particle of type t covers bps i to j, while a second particle of type s covers bps k to l. Using Equation (3.5) we can compute occupancy for each type of particles t and for each bp i,

$$\text{Occ}_t(i) = \sum_{k=i-a^t_{\max}+1}^{i}\ \sum_{l=\max(i,k+a^t_{\min})}^{k+a^t_{\max}-1} n_1^t(k,l). \tag{3.8}$$

In the following Sections, we focus on the one-particle density function $n_1^t(k,l)$.

3.3 Direct problem: Recursive solution for hard-core interactions

In this Section, we explain how to compute the nucleosome distribution by a recursive method. In Section 3.3.1 we discuss the general case of partially unwrapped particles, while in Section 3.3.2, we particularize our results to the case without unwrapping.

A straightforward application of Equations (3.5) and (3.6) is computationally intensive because of the manipulations with huge matrices which are required. Although they are sparse matrices, their typical dimension is one million × one million. Fortunately, for particles that interact only through hard-core repulsion, rather than long-range two-body interactions, the one-particle distribution can be computed recursively, and therefore much more efficiently.

3.3.1 General case

With multiple-type particles, Equation (3.5) can be rewritten as

$$n_1^t(i,j) = \frac{1}{Z} Z^-(i) \langle t, i|z|t, j\rangle Z^+(j),$$

where $Z^-(i)$ and $Z^+(j)$ are the partition functions for the domains $[1, i)$ and $(j, L]$, respectively. Note that in the case of hard-core interactions alone, $Z^-(i)$ and $Z^+(j)$ do not depend on the type of the particle occupying positions i through j.

In the case of steric exclusion alone, these partial partition functions satisfy the following recursion relations:

$$Z^-(i) = Z^-(i-1) + \sum_s \sum_{i-a^s_{\max} \le j \le i-a^s_{\min}} Z^-(j) \langle s, j|z|s, i-1\rangle, \tag{3.9}$$

and

$$Z^+(i) = Z^+(i+1) + \sum_s \sum_{i+a^s_{\min} \le j \le i+a^s_{\max}} \langle s, i+1|z|s, j\rangle Z^+(j). \tag{3.10}$$

Here each particle type s has two characteristic lengths, corresponding to its minimum and maximum DNA footprints, $a^s_{\min}$ and $a^s_{\max}$, respectively. The boundary conditions are

$$\begin{cases} Z^-(1) = 1, \\ Z^+(L) = 1. \end{cases}$$

The full partition function is given by

$$Z = Z^-(L+1) = Z^+(0).$$

Note that all unphysical terms for which bound particles run off the lattice, automatically vanish from Equations (3.9) and (3.10). To avoid numeric instabilities, the recursion is done in log space. Let

$$F(i) = \ln Z^-(i),$$
$$R(i) = \ln Z^+(i).$$

With this notation, Equations (3.9) and (3.10) become

$$\begin{aligned} F(i) &= F(i-1) \\ &\quad + \ln\left\{1 + \sum_s \sum_{i-a^s_{\max} \le j \le i-a^s_{\min}} e^{F(j)-F(i-1)+\beta[\mu_s - u_s(j,i-1)]}\right\}, \\ R(i) &= R(i+1) \\ &\quad + \ln\left\{1 + \sum_s \sum_{i+a^s_{\min} \le j \le i+a^s_{\max}} e^{R(j)-R(i+1)+\beta[\mu_s - u_s(i+1,j)]}\right\}, \end{aligned} \tag{3.11}$$

with the boundary conditions

$$\begin{cases} F(1) = 0, \\ R(L) = 0. \end{cases}$$

Then the one-particle distribution function is

$$n_1^t(i,j) = e^{F(i)+R(j)-\ln Z+\beta[\mu_t - u_t(i,j)]},$$

where

$$\ln Z = F(L+1) = R(0).$$

The two-particle distribution can be computed similarly. The only new ingredient in Equation (3.6) is the partition function for the box with walls at two arbitrary positions,

$$Z(t,j,s,k) = \langle t,j|w(I-zw)^{-1}|s,k\rangle.$$

This partition function can be computed recursively, exactly as the partial partition functions $Z^{\pm}$ discussed above.

3.3.2 Special case: No unwrapping

The special case in which all particles are fully attached to their DNA sites (i.e., there is no DNA unwrapping) can be easily obtained from our general formalism. Indeed, when we restrict in Equation (3.11)

$$\begin{cases} a^s_{\min} = a^s, \\ a^s_{\max} = a^s, \end{cases}$$

We obtain

$$F(i) = F(i-1) + \ln\Big\{1 + \sum_s e^{F(i-a^s)-F(i-1)+\beta[\mu_s - u_s(i-a^s,i-1)]}\Big\},$$
$$R(i) = R(i+1) + \ln\Big\{1 + \sum_s e^{R(i+a^s)-R(i+1)+\beta[\mu_s - u_s(i+1,i+a^s)]}\Big\}.$$

As before, the boundary conditions are

$$\begin{cases} F(1) = 0, \\ R(L) = 0, \end{cases}$$

and the one-particle distribution is given by

$$n^t_1(i, i+a^t-1) = e^{F(i)+R(i+a^t-1)-\ln Z+\beta\left[\mu_t - u_t(i,i+a^t-1)\right]}.$$

3.4 Inverse problem: Recursive solution for hard-core interactions

In the previous Section, we solved the direct problem: given the binding energies for all particle types, we compute s-particle distributions. However, typically it is particle distributions that are observed experimentally, and the energetics of particle-DNA interactions need to be inferred. Here we solve the inverse problem recursively for the case of systems with multiple-type particles, partial unwrapping (variable footprints), and steric exclusion. The recursive solution is efficient enough to be employed on the genome-wide scale.

As in Section 3.3, here we explain how to compute the particle distribution by a recursive method. In Section 3.4.1, we treat the general case of partially unwrapped particles, while in Section 3.4.2, we consider the case without unwrapping.

3.4.1 General case

Using Equations (3.5), (3.9) and (3.10), we obtain:

$$\begin{aligned} Z^-(i) &= Z^-(i-1)\left[1 + \sum_{\substack{t,\\ i-a^t_{\max}\le j\le i-a^t_{\min}}} \frac{Z}{Z^-(i-1)Z^+(i-1)} n_1^t(j,i-1)\right] \\ &= Z^-(i-1)\left[1+\frac{N^R(i-1)}{\xi(i-1)}\right], \end{aligned} \tag{3.12}$$

where

$$N^R(i) = \sum_t \sum_{i-a^t_{\max}+1\le j\le i-a^t_{\min}+1} n_1^t(j,i)$$

represents the probability of finding a particle of any type with the right edge at bp i, and

$$\xi(i) = \frac{Z^-(i)Z^+(i)}{Z}.$$

The partition function Z^+ satisfies a similar recursive relation,

$$Z^+(i) = Z^+(i+1)\left[1+\frac{N^L(i+1)}{\xi(i+1)}\right], \tag{3.13}$$

where $N^L(i)$ is the probability of finding a particle of any type with the left edge at bp i,

$$N^L(i) = \sum_t \sum_{i+a^t_{\min}-1\le j\le i+a^t_{\max}-1} n_1^t(i,j).$$

The quantity $\xi(i)$ satisfies

$$\begin{aligned} \xi(i+1)-\xi(i) =& \frac{1}{Z}\left[Z^-(i+1)Z^+(i+1) - Z^-(i)Z^+(i)\right] \\ =& \frac{1}{Z}\Big\{Z^-(i+1)\left[Z^+(i+1)-Z^+(i)\right] \\ &+ Z^+(i)\left[Z^-(i+1)-Z^-(i)\right]\Big\} \\ =& N^R(i) - N^L(i+1), \end{aligned}$$

so that

$$\xi(i) = 1 + \sum_{k=0}^{i-1}\left[N^R(k) - N^L(k+1)\right], \tag{3.14}$$

with the boundary condition $\xi(0) = 1$. After we compute both Z^- and Z^+ using Equations (3.12) and (3.13), the total partition function is given by

$$Z = Z^-(L+1) = Z^+(0),$$

and the binding energy, for any particle of type t attached to the DNA, is given by

$$\beta\left[u_t(i,j)-\mu_t\right] = -\ln\left[n_1^t(i,j)\frac{Z}{Z^-(i)Z^+(j)}\right]. \tag{3.15}$$

3.4.2 Special case: No unwrapping

In the case of the all-or-none binding, all matrix elements $\langle i|n_1^t|j\rangle$ vanish unless

$$j = i + a^t - 1,$$

where a^t is the length of the binding site for the particle of type t. Thus, we obtain

$$N^L(i) = \sum_t n_1^t(i, i+a^t-1),$$
$$N^R(i) = \sum_t n_1^t(i-a^t+1, i).$$

Using these expressions, we employ Equations (3.12), (3.13) and (3.14) to compute Z^+ and Z^- in log space. Finally, Equation (3.15) is used to compute the binding energies. If all particles are of the same type, the quantity ξ can be simplified further:

$$\xi(i) = 1 - \sum_{k=i-a+1}^{i} N^L(k) = 1 - \mathrm{Occ}(i),$$

where $\mathrm{Occ}(i)$ is the probability that bp i is covered by a particle. Thus, in this limit, $\xi(i)$ is simply the probability that bp i is not occupied by any particles.

The recursion relations for Z^- and Z^+ become

$$Z^-(i+1) = Z^-(i)\left[1+\frac{N^L(i-a+1)}{1-\mathrm{Occ}(i)}\right],$$
$$Z^+(i) = Z^+(i+1)\left[1+\frac{N^L(i+1)}{1-\mathrm{Occ}(i+1)}\right].$$

These expressions are equivalent to those previously obtained in [76].

3.4.3 Sequence-specific nucleosome formation energies

The binding energy of a nucleosome is the sum of two components. One is the electrostatic energy of the negatively charged DNA wrapped around the positively charged histone octamer. This energy is negative, that is favorable for nucleosome formation.

The other component is the elastic energy required to bend the DNA polymer around the histone in about 1.67 turns. This energy is positive, that is disfavorable for nucleosome formation. The absolute value of the electrostatic energy is greater than the elastic energy such that the total energy of a nucleosome is negative and histones are able to bind to any DNA sequence and form nucleosomes.

We want to model the total formation energy of the nucleosomes, as a function of the DNA sequence which is wrapped around the histone octamer. Suppose a histone octamer is bound to a DNA sequence of length N,

$$S(N) = S_1 S_2 \cdots S_N,$$

where S_i represents the nucleotide from position i of the given DNA sequence. Here N can be different than 147 bp, because we want our model to be applicable for any degree of nucleosome unwrapping, and even for the cases when more than 147 bp of DNA are in contact with the histones, e.g. when a linker histone is present and the effective number of bps of nucleosomal DNA is greater than 147.

Let us denote the binding energy of a nucleosome containing this DNA sequence by $u_{S(N)}$. DNA sequences have different bending rigidities depending on their nucleotide compositions. As a consequence the nucleosome formation energies, $u_{S(N)}$, will vary among different DNA sequences of length N. Let us denote the average energy of all genomic sequences of length N by $\langle u_{S(N)} \rangle$, and the deviation of the energy of a sequence $S(N)$ from this average by $\delta u_{S(N)} = u_{S(N)} - \langle u_{S(N)} \rangle$. Obviously, we have the identity

$$\begin{aligned} u_{S(N)} &= \langle u_{S(N)} \rangle + \delta u_{S(N)} \\ &= u_N^{SI} + u_{S(N)}^{SD} \end{aligned}$$

where $u_N^{SI} = \langle u_{S(N)} \rangle$ and $u_{S(N)}^{SD} = \delta u_{S(N)} = u_{S(N)} - \langle u_{S(N)} \rangle$ represent the sequence-independent and sequence-dependent parts of the binding energy, respectively. Also, by definition we have that

$$\langle u_{S(N)}^{SD} \rangle = 0. \tag{3.16}$$

We model the sequence dependent part of the binding energy, and assume that this

depends only on the mono- and dinucleotide counts in the nucleosomal DNA,

$$u^{SD}_{S(N)} = \sum_{i=1}^{N} \epsilon_{S_i} + \sum_{i=1}^{N-1} \epsilon_{S_i S_{i+1}},$$

where ϵ_{S_i} and $\epsilon_{S_i S_{i+1}}$ are the contributions from the mononucleotide S_i and dinucleotide $S_i S_{i+1}$, respectively.

Because of the complementary base pairing, we only have two unique paired mononucleotides ($A/T, C/G$) and ten unique paired dinucleotides (AA/TT, AC/GT, AG/CT, AT/AT, CA/TG, CC/GG, CG/CG, GA/TC, GC/GC, TA/TA), where each dinucleotide is written in the 5' to 3' order. For this reason the sequence-dependent part of the binding energy is parametrized by twelve unique parameters: $\epsilon_{A/T}$, $\epsilon_{C/G}$, $\epsilon_{AA/TT}$, $\epsilon_{AC/GT}$, $\epsilon_{AG/CT}$, $\epsilon_{AT/AT}$, $\epsilon_{CA/TG}$, $\epsilon_{CC/GG}$, $\epsilon_{CG/CG}$, $\epsilon_{GA/TC}$, $\epsilon_{GC/GC}$, $\epsilon_{TA/TA}$.

The sequence-dependent energy corresponding to a histone attached to bps i to j of the DNA is

$$u^{SD}(i,j) = \sum_{k=i}^{j} \epsilon_{S_k} + \sum_{k=i}^{j-1} \epsilon_{S_k S_{k+1}},$$

where by ϵ_{S_k} we understand the energy contribution of the mononucleotide pair $S_k/\widetilde{S}_k$, and $\epsilon_{S_k S_{k+1}}$ represents the energy contribution from the paired dinucleotide $S_k S_{k+1}/\widetilde{S}_{k+1}\widetilde{S}_k$, with $\widetilde{S}_k$ being the nucleotide complementary to S_k.

The twelve parameters which parametrize the nucleosome formation energies are not all independent. Equality (3.16) imposes the following constraint on the parameters

$$\langle \sum_{k=i}^{j} \epsilon_{S_k} + \sum_{k=i}^{j-1} \epsilon_{S_k S_{k+1}} \rangle = 0,$$

which is equivalent to

$$(j-i+1)\langle \epsilon_{S_k} \rangle + (j-i)\langle \epsilon_{S_k S_{k+1}} \rangle = 0.$$

This has to be true for all sequence lengths $j-i+1$, so that we must have

$$\begin{cases} \langle \epsilon_{S_k} \rangle = 0, \\ \langle \epsilon_{S_k S_{k+1}} \rangle = 0, \end{cases}$$

or equivalently,

$$\begin{cases} \epsilon_{A/T} f(A/T) + \epsilon_{C/G} f(C/G) = 0, & (3.17) \\ \epsilon_{AA/TT} f(AA/TT) + \epsilon_{AC/GT} f(AC/GT) + \ldots + \epsilon_{TA/TA} f(TA/TA) = 0, & (3.18) \end{cases}$$

where $f(S_i/\widetilde{S}_i)$ and $f(S_iS_{i+1}/\widetilde{S}_{i+1}\widetilde{S}_i)$ represent the genomic frequencies of mononucleotide pair $S_i/\widetilde{S}_i$ and dinucleotide pair $S_iS_{i+1}/\widetilde{S}_{i+1}\widetilde{S}_i$, respectively. Genomic frequencies are different from one organism to another, and in the case of *S. cerevisiae* these frequencies are:

Sequence	Frequency
A/T	0.6170
C/G	0.3830
AA/TT	0.2161
AC/GT	0.1054
AG/CT	0.1168
AT/AT	0.0894
CA/TG	0.1297
CC/GG	0.0779
CG/CG	0.0294
GA/TC	0.1247
GC/GC	0.0375
TA/TA	0.0733

The inference of the parameters which generate the sequence-dependent and the sequence-independent components of the nucleosome binding energy is done in two steps.

In the *first step*, we estimate the sequence-independent part of the nucleosome formation energy, $u^{SI}(i,j)$. Because $\langle u^{SD}(i,j)\rangle = 0$, in the first approximation, we neglect the contributions from the sequence-dependent part. We test eight different models of sequence-independent energies, parametrized as described in Models A-H from Appendix B. Each set of parameters predicts a nucleosome distribution, which is used to compute the distribution of nucleosome footprint sizes and the distribution of inter-dyad distances. We find the optimal set of parameters, for all models, by minimizing the error between the histograms of lengths predicted by the model and that obtained from the experiments. In the case of data from Brogaard et al. [92], the paired-end DNA fragments generates the histogram of inter-dyad distances, while in the case of a

typical MNase-seq experiment, the paired-end reads generates the histogram of MNase-protected nucleosomal footprints. Any of these two types of data sets can be used to fit the parameters of the Models A-H (Appendix B).

In particular, for the Brogaard et al. data, we use the histogram of inter-dyad lengths obtained from the paired-end DNA fragments. With the aid of the genetic algorithm optimization function `ga` from the MATLAB Global Optimization toolbox, we minimize the objective function

$$\text{O.F.} = \begin{cases} RMS & \text{if } RMS \geq 10^{-3}, \\ RMS - r_{\text{osc}} \simeq -r_{\text{osc}} & \text{if } RMS < 10^{-3}, \end{cases}$$

where RMS is the root-mean-square deviation between predicted and observed inter-dyad distributions, and r_{osc} is the linear correlation between observed and predicted oscillations after the smooth background has been subtracted from the inter-dyad distributions, as in Figure 1D. In this way, the parameters are initially optimized such that the overall shape of the histogram is well approximated, i.e. the RMS decreases bellow a threshold (10^{-3}). Once this is achieved, the objective function is replaced by r_{osc}, and the fine oscillations of the histogram are fitted. The optimized parameters for all models are given in Appendix B.

In the *second step* of the optimization procedure, we compute the sequence-dependent part of the nucleosome binding energy corresponding to each DNA sequence, $u^{SD}(i,j)$, by subtracting the sequence-independent part, $u^{SI}(i,j)$, from the total binding energy, $u(i,j)$, given by Equation (3.15). Thus we obtain the following system of equations:

$$\begin{aligned} u^{SD}(i,j) - \mu &= \sum_{k=i}^{j} \epsilon_{S_k} + \sum_{k=i}^{j-1} \epsilon_{S_k S_{k+1}} - \mu \\ &= \begin{pmatrix} m_{A/T} & m_{C/G} & m_{AA/TT} & \cdots & m_{TA/TA} & -1 \end{pmatrix} \begin{pmatrix} \epsilon_{A/T} \\ \vdots \\ \epsilon_{TA/TA} \\ \mu \end{pmatrix}, \end{aligned} \tag{3.19}$$

where $m_{X/\widetilde{X}}$ and $m_{XY/\widetilde{Y}\widetilde{X}}$ are the counts of mono- and dinucleotide pairs $X/\widetilde{X}$ and

$XY/\widetilde{Y}\widetilde{X}$ in the sequence, respectively.

Using all possible combinations of pairs (i, j), where a nucleosome can form, we obtain a large number, P, of equations of the type

$$E - \mu \;=\; M \begin{pmatrix} \epsilon \\ \mu \end{pmatrix}. \tag{3.20}$$

Here, $E-\mu$ is a column vector of dimension P, where each row contains one $u^{SD}(i,j)-\mu$ element from Equation (3.19). $\begin{pmatrix} \epsilon \\ \mu \end{pmatrix}$ is the column vector from Equation (3.19), and M is a $P \times 13$ matrix with mono- and dinucleotide counts and -1's in the last column. From Equation (3.20), we derive the parameters ϵ and μ by a least squares fit.

Because in every DNA sequence the number of mononucleotides is equal to the length of the sequence, and the number of dinucleotides is equal to the length of the sequence minus 1, the columns of the matrix M are not linearly independent. Indeed, the column vector

$$|V\rangle = \begin{pmatrix} 1 \\ 1 \\ -1 \\ \vdots \\ -1 \\ 1 \end{pmatrix}$$

is the only linearly independent vector from the kernel of M: $M|V\rangle = 0$, i.e. the kernel of M is spanned by $|V\rangle$. Thus the rank of matrix M is 12 which is greater than the number of independent parameters, 11. We have 10 independent ϵ parameters [2 out of 12 are fixed by Equation (3.16)], and the 11-th parameter is μ. This means that a least square fit with 2 constrains [Equation (3.16)] will result in a unique set of parameters. Constrained linear least-squares problems are solved in MATLAB using the function `lsqlin` from the Optimization toolbox.

This completes the description of the two-step optimization procedure.

3.5 Applications

Next, we present the results and few applications of our new model.

3.5.1 Nucleosome unwrapping potential

We use a high-resolution *in vivo* map of nucleosome dyad positions, measured using a new method developed by Brogaard et al. [92], based on chemical modification of engineered histones and DNA backbone cleavage by hydroxyl radicals. Data provided by Brogaard et al. gives a direct measurement of both dyad positions and distances between adjacent dyads. Although superior to methods based on MNase digestion whose accuracy is affected by MNase sequences preferences and its tendency to over- or under-digest DNA [149, 150, 151], the map provided by Brogaard et al. [92] is biased by unknown hydroxyl radical cutting preferences for two alternate sites at each DNA strand [92].

From the paired-end reads deposited by Brogaard et al. in the GEO database (GEO accession GSM880651) we obtain that 38.7% of the distances between neighboring nucleosomes are less than 147 bp (blue line in Figure 3.9B). This means that many nucleosomes are partially unwrapped in yeast, *in vivo*. Models which disregard nucleosome unwrapping simply do not allow inter-dyad distances which are less than 147 bp (Figure 3.4).

In order to study the energetics of unwrapping, we introduce a simple model for the sequence-independent part of the nucleosome binding energy, u_N^{SI}. This is based on the 10-11 bp periodic pattern of histone contacts with the minor groove of the nucleosomal DNA [154, 152] (Figure 3.9A). As DNA is peeled off each contact patch, its free energy increases because hydrogen bonds and favorable electrostatic contacts between histone side chains and the DNA phosphate backbone are lost. However, once DNA breaks free from the contact patch, it may adopt multiple conformations, which allows it to increase its entropy and thus lower its total free energy. The favorable entropic term grows with the extent of unwrapping until the next contact patch is reached, completing one cycle in the oscillatory energy profile. The oscillations are superimposed on a straight line

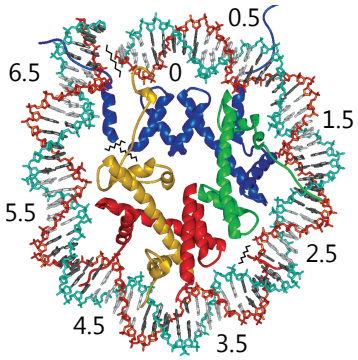


Figure 3.8: View of the NCP147 crystal structure [152] down the DNA superhelix axis showing the major groove-inward (grey DNA bases) and minor groove-inward (white DNA bases) facing regions for approximately one-half of the particle. Numbers correspond to double-helical turns from the nucleosome centre (0), which coincides with the central base pair at the particle pseudo dyad axis where the major groove directly faces the histone octamer. The phosphodiester backbone of the DNA strands appears as cyan and orange. Histone proteins are colored gold for H2A, red for H2B, blue for H3, and green for H4. Reprinted by permission from Oxford University Press, copyright (2010) [153].

whose slope equals the average free energy cost per bp of histone-DNA contact formation minus the average cost of DNA bending. Additional details of the potential construction can be found in Appendix B, Model A. The histone-DNA potential constructed in this way has no sequence specificity. All the sequence-dependent corrections are included in the term $u^{SD}_{S(N)}$, as discussed in Section 3.4.3.

We aim to reproduce the observed distribution of inter-dyad distances with a model in which nucleosome energetics is sequence-independent but transient unwrapping is allowed. To predict the distribution of inter-dyad lengths, we compute the conditional probability of having a nucleosome with the dyad at bp $c+d$, given that the adjacent upstream nucleosome has the dyad at bp c,

$$P(c+d|c) = \frac{N_2(c, c+d)}{N_1(c)}, \tag{3.21}$$

where the probability distributions of the nucleosome centers can be computed using

Equations (3.5) and (3.6):

$$N_1(c) = \sum_{\Delta_1} n_1^{\text{nuc}}(c - \Delta_1, c + \Delta_1),$$
$$N_2(c, c+d) = \sum_{\Delta_1,\Delta_2} \overline{n}_2^{\text{nuc,nuc}}(c - \Delta_1, c + \Delta_1; c + d - \Delta_2, c + d + \Delta_2).$$

Here, $2\Delta_{1,2} + 1$ are the lengths of the particles centered at bp c and $c + d$, respectively. To estimate $P(c + d|c)$, we use $c = 5$ kbp and a box of length $L = 10$ kbp, so that the boundaries of the box are far away.

Hydroxyl radicals that cleave DNA near the nucleosome dyad have two preferred cutting sites, at positions -1 bp and +6 bp with respect to the dyad [92]. If DNA is cut at these positions with frequencies f and $1 - f$, the distance between two consecutive cuts, one on the Watson and one on the Crick strand, is given by

$$d_{\text{cuts}} = d_{\text{dyads}} + b.$$

Above, d_{dyads} is the distance between two neighboring dyads, and the bias b is

$$b = \begin{cases} -12 \\ -5 \\ 2 \end{cases}, \text{ with probability } \begin{cases} (1-f)^2 \\ 2f(1-f) \\ f^2 \end{cases}.$$

We assumed here that the two cutting events are independent and the joint probability of obtaining both cuts at the same time is simply the product of individual probabilities. The cleavage bias has to be taken into account by convolving the predicted inter-dyad distance probability $P(c + d|c)$ with a kernel, F, corresponding to this bias,

$$F(x) = \begin{cases} (1-f)^2 & \text{for } x = -12, \\ 2f(1-f) & \text{for } x = -5, \\ f^2 & \text{for } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

We convolve $P(c + d|c)$ with this kernel to account for site-specific chemical cleavage bias, and fit the model parameters such that the predicted distribution of inter-cut lengths reproduces the observed distribution

Since inter-dyad distances cannot be used to distinguish between symmetric and asymmetric unwrapping, we assume the former for simplicity. The model is fit to the observed distribution of inter-dyad distances (Appendix B). The free parameters of the model include the amplitude of the oscillations, the slope of the free energy profile and $a_{\mathrm{min(max)}}$, the minimum (maximum) effective length of the nucleosome particle (Appendix B, Model A). The maximum extent of nucleosome unwrapping is controlled by a_{min}, while a_{max} is allowed to exceed 147 bp in order to account for the effects of higher-order chromatin structure and linker histone deposition. We also fit the relative frequency of hydroxyl radical DNA cleavage at the -1 position with respect to the nucleosome dyad, f, and the chemical potential of histone octamers, μ. Our model reproduces both the overall shape and fine oscillatory structure of the observed inter-dyad distance distribution (Figure 3.9B, C). In contrast, models without unwrapping are unable to capture even the overall shape of the observed inter-dyad distribution (gray line in Figure 3.9B).

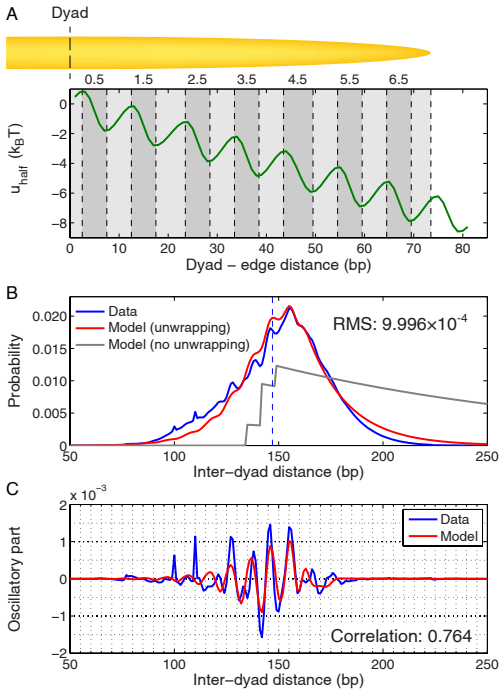


Figure 3.9: **Genome-wide distribution of inter-dyad distances.** (A) Nucleosome energy profile. The energy of a nucleosome that covers $2x + 1$ bps, and is symmetrically unwrapped, is given by $u_{\text{nuc}}^{SI} = 2u_{\text{half}}(x)$. The minima and maxima of the energy landscape are based on information from the crystal structure of the nucleosome core particle [154, 152]. Dark gray bars show where the histone binding motifs interact with the DNA minor groove in the structure. Light gray bars show where the DNA major groove faces the histones. The energy profile is obtained by a polynomial fit as described in Appendix B, Model A. (B) The inter-dyad distance distribution from a high-resolution nucleosome map [92] (blue), and from the model with (red) and without unwrapping (gray). In the model without unwrapping, $a_{\text{min}} = a_{\text{max}} = 147$ bp and the fitting parameters are E_b, μ and f (Appendix B, Model A). RMS represents the total root-mean-square deviation between the model and the data. (C) Oscillations in the observed (blue) and predicted (red) inter-dyad distance distributions, obtained by subtracting a smooth background from the data and the model with unwrapping in (B). The smooth background is found by applying a Savitzky-Golay filter of polynomial order 3 with 31 bp length (using the `sgolayfilt` function from the Signal Processing Toolbox of MATLAB). Correlation refers to r_{osc}, the linear correlation coefficient between measured and predicted oscillations.

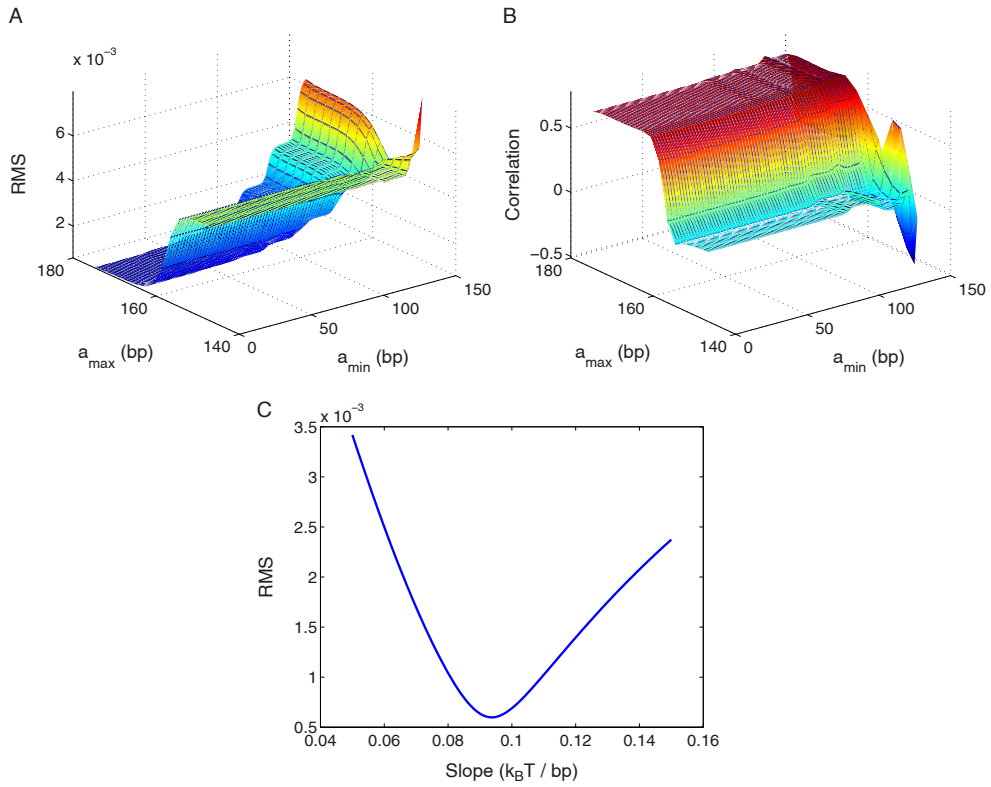


Figure 3.10: **Sensitivity of the predicted inter-dyad distribution to the parameters of the unwrapping potential based on nucleosome crystal structures.** (A) Root-mean-square error (RMS) of the inter-dyad distribution predicted using the model in Figure 3.9A, as a function of $a_{\min}$ and $a_{\max}$. (B) The linear correlation coefficient between oscillations in the predicted and observed inter-dyad distributions, r_{osc}, as a function of $a_{\min}$ and $a_{\max}$. The oscillations are obtained by subtracting the smooth background from inter-dyad distributions, as described in the caption of Figure 3.9. (C) Variation of the RMS with the slope of the unwrapping potential in Figure 3.9A. In all panels, model parameters that are not varied, are kept fixed at their best-fit values given in Appendix B, Model A.

3.5.2 Higher-order chromatin structure and linker histone energetics

The effective length of the particle that we found in the fit, $a_{\max} = 163$ bp, is greater than 147 bp, the length of the DNA in the nucleosome core [152]. The maximum particle length, $a_{\max} = 147$ bp is incompatible with the observed inter-dyad distribution (Figure 3.10A, B). The model is less sensitive to the minimum length of the particles, $a_{\min}$, because extensively unwrapped particles are energetically unfavorable and therefore are not frequently seen in the data. The overall shape of the inter-dyad distribution is also sensitive to the slope of the energy profile in Figure 3.9A. This provided a robust

fit of the average nucleosome binding energy per bp of the nucleosomes (Figure 3.10C). The fitted slope yields a value of 14.4 k_BT for the histone-DNA interaction energy in a fully wrapped nucleosome, that is, 147 bp of DNA wrapped around the histone octamer.

Thus the energy profile in Figure 3.9A describes both DNA interactions with the histone octamer core (up to 73 bp from the dyad) and the effects of higher-order chromatin structure, including, potentially, the attachment of Hho1p, the H1 linker histone of *S. cerevisiae*, to the DNA immediately outside of the nucleosome core [155, 156, 157]. Although Hho1p is less abundant in yeast than in higher eukaryotes, it is involved in higher-order chromatin organization, including chromatin compaction in stationary phase [158, 157]. Relatively little is known about the molecular mechanism of H1 binding. There is no consensus yet whether the binding is symmetric or asymmetric, or even what the extent of the H1 footprint is [155, 156]. H1 binding and other factors that mediate chromatin folding into higher-order structures cause linker lengths to be discretized [159, 97]. Linker length discretization can be described by a periodic, decaying two-body effective potential between neighboring nucleosomes, with the first minimum occurring approximately 5 bp away from the nucleosome edge [159, 98, 94].

Based on these observations, we construct two models for the energy profile outside of the nucleosome core region. The first model is a polynomial fit that extends the quasiperiodic profile of the unwrapping energy through another cycle (Figure 3.9A; Appendix B, Model A). The depth and the position of the first minimum outside of the nucleosome core are additional free parameters. As can be seen in Figure 3.11B, our fit robustly predicts the first minimum to be positioned 5-6 bp outside of the nucleosome core, in agreement with previous studies [159, 98, 94]. The depth of this minimum is comparable to the depth of the unwrapping minima (Figure 3.11, Appendix B, Model A).

The second model represents the energy profile outside of the nucleosome core by a linear function (Figure 3.12A; Appendix B, Model B). The two free parameters are the slope and the range of the linear function, which are related to the H1-DNA interaction energy and the H1 footprint, respectively. This model assumes the H1 histone being gradually detached from its DNA site immediately outside of the nucleosome core. This

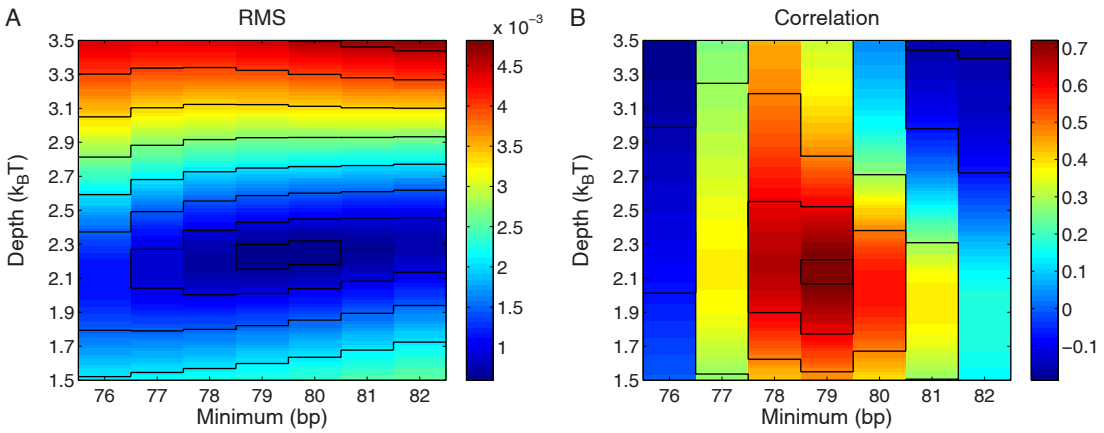


Figure 3.11: **Sensitivity of the predicted inter-dyad distribution to model parameters describing higher-order chromatin structure.** The unwrapping potential is based on nucleosome crystal structures (Appendix B, Model A). (A) Root-mean-square error (RMS) of the predicted inter-dyad distribution, as a function of the position and the depth of the first minimum outside of the nucleosome core (Figure 3.9A). The depth of the first minimum is computed with respect to $u_{\mathrm{half}}(x) = 73$ bp. (B) The linear correlation coefficient between oscillations in the predicted and observed inter-dyad distributions, r_{osc}, as a function of the position and the depth of the first minimum outside of the nucleosome core (Figure 3.9A). The oscillations are obtained by subtracting the smooth background from inter-dyad distributions, as described in the Figure 3.9 caption. In both panels, model parameters that are not varied are kept fixed at their best-fit values (Appendix B, Model A).

alternative scenario, although likely oversimplified, can be used to check the sensitivity of our results toward a particular energy profile outside of the core region. We find that the linear profile fits the overall shape of the inter-dyad distribution somewhat less well than the oscillatory one (compare the RMS values in Figures 3.9B and 3.12B), although the 10-11 bp periodic fine structure is reproduced in both cases (Figures 3.9C, 3.12C). The optimal linear profile is 7 bp long, yielding a symmetric H1 footprint with two 7 bp half-sites (Figure 3.12D), and the H1-DNA interaction energy of approximatively 5 k_BT (Figure 3.12E).

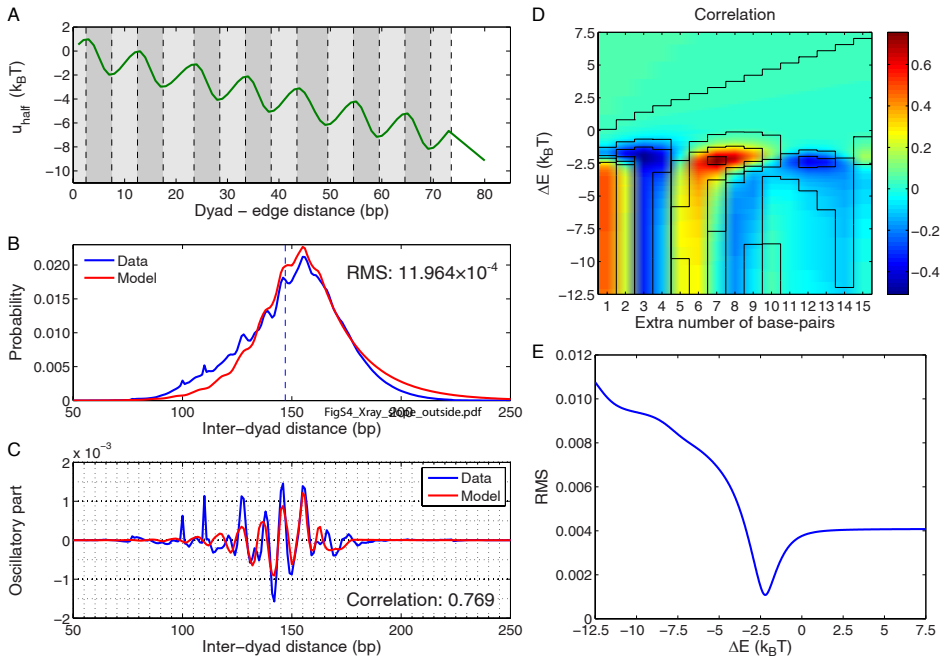


Figure 3.12: **Crystal structure-based model augmented by a linear potential outside of the nucleosome core.** (A) The energy profile fitted to reproduce the inter-dyad distance distribution shown in (B). All fitting parameters are listed in Appendix B, Model B. Under the symmetric unwrapping assumption, the energy of a nucleosome which covers $2x+1$ bps is given by $2u_{\mathrm{half}}(x)$. (B) The inter-dyad distance distribution observed in a high-resolution nucleosome map [92] (blue line), and predicted using Model B in Appendix B (red line). RMS - root-mean-square deviation between the model and the data. Note that in this model RMS below 10^{-3} could not be achieved, and thus optimization was switched to maximize the correlation coefficient r_{osc} once RMS reached 1.2×10^{-3} (see Appendix B for details). (C) Oscillations in the observed (blue line) and predicted (red line) inter-dyad distributions. The oscillations were obtained by subtracting the smooth background from the data and the model in (B), as described in the Figure 3.9 caption. Correlation refers to r_{osc}, the linear correlation coefficient between measured and predicted oscillations. (D) Heatmap with superimposed contour lines of the r_{osc} dependence on the two parameters of the linear potential outside of the nucleosome core: $\Delta x = x_{\mathrm{last}} - 73$ bp and $\Delta E = u_{\mathrm{half}}(x_{\mathrm{last}}) - u_{\mathrm{half}}(73)$, where $[1, x_{\mathrm{last}}]$ is the range of the energy profile (Appendix B, Model B). Note that the best fit corresponds to $\Delta x = 7$ bp. (E) The dependence of the RMS on ΔE for the best-fit value of $\Delta x = 7$ bp. All parameters not explicitly varied in (D) and (E) were kept fixed at their best-fit values (Appendix B, Model B).

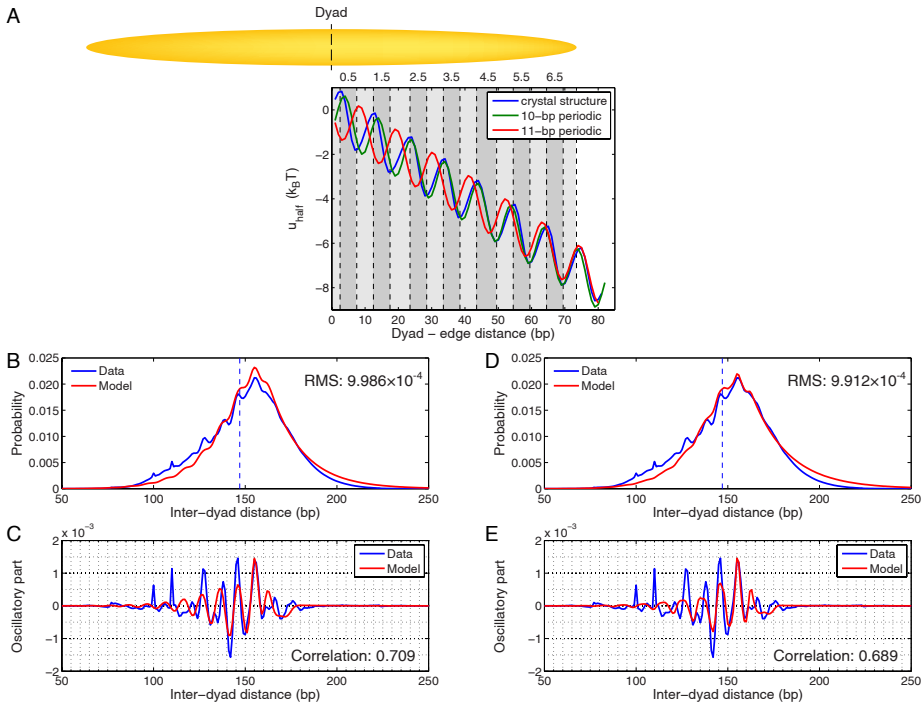


Figure 3.13: **Strictly periodic models of nucleosome unwrapping.** (A) Nucleosome unwrapping/higher-order structure potential energy profiles. Under the symmetric unwrapping assumption, the energy of a nucleosome that covers $2x+1$ bps is given by $2u_{\mathrm{half}}(x)$. The minima and maxima of the energy landscape are either based on the crystal structures of the nucleosome core particle as in Figure 3.9 (blue), or else are 10 (green) and 11 (red) bp-periodic oscillations with fitted initial phase (Appendix B, Models C and D). Dark gray bars show where the histone binding motifs interact with the DNA minor groove. Light gray bars indicate where the DNA major groove faces the histones. (B) The inter-dyad distance distribution from a high-resolution nucleosome map [92] (blue line), and from the 10 bp-periodic model (red line). All model parameters are listed in Appendix B, Model C. RMS - root-mean-square deviation between the model and the data. (C) Oscillations in the observed (blue line) and predicted (red line) inter-dyad distributions. The oscillations are obtained by subtracting a smooth background from the data and the model in (B), as described in the Figure 3.9 caption. Correlation refers to r_{osc}, the linear correlation coefficient between measured and predicted oscillations. (D) Same as (B), for the 11 bp-periodic model. All model parameters are listed in Appendix B, Model D. (E) Same as (C), for the 11 bp-periodic model.

3.5.3 Alternative models of nucleosome unwrapping

We next test the sensitivity of our fits to the analytical form of the unwrapping free energy profile. Although our primary model follows nucleosome crystal structures in creating a quasi-periodic energy profile with both 10 and 11 bp modes, strictly periodic 10 or 11 bp sinusoidal profiles yield nearly the same quality of fit (Figure 3.13; Appendix B, Models C, D). Since the initial phase of the oscillations is not determined by the crystal structure, it becomes another fitting parameter. The fitted initial phases in the 10 and 11-bp models make the periodic curves match the crystal structure further away from the dyad, where most of the observed unwrapping takes place (Figure 3.13A). The phases diverge closer to the dyad, where they are not as strongly constrained by the data. The root-mean-square deviation, RMS, is less sensitive to the initial phase than to the linear correlation, r_{osc}, between predicted and observed oscillations in the inter-dyad histograms (Figure 3.14). The primary peak in the dependence of the correlation coefficient on the initial phase matches the crystal structure. There is also a secondary peak corresponding to the 5 bp shift in the unwrapping energy profile, which in turn leads to the 10 bp, in-phase shift in the distribution of inter-dyad oscillations (Figure 3.14B).

Since the inter-dyad distance distribution has a distinct oscillatory component, it is not surprising that a purely linear model of unwrapping energy does not fit the data as well, although it does match its overall shape (Figure 3.15A; Appendix B, Model E). Less trivially, it was suggested on the basis of single-nucleosome unzipping experiments that nucleosome unwrapping proceeds with 5-bp periodicity because histones interact with each DNA strand separately where the DNA minor groove faces the histone octamer surface, creating two distinct contact "subpatches" [142]. This single-molecule data was fit to a model with a step-wise unwrapping free energy profile [114]. Each step in the profile corresponds to breaking a point histone-DNA contact, and the steps occur every 5.25 bp on average. We do not find any evidence for 5 bp periodicity of nucleosome unwrapping in the genomic data. Indeed, both 5 bp step-wise and 5 bp periodic sinusoidal profiles fit the data poorly, about as well as the linear model (Figure 3.15B,

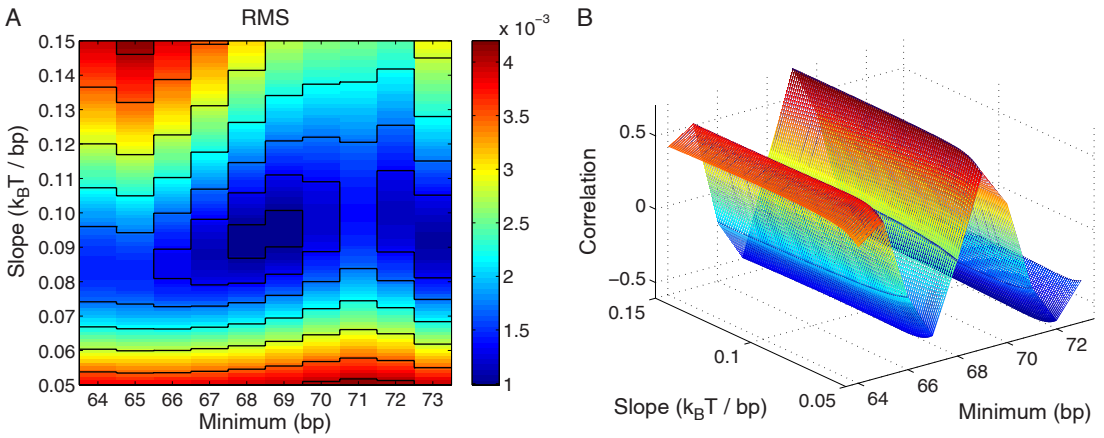


Figure 3.14: **Sensitivity of the predicted inter-dyad distribution to parameters of the 10 bp-periodic model.** (A) Heatmap with superimposed contour lines of the RMS dependence on the slope of the energy profile and the position of the last minimum within the nucleosome core. RMS - root-mean-square deviation between the model and the data. (B) The linear correlation coefficient, r_{osc}, between oscillations in the predicted and observed inter-dyad distributions, as a function of the overall slope of the energy profile and the position of the last minimum within the nucleosome core. All parameters not explicitly varied are kept fixed at their best-fit values (Appendix B, Model C).

C). Even the 10-bp step-wise unwrapping profile, while clearly having the right periodicity, does not fit the data as well as the structure-based model (Figure 3.15D). This observation suggests that the picture of gradual loss of favorable finite-range histone-DNA interactions, followed by gain in DNA conformational entropy, is closer to reality than abrupt disruption of short-range histone-DNA contacts. A direct comparison of single-molecule and genome-wide energy profiles is unfortunately obscured by the fact that the reported single-nucleosome unzipping experiments are specific to the 601 nucleosome-forming sequence [71], in contrast to our methodology which provides the average, sequence-independent picture of unwrapping energetics.

3.5.4 Genome-wide organization of nucleosome unwrapping states

Figure 3.16A, in which genes are sorted by the promoter length and aligned by the TSS, shows a canonical picture of nucleosomes depleted in promoters and well-positioned over coding regions [60, 87]. Interestingly, promoter nucleosomes have shorter inter-dyad distances and are therefore more unwrapped (Figure 3.16B). When averaged over all

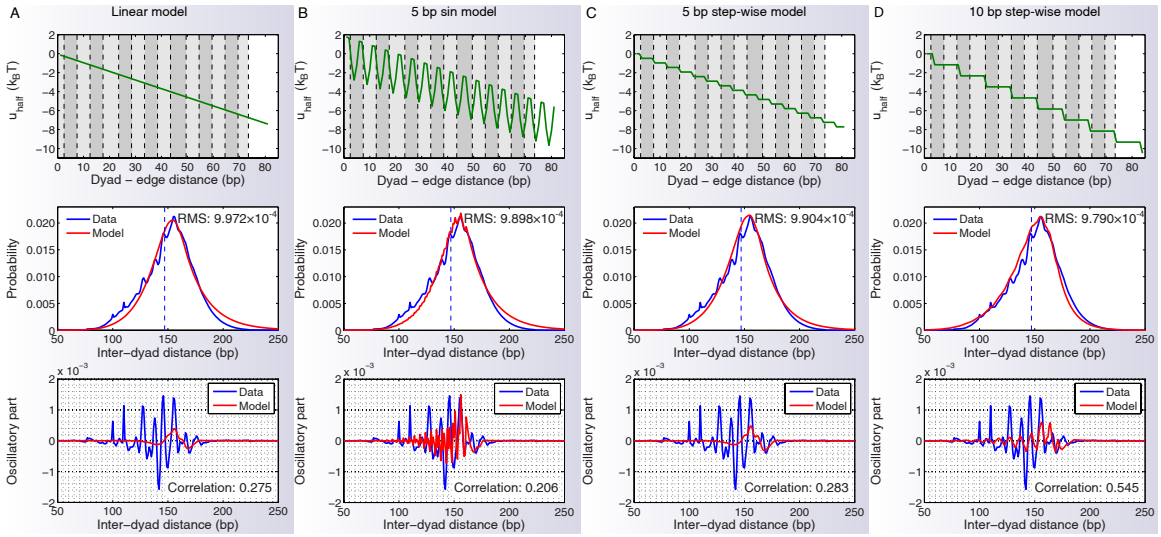


Figure 3.15: **Alternative models of nucleosome unwrapping.** (A) Linear model (Appendix B, Model E). (B) 5-bp periodic model (Appendix B, Model F). (C) 5-bp step-wise model (Appendix B, Model G). (D) 10-bp step-wise model (Appendix B, Model H). In each column, the upper panel shows the nucleosome unwrapping/higher-order structure potential energy profile (as in Figure 3.9A), the middle panel shows the comparison of experimental and predicted inter-dyad distance distributions (as in Figure 3.9B), and the lower panel shows observed and predicted oscillations in the inter-dyad distance distributions (as in Figure 3.9C).

genes, the number of dyads at a given bp and the average inter-dyad distance at that bp are strongly correlated (compare blue and red lines in Figure 3.16C). The profile of average inter-dyad distances is also correlated with the distribution of DNA fragment lengths in an MNase assay which mapped both nucleosomes and subnucleosome-size particles by paired-end sequencing (Figure 3.17A, green line in Figure 3.16C) [146]. The two profiles do not coincide completely because inter-dyad distances also depend on the distribution of linker lengths. The observed behavior is opposite of the naive expectation that unwrapping increases with occupancy due to nucleosome crowding. This behavior is also reproduced in a simple sequence-independent model in which nucleosomes phase off a potential barrier placed in the promoter region (Figure 3.18).

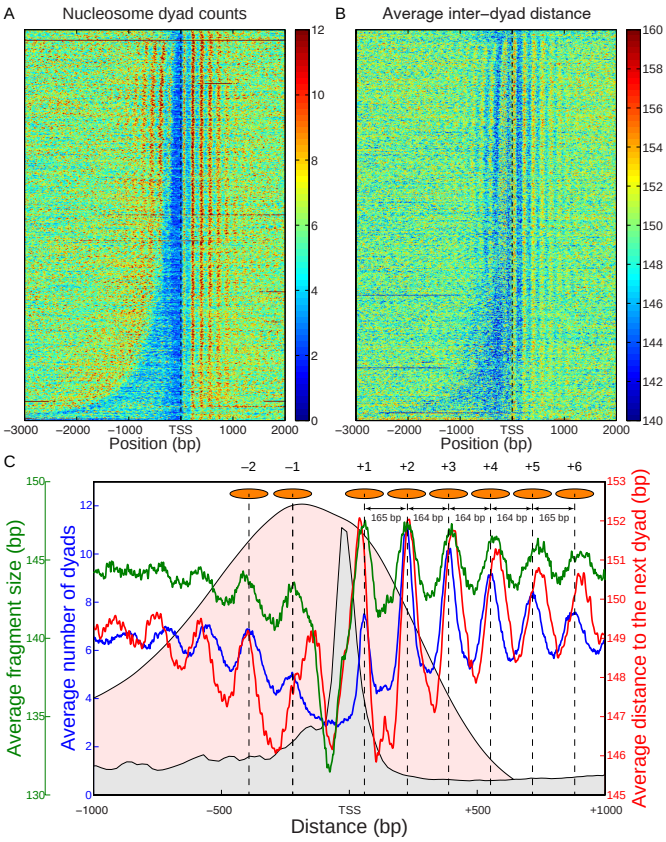


Figure 3.16: **Nucleosome unwrapping in the vicinity of transcription start sites.** (A) Distribution of nucleosome dyad counts [92] near the TSS. 4763 verified *S.cerevisiae* open reading frames (ORFs) are aligned by their TSS and sorted by promoter lengths. Each horizontal line corresponds to one ORF. (B) Distribution of the average distance between neighboring dyads. For each bp, the distances between a dyad at that bp and all neighboring dyads are averaged. ORFs are sorted as in (A). In (A) and (B), values at bps without dyads are obtained by interpolation, and heatmaps are smoothed using a 2D Gaussian kernel with $\sigma = 3$ pixels. (C) Data in (A), (B) and Figure 3.17A-C is averaged over all genes. Blue: nucleosome dyad counts, red: average distance between neighboring dyads, green: average length of DNA-bound particles mapped by MNase digestion [146] (see Figure 3.17A for details). Curve with light gray background: combined occupancy of 9 PICs (TBP, TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, TFIIK, PolII) [160], curve with light pink background: average histone turnover rate [161]. The peaks in the dyad count profile (blue) are marked with orange ovals representing nucleosomes, and peak-to-peak distances are shown.

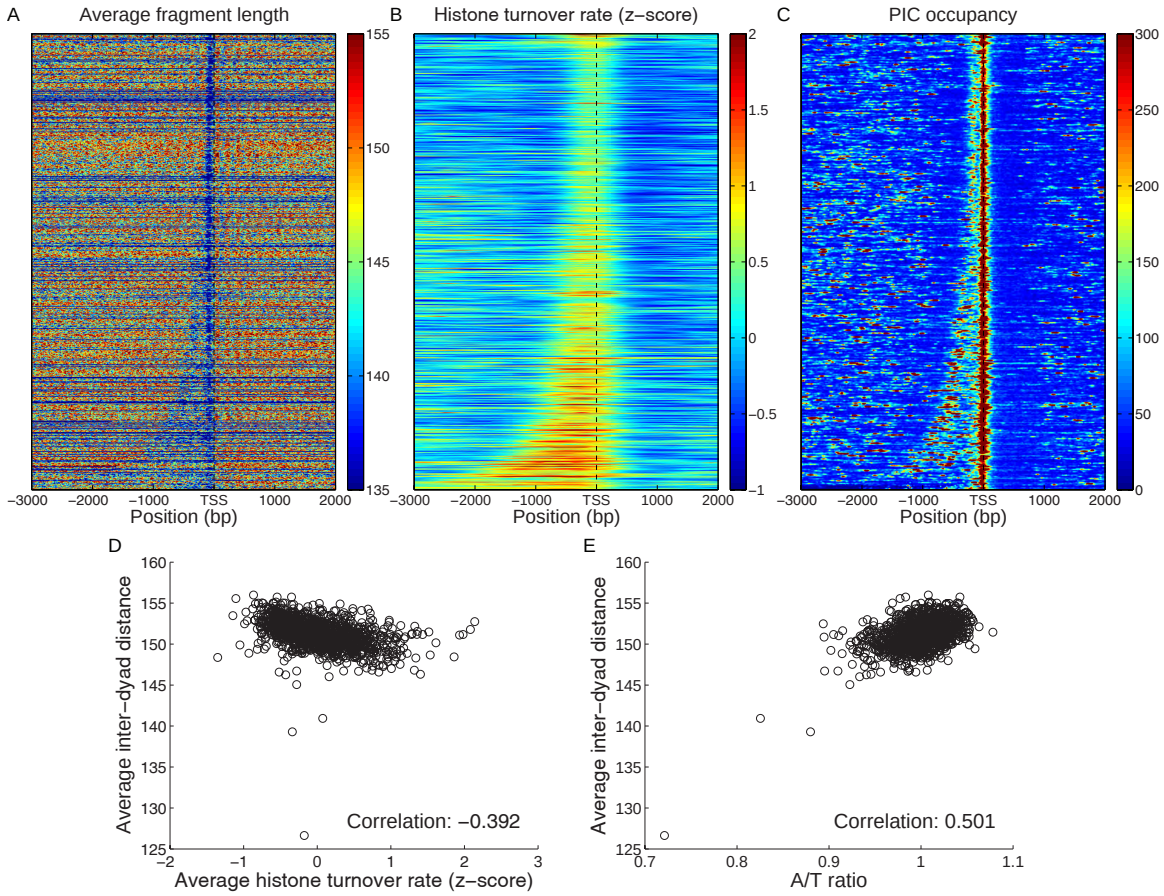


Figure 3.17: **Genome-wide distribution of nucleosome lengths, histone turnover rates, and transcription pre-initiation complexes.** (A) Distribution of average lengths of DNA-bound particles mapped by MNase digestion [146] in the vicinity of TSS. We consider particles with sizes between 80 and 200 bp, and assign particle lengths to the mid-point of each particle. Values for bps without dyads are obtained by interpolation. (B) Distribution of histone turnover rates [161] in the vicinity of TSS. (C) Distribution of the combined occupancy of 9 transcription pre-initiation complexes (PICs) [160] in the vicinity of TSS. PIC occupancies provided at 20 bp intervals in [160] are interpolated. In panels (A)-(C), the genes' order is as in Figure 3.9B, and the heatmaps are smoothed using a 2D Gaussian kernel with $\sigma = 3$ pixels. (D) Correlation between inter-dyad distances and histone turnover rates averaged over 10 kbp windows tiling the yeast genome. (E) Correlation between average inter-dyad distances and the A/T ratio in 10 kbp windows tiling the yeast genome. A/T ratio is the fraction of A/T nucleotides in the window, divided by the genome-wide A/T fraction. Correlation in (D) and (E) refers to the linear correlation coefficient.

Partially unwrapped nucleosomes tend to have elevated histone turnover rates [161], both around TSS and genome-wide (Figures 3.16C, 3.17B, 3.17D). We find that nucleosomes at loci enriched in PICs [160] are also more unwrapped (Figures 3.16C, 3.17C). Finally, inter-dyad distances tend to increase with the fraction of A/T nucleotides, indicating that nucleosomes occupying A/T-rich sequences have longer footprints genome-wide (Figure 3.17E). We note that it is misleading to equate inter-nucleosome distances with peak-to-peak distances in the average profile of nucleosome dyad counts (blue line in Figure 3.16C). The peak-to-peak distances are 164-165 bp, while the average inter-dyad distance for the nucleosomes in the [TSS, TSS+1000] region is 149.6 bp. Thus nucleosome unwrapping is much more common than could be predicted by mapping single-nucleosome positions alone.

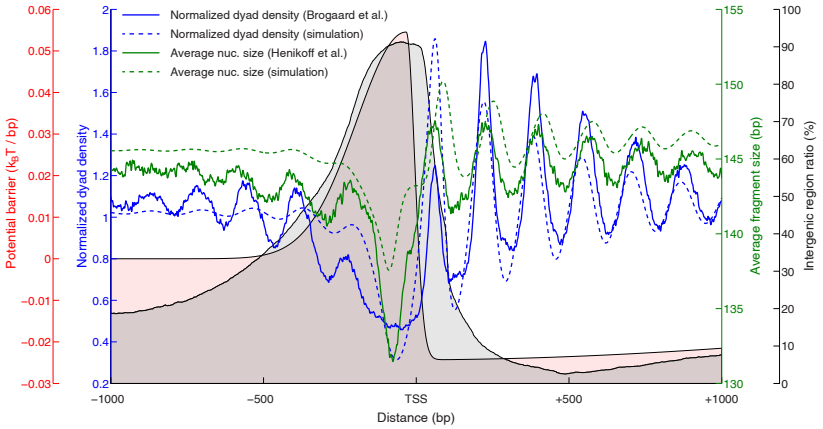


Figure 3.18: **Modeling distributions of nucleosome lengths and dyad positions in the vicinity of TSS.** As in Figure 3.16 we align all yeast genes by their TSS and for each bp, we compute the fraction of times a fixed bp is found in an intergenic region, as opposed to the ORF of a neighboring gene (grey background curve). We use the shape of the intergenic ratio as a guide for constructing an energy barrier for *in vivo* histone deposition (pink background curve). The barrier is composed of three half-Gaussians: $B(x) = H \exp\left[-\frac{(x-c)^2}{2\sigma_1^2}\right]$ $(x \leq c)$, $(H+D)\left\{\exp\left[-\frac{(x-c)^2}{2\sigma_2^2}\right]-1\right\} - D\exp\left[-\frac{(x-c)^2}{2\sigma_3^2}\right]$ $(x > c)$. The free parameters of the barrier are fit to maximize the sum of two correlations between observed and predicted normalized dyad counts [92] (solid and dashed blue lines, respectively), and between observed and predicted average nucleosome DNA lengths [146] (solid and dashed green lines, respectively). Normalized dyad counts are computed as the total number of dyads at a given bp divided by the average around the TSS. Average DNA lengths are computed for all nucleosomes with a midpoint at a given bp, for all genes. The fitted parameters are: $H = 0.0545$ $\mathrm{k_BT}, D = 0.0243$ $\mathrm{k_BT}, c = x_{\mathrm{TSS}} - 32$ bp, $\sigma_1 = 162.7$ bp, $\sigma_2 = 28.0$ bp, $\sigma_3 = 2090.9$ bp, where x_{TSS} is the absolute position of the TSS in the box, c is the center of the 3 Gaussians, H is the height of the first Gaussian, D is the depth of the third Gaussian, and $\sigma_1, \sigma_2, \sigma_3$ are the standard deviations of the three Gaussian distributions. The simulations are done in a 15 kbp box with the barrier placed at its center to eliminate the boundary effects. Unwrapping is assumed to be symmetric and the nucleosome structure-based unwrapping potential (Appendix B, Model A) is used. The total free energy, $u_{\mathrm{nuc}}(k,l)$, of a nucleosome occupying bps $k,\ldots,l$ is a sum of u^{SI}_{nuc} and $u^{SD}_{\mathrm{nuc}} = \sum_{j=k}^{l} \epsilon_j$, where ϵ_j is the value of the barrier at bp j.

3.5.5 Accessibility of nucleosomal DNA to factor binding

Partial unwrapping of nucleosomal DNA results in differential accessibility of factor binding sites with respect to their position inside the nucleosome – sites on the edges are more accessible than those closer to the dyad. In contrast, all-or-none nucleosome formation should not be sensitive to the binding site position – a nucleosome, once unfolded, liberates its entire site. Polach and Widom [28] studied differential accessibility of six restriction enzymes to their target sites. The sites were placed at various positions throughout the 5S rRNA nucleosomal sequence (Figure 3.19A). A later study used the 601 sequence and an extended set of eleven restriction enzymes (Figure 3.19B) [39]. These studies measured equilibrium constants for site exposure, $K_{\rm eq}^{\rm conf}$, which are related to the probability for a site to be accessible for binding by the equation $p_{\rm open} = K_{\rm eq}^{\rm conf}/(1 + K_{\rm eq}^{\rm conf}) \approx K_{\rm eq}^{\rm conf}$ [48].

We use our crystal structure-based unwrapping model (Figure 3.9A; Appendix B, Model A) to fit the data on site accessibility [28, 39]. Here the system consists of a single nucleosome and asymmetric unwrapping is allowed. We assume that a site becomes accessible for the enzyme only after an additional number of bps, d, have been unwrapped from the histone octamer surface [48]. We also assume that once the dyad is unwrapped from either end, the entire nucleosome is unfolded. The probability for a binding site to be accessible is given by

$$p_{\rm open}(x) = \begin{cases} 1 - {\rm Occ}_{\rm nuc}(x+d) & \text{for } x < x_{\rm d} - d, \\ 1 - {\rm Occ}_{\rm nuc}(x_{\rm d}) & \text{for } x_{\rm d} - d \le x \le x_{\rm d} + d, \\ 1 - {\rm Occ}_{\rm nuc}(x-d) & \text{for } x > x_{\rm d} + d, \end{cases}$$

where $x \in [1, 147]$ bp, $x_{\rm d} = 74$ bp is the position of the dyad, and the nucleosome occupancy is given by Equation (3.8).

Besides d, the fitting parameters of the model are the overall slope of the binding energy, ϵ, and the histone chemical potential, μ. All other parameters are as in Appendix B, Model A, with the exception of $a_{\rm min} = 1$ bp, and $a_{\rm max} = 147$ bp. For the 5S rRNA measurements [28], we obtain $\epsilon^{\rm 5S} = -0.13$ ${\rm k_B T}$/bp, $\mu = -17.5$ ${\rm k_B T}$, and $d = 23$ bp. For the 601 measurements [39], we obtain $\epsilon^{601} = -0.16$ ${\rm k_B T}$/bp,

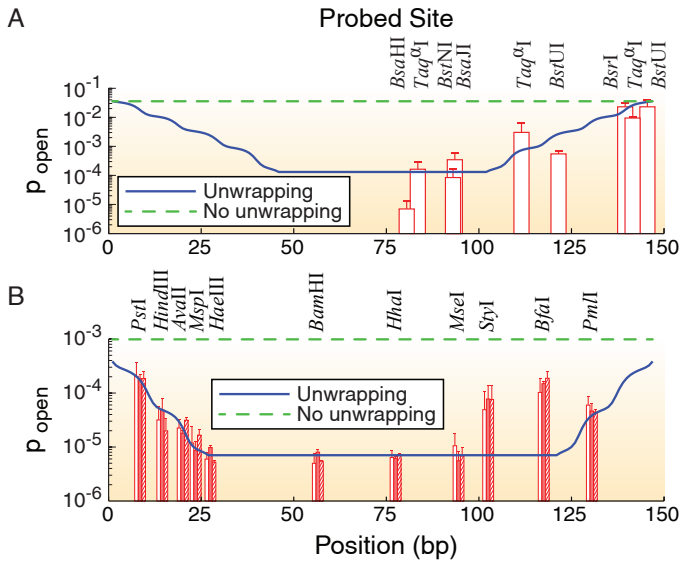


Figure 3.19: **Probability of binding site exposure within a nucleosome.** The solid blue and dashed green lines represent model predictions with and without unwrapping, respectively. In the latter case, $a_{\min} = a_{\max} = 147$ bp and all other parameters are adopted from the model with unwrapping. The dyad is fixed at bp 74. (A) Restriction enzyme sites inserted into the 5S rRNA sequence at locations indicated by the centers of vertical red bars [28]. (B) Restriction enzyme sites inserted into the 601 sequence at locations indicated by the centers of the vertical red bars in the middle of each group [39]. Each group of three bars corresponds to independent measurements in which the 601 sequence was flanked by different DNA sequences. In (A) and (B), the height of each bar is the equilibrium constant for site exposure averaged over multiple experiments (error bars show standard deviation).

$\mu = -16.4$ k_BT, and $d = 45$ bp. As expected, the nucleosome formation energy of the 601 sequence is $147 \times (\epsilon^{5S} - \epsilon^{601}) = 4.4$ k_BT more favorable than that of the 5S sequence, in agreement with the experimentally measured difference of 4.9 k_BT [54]. The nucleosome formation energy of the 601 sequence is 24.1 k_BT, close to the 23.8 k_BT estimate made on the basis of 601 unzipping experiments [114]. Interestingly, the 601 DNA has to unwrap more extensively past the binding site to allow access to restriction enzymes.

Overall, our model reproduces the observed differential accessibility of restriction enzyme binding sites with respect to the nucleosome dyad (Figure 3.19). The only outliers are *Sty*I and *Bfa*I binding sites in the 601 series which are not used in the fit and which, cannot be more open than the *Pml*I site located further away from the dyad, if unwrapping proceeds from the ends. It is possible that *Sty*I and *Bfa*I require

less extensive unwrapping in order to bind to their target sites and cleave DNA.

3.6 Nucleosome-induced cooperativity

If multiple biding sites reside within a single nucleosome, binding of one factor makes the other sites more accessible, in a phenomenon known as nucleosome-induced cooperativity [27, 40, 47]. The cooperativity disappears in the absence of nucleosomes and reduces in extent with the distance between consecutive sites [27]. Moreover, the cooperativity is not observed if the two sites are on the opposite sides of the nucleosome dyad [50].

We can use our model of nucleosome unwrapping (Appendix B, Model A with $a_{\min} = 1$ bp, and $a_{\max} = 147$ bp) to capture all these aspects of nucleosome-induced cooperativity (Figure 3.20). Specifically, for sites located more than 40 bp away from the dyad site accessibility is strongly enhanced if DNA unwrapping is allowed (Figure 3.20A). Interestingly, cooperativity between two TFs bound on the same side of the dyad is observed both with and without unwrapping (Figure 3.20B). However, without unwrapping it is impossible to show that binding on the opposite sides of the dyad is not cooperative, as observed in experiments [50] (Figure 3.20C). Furthermore, the decrease of cooperativity with distance [27] cannot be reproduced (Figure 3.20D). Thus modeling transient nucleosome unwrapping is necessary for understanding how TFs and other DNA-binding proteins gain access to their nucleosome-covered sites.

3.6.1 Sequence-dependent nucleosome positioning and unwrapping

We now focus on the sequence-dependent correction to the average free energy of nucleosome formation, $u^{SD}_{S(N)}$. We assume that $u^{SD}_{S(N)}$ depends only on the number of mono- and dinucleotides in the nucleosomal DNA, as discussed in Section 3.4.3. We consider three *in vivo* nucleosome maps in *S.cerevisiae* based on paired-end sequencing [146, 147, 148], and an *in vitro* map in which nucleosomes were assembled on yeast genomic DNA and sequenced using single-end reads [57]. In the latter case, we assume

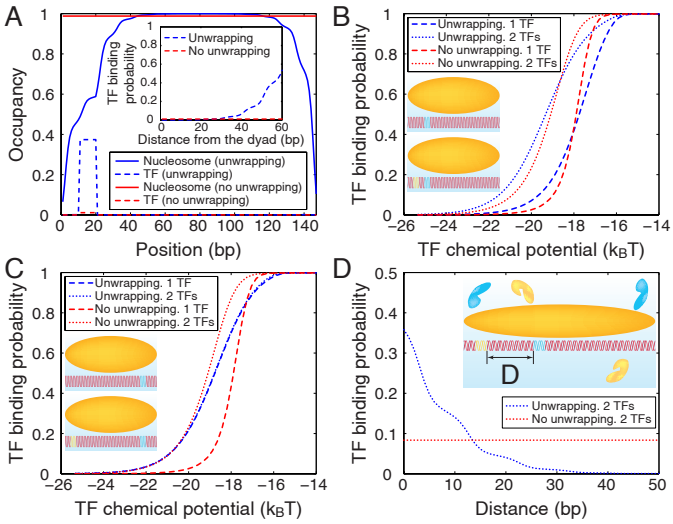


Figure 3.20: **Modes of nucleosome-induced cooperativity.** (A) TF and nucleosome occupancy with and without unwrapping. The TF binding site occupies bps 11–20. Inset: TF binding probability as a function of the distance between the nucleosome dyad and the proximal edge of the TF site, with and without unwrapping. (B) TF titration curves for one TF site vs. two TF sites located on the same side of the dyad. Site 1 occupies bps 11–20, site 2 occupies bps 31–40. Inset: Binding site locations. (C) Same as (B), but with the two TF sites located on the opposite sides of the dyad. Site 1 occupies bps 11–20, site 2 occupies bps 117–126. Inset: Binding site locations. (D) Nucleosome-induced cooperativity as a function of the distance between two TF binding sites. The binding probability of the second TF is shown. Site 1 occupies bps 11-20, while the position of the second site is variable. Inset: Definition of the distance between the two binding sites. In all panels, the free energy of a fully wrapped nucleosome is $-\ln(10^9)$ k_BT; the histone chemical potential is $\ln(10^{-6})$ k_BT; the TF binding energy is $-\ln(10^{10})$ k_BT to cognate sites, and $-\ln(10^6)$ k_BT to all other sites; the TF chemical potential is $\ln(10^{-9})$ k_BT unless varied. Asymmetric unwrapping is allowed; in the model without unwrapping, $a_{\min} = a_{\max} = 147$ bp, and all other parameters are the same as in the model with unwrapping.

that all nucleosomes have a canonical length of 147 bp. All four nucleosome mapping experiments used MNase digestion to isolate mononucleosomes. We first compute the total energy of nucleosome formation $u_{S(N)}$ using Equation (3.15). The sequence-independent part of the binding energy, u_N^{SI}, is obtained from the high-resolution map of inter-dyad distances [92] by fitting the parameters of Model A, from Appendix B, as described in Section 3.5.1. Subtracting the sequence-independent part, u_N^{SI}, from the total nucleosome energy, $u_{S(N)}$, we obtain the sequence-dependent correction, $u_{S(N)}^{SD}$. As described in Section 3.4.3, we fit the sequence-dependent model to $u_{S(N)} - u_N^{SI}$, and obtain the parameters which give the energetic contributions of all mono- and

Table 3.1: **Table of energy parameters inferred from four nucleosome maps** Energy parameters were predicted using three *in vivo* nucleosome maps based on paired-end reads from Henikoff et al. [146], Nagarajavel et al. [148], and Cole et al. [147], and one *in vitro* nucleosome map based on single-end reads from Kaplan et al. [57]. In the Kaplan et al. map, each single-end sequence read was extended to the canonical nucleosome length of 147 bp. For each nucleosome map, we obtain an estimate of $n_1^{\mathrm{nuc}}(i,j)$ by normalizing raw read counts, that is the number of nucleosomes of any length that start at a given bp so that the maximum nucleosome occupancy is 1.0 for each chromosome. We compute the total nucleosome formation energy $u(i,j)$ from $n_1^{\mathrm{nuc}}(i,j)$ using Equation (3.15). Next we subtract the sequence-independent part, $u^{\mathrm{SI}}(i,j)$, predicted using Brogaard et al. data [92] (Appendix B, Model A), and fit the parameters of the sequence-dependent correction, $u^{\mathrm{SD}}(i,j)$, as described in Section 3.4.3. Last row indicated the number of reads per bp in each dataset, N_{rpbp}.

	Kaplan et al.	Henikoff et al.	Nagarajavel et al.	Cole et al.
$\epsilon_{A/T}$	-0.180	-0.081	-0.200	-0.195
$\epsilon_{C/G}$	0.290	0.130	0.322	0.314
$\epsilon_{AA/TT}$	0.221	0.069	0.210	0.210
$\epsilon_{AC/GT}$	-0.076	-0.031	-0.055	0.017
$\epsilon_{AG/CT}$	-0.068	-0.010	-0.123	-0.130
$\epsilon_{AT/AT}$	0.201	0.089	0.141	0.198
$\epsilon_{CA/TG}$	-0.092	-0.003	-0.085	-0.166
$\epsilon_{CC/GG}$	-0.305	-0.138	-0.324	-0.306
$\epsilon_{CG/CG}$	-0.319	-0.169	-0.396	-0.462
$\epsilon_{GA/TC}$	-0.054	-0.023	0.013	0.016
$\epsilon_{GC/GC}$	-0.315	-0.162	-0.253	-0.148
$\epsilon_{TA/TA}$	0.189	0.090	0.246	0.173
N_{rpbp}	1.02	5.53	0.81	2.71

dinucleotides to this correction. The obtained parameters are summarized in Table 3.1.

The number of unwrapped nucleosome species may be as high as several thousand, depending on the maximum extent of unwrapping, and the available levels of read coverage, that is the mean number of reads starting at a bp, are relatively low, see Table 3.1. In the absence of high-resolution, high-coverage experimental data, we have tested our ability to predict nucleosome unwrapping energetics using a realistic model system, with a limited read coverage. We tested a series of read coverages of 1, 10, and 100 reads per bp. Specifically, we assume that the sequence-depending part of the binding energy, $u_{S(N)}^{SD}$, is given by the energy parameters inferred from the Henikoff et al. dataset [146] (Table 3.1), and the sequence-independent part, u_N^{SI}, is defined as in

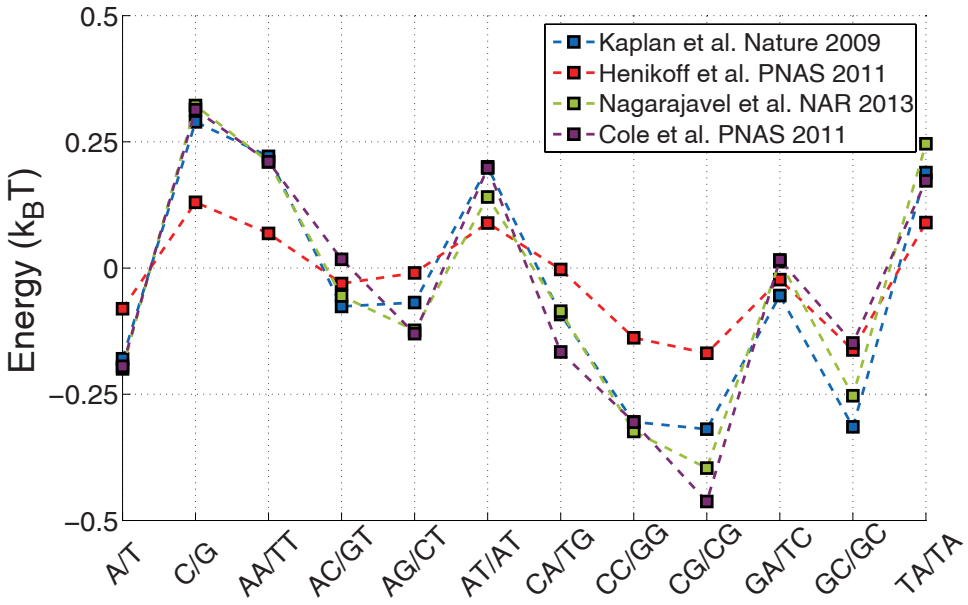


Figure 3.21: **Energy parameters inferred from four nucleosome maps.** Energy parameters obtained using three *in vivo* nucleosome maps based on paired-end reads from Henikoff et al. [146], Nagarajavel et al. [148], and Cole et al. [147], and one *in vitro* nucleosome map based on single-end reads from Kaplan et al. [57]. The WW dinucleotides, W denoting either an A or a T nucleotide, have the highest energies within the set of ten unique dinucleotides. Also, the SS dinucleotides, with S denoting either a C or a G nucleotide, have the lowest energies, while the mixed dinucleotides containing a W and a S nucleotide, have intermediate energies.

Appendix B, Model A. Using Equation (3.3), with the chemical potential $\mu = -13$ k_BT, we compute the exact nucleosome distribution $n_1^{\text{nuc}}(k,l)$ for the *S.cerevisiae* chromosome I. We sample paired-end nucleosomal reads (k,l) from the exact distribution, $n_1^{\text{nuc}}(k,l)$, until a desired level of read coverage is reached. From this finite sample, we construct a histogram of nucleosome lengths, $P(N)$, and use it to optimize the parameters of the unwrapping potential u_N^{SI}. Next, we use Equation (3.15) to predict the total binding energy, $u_{S(N)}$, from the same sample, and fit the sequence-dependent correction $u_{S(N)} - u_N^{SI}$ as described in Section 3.4.3, assuming that the dyad is at the mid-point of each particle. Finally, using the fitted binding energy components u_N^{SI} and $u_{S(N)}^{SD}$, we compute the predicted nucleosome dyad distribution and coverage, which are then compared with the exact distributions.

We find that we are able to infer the unwrapping potential even at modest levels of

read coverage (Figure 3.22A, B). The overall slope of the potential is slightly overestimated, likely because the histogram of particle lengths is affected by well-positioned nucleosomes with negative formation energies. The average of these energies may bias the slope. Nucleosome occupancies and dyad positions are reproduced reasonably well using at minimum of 10 reads per bp, but at least 100 reads per bp are required to recover the energy parameters (Figure 3.22C).

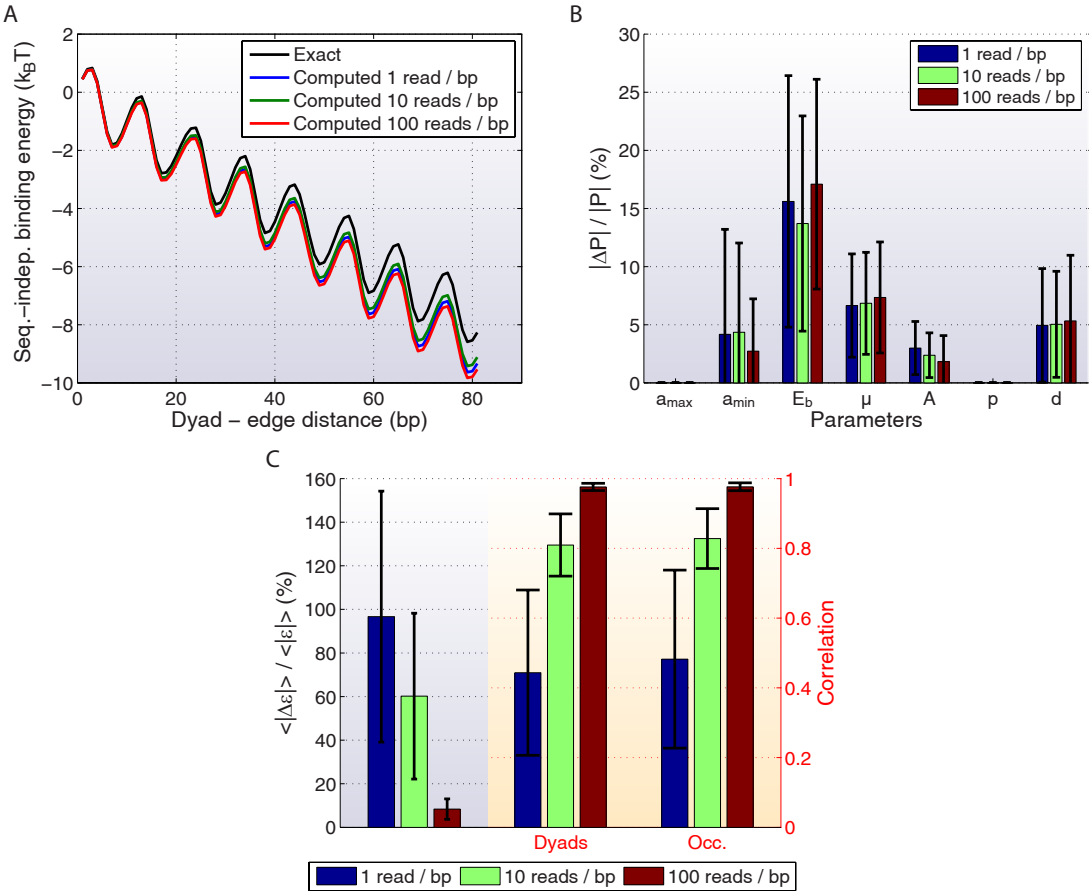


Figure 3.22: **Inference of the unwrapping potential and sequence-specific nucleosome formation energies in a model system.** (A) The exact and predicted unwrapping potentials at three levels of sequence coverage. All calculations are done using the DNA sequence from chromosome I of *S. cerevisiae*. $M \times L$ reads are randomly sampled from the exact nucleosome distribution, n_1^{nuc}, corresponding to the total binding energy $u_{\text{nuc}} = u_{\text{nuc}}^{\text{SI}} + u_{\text{nuc}}^{\text{SD}}$ [Equation 3.3], where $u_{\text{nuc}}^{\text{SI}}$ is given by Model A in AppendixB, $u_{\text{nuc}}^{\text{SD}}$ is defined by a set of energy parameters inferred from the Henikoff et al. nucleosome map [146] (Table 3.1), and $M \in \{1, 10, 100\}$ is the desired level of read coverage per bp. Sampled reads are used to compile a chromosome-wide histogram of nucleosome DNA lengths, to which the unwrapping potential in Model A is fit by using a genetic algorithm optimization function `ga` from the MATLAB Global Optimization toolbox to minimize the root-mean-square error of the predicted distribution of nucleosome lengths. (B) Relative errors between predicted and exact parameters of the unwrapping potential, described in Appendix B, Model A, and predicted and exact chemical potential at three levels of sequence coverage. P denotes any parameter on the horizontal axis. (C) Relative errors between predicted and exact energy parameters (Table 3.1, Henikoff et al. nucleosome map [146]) (light blue background). Linear correlation coefficients, between predicted and exact distributions of dyad positions and nucleosome occupancy (light pink background). The height of each bar in (B) and (C) represents the mean relative error for the corresponding parameter or the mean correlation coefficient, obtained by averaging the results of a hundred random sampling experiments. The uncertainty intervals represent standard deviations.

Chapter 4

Other joint projects

In this chapter I present a short summary of the other projects in which I have been involved during my PhD studies. These projects are collaborations with three labs: James Broach's lab from Princeton University (now at Penn State University), Yuri Moshkin's lab from Erasmus University, Rotterdam, The Netherlands, and Stefan Björklund's lab from Umeå University, Umeå, Sweden.

This Chapter is based on our work which was published in [162], and on two ongoing research projects [163, 78, 52].

4.1 Msn2 signaling

In order to survive, the yeast *S. cerevisiae* needs to sense and respond to various environmental conditions, such as nutrient availability, osmolarity, and temperature. Numerous signaling pathways responsive to environmental conditions concentrate on the general stress response pathway, mediated predominantly by the transcription factor Msn2. In favorable growth conditions, Msn2 remains in the cytoplasm, but upon a stress it moves into the nucleus and activates many genes that help to protect cells from a variety of stresses. Because Msn2 is a node at which many signaling pathways converge, it can serve as a useful model for understanding how a cell integrates information from multiple, and possibly conflicting inputs.

In our study [162] we carry out experimental and computational analyses in order to reveal the principles which dictate the diverse behaviors of genetically identical cells. Specifically, we provide mechanistic insights into the recently described "bursting" behavior of cellular transcription factors [164], a process that has been proposed to allow coherent transcription of many genes, but for which little molecular explanations are

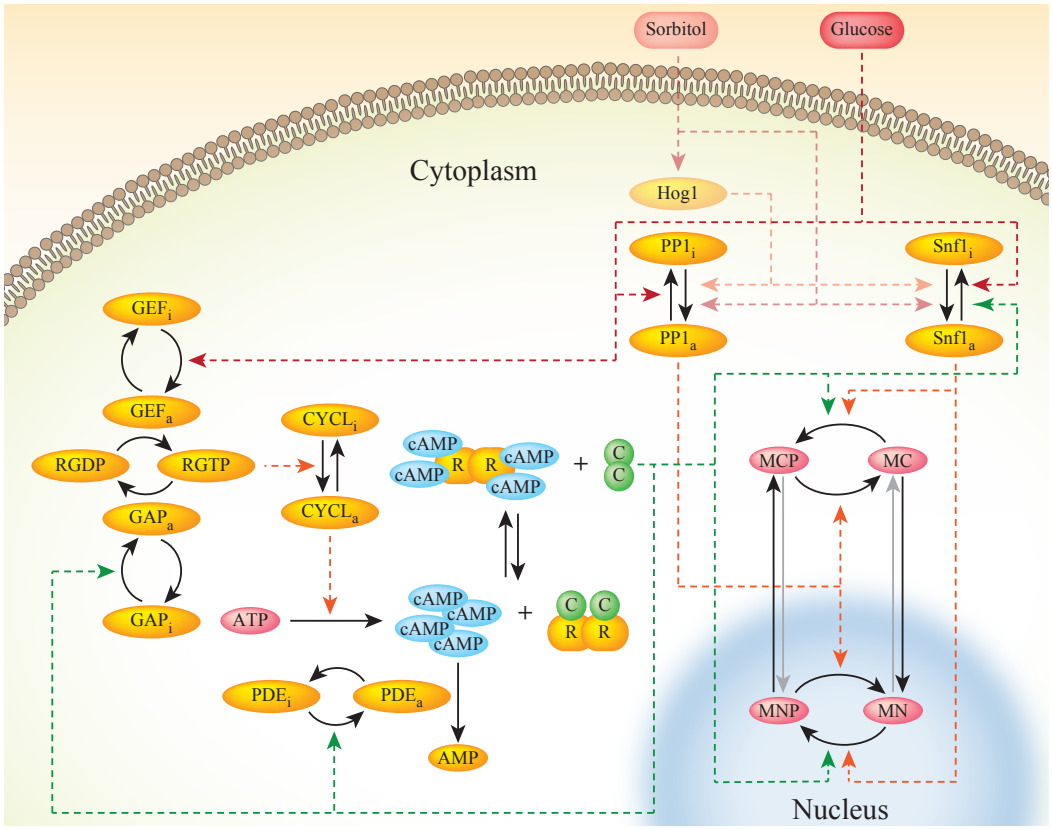


Figure 4.1: Our proposed Msn2 signaling network [162], showing the Ras/PKA branch, Snf1, PP1, and Hog1. MCP/MNP: phosphorylated cytoplasmic/nuclear Msn2; MC/MN: unphosphorylated cytoplasmic/nuclear Msn2; C: catalytic subunit of PKA; R: regulatory subunit of PKA; RGTP and RGDP: Ras bound to GTP and GDP, respectively; CYCL: adenylyl cyclase; GAP: GTPase activating proteins Ira1/2; GEF: GDP/GTP exchange factor Cdc25; PDE: phosphodiesterase. The subscripts "a" and "i" indicate active and inactive forms of proteins, respectively. Dashed lines show catalytic reactions, while solid lines mean physical transitions.

available.

Single-cell studies have revealed that Msn2 responds to many stresses in an unusual behavior of irregular and cell-autonomous oscillations into and out of the nucleus, referred to as "bursting". The frequency, duration and amplitude of Msn2 oscillations have vary as a result of differing stresses [165], but what causes these oscillations is still unknown. Similar translocations between the nucleus and cytoplasm were previously observed in the case of the calcium-sensitive transcription factor Crz1, which changes the frequency of these oscillations depending on the strength of the calcium signal [164].

The important role of the stochastic noise in the biology of cells has become evident in the recent years. Observations that genetically identical cells, including cancer cells, can display distinct behaviors, leading to different cell fates [166] has drawn attention of the scientific community. We note that the stress response of Msn2 has both deterministic and stochastic components, further making it a useful model for studying signaling and transcription in cells. This stochastic behavior of Msn2 allows identical yeast cells to respond in different ways to identical stimuli, providing a novel mechanism for yeast as a community to survive in an uncertain environment.

To study the convergence of signals on Msn2 and how they affect its behavior, a flow-chamber was used to monitor fluorescently labeled Msn2 in individual live cells in real time using fluorescence microscopy. The flow-chamber allowed an almost instantaneous switch between different experimental media, which acted as the signal for Msn2 in this system [162]. We find that Msn2 responds to most stresses, such as a limitation in glucose or nitrogen in the media, by an initial coherent displacement from the cytoplasm into the nucleus, followed by random translocations in and out of the nucleus, the pattern of which differs from cell to cell even in a genetically identical population. We determine that this behavior is caused by the interplay between several signaling pathways (Figure 4.1), including Ras/Protein kinase A, AMP activated kinase, the HOG map kinase pathway, and Protein Phosphatase 1. In addition, we show that noise in the regulation of Msn2 results in diverse behaviors of genetically identical cells. Using stochastic modeling, we reproduce through computer simulations the responses of Msn2 to different stresses and demonstrate that the noisy cycling in and out of the nucleus arises from the small number of Msn2 molecules in the cell. The resulting diversity in the behavior of genetically identical cells may allow cell populations to optimize their responses to an unpredictable environment.

For this project, apart from analysing the behavior of the signaling pathways responsible for the dynamics of Msn2, I also built several graphical user interfaces (GUIs) (Figure 4.2) which allowed my collaborators to track yeast cells in live cell videos and to record the nuclear localization of fluorescently tagged Msn2 in an automated way. Using these GUIs, the process of obtaining quantitative data from many experiments

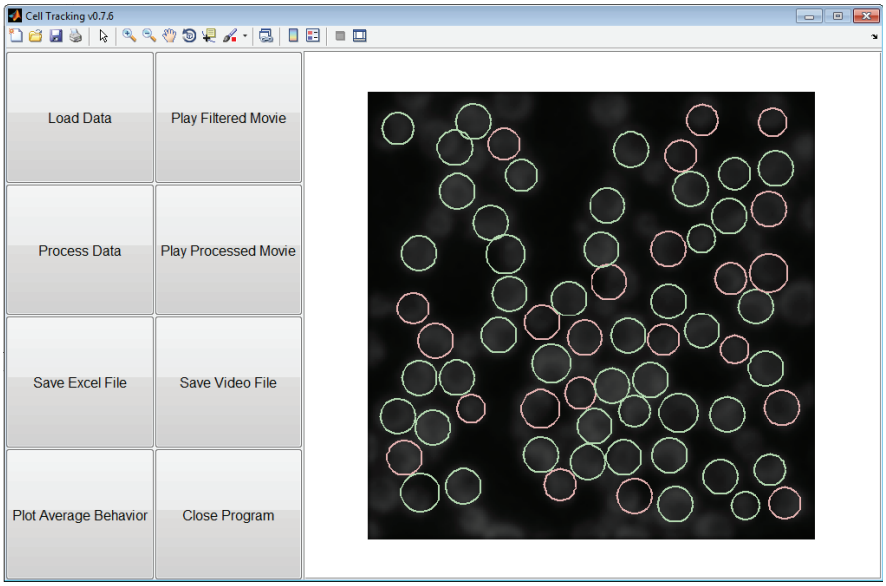


Figure 4.2: Snapshot of the cell tracking program user interface. The program takes as input a live cell video and it tracks frame-by-frame the motion of the yeast cells in the field of view. Some cells remain in the field of view during the entire experiment, and these are marked in green. The cells which disappear from the field of view during the experiment, or which are undetectable in some frames of the movie, are marked in red and are discarded from further analysis. The program records the fluorescence intensity levels of all the pixels corresponding to each cell, and also computes the ratio of fluorescently labeled Msn2 protein in the nucleus and in the cytoplasm at each time point. At the end of the image processing, the user is allowed to plot and save the results in an Excel file.

was made more efficient.

This work was done in collaboration with Natalia Petrenko, Megan McClean, James Broach and Alexandre Morozov.

4.2 Msn2-Mediator-nucleosome interplay

In order to increase gene transcription, activator proteins need to recruit other factors refered to as coactivators. Mediator is one such coactivator, a multiprotein complex which acts as a bridge between the activators and the RNA polymerase II (RNAP II) transcription machinery. In *S. cerevisiae*, the Mediator complex contains 21 subunits and it interacts directly with the carboxy terminal domain (CTD) of the largest subunit of RNAP II [168]. Activators, such as Msn2, and coactivators, such as Mediator, can

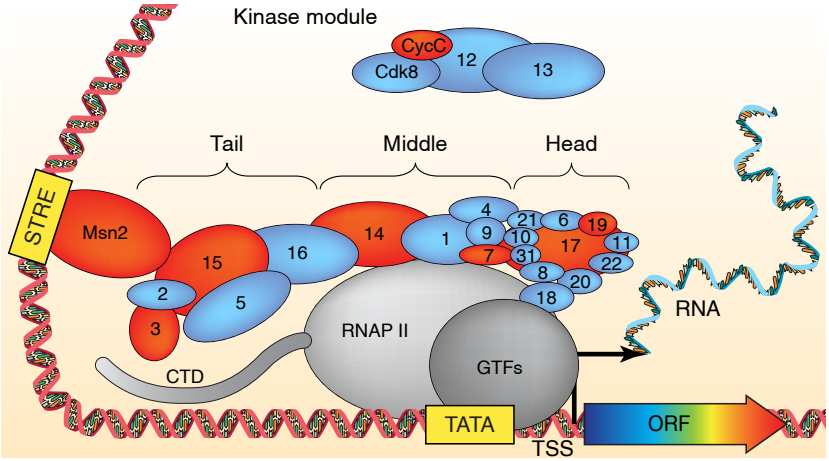


Figure 4.3: Mediator functions as a bridge between Msn2 and the general RNAP II transcription machinery at the promoter. Msn2 interacts with the tail region of Mediator, while RNAP II interacts with the head and middle regions. A subgroup of Mediator components (Med12, Med13, Cdk8 and CycC) forms a kinase module (Srb811) that is involved in negative regulation of transcription. Only Mediator lacking the Srb811 module can associate with RNAP II [167]. GTFs represent the general transcription factors, CTD is the carboxy terminal domain of RNAP II, ORF denotes the open reading frame and TSS the transcription start site. Proteins labelled in red represent those that were tagged with 13xMyc tags for the ChIP-sequencing assays.

facilitate access of the transcriptional machinery to the DNA by causing the removal of nucleosomes. Alternatively, nucleosomes can block DNA binding by activators in the absence of the appropriate signal. The interaction between activators and nucleosomes has been studied in the cases of several individual genes but is still poorly understood on a genome-wide scale, especially in dynamic conditions where cells are growing or responding to environmental stimuli.

An ongoing project in which I have been involved is to study the changes in DNA binding of several transcription related proteins (Msn2, Mediator, RNAP II) in response to a transcriptional switch, and the corresponding changes in nucleosome organization and gene expression. [78, 163]. To study the genome-wide transcriptional regulation in budding yeast *S. cerevisiae*, we carry out a glucose to glycerol media switch where about half of the yeast genes change their expression levels twofold or more within 15 minutes. Before and after this switch, we measure the global transcription levels, monitor the nuclesome occupancy profiles and the chromatin immunoprecipitation (ChIP) profiles

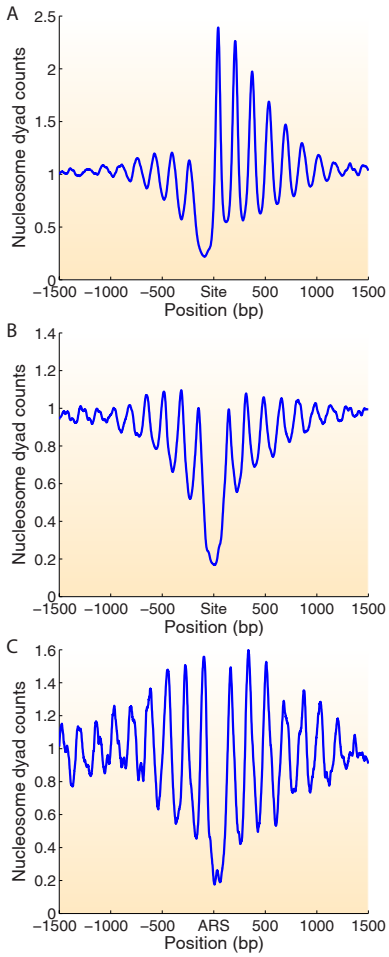


Figure 4.4: (A) The canonical nucleosome phasing near the TSS. (B) Nucleosome phasing near the binding sites of the transcription factors Gal4, Phd1, Rap1, and Reb1, obtained from [169]. (C) Nucleosome phasing near the origins of replication, obtained from [170].

of Msn2, as well as several subunits of Mediator (Med3, Med7, Med14, Med15, Med17, Med19 and CycC) and RNAP II (CTD) (Figure 4.3).

We find that nucleosome arrays around promoters exist in several distinct classes of patterns and that nucleosome rearrangements in response to stress are correlated with the expression change of the corresponding genes. For example, the class of promoters with the most extended nucleosome depleted region is highly enriched for genes involved in different stress response pathways, and it is also enriched for Mediator and Msn2 binding. We also present evidence for both activating and repressive functions of

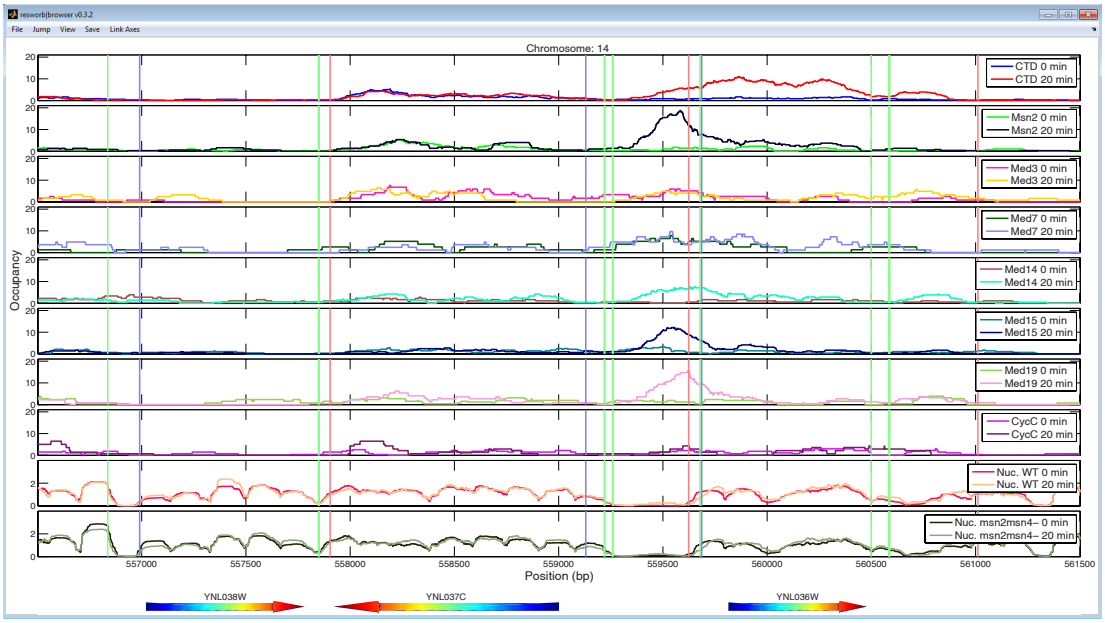


Figure 4.5: The user interface of the genome browser. We use this browser to display in parallel the occupancy profiles and compare the dynamics of different DNA-binding proteins. The browser displays the locations of the ORFs (colored arrows), TSS (blue vertical line), TATA boxes (green vertical line), and STRE elements (red vertical line). In this example, we show the ChIP profiles corresponding to 0 and 20 min after the glucose removal. The individual panels show the binding profiles corresponding to CTD, Msn2, Med3, Med7, Med14, Med15, Med19, CycC, nucleosomes in WT cells and nucleosomes in *msn2msn4* double deletion mutant, respectively.

Mediator and Msn2. Finally, we argue for the existence of several Mediator subcomplexes, composed of different subunits and with distinct roles in transcription [163].

Having paired-end sequencing data, we are able to map with high precision the nucleosome organization in *S. cerevisiae*. For example, we detect important nucleosome phasing, not only near the TSSs, but also near the binding sites of different TFs and near the autonomously replicating sequences (ARS) (Figure 4.4). ARS are the locations where the origin recognition complex (ORC) and the mini chromosome maintenance (MCM) protein complex bind before DNA starts to be replicated. We obtained the locations of the origins of replication from [170]. From the high-resolution ChIP-exo experiments by Rhee and Pugh [169], we use the binding sites of Gal4, Phd1, Rap1, and Reb1. The binding of these transcription factors to the DNA creates important potential barriers for the histones and these become phased, organizing in regular arrays.

For this project, I also developed several GUIs which allow the user to scan along chromosomes and visualize the distribution of different DNA-binding proteins, together with important genomic features, as for example, open reading frames, transcription start/termination sites, TATA boxes, and STRE elements (Figure 4.5). Having a simple tool which permits an easier visualization of all the ChIP profiles in parallel, we study important loci along the yeast genome on an individual basis, in a time efficient way.

This work was done in collaboration with Nils Elfving, Alexandre Morozov, James Broach and Stefan Björklund.

4.3 Fragile nucleosomes

Micrococcal nuclease (MNase) digestion is a widespread method for mapping nucleosome positions. Chromatin is cross-linked, and MNase is added to digest DNA unprotected by nucleosomes. The protected DNA is subsequently recovered and analyzed. However, it is also well known that this technique is subject to some bias, due to intrinsic MNase sequence preference, as well as some variation depending on the extent of MNase digestion performed. Nevertheless, this method gives clean and reproducible nucleosome profiles and, if properly understood, can continue to be a valuable one.

Using differential MNase digestion of chromatin, Xi et al. [51] identified throughout the yeast genome a special group of nucleosomes termed "fragile" nucleosomes. About 1000 of these unstable nucleosomes were detected at locations previously believed to be free of nucleosomes. In Figure 4.6 we show the nucleosome organization that was obtained in [51]. It is clear that in the complete digestion experiment many "fragile" nucleosomes from the inter-genic regions were lost and not properly mapped to the yeast genome.

We study the positioning of "stable" (MNase-resistant) and "fragile" (MNase-sensitive) nucleosomes in Drosophila Schneider 2 (S2) cells [52] by using two concentrations of MNase for chromatin digestion. We find that the nucleosome density is similar in the genic and intergenic regions (Figure 4.7, right panel). Although intergenic regions appear to be nucleosome free when the typical concentration of MNase is used for DNA

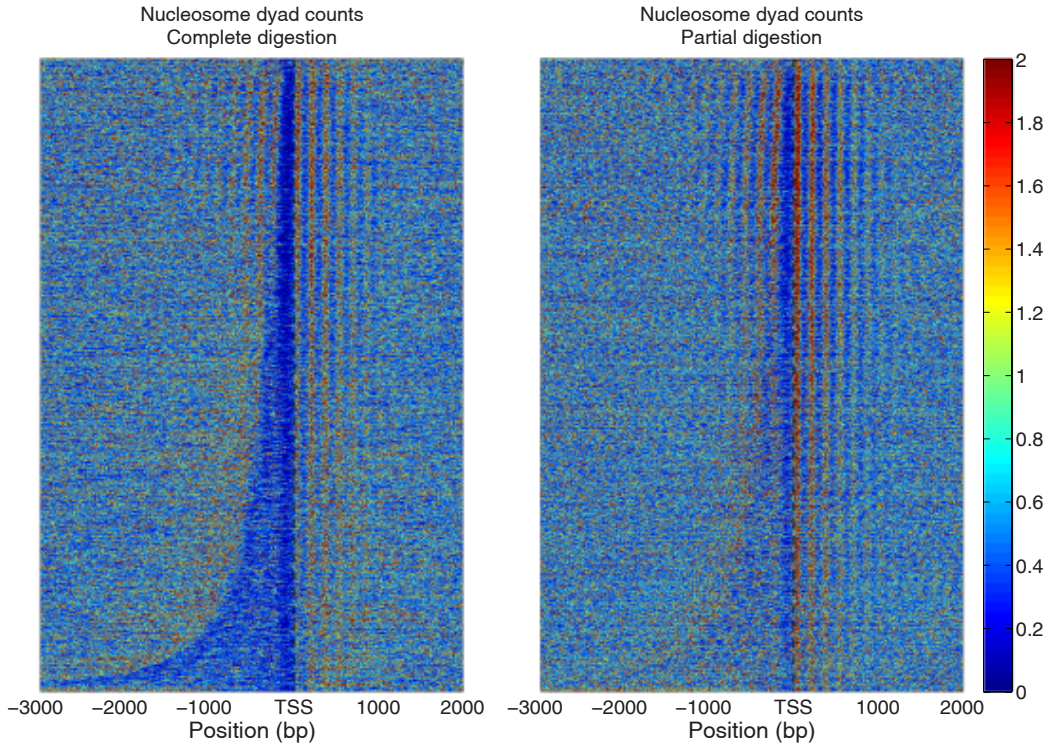


Figure 4.6: Centers of mononucleosomal DNA fragments recovered at two time points during the MNase digestion. A partial digestion sample was obtained at an early stage of the digestion, when only a minor portion of the chromatin (∼10%) was reduced to mononucleosomes. A complete digestion sample was obtained at a later time point when nearly all chromatin was reduced to mononucleosomes [51]. The inter-genic regions which seem to have a lower nucleosome density in the completely digested sample (left panel) are actually occupied by unstable nucleosomes (seen in the partially digested sample, right panel), which are easily detached from the DNA by the activity of MNase. This figure is based on data from Xi et al. [51].

digestion (Figure 4.7, left panel), when we use a 20-fold smaller concentration of MNase, many new nucleosomes are detected in the "nucleosome depleted regions" (Figure 4.7, right panel).

The heat maps in Figure 4.7 contain the Drosophila genes clustered in two groups, according to gene expression. The top clusters contain the active genes, while the bottom cluster contain the inactive genes. We notice that only the active genes contain highly phased arrays of nucleosome. In the inactive genes, there is a lower nucleosome density in the intergenic region (probably because of the competition with different

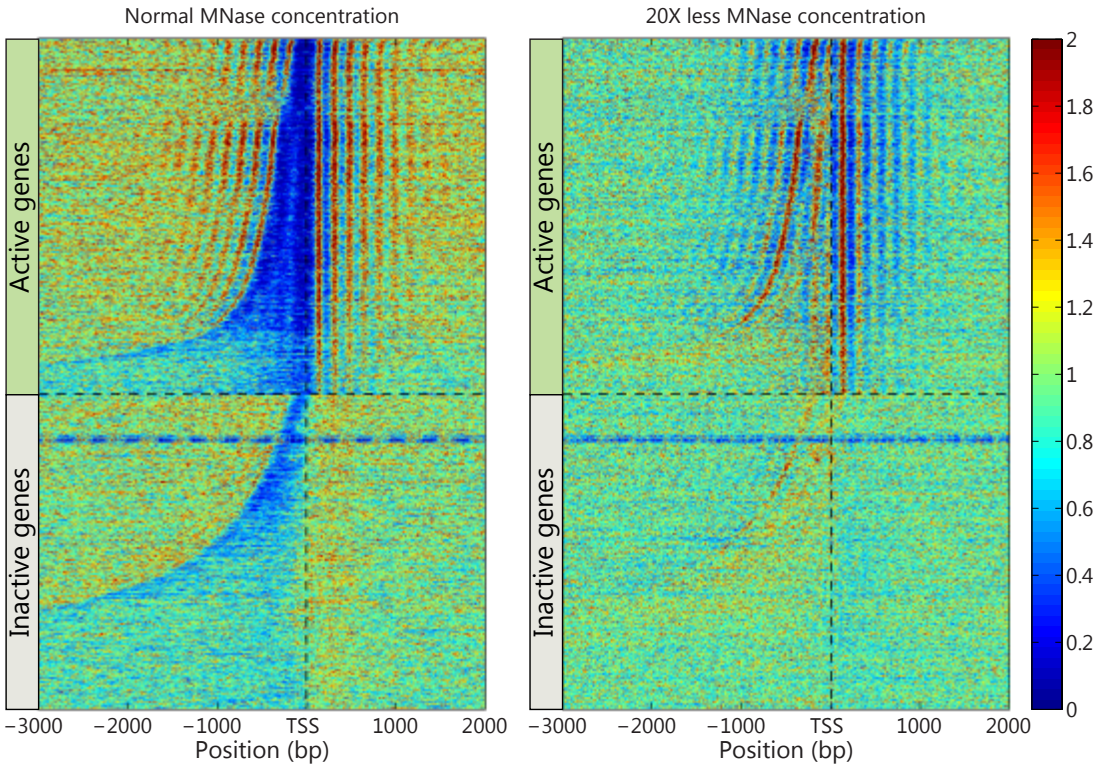


Figure 4.7: Nucleosome distribution in Drosophila S2 cells. The left panel shows the mapped mononucleosomes which are obtained by MNase digestion with the typical MNase concentration, while the right panel shows the nucleosome organization which is obtained when using 20-fold less MNase concentration to digest the chromatin. Many loose nucleosome which are found in the intergenic regions are lost after the complete digestion (left panel). The genes are clustered by the expression level and sorted in the increasing order of the intergenic lengths in each cluster. Only the active genes present highly phased nucleosomes. Interestingly this situation is reversed in *S. cerevisiae*, where the most active genes have the least regular arrays of nucleosomes.

DNA-binding proteins which try to gain access to their target sites), and the nucleosome density is almost constant downstream TSS, lacking the typical oscillatory pattern. When genes become active, different transcription pre-initiation factors attach to the corresponding promoters, and these create strong potential barriers which help to create the regular nucleosome arrays by statistical positioning. We model the nucleosome distributions from the active and inactive genes. The average nucleosome density from the experiment using the normal concentration of MNase, is shown in Figure 4.8 A by

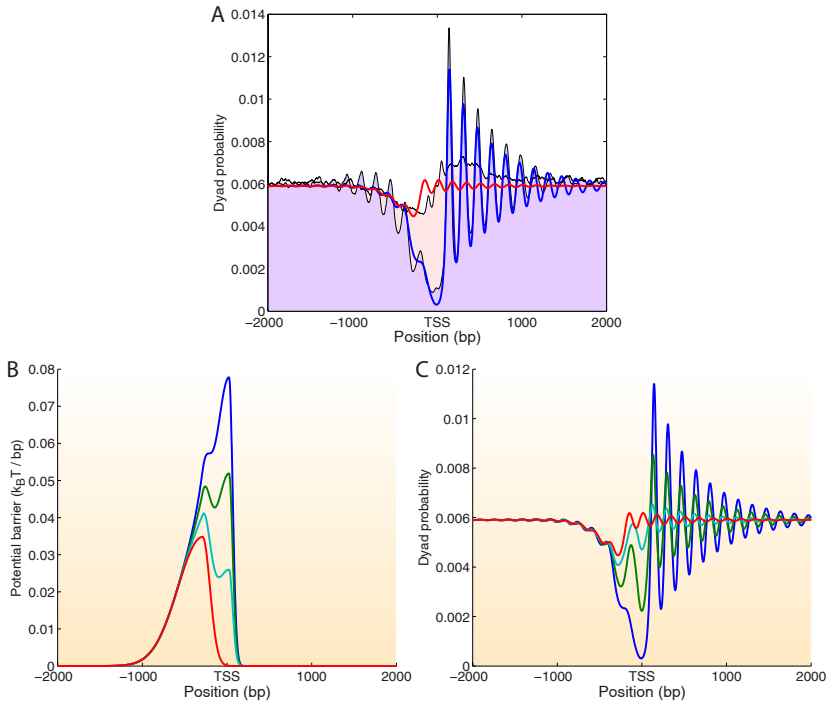


Figure 4.8: (A) The nucleosome dyad distribution near TSS in active (blue background) and inactive (pink background) genes in Drosophila S2 cells. Similar distributions can be obtained by simulating the distribution of hard rods of length 147 bp near a potential barrier positioned near TSS. We use 4 different potential barriers (B) to simulate the corresponding particle distribution in a window centered around the TSS. The potential barriers have 2 components: one barrier with fixed height and one additional barrier with variable height. The resulting potential energy is shown in (B) and the corresponding predicted particle distributions are shown in (C). Increasing potential barriers generate better phasing of the nearby nucleosome arrays. The measured nucleosome distributions are reproduced reasonably well in this simple model (blue and red lines in panel A).

the blue and pink background, corresponding to the active and inactive genes, respectively. The distribution of the nucleosomes in the inactive genes can be obtained by simulating the distribution of a system of hard rods near a potential barrier as the one shown in Figure 4.8 B (red line). This potential barrier might be generated by DNA sequences from the promoters which are unfavorable to nucleosome formation, or by the competition with other DNA-binding proteins which have their target sites upstream TSS. The obtained nucleosome distribution is shown in Figures 4.8 A, C (red lines). If the concentration of transcription pre-initiation complexes increases in the promoters of

the active genes, an additional potential barrier appears (Figure 4.8 B), and the full potential barrier (Figure 4.8 B, blue line) can generate the typical oscillatory nucleosome distribution as shown in Figures 4.8 A, C (blue lines). Increasing potential barriers (Figure 4.8 B) generate increasing degrees of nucleosome phasing (Figure 4.8 C), as expected.

The potential barrier which reproduces a similar nucleosome distribution to the one observed in the inactive genes (red line in Figure 4.8 B) has the analytic form

$$B_{\text{inactive}}(x) = \begin{cases} H_1 \exp\left[-\frac{(x-c_1)^2}{2\sigma_{L1}^2}\right] & \text{if } x \leq c_1, \\ H_1 \exp\left[-\frac{(x-c_1)^2}{2\sigma_{R1}^2}\right] & \text{if } x > c_1, \end{cases}$$

with $H_1 = 0.035$ $\text{k}_\text{B}\text{T}$, $c_1 = x_{\text{TSS}} - 288$ bp, $\sigma_{L1} = 292$ bp, and $\sigma_{R1} = 87$ bp. In order to reproduce the distribution of the nucleosome distribution in the active genes, we use the potential barrier (blue line in Figure 4.8 B)

$$B_{\text{active}}(x) = B_{\text{inactive}}(x) + \begin{cases} H_2 \exp\left[-\frac{(x-c_2)^2}{2\sigma_{L2}^2}\right] & \text{if } x \leq c_2, \\ H_2 \exp\left[-\frac{(x-c_2)^2}{2\sigma_{R2}^2}\right] & \text{if } x > c_2, \end{cases}$$

where the parameters of the additional barrier are $H_2 = 0.078$ $\text{k}_\text{B}\text{T}$, $c_2 = x_{\text{TSS}} + 24$ bp, $\sigma_{L2} = 182$ bp, and $\sigma_{R2} = 43$ bp.

A long-standing problem in nucleosome positioning was to understand the mechanisms that allow cells to both organize their genomes into compact chromatin fibers, and to also keep the DNA accessible to many factors and enzymes. The discovery of ATP-dependent nucleosome remodeling complexes has been a big step forward. The remodelers weaken the histone-DNA interactions, and facilitate the nucleosome sliding along the DNA, thereby increasing the accessibility of proteins to their binding sites.

Because histones and DNA are held together by a large number of interactions which have to be disrupted in order to reposition the nucleosomes, the remodelers require energy which is obtained from ATP hydrolysis. The common component of all nucleosome remodelling factors is a dedicated ATPase domain which catalyzes the decomposition of ATP into ADP and a free phosphate ion, releasing energy which is further used by remodelers to slide the nucleosomes.

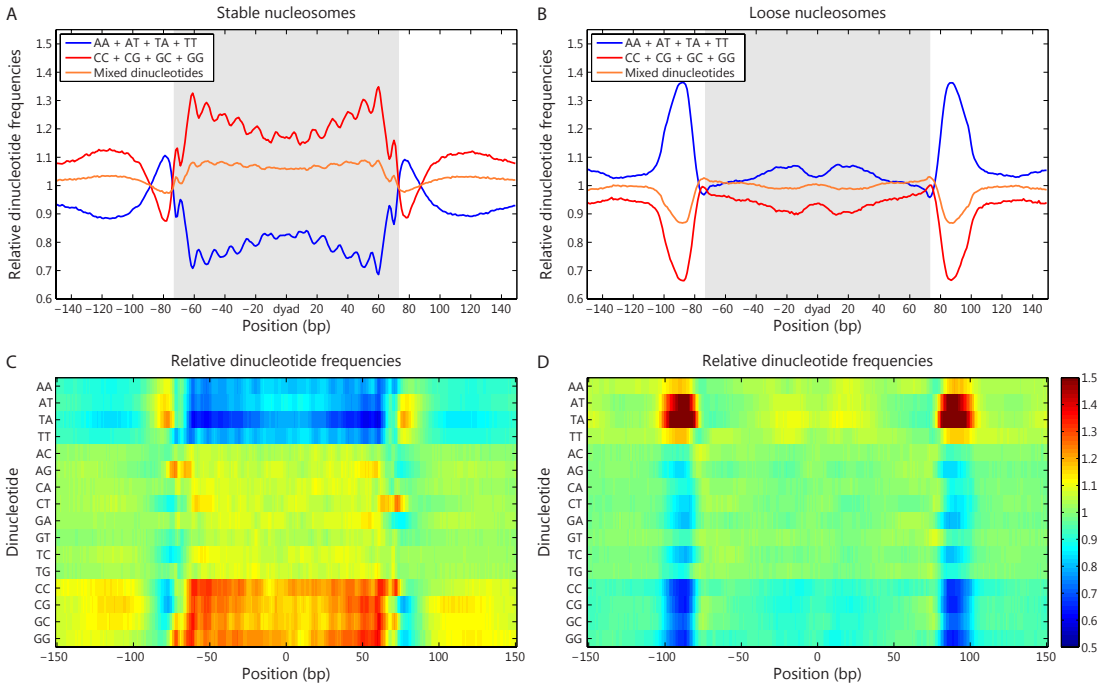


Figure 4.9: Dinucleotide frequencies in the sequences occupied by stable (A, C) and loose (B, D) nucleosomes. The stable nucleosomes are formed on G/C reach sequences, which also have oscillatory dinucleotide distributions (A, C). These nucleosome positioning signals are absent from the sequences where loose nucleosomes are detected (B, D). In panels A and B we show the relative dinucleotide frequencies corresponding to WW (blue lines), SS (red lines) and WS or SW (orange lines) dinucleotides. These frequencies are relative to the genome-wide dinucleotide frequencies in the *Drosophila melanogaster* genome. In panels C and D, we show the individual dinucleotide frequencies for all 16 cases, as heat maps. Each row corresponds to a specific dinucleotide, and the windows contain 300 bp, centered at the nucleosome dyads.

We study the action of two nucleosome remodeling complexes: ISWI (imitation switch, for a review, see [171]) and NuRD (nucleosome remodeling and deacetylation, for a review, see [172]). We find that the remodelers ISWI and NuRD have different actions on the nucleosomes arrays – ISWI is increasing the spacing between neighboring nucleosomes, while NuRD does the opposite.

In order to account for the behavior of the two different classes of nucleosomes, we compute the dinucleotide frequencies in the DNA sequences which are occupied by the stable and loose nucleosomes. We see that the stable nucleosomes occupy G/C rich sequences, unlike the loose nucleosomes (Figure 4.9 A, B). The heatmaps from

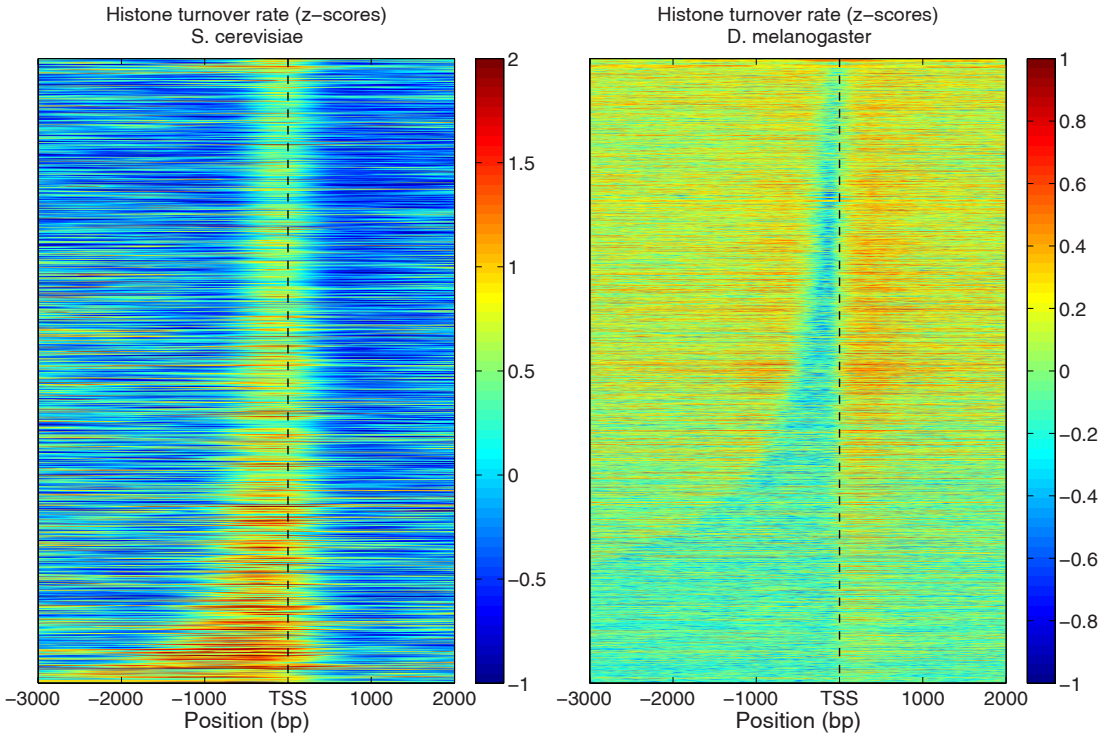


Figure 4.10: Histone turnover rates, normalized as z-scores, in *S. cerevisiae* (left panel) and *Drosophila* (right panel). This Figure is based on data from Deal et al. [174] and Dion et al. [161].

Figure 4.9 C, D show the distributions of all 16 dinucleotides, while Figure 4.9 A, B show the distributions of groups of similar dinucleotides. For the stable nucleosomes, we see oscillating dinucleotide distributions which are believed to be important nucleosome positioning signals [173].

We also study the dynamics of the histones, and we note that in *Drosophila*, the histone turnover rates are reduced in the intergenic regions [174] , as opposed to the situation in *S. cerevisiae* [161]. As we show in Figure 4.10, in *S. cerevisiae*, the most dynamic nucleosomes are the ones which occupy the intergenic regions. We saw in Chapter 3 that these nucleosomes are also more unwrapped, which is another indication that the corresponding DNA information must be readily accessible in *S. cerevisiae*. The yeast cells resolved this problem by positioning in that region nucleosomes which are less tightly bound, as found by Xi et al. [51] (see Figure 4.6).

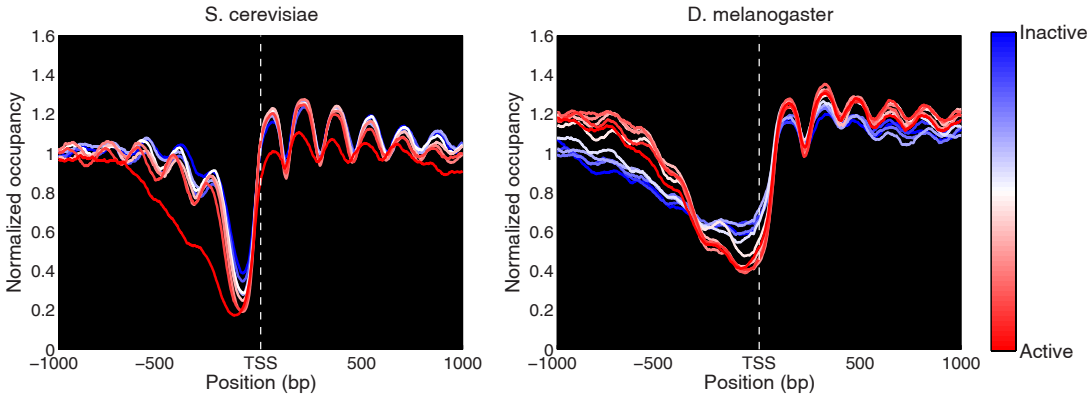


Figure 4.11: Comparison of average nucleosome occupancy in clusters of 500 genes and 1000 genes from *S. cerevisiae* and *D. melanogaster*, respectively. The genes were grouped according to their expression in wild-type cells. The active genes are represented by red lines and the inactive genes are represented by blue lines.

Another major difference between the nucleosome organization in these two organisms is that the degrees of localization in active versus inactive genes are reversed. In *S. cerevisiae*, the inactive genes have better phased nucleosomes and these regular arrays are disrupted in the most active genes. On the contrary, in *Drosophila*, the active genes contain the best phased nucleosomes and the inactive genes display a more uniform nucleosome density.

This project is a collaboration with Tsung-Wai Kan, Victor Guryev, Alexandre Morozov, and Yuri Moshkin.

Chapter 5

Conclusions

The packaging of eukaryotic genomes into chromatin fibers with the aid of nucleosomes has a major impact on all biological processes which take place in living cells and involve DNA. The recent advancements of the DNA sequencing technologies resulted in the availability of genome-wide nucleosome maps for many organisms in both wild type and mutant cells. This abundance of data revolutionized our understanding of the factors which affect the nucleosome organization and the influence of this on the regulation of different processes, as gene expression, DNA replication, repair, and recombination, among others.

In Chapter 1 we present the general problem of nucleosome positioning and the importance of understanding the genome-wide organization of the nucleosomes. We outline a short summary of the previous experimental studies concerning nucleosome organization which serve as motivation for our research.

As the only nucleosome maps that were available at the beginning of this study, contained only single-end reads of the sequenced nucleosomal DNA fragments, our first model assumed a fixed size of the nucleosomal DNA. In other words, initially we assumed that all nucleosomes are fully wrapped by 147 base-pairs of DNA. This model is described in Chapter 2, where we present a statistical mechanics formalism for studying single-type particles which are confined in a one dimensional lattice, and some of the applications of this model.

In the last couple of years, paired-end sequencing replaced single-end sequencing experiments, and higher quality nucleosome maps appeared. It became evident that, in order to explain the new experimental observations, the partial nucleosome unwrapping had to be considered. In Chapter 3, we present a rigorous statistical mechanics

treatment of the nucleosome unwrapping, which explains the recent experimental observations, and we outline some of the applications of our model.

In parallel with my theoretical study of the nucleosome positioning problem, during my PhD research, I had the pleasure to collaborate with three laboratories. In Chapter 4, we present a short overview of my contributions to these joint projects and some of the interesting experimental observations. We show interesting differences between the nucleosome organization in *Saccharomyces cerevisiae* and *Drosophila melanogaster*.

References

[1] A. Annunziato, "DNA packaging: Nucleosomes and chromatin," *Nature Education*, vol. 1, no. 1, 2008.

[2] K. Luger *et al.*, "Crystal structure of the nucleosome core particle at 2.8 Å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Science, 2008.

[4] A. L. Olins and D. E. Olins, "Spheroid chromatin units (v bodies)," *Science*, vol. 183, no. 4122, pp. 330–332, 1974.

[5] R. Kornberg, "Chromatin structure: a repeating unit of histones and dna.," *Science*, vol. 184, no. 4139, pp. 868–871, 1974.

[6] C. L. Woodcock, J. P. Safer, and J. E. Stanchfield, "Structural repeating units in chromatin. I. Evidence for their general occurrence.," *Exp. Cell Res.*, vol. 97, pp. 101–110, 1976.

[7] G. Felsenfeld and M. Groudine, "Controlling the double helix," *Nature*, vol. 421, pp. 448–453, 2003.

[8] Y. Lorch, J. W. LaPointe, and R. D. Kornberg, "Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones," *Cell*, vol. 49, no. 2, pp. 203–210, 1987.

[9] M. Han and M. Grunstein, "Nucleosome loss activates yeast downstream promoters in vivo," *Cell*, vol. 55, no. 6, pp. 1137–1145, 1988.

[10] K. Struhl, "Fundamentally different logic of gene regulation in eukaryotes and prokaryotes," *Cell*, vol. 98, no. 1, pp. 1–4, 1999.

[11] J. Mellor, "Dynamic nucleosomes and gene transcription," *Trends Genet.*, vol. 22, no. 6, pp. 320–329, 2006.

[12] B. Li, M. Carey, and J. L. Workman, "The role of chromatin during transcription," *Cell*, vol. 128, no. 4, pp. 707–719, 2007.

[13] L. Bai and A. V. Morozov, "Gene regulation by nucleosome positioning," *Trends Genet.*, vol. 26, no. 11, pp. 476–483, 2010.

[14] A. Groth, W. Rocha, A. Verreault, and G. Almouzni, "Chromatin challenges during DNA replication and repair," *Cell*, vol. 128, no. 4, pp. 721–733, 2007.

[15] M. L. Eaton, K. Galani, S. Kang, S. P. Bell, and D. M. MacAlpine, "Conserved nucleosome positioning defines replication origins," *Genes Dev.*, vol. 24, no. 8, pp. 748–753, 2010.

[16] N. L. Adkins, H. Niu, P. Sung, and C. L. Peterson, "Nucleosome dynamics regulates DNA processing," *Nat. Struct. Mol. Biol.*, vol. 20, no. 7, pp. 836–842, 2013.

[17] S. Bevington and J. Boyes, "Transcription-coupled eviction of histones H2A/H2B governs V(D)J recombination," *EMBO J*, vol. 32, no. 10, pp. 1381–1392, 2013.

[18] C. Hodges *et al.*, "Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II," *Science*, vol. 325, no. 5940, pp. 626–628, 2009.

[19] L. S. Churchman and J. S. Weissman, "Nascent transcript sequencing visualizes transcription at nucleotide resolution," *Nature*, vol. 469, no. 7330, pp. 368–373, 2011.

[20] S. Henikoff, T. Furuyama, and K. Ahmad, "Histone variants, nucleosome assembly and epigenetic inheritance," vol. 20, no. 7, pp. 320–326, 2004.

[21] D. E. Olins and A. L. Olins, "Chromatin history: our view from the bridge," *Nat. Rev. Mol. Cell Biol.*, vol. 4, no. 10, pp. 809–814, 2003.

[22] M. Vignali *et al.*, "ATP-dependent chromatin-remodeling complexes," *Mol. Cell. Biol.*, vol. 20, no. 6, pp. 1899–1910, 2000.

[23] A. Saha, J. Wittmeyer, and B. R. Cairns, "Chromatin remodelling: the industrial revolution of DNA around histones," *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 6, pp. 437–447, 2006.

[24] C. R. Clapier and B. R. Cairns, "The biology of chromatin remodeling complexes," *Annu. Rev. Biochem.*, vol. 78, no. 1, pp. 273–304, 2009.

[25] G. D. Bowman, "Mechanisms of ATP-dependent nucleosome sliding," *Curr. Opin. Struct. Biol.*, vol. 20, no. 1, pp. 73–81, 2010.

[26] A. Flaus and T. Owen-Hughes, "Mechanisms for ATP-dependent chromatin remodelling: the means to the end," *FEBS Journal*, vol. 278, no. 19, p. 3579–3595, 2011.

[27] C. C. Adams and J. L. Workman, "Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative," *Mol. Cell. Biol.*, vol. 15, pp. 1405–1421, 1995.

[28] K. Polach and J. Widom, "Mechanism of protein access to specific DNA sequences in chromatin: A dynamic equilibrium model for gene regulation," *J. Mol. Biol.*, vol. 254, no. 2, pp. 130–149, 1995.

[29] K. Polach and J. Widom, "A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites," *J. Mol. Biol.*, vol. 258, no. 5, pp. 800–812, 1996.

[30] S. Chávez and M. Beato, "Nucleosome-mediated synergism between transcription factors on the mouse mammary tumor virus promoter," *Proc. Natl. Acad. Sci. USA*, vol. 94, no. 7, pp. 2885–2890, 1997.

[31] R. U. Protacio, K. J. Polach, and J. Widom, "Coupled-enzymatic assays for the rate and mechanism of DNA site exposure in a nucleosome," *J. Mol. Biol.*, vol. 274, no. 5, pp. 708–721, 1997.

[32] S. Vashee, K. Melcher, W. Ding, S. A. Johnston, and T. Kodadek, "Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein–protein interactions," *Curr. Biol.*, vol. 8, no. 8, pp. 452–458, 1998.

[33] S. Vashee, J. Willie, and T. Kodadek, "Synergistic activation of transcription by physiologically unrelated transcription factors through cooperative DNA-binding," *Biochem. Biophys. Res. Commun.*, vol. 247, no. 2, pp. 530–535, 1998.

[34] K. Polach, J. Widom, and A. P. W. Paul M. Wassarman, "Restriction enzymes as probes of nucleosome stability and dynamics," in *Chromatin*, vol. Volume 304, pp. 278–298, Academic Press, 1999.

[35] J. D. Anderson and J. Widom, "Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites," *J. Mol. Biol.*, vol. 296, pp. 979–987, 2000.

[36] K. Polach, P. Lowary, and J. Widom, "Effects of core histone tail domains on the equilibrium constants for dynamic dna site accessibility in nucleosomes," *J. Mol. Biol.*, vol. 298, no. 2, pp. 211–223, 2000.

[37] J. Anderson, P. Lowary, and J. Widom, "Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites," *J. Mol. Biol.*, vol. 307, no. 4, pp. 977–985, 2001.

[38] J. D. Anderson and J. Widom, "Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites," *Mol. Cell. Biol.*, vol. 21, no. 11, pp. 3830–3839, 2001.

[39] J. D. Anderson, A. Thåström, and J. Widom, "Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation," *Mol. Cell. Biol.*, vol. 22, no. 20, pp. 7147–7157, 2002.

[40] J. A. Miller and J. Widom, "Collaborative competition mechanism for gene activation in vivo," *Mol. Cell. Biol.*, vol. 23, pp. 1623–1632, 2003.

[41] G. Li, M. Levitus, C. Bustamante, and J. Widom, "Rapid spontaneous accessibility of nucleosomal DNA," *Nat. Struct. Mol. Biol.*, vol. 12, no. 1, pp. 46–53, 2004.

[42] G. Li and J. Widom, "Nucleosomes facilitate their own invasion," *Nat. Struct. Mol. Biol.*, vol. 11, no. 8, pp. 763–769, 2004.

[43] A. Bucceri, K. Kapitza, and F. Thoma, "Rapid accessibility of nucleosomal DNA in yeast on a second time scale," *EMBO J*, vol. 25, no. 13, pp. 3123–3132, 2006.

[44] H. S. Tims and J. Widom, "Stopped-flow fluorescence resonance energy transfer for analysis of nucleosome dynamics," *Methods*, vol. 41, no. 3, pp. 296–303, 2007.

[45] M. G. Poirier *et al.*, "Spontaneous access to DNA target sites in folded chromatin fibers," *J. Mol. Biol.*, vol. 379, no. 4, pp. 772–786, 2008.

[46] M. G. Poirier, E. Oh, H. S. Tims, and J. Widom, "Dynamics and function of compact nucleosome arrays," *Nat. Struct. Mol. Biol.*, vol. 16, pp. 938–945, 2009.

[47] L. A. Mirny, "Nucleosome-mediated cooperativity between transcription factors," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 52, pp. 22534–22539, 2010.

[48] P. Prinsen and H. Schiessel, "Nucleosome stability and accessibility of its DNA to proteins," *Biochimie*, vol. 92, no. 12, pp. 1722–1728, 2010.

[49] H. S. Tims, K. Gurunathan, M. Levitus, and J. Widom, "Dynamics of nucleosome invasion by DNA binding proteins," *J. Mol. Biol.*, vol. 411, no. 2, pp. 430–448, 2011.

[50] G. Moyle-Heyrman, H. S. Tims, and J. Widom, "Structural constraints in collaborative competition of transcription factors against the nucleosome," *J. Mol. Biol.*, vol. 412, no. 4, pp. 634–646, 2011.

[51] Y. Xi, J. Yao, R. Chen, W. Li, and X. He, "Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation," *Genome Res.*, vol. 21, no. 5, pp. 718–724, 2011.

[52] R. V. Chereji, T.-W. Kan, V. P. Guryev, A. V. Morozov, and Y. M. Moshkin, "The positioning of stable and fragile nucleosomes depends on distinct sequence rules and ATP-dependent chromatin remodelers," In preparation.

[53] B. L. Kidder, G. Hu, and K. Zhao, "ChIP-Seq: technical considerations for obtaining high-quality data," *Nat. Immunol.*, vol. 12, no. 10, pp. 918–922, 2011.

[54] A. Thåström, L. Bingham, and J. Widom, "Nucleosomal locations of dominant DNA sequence motifs for histone–DNA interactions and nucleosome positioning," vol. 338, no. 4, pp. 695–709, 2004.

[55] D. V. Fyodorov and J. T. Kadonaga, "Chromatin assembly in vitro with purified recombinant ACF and NAP-1," in *Methods in Enzymology* (Sankar L.Adhya and Susan Garges, ed.), vol. Volume 371, pp. 499–515, Academic Press, 2003.

[56] R. D. Kornberg and L. Stryer, "Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism," *Nucleic Acids Res.*, vol. 16, no. 14, pp. 6677–6690, 1988.

[57] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, "The dna-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, pp. 362–366, 2009.

[58] C.-K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb, "Evidence for nucleosome depletion at active regulatory regions genome-wide," *Nat. Genet.*, vol. 36, no. 8, pp. 900–905, 2004.

[59] B. E. Bernstein, C. L. Liu, E. L. Humphrey, E. O. Perlstein, and S. L. Schreiber, "Global nucleosome occupancy in yeast," vol. 5, no. 9, p. R62, 2004.

[60] G. C. Yuan, 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*, vol. 309, pp. 626–630, 2005.

[61] E. A. Sekinger, Z. Moqtaderi, and K. Struhl, "Intrinsic histone-dna interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast," *Mol. Cell*, vol. 18, no. 6, pp. 735–748, 2005.

[62] R. Skloot, *The immortal life of Henrietta Lacks*. Random House Digital, Inc., 2010.

[63] W. Lee, 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," *Nat. Genet.*, vol. 39, no. 10, pp. 1235–1244, 2007.

[64] I. Whitehouse, O. J. Rando, J. Delrow, and T. Tsukiyama, "Chromatin remodelling at promoters suppresses antisense transcription," *Nature*, vol. 450, no. 7172, pp. 1031–1035, 2007.

[65] F. Wyers, M. Rougemaille, G. Badis, J.-C. Rousselle, M.-E. Dufour, J. Boulay, B. Régnault, F. Devaux, A. Namane, B. Séraphin, *et al.*, "Cryptic pol ii transcripts are degraded by a nuclear quality control pathway involving a new poly (a) polymerase," *Cell*, vol. 121, no. 5, pp. 725–737, 2005.

[66] I. Albert, 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*, vol. 446, no. 7135, pp. 572–576, 2007.

[67] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh, "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome," *Genome Res.*, vol. 18, pp. 1073–1083, 2008.

[68] W. Möbius and U. Gerland, "Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites," *PLoS Comput. Biol.*, vol. 6, no. 8, p. e1000891, 2010.

[69] M. Radman-Livaja and O. J. Rando, "Nucleosome positioning: How is it established, and why does it matter?," *Dev. Biol.*, vol. 339, no. 2, pp. 258–266, 2010.

[70] S. Shivaswamy, 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*, vol. 6, no. 3, p. e65, 2008.

[71] A. Thastrom, P. T. Lowary, H. R. Widlund, H. Cao, M. Kubista, and J. Widom, "Sequence motifs and free energies of selected natural and non-natural nucleosom e positioning DNA sequences," *J. Mol. Biol.*, vol. 288, pp. 213–229, 1999.

[72] Y. Field, N. Kaplan, Y. 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.*, vol. 4, no. 11, p. e1000216, 2008.

[73] V. Miele, C. Vaillant, Y. d'Aubenton Carafa, C. Thermes, and T. Grange, "DNA physical properties determine nucleosome occupancy from yeast to fly," *Nucleic Acids Res.*, vol. 36, no. 11, pp. 3746–3756, 2008.

[74] A. V. Morozov, 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.*, vol. 37, pp. 4707–4722, 2009.

[75] Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl, "Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo," *Nat. Struct. Mol. Biol.*, vol. 16, no. 8, pp. 847–853, 2009.

[76] G. Locke, 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*, vol. 107, pp. 20998–21003, 2010.

[77] A. Yarragudi, T. Miyake, R. Li, and R. H. Morse, "Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast Saccharomyces cerevisiae," *Mol. Cell. Biol.*, vol. 24, no. 20, pp. 9152–9164, 2004.

[78] N. Elfving, R. V. Chereji, A. V. Morozov, S. Björklund, and J. R. Broach, "A dynamic interplay of nucleosome and Msn2 binding regulates activation and repression of gene expression following stress," In preparation.

[79] P. D. Hartley and H. D. Madhani, "Mechanisms that specify promoter nucleosome location and identity," vol. 137, no. 3, pp. 445–458, 2009.

[80] A. Jansen and K. J. Verstrepen, "Nucleosome positioning in saccharomyces cerevisiae," vol. 75, no. 2, pp. 301–320, 2011.

[81] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow, "Determinants of nucleosome organization in primary human cells," vol. 474, no. 7352, pp. 516–520, 2011.

[82] L. R. Racki, J. G. Yang, N. Naber, P. D. Partensky, A. Acevedo, T. J. Purcell, R. Cooke, Y. Cheng, and G. J. Narlikar, "The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes," *Nature*, vol. 462, no. 7276, pp. 1016–1021, 2009.

[83] Z. Zhang, 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*, vol. 332, no. 6032, pp. 977–980, 2011.

[84] T. Gkikopoulos, P. Schofield, V. Singh, M. Pinskaya, J. Mellor, M. Smolle, J. L. Workman, G. J. Barton, and T. Owen-Hughes, "A role for snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization," *Science*, vol. 333, no. 6050, pp. 1758–1760, 2011.

[85] W. Olson, A. Gorin, X.-J. Lu, L. Hock, and V. Zhurkin, "Dna sequence-dependent deformability deduced from proteindna crystal complexes," *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 19, pp. 11163–11168, 1998.

[86] P. T. Lowary and J. Widom, "New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning," *J. Mol. Biol.*, vol. 276, no. 1, pp. 19–42, 1998.

[87] T. N. Mavrich, C. Z. Jiang, I. P. Ioshikhes, X. Y. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert, and B. F. Pugh, "Nucleosome organization in the Drosophila genome," *Nature*, vol. 453, pp. 358–362, 2008.

[88] J. L. Workman, "Nucleosome displacement in transcription," *Genes Dev.*, vol. 20, pp. 2009–2017, 2006.

[89] L. O. Barrera and B. Ren, "The transcriptional regulatory code of eukaryotic cells insights from genome-wide analysis of chromatin organization and transcription factor binding," *Curr. Opin. Cell Biol.*, vol. 18, no. 3, pp. 291–298, 2006.

[90] G. R. Schnitzler, "Control of nucleosome positions by dna sequence and remodeling machines," *Cell Biochem. Biophys.*, vol. 51, no. 2-3, pp. 67–80, 2008.

[91] K. Struhl and E. Segal, "Determinants of nucleosome positioning," *Nat. Struct. Mol. Biol.*, vol. 20, no. 3, pp. 267–273, 2013.

[92] K. Brogaard *et al.*, "A map of nucleosome positions in yeast at base-pair resolution," *Nature*, vol. 486, pp. 496–501, 2012.

[93] G. Locke, D. Haberman, S. M. Johnson, and A. V. Morozov, "Global remodeling of nucleosome positions in c. elegans," vol. 14, no. 1, p. 284, 2013.

[94] R. V. Chereji and A. V. Morozov, "Statistical mechanics of nucleosomes constrained by higher-order chromatin structure," *J. Stat. Phys.*, vol. 144, no. 2, pp. 379–404, 2011.

[95] R. V. Chereji *et al.*, "Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions," *Phys. Rev. E*, vol. 83, no. 5, p. 050903, 2011.

[96] L. E. Ulanovsky and E. N. Trifonov, *Biomolecular Stereodynamics III*, pp. 35–44. Schenectady, NY: Adenine Press, 1986.

[97] J. Widom, "A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells," *Proc. Natl. Acad. Sci. USA*, vol. 89, pp. 1095–1099, 1992.

[98] J. P. Wang, Y. Fondufe-Mittendorf, L. Xi, G. F. Tsai, E. Segal, and J. Widom, "Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae," *PLoS Comput. Biol.*, vol. 4, p. e1000175, 2008.

[99] W. Möbius, R. A. Neher, and U. Gerland, "Kinetic accessibility of buried DNA sites in nucleosomes," *Phys. Rev. Lett.*, vol. 97, no. 20, p. 208102, 2006.

[100] M. Engeholm *et al.*, "Nucleosomes can invade DNA territories occupied by their neighbors," *Nat. Struct. Mol. Biol.*, vol. 16, no. 2, pp. 151–158, 2009.

[101] W. J. A. Koopmans, R. Buning, T. Schmidt, and J. v. Noort, "spFRET using alternating excitation and FCS reveals progressive DNA unwrapping in nucleosomes," *Biophys. J.*, vol. 97, no. 1, pp. 195–204, 2009.

[102] J. J. Otterstrom and A. M. v. Oijen, "Nudging through a nucleosome," *Science*, vol. 325, no. 5940, pp. 547–548, 2009.

[103] L. S. Shlyakhtenko, A. Y. Lushnikov, and Y. L. Lyubchenko, "Dynamics of nucleosomes revealed by time-lapse atomic force microscopy," *Biochemistry*, vol. 48, no. 33, pp. 7842–7848, 2009.

[104] M.-R. Duan and M. J. Smerdon, "UV damage in DNA promotes nucleosome unwrapping," *J. Biol. Chem.*, vol. 285, no. 34, pp. 26295–26303, 2010.

[105] V. B. Teif and K. Rippe, "Statistical–mechanical lattice models for protein–DNA binding in chromatin," *J. Phys.: Condens. Matter*, vol. 22, no. 41, p. 414105, 2010.

[106] M. Simon, J. A. North, J. C. Shimko, R. A. Forties, M. B. Ferdinand, M. Manohar, M. Zhang, R. Fishel, J. J. Ottesen, and M. G. Poirier, "Histone fold modifications control nucleosome unwrapping and disassembly," *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 31, p. 12711, 2011.

[107] A. Miyagi, T. Ando, and Y. L. Lyubchenko, "Dynamics of nucleosomes assessed with time-lapse high-speed atomic force microscopy," *Biochemistry*, vol. 50, no. 37, pp. 7901–7908, 2011.

[108] R. Blossey and H. Schiessel, "The dynamics of the nucleosome: thermal effects, external forces and ATP," *FEBS Journal*, vol. 278, no. 19, pp. 3619–3632, 2011.

[109] K. Voltz, J. Trylska, N. Calimet, J. C. Smith, and J. Langowski, "Unwrapping of nucleosomal DNA ends: A multiscale molecular dynamics study," *Biophys. J.*, vol. 102, no. 4, pp. 849–858, 2012.

[110] I. C. Taylor, J. L. Workman, T. J. Schuetz, and R. E. Kingston, "Facilitated binding of GAL4 and heat shock factor to nucleosomal templates: differential function of DNA-binding domains.," *Genes Dev.*, vol. 5, no. 7, pp. 1285–1298, 1991.

[111] T. Raveh-Sadka, M. Levo, and E. Segal, "Incorporating nucleosomes into thermodynamic models of transcription regulation," *Genome Res.*, vol. 19, no. 8, pp. 1480–1496, 2009.

[112] J. Zlatanova, T. C. Bishop, J.-M. Victor, V. Jackson, and K. van Holde, "The nucleosome family: Dynamic and growing," *Structure*, vol. 17, no. 2, pp. 160–171, 2009.

[113] A. Gansen, A. Valeri, F. Hauger, S. Felekyan, S. Kalinin, K. Tth, J. Langowski, and C. A. M. Seidel, "Nucleosome disassembly intermediates characterized by

single-molecule FRET," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 15308–15313, 2009.

[114] R. Forties, J. North, S. Javaid, O. Tabbaa, R. Fishel, M. Poirier, and R. Bundschuh, "A quantitative model of nucleosome dynamics," *Nucleic Acids Res.*, vol. 39, pp. 8306–8313, 2011.

[115] V. B. Teif and K. Rippe, "Nucleosome mediated crosstalk between transcription factors at eukaryotic enhancers," *Phys. Biol.*, vol. 8, no. 4, p. 044001, 2011.

[116] V. B. Teif and K. Rippe, "Calculating transcription factor binding maps for chromatin," *Brief. Bioinform.*, vol. 13, no. 2, pp. 187–201, 2012.

[117] H. Takahashi, "A simple method for treating the statistical mechanics of one-dimensional substances," in *Proc. Phys.-Math. Soc. Japan*, vol. 24, p. 60, 1942.

[118] H. L. Frisch and J. L. Lebowitz, *The equilibrium theory of classical fluids: a lecture note and reprint volume.* WA Benjamin, 1964.

[119] J. K. Percus, "Equilibrium state of a classical fluid of hard rods in an external field," *J. Stat. Phys.*, vol. 15, no. 6, pp. 505–511, 1976.

[120] J. K. Percus, "Entropy of a non-uniform one-dimensional fluid," *J. Phys.: Condens. Matter*, vol. 1, no. 17, pp. 2911–2922, 1989.

[121] F. Gürsey, "Classical statistical mechanics of a rectilinear assembly," in *Math. Proc. Cambridge Philos. Soc.*, vol. 46, pp. 182–194, 1950.

[122] Z. Salsburg, R. Zwanzig, and J. Kirkwood, "Molecular distribution functions in a one-dimensional fluid," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1098–1107, 1953.

[123] I. Z. Fisher and T. M. Switz, *Statistical theory of liquids.* Chicago, IL: University of Chicago Press, 1964.

[124] L. Tonks, "The complete equation of state of one, two and three-dimensional gases of hard elastic spheres," *Phys. Rev.*, vol. 50, no. 10, pp. 955–963, 1936.

[125] D. J. Schwab, R. F. Bruinsma, J. Rudnick, and J. Widom, "Nucleosome switches," *Phys. Rev. Lett.*, vol. 100, no. 22, p. 228105, 2008.

[126] E. Segal *et al.*, "A genomic code for nucleosome positioning," *Nature*, vol. 442, pp. 772–778, 2006.

[127] K. A. Zawadzki, A. V. Morozov, and J. R. Broach, "Chromatin-dependent transcription factor accessibility rather than nucleosome remodeling predominates during global transcriptional restructuring in Saccharomyces cerevisiae," *Mol. Biol. Cell*, vol. 20, no. 15, pp. 3503–3513, 2009.

[128] G. Chevereau, L. Palmeira, C. Thermes, A. Arneodo, and C. Vaillant, "Thermodynamics of intragenic nucleosome ordering," *Phys. Rev. Lett.*, vol. 103, no. 18, p. 188103, 2009.

[129] A. Weiner, A. Hughes, M. Yassour, O. Rando, and N. Friedman, "High-resolution nucleosome mapping reveals transcription-dependent promoter packaging," *Genome Res.*, vol. 20, pp. 90–100, 2010.

[130] R. V. Chereji and A. V. Morozov, "Ubiquitous nucleosome crowding and unwrapping in the yeast genome," Submitted.

[131] G. Li *et al.*, "Rapid spontaneous accessibility of nucleosomal DNA," *Nat. Struct. Mol. Biol.*, vol. 12, no. 1, pp. 46–53, 2005.

[132] T. Chou, "An exact theory of histone-DNA adsorption and wrapping," *Europhys. Lett.*, vol. 62, no. 5, pp. 753–759, 2003.

[133] V. B. Teif and K. Rippe, "Calculating transcription factor binding maps for chromatin," *Brief. Bioinform.*, vol. 13, no. 2, pp. 187–201, 2012.

[134] R. T. Simpson, "Mechanism of a reversible, thermally induced conformational change in chromatin core particles.," *J. Biol. Chem.*, vol. 254, no. 20, pp. 10123–10127, 1979.

[135] T. D. Yager, C. T. McMurray, and K. E. Van Holde, "Salt-induced release of DNA from nucleosome core particles," *Biochemistry*, vol. 28, no. 5, pp. 2271–2281, 1989.

[136] N. L. Marky and G. S. Manning, "A theory of DNA dissociation from the nucleosome," vol. 254, no. 1, pp. 50–61, 1995.

[137] H. S. Tims *et al.*, "Dynamics of nucleosome invasion by DNA binding proteins," *J. Mol. Biol.*, vol. 411, no. 2, pp. 430–448, 2011.

[138] M. L. Bennink, S. H. Leuba, G. H. Leno, J. Zlatanova, B. G. de Grooth, and J. Greve, "Unfolding individual nucleosomes by stretching single chromatin fibers with optical tweezers," *Nat. Struct. Mol. Biol.*, vol. 8, no. 7, pp. 606–610, 2001.

[139] B. D. Brower-Toland, 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*, vol. 99, no. 4, pp. 1960–1965, 2002.

[140] J. J. Hayes and J. C. Hansen, "New insights into unwrapping DNA from the nucleosome from a single-molecule optical tweezers method," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 4, pp. 1752–1754, 2002.

[141] A. H. Mack, D. J. Schlingman, R. P. Ilagan, L. Regan, and S. G. Mochrie, "Kinetics and thermodynamics of phenotype: Unwinding and rewinding the nucleosome," *J. Mol. Biol.*, vol. 423, no. 5, pp. 687–701, 2012.

[142] M. A. Hall, 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," vol. 16, no. 2, pp. 124–129, 2009.

[143] M. Li and M. D. Wang, "Chapter two - unzipping single DNA molecules to study nucleosome structure and dynamics," in *Methods in Enzymology* (C. Wu and C. D. Allis, eds.), vol. Volume 513, pp. 29–58, Academic Press, 2012.

[144] J. A. North, J. C. Shimko, S. Javaid, A. M. Mooney, M. A. Shoffner, S. D. Rose, R. Bundschuh, R. Fishel, J. J. Ottesen, and M. G. Poirier, "Regulation of the nucleosome unwrapping rate controls DNA accessibility," *Nucleic Acids Res.*, 2012.

[145] D. Hasson, T. Panchenko, K. J. Salimian, M. U. Salman, N. Sekulic, A. Alonso, P. E. Warburton, and B. E. Black, "The octamer is the major form of CENP-A nucleosomes at human centromeres," *Nat. Struct. Mol. Biol.*, vol. advance online publication, 2013.

[146] J. G. Henikoff, J. A. Belskyb, K. Krassovsky, D. MacAlpine, and S. Henikoff, "Epigenome characterization at single base-pair resolution," *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 18318–18323, 2011.

[147] H. A. Cole, B. H. Howard, and D. J. Clark, "Activation-induced disruption of nucleosome position clusters on the coding regions of Gcn4-dependent genes extends into neighbouring genes," *Nucleic Acids Res.*, vol. 39, pp. 9521–9535, 2011.

[148] V. Nagarajavel, J. R. Iben, B. H. Howard, R. J. Maraia, and D. J. Clark, "Global 'bootprinting' reveals the elastic architecture of the yeast TFIIIB-TFIIIC transcription complex in vivo," *Nucleic Acids Res.*, 2013.

[149] C. Dingwall, G. P. Lomonosoff, and R. A. Laskey, "High sequence specificity of micrococcal nuclease," *Nucleic Acids Res.*, vol. 9, pp. 2659–2673, 1981.

[150] J. D. McGhee and G. Felsenfeld, "Another potential artifact in the study of nucleosome phasing by chromatin digestion with micrococcal nuclease," *Cell*, vol. 32, pp. 1205–1215, 1983.

[151] H.-R. Chung, I. Dunkel, F. Heise, C. Linke, S. Krobitsch, A. E. Ehrenhofer-Murray, S. R. Sperling, and M. Vingron, "The effect of micrococcal nuclease digestion on nucleosome positioning data," *PLoS ONE*, vol. 5, no. 12, p. e15754, 2010.

[152] T. J. Richmond and C. A. Davey, "The structure of DNA in the nucleosome core," *Nature*, vol. 423, pp. 145–150, 2003.

[153] G. E. Davey, B. Wu, Y. Dong, U. Surana, and C. A. Davey, "DNA stretching in the nucleosome facilitates alkylation by an intercalating antitumour agent," *Nucleic Acids Res.*, vol. 38, no. 6, pp. 2081–2088, 2010.

[154] C. Davey, D. Sargent, K. Luger, A. Maeder, and T. Richmond, "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 åresolution," *J. Mol. Biol.*, vol. 319, pp. 1097–1113, 2002.

[155] J. Zlatanova, C. Seebart, and M. Tomschik, "The linker-protein network: control of nucleosomal DNA accessibility," *Trends Biochem. Sci.*, vol. 33, pp. 247–253, 2008.

[156] S. H. Syed *et al.*, "Single-base resolution mapping of H1–nucleosome interactions and 3D organization of the nucleosome," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 21, pp. 9620–9625, 2010.

[157] M. Georgieva, A. Roguev, K. Balashev, J. Zlatanova, and G. Miloshev, "Hho1p, the linker histone of *saccharomyces cerevisiae*, is important for the proper chromatin organization *in vivo*," *Biochim. et Biophys. Acta*, vol. 1819, pp. 366–374, 2012.

[158] G. Schafer, C. McEvoy, and H.-G. Patterton, "The *saccharomyces cerevisiae* linker histone Hho1p is essential for chromatin compaction in stationary phase and is displaced by transcription," *Proc. Natl. Acad. Sci. USA*, vol. 105, pp. 14838–14843, 2008.

[159] K. E. van Holde, *Chromatin*. New York: Springer, 1989.

[160] H. S. Rhee and B. F. Pugh, "Genome-wide structure and organization of eukaryotic pre-initiation complexes," *Nature*, vol. 483, pp. 295–301, 2012.

[161] M. F. Dion, T. Kaplan, M. Kim, S. Buratowski, N. Friedman, and O. J. Rando, "Dynamics of replication-independent histone turnover in budding yeast," *Science*, vol. 315, pp. 1405–1408, 2007.

[162] N. Petrenko, R. V. Chereji, M. McClean, A. V. Morozov, and J. R. Broach, "Noise and interlocking signaling pathways promote distinct transcription factor dynamics in response to different stresses," *Mol. Biol. Cell*, vol. 24, no. 12, pp. 2045–2057, 2013.

[163] N. Elfving, R. V. Chereji, M. Larsson, A. V. Morozov, J. R. Broach, and S. Björklund, "Mediator exists in multiple forms and is predominantly associated to promoters with low nucleosome density," In preparation.

[164] L. Cai, C. K. Dalal, and M. B. Elowitz, "Frequency-modulated nuclear localization bursts coordinate gene regulation," *Nature*, vol. 455, no. 7212, pp. 485–490, 2008.

[165] N. Hao and E. K. O'Shea, "Signal-dependent dynamics of transcription factor translocation controls gene expression," *Nat. Struct. Mol. Biol.*, vol. 19, no. 1, pp. 31–39, 2012.

[166] D. Iliopoulos, H. A. Hirsch, G. Wang, and K. Struhl, "Inducible formation of breast cancer stem cells and their dynamic equilibrium with non-stem cancer cells via IL6 secretion," *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 4, pp. 1397–1402, 2011.

[167] S. Björklund and C. M. Gustafsson, "The yeast Mediator complex and its regulation," *Trends Biochem. Sci.*, vol. 30, no. 5, pp. 240–244, 2005.

[168] L. C. Myers, C. M. Gustafsson, D. A. Bushnell, M. Lui, H. Erdjument-Bromage, P. Tempst, and R. D. Kornberg, "The med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain," *Genes Dev.*, vol. 12, no. 1, pp. 45–54, 1998.

[169] H. Rhee and B. Pugh, "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution," vol. 147, no. 6, pp. 1408–1419, 2011.

[170] C. A. Nieduszynski, Y. Knox, and A. D. Donaldson, "Genome-wide identification of replication origins in yeast by comparative genomics," *Genes Dev.*, vol. 20, no. 14, pp. 1874–1879, 2006.

[171] B. R. Cairns, "Chromatin remodeling machines: similar motors, ulterior motives," *Trends Biochem. Sci.*, vol. 23, no. 1, pp. 20–25, 1998.

[172] P. S. Knoepfler and R. N. Eisenman, "Sin meets NuRD and other tails of repression," *Cell*, vol. 99, no. 5, pp. 447–450, 1999.

[173] T. v. d. Heijden, J. J. F. A. v. Vugt, C. Logie, and J. v. Noort, "Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy," *Proc. Natl. Acad. Sci. USA*, 2012.

[174] R. B. Deal, J. G. Henikoff, and S. Henikoff, "Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones," *Science*, vol. 328, no. 5982, pp. 1161–1164, 2010.

Appendix A

The z-transform formalism

In this Appendix we present the z-transform formalism which is used to compute the partition function for a system of hard-rods in a 1D lattice. From the partition function, we obtain simple relationships for the chemical potential and for the pressure in the system. After we present the general formalism, we apply this to compute the chemical potential and the pressure in two systems, ideal gas and Tonks gas. We also compare the results between the continuous and the discrete case, that is gas in a lattice.

A.1 General method

Consider a system of N particles distributed on a segment of length L bp. We assume that the particles interact with each other through short-range nearest-neighbor interactions, which include steric exclusion if the particles have a finite size of a bp, and the total interaction energy is

$$U(x_1, x_2, \ldots, x_N) = \Phi(x_2 - x_1) + \Phi(x_3 - x_2) + \ldots + \Phi(x_N - x_{N-1}) + U_b,$$

where U_b is the boundary term which describes interaction between the walls and the first and last particles, and x_i represents the position of the i-th particle in the lattice.

For simplicity, we assume that the boundary conditions are enforced by two additional particles of the same kind, fixed at $x = 0$ and $x = L$, so that the boundary term is given by

$$U_b = \Phi(x_1) + \Phi(L - x_N).$$

The exact form of boundary conditions is not essential in the thermodynamic limit.

The canonical partition function of this system of N particles is

$$\begin{aligned} Q_N(L) &= \sum_{0\le x_1\le x_2\le \ldots \le x_N\le L} e^{-\beta\Phi(x_1-0)} e^{-\beta\Phi(x_2-x_1)} \ldots e^{-\beta\Phi(L-x_N)} \\ &= \sum_{x_N=0}^{L} \sum_{x_{N-1}=0}^{x_N} \ldots \sum_{x_1=0}^{x_2} e^{-\beta\Phi(x_1-0)} e^{-\beta\Phi(x_2-x_1)} \ldots e^{-\beta\Phi(L-x_N)} \end{aligned}$$

Denoting $f(x) \equiv e^{-\beta\Phi(x)}$, we obtain

$$Q_N(L) = \sum_{x_N=0}^{L} \sum_{x_{N-1}=0}^{x_N} \ldots \sum_{x_1=0}^{x_2} f(x_1-0) f(x_2-x_1) \ldots f(L-x_N).$$

Note that this represents the convolution of $N+1$ functions f, that is

$$Q_N(L) = (\underbrace{f * f * \ldots * f}_{N+1 \text{ functions}})(L).$$

The partition function can be computed using the z transform method. Let $\widetilde{Q}(z)$ be the z transform of $Q_N(L)$,

$$\widetilde{Q}(z) = \sum_{n=0}^{\infty} Q_N(n) z^{-n}.$$

From the convolution theorem, we have that

$$\widetilde{Q}(z) = \left[\widetilde{F}(z)\right]^{N+1},$$

where $\widetilde{F}(z)$ is the z transform of $f(n)$,

$$\widetilde{F}(z) = \sum_{n=0}^{\infty} f(n) z^{-n}.$$

The partition function can be recovered using the inverse z transform,

$$Q_N(L) = \frac{1}{2\pi i} \oint_{\Gamma} \left[\widetilde{F}(z)\right]^{N+1} z^{L-1} \mathrm{d}z.$$

The contour of integration, Γ, is any simple closed curve enclosing $|z| = R$, where $|z| > R$ is the region of convergence.

Let us define the function

$$h(z) = (N+1) \ln \widetilde{F}(z) + (L-1) \ln z.$$

With this notation, we obtain

$$Q_N(L) = \frac{1}{2\pi i} \oint_{\Gamma} e^{h(z)} \mathrm{d}z.$$

This integral can be computed by the saddle point method [123]. Expanding $h(z)$ around the saddle point z_0, we obtain

$$Q_N(L) \approx e^{h(z_0)} \frac{1}{2\pi i} \int e^{\frac{1}{2}h''(z_0)(z-z_0)^2} \mathrm{d}z.$$

Integration along the path of steepest descent yields a contribution from the Gaussian integral of order $O\left([h''(z_0)]^{-1/2}\right) = O\left(N^{-1/2}\right)$. Since we need to estimate $\ln Q_N(L)$, in order to compute the macroscopic quantities, and because in the thermodynamic limit the terms of order $O(\ln N)$ are not important, we can approximate the partition function by

$$\begin{aligned} Q_N(L) &\approx e^{h(z_0)} \\ &\approx z_0^L \left[\widetilde{F}(z_0)\right]^N, \end{aligned} \tag{A.1}$$

where z_0 is the saddle point, satisfying the equation

$$\left.\frac{\mathrm{d}h}{\mathrm{d}z}\right|_{z=z_0} \approx \frac{L}{z_0} + N\frac{\widetilde{F}'(z_0)}{\widetilde{F}(z_0)} = 0. \tag{A.2}$$

We can compute the chemical potential for the interacting hard rods by taking the derivative of the free energy, F, with respect to the number of particles in the system

$$\mu = \frac{\partial F}{\partial N} = -k_B T \frac{\partial \ln Q_N}{\partial N} = -k_B T \ln \widetilde{F}(z_0). \tag{A.3}$$

The pressure of the gas is given by the derivative of the free energy with respect to the length of the system. We denote the length of a base pair by b, such that the real length of the system is Lb. We obtain that the pressure of the gas is given by

$$\begin{aligned} p &= -\frac{1}{b}\frac{\partial F}{\partial L} \\ &= \frac{k_B T}{b}\frac{\partial \ln Q_N(L)}{\partial L} \\ &= \frac{k_B T}{b} \ln z_0, \end{aligned} \tag{A.4}$$

A.2 Applications: ideal gas, Tonks gas

The formalism presented in the preceding Section is used to obtain the equation of state and chemical potential for a system of N hard-rods, interacting through any generic

interaction $\Phi(x)$. Let us consider two simple cases – the 1D ideal lattice gas and the 1D Tonks lattice gas, which is characterized by the hard-core interaction

$$\Phi(x) = \begin{cases} \infty & \text{if } x < a, \\ 0 & \text{if } x \geq a. \end{cases}$$

For the ideal gas, the z transform of $e^{-\beta\Phi}$ and the saddle point z_0, obtained from Equation (A.2), are given by:

$$\widetilde{F}(z) = \frac{z}{z-1},$$
$$z_0 = \frac{L+N}{N}.$$

Using these expressions, together with Equation (A.1), we can compute the logarithm of the partition function

$$\ln Q_N(L) = L \ln\left(\frac{L+N}{L}\right) + N \ln\left(\frac{L+N}{N}\right) + O(\ln N),$$

which gives the pressure and the chemical potential for the ideal lattice gas

$$\beta p_1^{\mathrm{id}} = \frac{1}{b} \ln\left(1 + \frac{N}{L}\right), \tag{A.5}$$

$$\beta \mu_1^{\mathrm{id}} = \ln\left(\frac{N}{L+N}\right). \tag{A.6}$$

In the case of the Tonks lattice gas, we obtain

$$\widetilde{F}(z) = \frac{z^{1-a}}{z-1},$$
$$z_0 = \frac{L - Na + N}{L - Na},$$

while the pressure and the chemical potential are then given by

$$\beta p_1^{\mathrm{T}} = \frac{1}{b} \ln\left(1 + \frac{N}{L - Na}\right), \tag{A.7}$$

$$\beta \mu_1^{\mathrm{T}} = a \ln\left(\frac{L - Na + N}{L - Na}\right) + \ln\left(\frac{N}{L - Na + N}\right). \tag{A.8}$$

A.3 Comparison between lattice and continuous 1D fluids

It is useful to compare these results with the corresponding results for continuous one-dimensional gases. Denoting the physical length of the particles by $\mathcal{A}$, the length of the

box by $\mathcal{L}$, and the Laplace transform of the function $e^{-\beta\Phi}$ by

$$\varphi(s) = \int_0^\infty s^{-sx} e^{-\beta\Phi(x)} \mathrm{d}x,$$

we express the canonical partition function as an inverse Laplace transform,

$$Q_{\mathrm{cont}}(N, \mathcal{L}, T) = \frac{1}{\lambda(T)^N} \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{s\mathcal{L}} \left[\varphi(s)\right]^N \mathrm{d}s \approx e^{s_0 \mathcal{L}} \left[\frac{\varphi(s_0)}{\lambda(T)}\right]^N.$$

In the preceding identity, we have that

$$\lambda(T) = h/\sqrt{2\pi m k_B T},$$

where $\lambda(T)$ is the thermal de Broglie wavelength, and s_0 is the saddle point, which is a solution of the equation

$$\mathcal{L} + N \frac{\varphi'(s_0)}{\varphi(s_0)} = 0.$$

Recall that in the case of the ideal gas, we have

$$\varphi(s) = \frac{1}{s}$$
$$s_0 = \frac{N}{\mathcal{L}}$$

Using the preceding two equations, we then find for the ideal gas that

$$\beta p_{\mathrm{c}}^{\mathrm{id}} = \frac{\partial \ln Q_{\mathrm{cont}}}{\partial \mathcal{L}} = \frac{N}{\mathcal{L}}, \tag{A.9}$$

$$\beta \mu_{\mathrm{c}}^{\mathrm{id}} = -\frac{\partial \ln Q_{\mathrm{cont}}}{\partial N} = \ln \frac{N\lambda(T)}{\mathcal{L}}. \tag{A.10}$$

Recall that in the case of the Tonks gas, we have

$$\varphi(s) = \frac{e^{-\mathcal{A}s}}{s}$$
$$s_0 = \frac{N}{\mathcal{L} - N\mathcal{A}}$$

Similarly, for the Tonks gas we obtain

$$\beta p_{\mathrm{c}}^{\mathrm{T}} = \frac{N}{\mathcal{L} - N\mathcal{A}}, \tag{A.11}$$

$$\beta \mu_{\mathrm{c}}^{\mathrm{T}} = \frac{N\mathcal{A}}{\mathcal{L} - N\mathcal{A}} + \ln \left[\frac{N\lambda(T)}{\mathcal{L} - N\mathcal{A}}\right]. \tag{A.12}$$

To compare continuous and discrete results, we let the lattice constant b approach zero, while keeping the particle size $\mathcal{A} = ab$ and the box size $\mathcal{L} = Lb$ finite. We obtain

$$\lim_{b\to 0} \beta p_{\mathrm{l}}^{\mathrm{id}} = \lim_{b\to 0} \frac{1}{b} \ln\left(1 + \frac{Nb}{\mathcal{L}}\right) = \beta p_{\mathrm{c}}^{\mathrm{id}},$$
$$\lim_{b\to 0} \beta p_{\mathrm{l}}^{\mathrm{T}} = \lim_{b\to 0} \frac{1}{b} \ln\left(1 + \frac{Nb}{\mathcal{L} - N\mathcal{A}}\right) = \beta p_{\mathrm{c}}^{\mathrm{T}}.$$

Similarly, the chemical potentials for the ideal and Tonks lattice gases become asymptotically, as b tends to zero,

$$\beta\mu_{\mathrm{l}}^{\mathrm{id}} \sim \ln \frac{Nb}{\mathcal{L}},$$
$$\beta\mu_{\mathrm{l}}^{\mathrm{T}} \sim \frac{N\mathcal{A}}{\mathcal{L} - N\mathcal{A}} + \ln\left(\frac{Nb}{\mathcal{L} - N\mathcal{A}}\right).$$

These expressions are identical to the chemical potentials of the corresponding continuous gases [Equations (A.10) and (A.12)], with the length scale, $\lambda(T)$, replaced by the typical length scale of the lattice, b.

Appendix B

Alternative nucleosome unwrapping models

In this Appendix, we present eight alternative models that we used for the sequence-independent part of the binding energy of a partially unwrapped nucleosome, u^{SI}, as discussed in Chapter 3. Parameter fitting for all models was carried out in a two-stage procedure using the genetic algorithm optimization function `ga` from the MATLAB Global optimization toolbox. First, the objective function to be minimized was set equal to the root-mean-square deviation, RMS, between predicted and observed inter-dyad distributions. Once RMS decreased bellow 10^{-3}, the objective function was replaced by $RMS - r_{\rm osc} \simeq -r_{\rm osc}$, where $r_{\rm osc}$ is the linear correlation between observed and predicted oscillations after the smooth background has been subtracted from inter-dyad distributions, as in Figure 3.9C. We have found that the two-stage optimization allows us to effectively fit both the overall shape and the fine oscillatory structure in the data. The best-fit parameters for all models are given bellow.

The sequence-independent binding energy of a particle of length $a = 1 + x_1 + x_2$, where one bp corresponds to the dyad, and x_1 and x_2 correspond to the extra number of bps in contact with the histone octamer on each side of the dyad (Figure B.1), is given by

$$u^{SI} = u_{\rm half}(x_1) + u_{\rm half}(x_2), \tag{B.1}$$

where $u_{\rm half}(x)$ contains both the electrostatic interaction between the piece of x bp of negatively charged DNA, and the positively charged histone octamer, and the elastic energy necessary to bend this piece of DNA around the histone. In the remaining part of this Appendix, we present alternative models for the function $u_{\rm half}(x)$, and implicitly, for the sequence-independent part of the nucleosome formation energy.

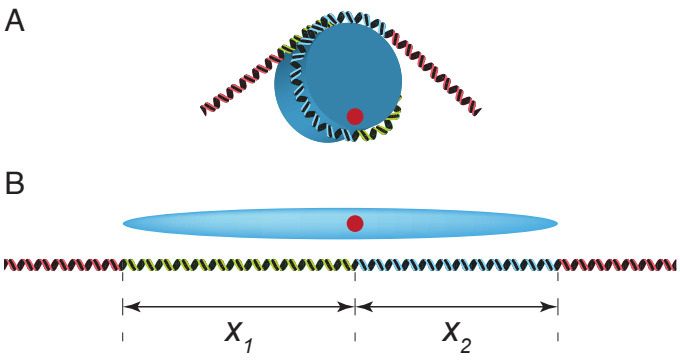


Figure B.1: Cartoons representing a partially unwrapped nucleosome, in 3D space (A), and in a reduced, 1D space. Our models use the 1D representation of the DNA and nucleosomes. Each nucleosome covers a numbers of bps in the 1D lattice, which represents the DNA. If the red circle denotes the dyad of the nucleosome, then the green and blue pieces of DNA represent the fragments that are in contact with the histone octamers on both sides of the dyad, with lengths x_1 and x_2, respectively. The sequence-independent binding energy corresponding to this configuration is given by Equation (B.1).

B.1 Model A: Crystal structure augmented with an additional well

Using information from the crystal structure of nucleosomes [152], we know the positions of the histone-DNA contacts. There are 14 contacts between DNA and a histone octamer, 7 on each side of the dyad. In this model, for each side of the nucleosomal DNA of length x (see Figure B.1), we use a sequence-independent binding energy, $u_{\text{half}}(x)$ which has the following expression,

$$u_{\text{half}}(x) = \texttt{interp1}(...) - \frac{E_b}{147}x.$$

Here, E_b is the binding energy of a fully wrapped particle in the absence of 10-11 bp oscillations. The MATLAB function `interp1(...)` is used to generate an oscillatory pattern by piecewise cubic Hermite interpolation using the data points from Table B.1. The oscillations are based on the crystal structure of nucleosomes [152], with the minima located at the positions of the histone-DNA contacts. The oscillatory pattern was superimposed onto a line with the slope of $-E_b/147$, which represents the average sequence-independent binding energy per bp.

x (Position)	f(x) (Energy)
-1	-A
3	A
7	-A
13	A
17	-A
24	A
28	-A
34	A
38	-A
44	A
49	-A
55	A
59	-A
65	A
69	-A
75	A
p	-d
85	A

Table B.1: Data points used for interpolation

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	3 bp
E_b	14.39 k_BT
μ	-14.51 k_BT
A	1.13 k_BT
f	0.51
p	79 bp
d	0.86 k_BT

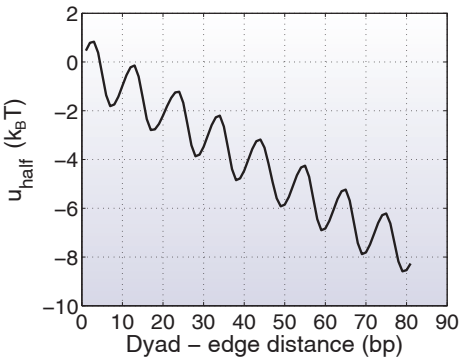


Table B.2: Fitted parameters for Model A. $a_{\max}$ and $a_{\min}$ are the maximum and minimum lengths of the nucleosome particle, μ is the histone octamer chemical potential, A is the amplitude of the oscillations, and f is the hydroxyl radical cutting frequency. p and d are the position and the depth of the first minimum outside of the nucleosome core particle, respectively. E_b is the binding energy of a fully wrapped particle in the absence of 10-11 bp oscillations.

We obtain the fit residuals:

$$RMS = 9.9958 \times 10^{-4},$$

$$r_{\text{osc}} = 0.764364,$$

$$RMS_{\text{osc}} = 1.8662 \times 10^{-4}.$$

RMS represents the root-mean-square error of the predicted inter-dyad distribution, while r_{osc} is the linear correlation between the oscillatory parts of the measured and predicted inter-dyad distributions. The oscillatory part is obtained by subtracting the smooth background from the full inter-dyad distribution. Smoothing is done by applying a Savitzky-Golay smoothing filter, also known as least-squares, or DISPO (Digital Smoothing Polynomial) filter, of polynomial order 3 and length 31 bp. By RMS_{osc} we denote the root-mean-square error of the oscillatory part of the predicted inter-dyad distribution.

B.2 Model B: Crystal structure augmented with a linear function

In this Model, we define $u_{\text{half}}(x)$ in the same way as in Model A, for x belonging to $[1, 73]$, while for $x \geq 74$, $u_{\text{half}}(x)$ is defined by a suitably chosen linear function as follows.

$$u_{\text{half}}(x) = \begin{cases} \texttt{interp1}(...) - \frac{E_b}{147}x & \text{for } x \in [1, 73], \\ \texttt{interp1}(...) - 73\frac{E_b}{147} - \frac{\Delta E}{\Delta X}(x - 73) & \text{for } x \in [74, 73 + \Delta X]. \end{cases}$$

Notice that $\Delta E/\Delta X$ is the slope of the linear function, where ΔE is the energy difference between the first and last points of the linear function, and ΔX is the cardinality of the range of the linear function.

Parameter	Value
$a_{\min}$	27 bp
E_b	14.66 k_BT
μ	-15.04 k_BT
A	1.28 k_BT
f	0.50
ΔE	-2.47 k_BT
ΔX	7 bp

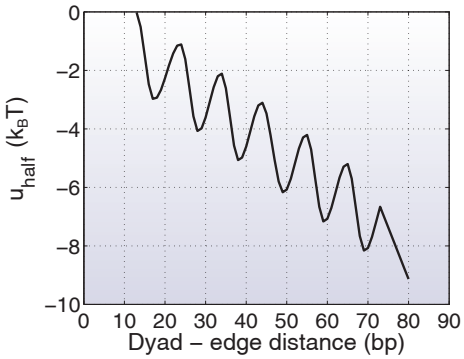


Table B.3: Fitted parameters for Model B. All parameters are as in Model A, except for ΔE and ΔX, which are defined above.

We obtain the fit residuals:

$$RMS = 0.0012 \ (RMS \text{ cannot decrease below } 10^{-3} \text{ in this model}),$$
$$r_{\text{osc}} = 0.769179,$$
$$RMS_{\text{osc}} = 1.8411 \times 10^{-4}.$$

All residuals are defined as in Model A.

B.3 Model C: 10-bp oscillations superimposed onto a linear function

We now let

$$u_{\text{half}}(x) = -A\cos\left(\frac{2\pi}{10}(x - x_0)\right) - \frac{E_b}{147}x,$$

where A is the amplitude of the oscillations, x_0 determines the phase of the oscillations, and E_b is the binding energy of a fully wrapped particle in the absence of oscillations.

Parameter	Value
$a_{\max}$	165 bp
$a_{\min}$	3 bp
E_b	14.43 k_BT
μ	-13.99 k_BT
A	1.06 k_BT
x_0	79 bp
f	0.50

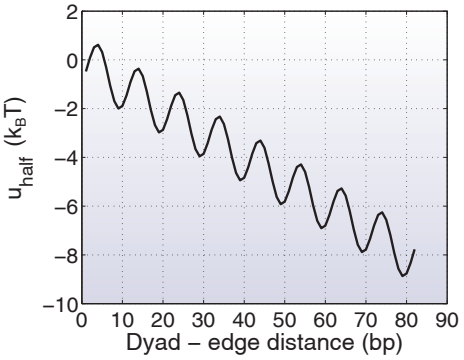


Table B.4: Fitted parameters for Model C. All parameters are as in Model A, except for x_0, which is defined above.

We obtain the fit residuals:

$$RMS = 9.9861 \times 10^{-4},$$

$$r_{\text{osc}} = 0.708620,$$

$$RMS_{\text{osc}} = 2.0203 \times 10^{-4}.$$

All residuals are defined as in Model A.

B.4 Model D: 11-bp oscillations superimposed onto a linear function

Similarly to Model C, we construct Model D using oscillations with the period of 11 bp instead of 10 bp, as before. Let

$$u_{\text{half}}(x) = -A\cos\left(\frac{2\pi}{11}(x - x_0)\right) - \frac{E_b}{147}x,$$

where A, x_0 and E_b have the same meaning as in Model C.

Parameter	Value
$a_{\max}$	161 bp
$a_{\min}$	25 bp
E_b	13.99 k_BT
μ	-14.30 k_BT
A	1.03 k_BT
x_0	80 bp
f	0.52

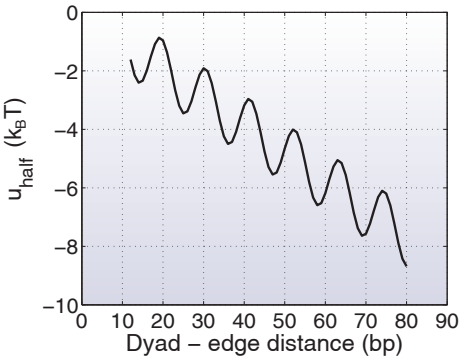


Table B.5: Fitted parameters for Model D. All parameters are as in Model C.

We obtain the fit residuals:

$$RMS = 9.9121 \times 10^{-4},$$

$$r_{\text{osc}} = 0.688838,$$

$$RMS_{\text{osc}} = 2.0735 \times 10^{-4}.$$

All residuals are defined as in Model A.

B.5 Model E: Uniform unwrapping

We also study the simplest model of sequence-independent binding energy, in which the energy necessary to unwrap a bp of DNA from the histone is constant. In this case, the function $u_{\text{half}}(x)$ has the form

$$u_{\text{half}}(x) = -\frac{E_b}{147}x,$$

where E_b is the binding energy of a fully wrapped nucleosome.

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	35 bp
E_b	13.40 k_BT
μ	-13.14 k_BT
f	0.58

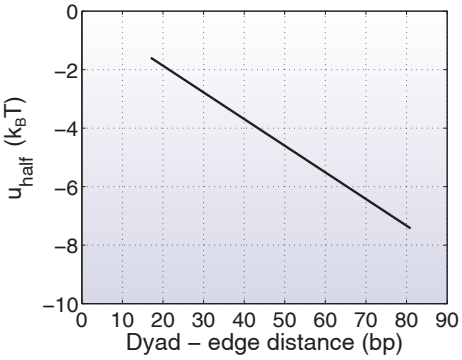


Table B.6: Fitted parameters for Model E. All parameters are as in Model A.

We obtain the fit residuals:

$$RMS = 9.9725 \times 10^{-4},$$

$$r_{\text{osc}} = 0.274786,$$

$$RMS_{\text{osc}} = 2.7507 \times 10^{-4}.$$

All residuals are defined as in Model A.

B.6 Model F: 5-bp oscillations superimposed onto a linear function

Similarly to Models C and D, we construct another model in which we change the periodicity of the oscillations to 5 bp. Let

$$u_{\text{half}}(x) = -A\cos\left(\frac{2\pi}{5}(x - x_0)\right) - \frac{E_b}{147}x,$$

where A, x_0 and E_b have the same meaning as in Model C.

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	39 bp
E_b	13.50 k_BT
μ	-16.13 k_BT
A	2.36 k_BT
x_0	74 bp
f	0.63

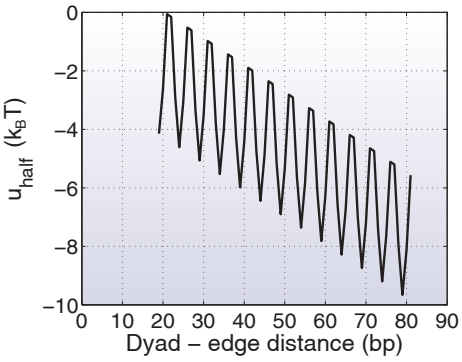


Table B.7: Fitted parameters for Model F. All parameters are as in Model C.

We obtain the fit residuals:

$$RMS = 9.8984 \times 10^{-4},$$

$$r_{\text{osc}} = 0.206240,$$

$$RMS_{\text{osc}} = 3.0554 \times 10^{-4}.$$

All residuals are defined as in Model A.

B.7 Model G: 5-bp stepwise unwrapping

We test two more alternative models for the sequence-independent binding energy of nucleosomes. We test two stepwise profiles, with the size of the steps of 5 bp and 10 bp, respectively. Model G considers the case of the 5 bp steps, and Model H the case of the 10 bp steps. Let

$$u_{\text{half}}(x) = -E_{\text{step}} \texttt{ceil}\left(\frac{x - x_0}{5}\right),$$

where E_{step} is the amount of energy lost in each step, and x_0 determines the phase of the stepwise profile.

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	39 bp
E_{step}	0.48 $k_B T$
μ	-12.83 $k_B T$
x_0	2 bp
f	0.63

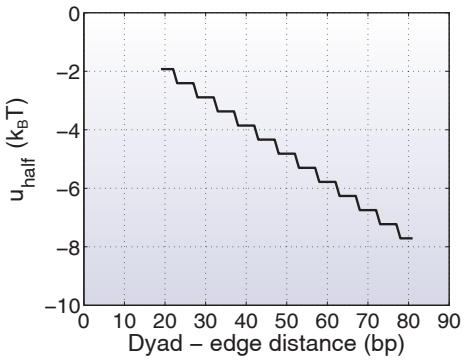


Table B.8: Fitted parameters for Model G. All parameters are as in Model A, except for E_{step} and x_0 defined above.

We obtain the fit residuals:

$$RMS = 9.9040 \times 10^{-4},$$

$$r_{\text{osc}} = 0.282593,$$

$$RMS_{\text{osc}} = 2.7421 \times 10^{-4}.$$

All residuals are defined as in Model A.

B.8 Model H: 10-bp stepwise unwrapping

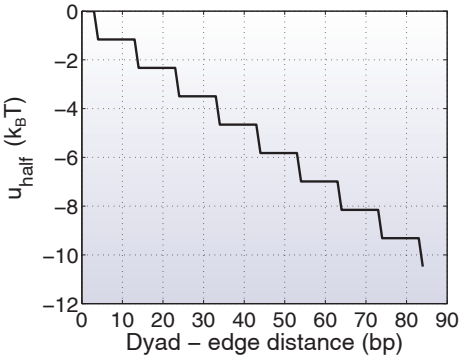
Model H is similar to Model G, but in this case we test a stepwise profile with the size of the steps of 10 bp. Let

$$u_{\text{half}}(x) = -E_{\text{step}} \, \texttt{ceil}\left(\frac{x - x_0}{10}\right),$$

where E_{step} is the amount of energy lost in each step, and x_0 determines the phase of the stepwise profile.

Parameter	Value
$a_{\max}$	169 bp
$a_{\min}$	3 bp
E_{step}	1.16 $k_B T$
μ	-12.04 $k_B T$
x_0	3 bp
f	0.62

Table B.9: Fitted parameters for Model E. All parameters are as in Model F.



We obtain the fit residuals:

$$RMS = 9.7895 \times 10^{-4},$$

$$r_{\text{osc}} = 0.545077,$$

$$RMS_{\text{osc}} = 2.4569 \times 10^{-4}.$$

All residuals are defined as in Model A.

Vita

Răzvan Viorel Chereji

Education

2013 Ph. D. in Physics, Rutgers University

2007 B. Eng. in Physics from Babeş-Bolyai University, Romania

Experience

2010-2013 Graduate assistant, Physics Department, Rutgers University

2008-2010 Teaching assistant, Physics Department, Rutgers University

2007-2008 Fellow, Physics Department, Rutgers University

Publications

1 Răzvan V. Chereji, Denis Tolkunov, George Locke, and Alexandre V. Morozov. Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions. Phys. Rev. E, 83(5):050903, May 2011.

2 Răzvan V. Chereji and Alexandre V. Morozov. Statistical mechanics of nucleosomes constrained by higher-order chromatin structure. J. Stat. Phys., 144(2):379–404, July 2011.

3 Natalia Petrenko, Răzvan V. Chereji, Megan McClean, Alexandre V. Morozov, and James R. Broach. Noise and interlocking signaling pathways promote distinct transcription factor dynamics in response to different stresses. Mol. Biol. Cell 24, 2045–2057 (2013)

4 Răzvan V. Chereji and Alexandre V. Morozov. Ubiquitous nucleosome crowding and unwrapping in the yeast genome. Submitted.

5 Nils Elfving*, Răzvan V. Chereji*, Alexandre V. Morozov, Stefan Björklund and James R. Broach. A dynamic interplay of nucleosome and Msn2 binding regulates activation and repression of gene expression following stress. In preparation.

6 Nils Elfving*, Răzvan V. Chereji*, Miriam Larsson, Alexandre V. Morozov, James R. Broach, and Stefan Björklund. Mediator exists in multiple forms and is predominantly associated to promoters with low nucleosome density. In preparation.

7 Răzvan V. Chereji, Tsung-Wai Kan, Victor P. Guryev, Alexandre V. Morozov, and Yuri M. Moshkin. The positioning of stable and fragile nucleosomes depends on distinct sequence rules and ATP-dependent chromatin remodelers. In preparation.

*These authors contributed equally