APPLICATIONS OF STATISTICAL MECHANICS TO NUCLEIC ACIDS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By
Robert A. Forties, B.S.
Graduate Program in Physics

The Ohio State University
2011

Dissertation Committee:
Ralf Bundschuh, Advisor
Michael G. Poirier
Mohit Randeria
Samir Mathur
David E. Symer
We investigate the use of physical modeling to extract mechanistic details from quantitative biological data, with a focus on the physical properties of nucleic acids. It is well understood that DNA stores genetic information, RNA acts as a carrier of this information, and that both must interact with a wide array of protein complexes in order to perform these functions. However, the physical mechanisms by which these interactions occur are much less clear.

For example, Protein-bound duplex DNA is often bent or kinked. Yet, quantification of intrinsic DNA bending that might lead to such protein interactions remains enigmatic. DNA cyclization experiments have indicated that DNA may form sharp bends more easily than predicted by the established worm-like chain (WLC) model. One proposed explanation suggests that local melting of a few base pairs introduces flexible hinges. We test this model for three sequences at temperatures from 23°C to 65°C. We find that small melted bubbles are significantly more flexible than double-stranded DNA and can alter DNA flexibility at physiological temperatures.

There are also many important proteins which bind single-stranded nucleic acids, such as the nucleocapsid protein in HIV and the RecA DNA repair protein in bacteria. The presence of such proteins can strongly alter the secondary structure of the nucleic acid molecules. Therefore, accurate modeling of the interaction between single-stranded nucleic acids and such proteins is essential to fully understanding many biological processes. We develop a model for predicting nucleic acid secondary structure in the presence of single stranded binding proteins, and implement it as an extension of the Vienna RNA Package. Using this model we are able to predict the probability of the protein binding at any position in the nucleic acid sequence, the impact of the protein on nucleic acid base pairing, the end-to-end distance distribution for the nucleic acid, and FRET distributions for fluorophores attached to the nucleic acid.

Eukaryotic DNA also interacts strongly with nucleosome protein complexes, which wrap and compact this DNA. The expression, replication and repair of DNA requires nucleosomes to be unwrapped and disassembled. We have developed a quantitative model of nucleosome dynamics and calibrated this model using results from high precision single molecule nu-
cleosome unzipping experiments. We then tested its predictions for experiments in which nucleosomes are disassembled by the DNA mismatch recognition complex hMSH2-hMSH6. We found that this calibrated model quantitatively describes hMSH2-hMSH6 induced disassembly rates of nucleosomes with two separate DNA sequences and four distinct histone modification states. In addition, this model provides mechanistic insight into nucleosome disassembly by hMSH2-hMSH6 and the influence of histone modifications on this disassembly reaction. We also found that this model accurately predicts the rate at which lexA is able to trap nucleosome unwrapping fluctuations. This model’s precise agreement with current experiments suggests that it can be applied more generally to provide important mechanistic understanding of the numerous nucleosome alterations that occur during DNA processing.
To my wife, for putting up with six years of this.
ACKNOWLEDGMENTS

I would like to thank numerous people for their help and guidance. My advisor, Prof. Ralf Bundschuh, has been extremely supportive throughout graduate school. He carefully selected interesting research topics for me, and provided me with all the tools I needed to obtain great results from my research. I cannot list all that I have learned from him, from techniques for modeling biological data to how to write better papers. He also has been very flexible, allowing me to also pursue experimental topics in collaboration with other research groups.

The most important of these collaborations has been with Prof. Michael Poirier’s group. I would like to thank them not only for including me in their group and allowing me to use their experimental facilities and equipment, but also for a huge amount of help teaching me how to perform biophysical experiments and interpret biological data. In particular, Justin North provided me with data for LexA binding to nucleosomes, and helped me analyze measurements of nucleosome interactions with hMSH2-hMSH6. Prof. Poirier has also personally advised me on all aspects of my experimental work, and provided me with all of the supplies I needed for these experiments.

I have also enjoyed numerous collaborations outside of the physics department. Prof. Richard Fishel’s group in the Comprehensive Cancer Center has provided me with both interesting data and sound advice. Sarah Javaid helped me to model her measurements of nucleosome displacement by hMSH2-hMSH6, and both Sarah Javaid and Chris Cook assisted me with procedures for creating mismatched DNA. Prof. Fishel has also provided valuable biological insights into all of my projects. Prof. Karin Musier-Forsyth of the Department of Chemistry has given me access to a Typhoon Trio gel scanner which was critical to the analysis of my cyclization experiments. Prof. Alexander Vologodskii of the New York University Department of Chemistry allowed me to include some of his measurements of DNA cyclization, and also discussed this data with me. I also had many helpful discussions with Prof. Paul Barbara of the University of Texas at Austin about of the behavior of single-stranded binding proteins.

In addition to my collaborators, I would also like to thank the many students who have worked with me. Sean Rose, Meagan Segbers, Cai Chen, Shayne Reichard and Malcolm
McCauley all helped me with the synthesis of DNA constructs and performed cyclization experiments. Omar Tabbaa calculated the force required to unzip naked DNA for comparison with nucleosome unzipping data. I greatly enjoyed working with all of them, and hope that they also enjoyed and learned from their experience.

I am also very grateful of all of the people who educated me throughout my life, and inspired me to pursue a career in science. In particular, my high school biology teacher Mrs. Margaret Feindel encouraged me to pursue scientific research, and helped me get accepted into two summer research programs during high school. My first research advisor, Dr. Frederic Marshall of the University of Rochester Laboratory for Laser Energetics, provided me with guidance in high school and throughout college. He taught me how to make careful measurements and how to write scientific papers. Prof. Robert Knox of the University of Rochester provided me with my first experiences in both biophysics and modeling, and helped me to decide to focus my research in these areas. I would also like to thank all of the excellent professors who taught me physics at Ohio State, and in particular Mohit Randeria and Samir Mathur, who in addition to being great classroom teachers also served on my candidacy and doctoral committees.

Finally, I would like to acknowledge all of the support I have received from my family. I am grateful to my mother for always encouraging me to read and learn about anything and everything that interested me throughout my life, and to my brother for sharing his vast knowledge of computers with me, and teaching me not to take myself too seriously. Most of all, I would like to thank my wife and best friend, for keeping me sane through all of this.

I have been funded directly by a National Science Foundation Graduate Research Fellowship and a Dean’s Distinguished University Fellowship from the Ohio State University, and also indirectly by the Petroleum Research Fund of the American Chemical Society [42555-G9] and the National Science Foundation [DMR-0706002] to Ralf Bundschuh and a Career Award in the Basic Biomedical Sciences from the Burroughs Wellcome Fund and the National Institutes of Health [R01 GM083055] to Michael Poirier.
VITA

August 26, 1982 ....................... Born—Binghamton, NY

May, 2005 ............................... B.S., University of Rochester, Rochester, NY

Autumn 2005–Spring 2008 ............... NSF Graduate Fellow, OSU, Columbus, OH

Autumn 2008–Spring 2011 ................ Dean’s Distinguished University Fellow, OSU, Columbus, OH

Publications


Fields of Study

Major Field: Physics

Studies in:

- biophysics theory Ralf Bundschuh
- biophysics experiment Michael Poirier
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td></td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td></td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td></td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td></td>
<td>xi</td>
</tr>
<tr>
<td>List of Tables</td>
<td></td>
<td>xiii</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td></td>
<td>xiv</td>
</tr>
</tbody>
</table>

## Chapters

1. **Introduction**
   - 1.1 The importance of nucleic acids
     - 1.1.1 DNA
     - 1.1.2 RNA
   - 1.2 Experimental tools for biophysics
     - 1.2.1 Magnetic tweezers
     - 1.2.2 Optical tweezers
     - 1.2.3 FRET
     - 1.2.4 Atomic force microscopy
     - 1.2.5 Biochemical techniques
   - 1.3 Statistical physics may be used to interpret biological data
     - 1.3.1 Boltzmann statistics
     - 1.3.2 Extension of a freely-jointed chain
   - 1.4 Outline

2. **The Flexibility of Locally Melted DNA**
   - 2.1 Extension of the worm-like chain to include melted bubbles within DNA
   - 2.2 Cyclization Experiments
   - 2.3 Extraction of bubble flexibilities from cyclization experiments
   - 2.4 Cyclization at high temperatures
   - 2.5 Cyclization of DNA containing single base pair mismatches
   - 2.6 Discussion

3. **Modeling the interplay of single-stranded binding proteins and nucleic acid secondary structure**
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Calculation of the partition function for nucleic acid secondary structure in the presence of single-stranded binding proteins</td>
<td>40</td>
</tr>
<tr>
<td>3.2</td>
<td>Observables</td>
<td>43</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Melting probability</td>
<td>43</td>
</tr>
<tr>
<td>3.2.2</td>
<td>End-to-end distance</td>
<td>44</td>
</tr>
<tr>
<td>3.2.3</td>
<td>FRET distribution</td>
<td>45</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Protein binding</td>
<td>47</td>
</tr>
<tr>
<td>3.3</td>
<td>Discussion</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>A quantitative model of nucleosome dynamics</td>
<td>52</td>
</tr>
<tr>
<td>4.1</td>
<td>Modeling nucleosome unwrapping dynamics</td>
<td>53</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Solution to the master equation for unwrapping dynamics</td>
<td>53</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Rates of nucleosome unzipping</td>
<td>54</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Calculation of dynamics in the presence of hMSH2-hMSH6</td>
<td>55</td>
</tr>
<tr>
<td>4.2</td>
<td>Results</td>
<td>58</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Mechanical Unzipping of Nucleosomes</td>
<td>58</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Nucleosome Disassembly by hMSH2-hMSH6</td>
<td>61</td>
</tr>
<tr>
<td>4.2.3</td>
<td>DNA Sequence Dependence of Nucleosome Disassembly by hMSH2-hMSH6</td>
<td>63</td>
</tr>
<tr>
<td>4.2.4</td>
<td>The influence of Histone Post-Translational Modifications of Nucleosome Disassembly by hMSH2-hMSH6</td>
<td>64</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Binding of LexA to the entry-exit region</td>
<td>65</td>
</tr>
<tr>
<td>4.3</td>
<td>Discussion</td>
<td>66</td>
</tr>
</tbody>
</table>

**Bibliography**

72

**Appendices**

A Calculation of cyclization J factors

B Fitting Bubble Flexibilities

C Twist Measurements

D Ligatable ends for high temperature ligases

E Experimental Methods

E.1 Cyclization of DNA fragments with T4 ligase

E.2 DNA precipitation with linear polyacrylamide

E.3 Procedure for generating single-stranded plasmid with M13 helper phage

E.4 BND cellulose extraction of single-stranded DNA

F Full partition function for nucleic acid secondary structure

107
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 DNA structure</td>
<td>2</td>
</tr>
<tr>
<td>1.2 RNA secondary structure</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Magnetic Tweezers</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Extension vs. force for a DNA molecule</td>
<td>7</td>
</tr>
<tr>
<td>1.5 Unzipping DNA using an applied force</td>
<td>10</td>
</tr>
<tr>
<td>1.6 Bending of DNA measured by AFM</td>
<td>12</td>
</tr>
<tr>
<td>1.7 Example electrophoresis gel</td>
<td>14</td>
</tr>
<tr>
<td>1.8 Diagram of PCR</td>
<td>15</td>
</tr>
<tr>
<td>1.9 Extension of a FJC polymer</td>
<td>18</td>
</tr>
<tr>
<td>2.1 Synthetic 116 bp DNA sequences</td>
<td>23</td>
</tr>
<tr>
<td>2.2 Rate of ligation versus concentration of T4 ligase</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Example gel showing a ligation time course</td>
<td>26</td>
</tr>
<tr>
<td>2.4 Cyclization $J$ factors as a function of temperature.</td>
<td>27</td>
</tr>
<tr>
<td>2.5 Linearity of high temperature ligases</td>
<td>30</td>
</tr>
<tr>
<td>2.6 Cyclization at high temperatures</td>
<td>32</td>
</tr>
<tr>
<td>3.1 Arch diagram showing the decomposition of the partition function</td>
<td>41</td>
</tr>
<tr>
<td>3.2 Predicted FRET distribution in the presence of binding proteins</td>
<td>45</td>
</tr>
<tr>
<td>3.3 Probability for protein to bind to a single-stranded nucleic acid</td>
<td>48</td>
</tr>
<tr>
<td>4.1 Model predictions compared with experimental measurements for nucleosome displacement by hMSH2-hMSH6.</td>
<td>56</td>
</tr>
<tr>
<td>4.2 Full simulations of nucleosome displacement compared to our approximate solution.</td>
<td>57</td>
</tr>
<tr>
<td>4.3 A free energy landscape consistent with measurements of mechanical unzipping of nucleosomes.</td>
<td>59</td>
</tr>
<tr>
<td>4.4 The force needed to unzip the DNA base pairing and stretch the DNA</td>
<td>60</td>
</tr>
<tr>
<td>4.5 A free energy landscape constrained to have steps separated by 5.25 bp</td>
<td>62</td>
</tr>
<tr>
<td>4.6 Experimental setup for probing nucleosome unwrapping fluctuations with LexA</td>
<td>66</td>
</tr>
<tr>
<td>4.7 Force exerted on the nucleosome by hMSH2-hMSH6.</td>
<td>67</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Persistence lengths for melted bubbles in DNA</td>
<td>29</td>
</tr>
<tr>
<td>4.1 Binding energy for modified nucleosomes relative to wild type nucleosomes</td>
<td>66</td>
</tr>
</tbody>
</table>
List of Abbreviations

AFM  atomic force microscopy. 1, 12, 20, 21

bp  base pair. 13, 20, 21, 24, 25, 27–29, 31, 33–37, 93–95, 100

dsDNA  double-stranded DNA. 9, 12, 13, 17, 21, 22, 27–29, 34–37, 94

FJC  freely-jointed chain. 7, 8, 17, 18

FRET  fluorescence resonance energy transfer. 1, 10, 11, 20, 39, 43–47, 49, 51, 53

LPA  linear polyacrylamide. 25, 97, 100, 101

LSEC  linear sub-elastic chain. 13

nt  nucleotide. 21

PCR  polymerase chain reaction. 14, 15, 34–36

SDS  Sodium dodecyl sulfate. 31

ssDNA  single-stranded DNA. 4, 9, 17, 23, 28, 29, 34–36, 105

ssNA  single-stranded nucleic acid. 19

tRNA  transfer RNA. 4

WLC  worm-like chain. 7, 8, 12, 13, 19, 29, 32, 33
Chapter 1
INTRODUCTION

The study of biology has been revolutionized over the past 25 years by the application of a number of technologies allowing detailed, quantitative measurement of biological systems at the microscopic level. Some of these new tools include optical [1, 2] and magnetic tweezers [3, 4], fluorescence resonance energy transfer (FRET) [5, 6], and atomic force microscopy (AFM) [7–9]. While these techniques provide detailed and accurate data, it is not always straightforward to relate this data to the underlying physical properties of biomolecules. Physical modeling is required to interpret measurements of force or distance, and extract mechanistic details about the underlying behavior of biological systems.

For example, by applying a polymer model [3, 10] to measurements of DNA stretching, one can determine the intrinsic flexibility of the DNA from measurements of force versus extension. Or, one can learn how a protein deforms DNA upon binding by modeling careful measurements of the force at which the protein binds to a DNA under tension [11]. Modeling becomes even more important when probing changes in single-stranded DNA, RNA, or proteins, as these polymers can have complex self-interactions, resulting in a huge ensemble of possible structures which must be considered when interpreting experimental measurements [12–17].

In this work we will discuss several applications of a combination of theoretical and experimental techniques to biological problems. In particular, we will focus on the structure and function of nucleic acids, as well as their interactions with proteins.

1.1 The importance of nucleic acids

DNA and RNA are vital to controlling the operation of cells. Almost all components of the cell are made from proteins, and the information needed to produce these proteins is contained on a single set of DNA within each cell. The information for a single protein is contained in a segment of the DNA called a gene. The cell needs to produce many copies of each protein simultaneously, so it creates many RNA copies of each gene from its DNA
which are then used by the cell’s production machinery to make these proteins.

Figure 1.1: The structure of a small DNA molecule determined using x-ray crystallography [18] and visualized using PyMOL [19]. Each nucleotide consists of a sugar-phosphate backbone, shown in black, which is the same in all nucleotides, and a base which is different in each of the four nucleotides. The four bases are color-coded, with adenine (A) shown in red, thymine (T) shown in blue, guanine (G) shown in green and cytosine (C) shown in yellow. All hydrogen atoms are shown in white. (a) Every base can pair with one other base by forming hydrogen bonds. As shown, A can pair with T, forming 2 hydrogen bonds (shown by black dashed lines), and G can pair with C, forming 3 hydrogen bonds. (b) The full crystal structure of the DNA molecule, showing the position of all atoms [18]. The nucleotides are all paired as described in (a), and the two strands of DNA twist around one another to form a double helix. (c) A cartoon representation of the same DNA molecule in the same orientation. The backbone of each DNA strand is shown as a thick black tube, and the bases are shown as thin colored rods. This cartoon emphasizes that the backbones of the two strands form a double helix, and that opposing bases are paired across the two strands.

1.1.1 DNA

Each DNA molecule is a polymer composed of two strands of nucleotide subunits, as shown in Fig. 1.1. The two strands are held together by hydrogen bonding between nucleotides on opposite strands. There are four different nucleotide subunits, which form a 4 base code. Inside of genes, every set of 3 consecutive DNA nucleotides codes for one amino acid, the basic subunit of proteins. There are 20 different amino acids, and three nucleotides can have
$4^3 = 64$ combinations, which is more than sufficient to code for all possible amino acids. The one-dimensional sequence of DNA nucleotides within a gene, and hence the sequence of amino acids within a protein, is sufficient to encode the three-dimensional structure of a protein [20, 21]. Many proteins can spontaneously form their native structure in solution starting from a denatured state, far away from this native structure [22], proving that their structure must be encoded by just their sequence.

Not only the sequence of DNA, but also its physical properties are important for it to perform its function as the information carrier for cells. It must be transcribed into RNA to allow protein production, replicated when the cell divides, and any damage to the DNA must be repaired to ensure fidelity of the genetic information. During all of these processes, large protein complexes must bind to the DNA. Many of these complexes interact with the DNA in interesting physical ways, such as by bending it or breaking apart the hydrogen bonding between the two strands. For example, the DNA repair complex hMSH2-hMSH6 sharply bends DNA when it binds [23] and the strands of DNA must be separated during DNA replication and RNA transcription [24]. Better understanding of the physical properties of DNA would help clarify the mechanism behind such interactions.

While it is quite well understood how DNA encodes protein sequences, the mechanism by which the cell controls which genes are active at any given time is much less clear. Every cell in a human contains identical DNA, yet these cells are highly differentiated based on which set of genes are active, and hence which proteins are produced. Defects in gene expression are implicated in many of human diseases, including heart disease and essentially all cancers [25–28]. A wide array of proteins bind with DNA in order to control gene expression, and many of these proteins also interact with the DNA physically. Proteins called transcription factors bind to the DNA and either promote or repress expression of genes [29, 30]. In one of the few systems where gene expression is well understood, the Lac operon in *E. coli*, a repressor protein deactivates a gene by binding to two distant sites within the DNA, and bringing them together to bend the DNA into a loop [31, 32]. This then prevents the gene from being expressed because the RNA polymerase, which translates the DNA into RNA, cannot physically fit on the DNA when it is bent into such a loop. In eukaryotes, such as humans, all DNA is tightly wrapped around protein complexes called histones to form nucleosomes which act like “spools” that compact the DNA so that it can fit inside the cell nucleus [33]. The positioning of these nucleosomes impacts gene expression, as DNA which is wrapped into a nucleosome cannot be translated as easily by a polymerase [34]. Studying these interactions could help us better understand and control gene expression, which has applications to a wide variety of human diseases.
Figure 1.2: A schematic representation of the secondary structure of a single-stranded RNA molecule. The individual nucleotides are represented by filled black circles, and the covalent bonds between them making up the RNA backbone by black lines. Since RNA is single-stranded, there is no complementary strand with which its bases can pair. It therefore bends back upon itself to allow bases within the same strand to pair (represented by red lines), as base pairing is very energetically favorable.

1.1.2 RNA

RNA is single-stranded, so instead of bases pairing with a complementary strand like in DNA, the strand bends back upon itself to form base pairings (Fig. 1.2). This base pairing within an RNA molecule is called its secondary structure. Single-stranded DNA (ssDNA) will also exhibit this behavior, pairing with itself to form a secondary structure.

As mentioned above, RNA is a copy of the genetic information for a gene which is used to make proteins. However, RNA is not merely a passive copy of the DNA, but rather actively controls protein production in many ways. RNA can form secondary structures which help promote initiation or termination of the translation of the RNA into a protein [35]. Most RNA is also heavily edited before the gene is translated, with non-coding regions called introns removed and the remaining RNA spliced back together [36, 37]. In some cases, the sequence of the RNA is altered by modifying one or more bases [38]. In all of these activities, the RNA structure plays an important role, by either helping to promote binding of other complexes, or even actively catalyzing reactions upon itself [39].

RNA performs many important activities within the cell apart from carrying genetic information. It is a major component of ribosomes, the complexes which produce proteins [40]. Small RNAs called transfer RNAs (tRNAs) are attached to amino acids, to allow them to be properly identified and incorporated into proteins by ribosomes [41]. RNAs also play many catalytic roles in the cell. In addition to the self-catalyzed splicing mentioned above, small RNAs help catalyze splicing or modification of other RNAs [42]. Other small RNAs decrease the activity of genes by binding and helping break down their RNAs, providing the cell with additional regulation of the expression of these genes [43].
These functions of RNA almost universally depend on its structure. Therefore, it is extremely important to understand this structure, and in particular the structure of RNA while bound to other molecules, such as proteins, DNA, or other RNAs.

1.2 Experimental tools for biophysics

In this section we will give a brief overview of the new experimental methods that allow measurements of biological systems such as DNA and RNA at the molecular scale, as well as some useful methods from classical biology. We will also provide examples of how modeling may be used to interpret data from some of these experiments.

1.2.1 Magnetic tweezers

![Diagram of magnetic tweezers system](image)

Figure 1.3: Experimental setup for an example magnetic tweezers system. A DNA molecule is attached at one end to a surface, and at the other end to a superparamagnetic bead [44]. A permanent magnet is placed above the bead, as shown, and the force on the bead may be adjusted by changing the distance between the magnet and the bead. In actual experiments, multiple magnets are often used to increase the magnetic field gradient, and therefore increase the force on the bead [3, 4].
In a magnetic tweezers apparatus, a biomolecule (usually DNA) is attached by one end to a superparamagnetic bead, and by the other end fixed to a surface, as shown in Fig. 1.3 [3, 4]. A superparamagnetic bead is made of a ferromagnetic material, but is small enough that it contains only a single magnetic domain, whose orientation fluctuates due to thermal excitations [44]. Under the influence of an external magnetic field, a superparamagnet will become polarized similarly to an ordinary paramagnet, but with much larger susceptibility [44].

A force is applied to the bead by placing it in a nonuniform magnetic field, generated by placing one or more permanent magnets near the bead [3, 4]. If the strength of the magnetic field applied to the bead is $B$ and we treat it as uniform across the bead, the induced magnetic dipole is [45, 46]

$$\mathbf{m}_i = \frac{V}{\mu} \left( \frac{\chi_m}{3 + \chi_m} \right) \mathbf{B}, \quad (1.1)$$

where $V$ is the volume of the bead, $\mu$ is the magnetic permeability of the solvent, and $\chi_m$ is the susceptibility of the bead. The force on the bead is then [45, 46]

$$\mathbf{F} = \nabla (\mathbf{m}_i \cdot \mathbf{B}) = \frac{V}{\mu} \left( \frac{\chi_m}{3 + \chi_m} \right) \nabla B^2. \quad (1.2)$$

In magnetic tweezers experiments, the external field $B$ is provided by permanent magnets [3, 4]. We will examine the simplest case where there is a single permanent magnetic dipole oriented vertically, and the superparamagnetic bead is placed along its axis (which we will call the $z$-axis). The magnetic field from a permanent dipole $\mathbf{m}_p$ is [45]

$$\mathbf{B}(\mathbf{x}) = \frac{\mu}{4\pi} \frac{3\mathbf{n}(\mathbf{n} \cdot \mathbf{m}_p) - \mathbf{m}_p}{|\mathbf{x}|^3}, \quad (1.3)$$

where $\mathbf{x}$ is the displacement from the center of the dipole and $\mathbf{n} = \mathbf{x}/|\mathbf{x}|$. Therefore, the $z$ component of the magnetic field along the $z$-axis is

$$B_z(z) = \frac{\mu}{2\pi} \frac{m_p}{z^3}. \quad (1.4)$$

If we insert Eq. 1.4 into Eq. 1.2, we obtain the vertical force on the bead

$$F_z = -V \frac{9\mu m_p^2}{2\pi^2} \frac{\chi_m}{3 + \chi_m} \frac{1}{z^7}. \quad (1.5)$$

Given the strong dependence of this force on the distance between the magnet and the bead, one can clearly obtain a wide range of forces simply by moving the magnet closer or farther from the bead. In actual magnetic tweezers experiments, multiple magnets are frequently used in order to obtain an even stronger magnetic field gradient [3, 4], but the mechanism by which the force is generated remains the same.
This approach allows one to accurately apply a constant force to the magnetic bead, while tracking it’s position using a microscope [3, 4]. It was used by Smith et al. to make precise measurements of the extension of double-stranded DNA molecules as a function of the applied force [3, 10]. Their data (Fig. 1.4) is accurate enough to show that the elasticity of DNA is not properly described by the freely-jointed chain (FJC) model, but rather requires application of the worm-like chain (WLC) model [10].

Figure 1.4: A plot from [10]. Reprinted with permission of AAAS. This plot shows extension vs. force for a DNA molecule obtained using magnetic tweezers. The DNA molecule used was a 97,000 base pair segment from bacteriophage λ, with total contour length 32.7 µm [10]. The data are shown by open squares, the best fit of the FJC model to the data (with contour length $L = 32.7\mu m$ and segment length $b = 100$ nm) is shown by the dashed line, and the best fit of the WLC model (with contour length $L = 32.8\mu m$ and persistence length $\lambda = 53.4$ nm) is shown by the solid line. The FJC clearly does not fit the data very well except at forces less than 100 fN, while the WLC model provides an excellent fit to the data, with only minor deviations at forces above 10 pN.

The FJC model gives the simplest possible description of a polymer [3]. The polymer is treated as series of rigid segments of equal length $b$, with no restriction on the relative orientation of neighboring segments. There is no energy cost to bend the polymer, and it therefore traces out a random walk in space. The expectation value for the end-to-end distance of the polymer under an applied force $F$ is (see Section 1.3.2) [3]

$$\langle x \rangle = L \left[ \coth \left( \frac{Fb}{k_B T} \right) - \frac{k_B T}{Fb} \right], \quad (1.6)$$
where $L$ is the contour length of the polymer, $k_B$ is Boltzmann’s constant and $T$ is the temperature. The FJC best fits the experimental data for extension of DNA with a segment length $b = 100$ nm and a contour length $L = 32.7 \mu m$, but the dashed line in Fig. 1.4 shows that a good fit is only obtained at very low forces, and the model predicts that the DNA should reach its full length extension at much lower forces than shown by the experimental data.

The WLC model treats a polymer as a uniform elastic rod whose stiffness is described by its persistence length $A$. The bending energy of a WLC polymer is

$$E_{WLC} = k_B T \int_0^L A \left| \frac{dt(s)}{ds} \right|^2 ds,$$

where $t(s)$ is the vector tangent to the polymer as a function of position $s$ along its contour length. The bending energy of the polymer at position $s$ is proportional to the square of its curvature at $s$, and its persistence length $A$. The force required to extend a WLC polymer to length $x$ must be calculated numerically, but it may be approximated by the interpolating function

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

The solid line in Fig.1.4 shows that this model provides an excellent fit to the data for DNA extension with a contour length $L = 32.8 \mu m$ and persistence length $A = 53.4$ nm.

This experiment provides a canonical example of how modeling can be used to interpret experimental data, and provide insight into the underlying physical properties of a biological system. The data clearly show that DNA is in fact a stiff polymer, as if one tries to model it with the FJC, the force needed to extend the DNA is too high at low extension, and too low at high extension. The FJC model overestimates the number of configurations available to the DNA at low force, as it neglects the stiffness of the polymer. It therefore predicts too large of an entropy loss, and hence too high of a force required to extend the DNA at low forces. The WLC model correctly models the stiffness of DNA, and therefore is able to reproduce its force response over a wide range. The stiffness of DNA was already known from bulk biochemical experiments, but this measurement provided an excellent confirmation at the single-molecule scale. Furthermore, an accurate model of the extension of naked DNA under force facilitates interpretation of experiments measuring the interaction of single DNA molecules with other biomolecules, as deviations from model predictions can be identified as interactions with the biomolecules.

### 1.2.2 Optical tweezers

Instead of using a magnetic bead to apply force to a biomolecule, one may instead use a bead trapped at the focus of a laser. A dielectric bead will refract any light passing through
it, and therefore due to conservation of momentum the light will exert a force on the bead. This force will act to draw the bead towards the focus of a laser with a Gaussian beam profile, trapping the bead [1, 49]. The force exerted on the bead will be proportional to its distance from the focus of the laser, allowing this force to be determined simply by measuring the position of the bead.

This apparatus may of course perform all the same types of measurements which are possible using magnetic tweezers. As with magnetic tweezers, one end of a biomolecule can be attached to the trapped bead, and the other end fixed. The main difference is that instead of directly controlling the force applied to the bead by adjusting the position of a magnet, one controls the extension of the attached biomolecule by adjusting the distance between the laser trap and the fixed end. One may still apply a constant force by creating a feedback loop which dynamically adjusts the trap position to maintain a constant force [2, 50]. The advantages of optical tweezers over magnetic tweezers are that they can apply larger forces, in excess of 100 pN [1, 49], and forces and positions may be measured more accurately [2, 50].

In particular, by attaching both ends of the biomolecule to dielectric beads placed in two separate optical traps, one decouples the system from any vibrations of the experimental apparatus, allowing displacements to be measured with nm resolution [2, 50]. Michelle Wang’s group attached separate dielectric beads to the two individual strands of a double-stranded DNA (dsDNA) molecule, and then pulled on them using two optical traps. This allowed them to measure the unzipping of the strands of DNA with single bp resolution [2, 50].

To model such unzipping data, one must consider both the free energy cost to break apart the DNA base pairing, and energy required to stretch both the single-stranded and double-stranded regions of the DNA [51, 52]. Fig. 1.5 shows that by using an appropriate polymer model for DNA stretching and fixed, sequence dependent free energy costs for breaking DNA base pairings one can accurately reproduce experimental unzipping data [51]. This model provides a mechanistic explanation for the sawtooth pattern found in the data. As the DNA is extended, the ssDNA is stretched and the force increases linearly, resulting in the sloped portion of the sawtooth. When the force becomes large enough, the base pairing of the next region of dsDNA is broken, the force drops rapidly and the ssDNA contracts again, allowing this cycle to be repeated. The strength of the DNA base pairing is not uniform, because pairing between G and C is significantly stronger than pairing between A and T. Therefore, the unzipping will pause at regions with many consecutive G and C base pairs, and then unzip all the way to the next GC rich region when the force is large enough to overcome this barrier.

This example demonstrates how multiple simple physical models may be combined to explain a complicated process, and reinforces the importance of having good models for simple biological systems, such as bare DNA. Without a good model for the behavior of a
Figure 1.5: A figure from [51] showing the applied force required to unzip DNA as a function of position. Experimental curves $E_1$ and $E_2$ show the measured force required to unzip the same segment of $\lambda$ DNA in both the forward and reverse directions. The unzipping is performed slowly enough to allow it to be accurately reproduced using equilibrium statistical mechanics, as shown by theoretical curves $T_1$ and $T_2$. Reprinted figure with permission from Bockelmann, U, Essevaz-Roulet, B, & Heslot, F. (1997) Molecular stick-slip motion revealed by opening DNA with piconewton forces. Phys. Rev. Lett. 79, 4489-4492. Copyright 1997 by the American Physical Society. Readers may view, browse, and/or download material for temporary copying purposes only, provided these uses are for noncommercial personal purposes. Except as provided by law, this material may not be further reproduced, distributed, transmitted, modified, adapted, performed, displayed, published, or sold in whole or part, without prior written permission from the American Physical Society.

biomolecule in isolation, it is impossible to understand its interactions with other molecules, or its behavior under extreme conditions. In Section 4.2.1 we will examine unzipping of DNA containing nucleosomes, using this model of unzipping bare DNA as a baseline.

1.2.3 FRET

While the tweezers techniques described above allow measurement of end-to-end distances in molecules under tension, FRET can report the distance between two small dye molecules attached to a biomolecule, allowing passive measurement of small distances. FRET was first described in 1959 [5], but the ability to easily conjugate appropriate dyes to biomolecules is relatively new [53, 54]. In FRET, a donor fluorophore absorbs a photon and becomes excited. In the absence of any other nearby fluorophores, this excitation will eventually decay, and the fluorophore will emit another photon. However, if there is a nearby acceptor fluorophore whose absorption spectrum significantly overlaps with the donor’s emission spectrum, the excitation can be transferred nonradiatively to the acceptor. The acceptor will then decay, emitting a photon. Assuming the emission wavelengths of the donor and acceptor are
significantly different, one can determine the efficiency with which FRET transfer occurs by monitoring the emission of both the donor and the acceptor [55, 56]

\[ E_{FRET} = \frac{I_A}{I_A + I_D}, \]  

(1.9)

where \( I_A \) is the emission intensity of the acceptor, and \( I_D \) is the emission intensity of the donor.

The strength of the interaction which drives this excitation transfer is due to the coupling between the transition dipole moments of the donor and the acceptor, which causes the distance dependence of this interaction to scale like [55]

\[ E_{FRET}(r) = \frac{1}{1 + (r/R_0)^6}, \]  

(1.10)

where \( r \) is the distance between the donor and the acceptor, and \( R_0 \) is a constant, called the FRET radius, which depends on the fluorophores used and solvent conditions. The distribution \( E_{FRET}(r) \) is sigmoidal in shape, with its strongest distance dependence at \( r = R_0 \). For the cyanine dyes Cy3 and Cy5 attached to DNA in aqueous solution, \( R_0 = 6 \text{nm} \) [55]. Therefore, these dyes are optimal for measuring distances on the scale of about 6 nm, which is perfect for measuring conformational changes in biomolecules.

FRET may either be measured in bulk solution or on single molecules, which may be isolated either using tweezers techniques as described above or simply by binding to a surface a low concentration [54]. Bulk FRET measurements only determine the average FRET efficiency, while measuring FRET for a collection of single molecules will result in a distribution of FRET efficiencies. If the molecules have multiple distinct conformations in which the distances between fluorophores are different, this will result in a structured FRET distribution whose peaks correspond to different conformations of the molecule. Careful analysis of such a FRET distribution, as will be described in Section 3.2.3, can allow one to learn about both the conformations of the underlying states of a biomolecule and their relative occurrences.

### 1.2.4 Atomic force microscopy

We may also measure the structure of biomolecules on the nm length scale by monitoring atomic forces between these molecules and a probe [57]. An atomic force microscope consists of a very sharp tip mounted on a flexible cantilever [7, 57, 58]. When the tip is moved close to a surface, the interaction between the tip and the surface is detected by monitoring the deflection of the cantilever. The tip may either be operated in contact mode, where the cantilever exerts a static force on the surface beneath it, and the force is proportional to the deflection of the cantilever, or non-contact mode, where the tip is placed near to but not touching the surface, and vibrated near its resonant frequency. Forces are then determined...
based on the damping of the vibrational oscillations due to interactions with the surface.

This device is capable of producing extremely precise images of surfaces, or molecules deposited onto a surface. Its resolution in the plane of the surface is limited by the width of the tip, which can be as small as a few nm [7, 57, 58], while the vertical resolution can be less than 1 Å [57]. This allows single molecules on the surface to be resolved readily.

This technique was used by Wiggins et al. to test the flexibility of dsDNA on short length scales [9]. They deposited dsDNA on mica, and used AFM to measure the configurations of the DNA molecules. Fig. 1.6a shows the distribution of the angle $\theta$ between the vectors tangent to two points on the DNA as a function of the contour length $L$ by which the points are separated. Fig. 1.6b shows the real-space distance $R$ between two points as a function of $L$. For points which are separated by large contour lengths or cases where bend angles are small, their measurements agree very well with the WLC. However, at short distances they find that sharp bends occur significantly more often than predicted by the WLC [9].

Wiggins et al. suggest that this discrepancy could be caused by the WLC overestimating the energy cost for DNA to make sharp bends. In the WLC model, the energy cost to bend a small segment of DNA is [47]

$$E_{WLC} = \frac{A}{2b} \theta^2 \times k_B T, \quad (1.11)$$

where $A$ is the persistence length of the DNA, $b$ is the length of the small segment, $\theta$ is the bend angle, $k_B$ is Boltzmann’s constant and $T$ is the temperature. They propose breaking
the DNA into segments of length $b = 2.5 \text{ nm}$, and replacing the energy function with \[9\]

$$E_{\text{LSEC}} = \alpha |\theta| \times k_B T,$$

(1.12)

where $\alpha$ is a dimensionless constant which depends on the choice of segment length. They call this model the linear sub-elastic chain (LSEC), as the increase in force with bend angle is slower than the quadratic potential of the WLC. They find that this model fits correlations at large distance with $\alpha = 6.8$. Fig 1.6 shows that this model also does a good job explaining bending at short distances, and is a significant improvement over the WLC for these conditions. This example demonstrates that part of the importance of having a good quantitative model is that it allows one to test the limits of the model, and provides a framework for further extension of the theory.

1.2.5 Biochemical techniques

In this section we will provide a brief overview of some of the biological techniques that will be applied in this work. These methods are frequently used in biophysics to supplement some of the new technologies described above, or to test quantitative predictions.

Gel electrophoresis

Gel electrophoresis is an extremely useful technique which allows separation and visualization of molecules based on their size and shape. The molecules are loaded into wells at the top of the gel, and then an electric field is applied. If the molecules themselves are charged (as is the case with DNA) or they are bound to another charged molecule, the electric field will cause them to migrate through the gel. Smaller molecules will be able to move faster, as they can more easily squeeze through the pores of the gel. In general, similar molecules will be separated just based upon their mass. For example, except for a few exotic sequences, dsDNA will separate just based on the number of base pairs (bps) within each DNA molecule, with longer DNAs moving slower and shorter ones moving faster. However, as will be discussed in detail in Section 2.2, dsDNA that is not linear, but which has its ends joined together to form a circle will move through a gel much more slowly than linear DNA of the same size. This is because DNA is quite stiff, and therefore it is difficult for it to bend enough so that the circles can squeeze through the pores, while a linear DNA can move through the pores relatively easily if it is aligned with its direction of motion. DNA bound by a protein will also migrate differently depending on the position of the protein. In the protein is bound at the middle of the DNA, it will run slower than if the DNA is bound at the end, as the protein in the middle has two DNA “tails”, which resist migration more than the single DNA “tail” created when the proteins is positioned at the end of the DNA [59].
Figure 1.7: An example electrophoresis gel showing the products of PCR cycle titrations for two DNA sequences. DNA is loaded into the wells at the top of the gel, and an electric field is applied. The field causes the DNA to migrate through the gel due to its strong negative charge. The leftmost lane shows a size standard ladder, containing DNA fragments of known size and concentration. The fragments increase in size in 100 bp increments, and each band is labeled with its size. The smaller DNA fragments run faster through the gel, and the larger ones run slower. The remaining lanes show PCR cycle titrations for a 416 bp DNA sequence and a 597 bp DNA sequence. In a cycle titration, aliquots are removed after every 3 PCR cycles and analyzed to optimize the reaction. The lanes are labeled with the number of PCR cycles after which each sample was removed. Because we know the size of the product that we expect, we can see that the PCR for the 416 bp DNA sequence appears well optimized, while the 597 bp DNA contains a significant amount of a short DNA impurity. The faint band just below the wells is the template DNA plasmid, which is circular and has a size of about 5000 bp.

As an example, gel electrophoresis may be used to analyze the results of a polymerase chain reaction (PCR) cycle titration (see Fig 1.7 and below for details of PCR). The approximate size of each PCR product may be determined by comparing with a standard consisting of DNA fragments of known size and concentration. Given that we know the expected size of the PCR products, we can determine the efficiency with which each is produced and the degree of contamination by looking at the intensity of bands corresponding to the desired size compared to any other bands present in the gel.
Figure 1.8: A schematic representation of PCR. (a) The initial DNA template is shown in black. (b) When dsDNA is heated to 95°C, its strands separate. (c) The DNA is cooled to about 65°C, allowing short single-stranded pieces of DNA (shown in red) to anneal at the ends of the template. (d) The DNA is heated 72°C activating a thermostable polymerase (represented by green ovals). (e) The polymerase fills in the rest of each strand (new DNA shown in blue), until we are left with two full copies of the initial DNA template. The cycle may then be repeated, using these new DNAs as templates.

**PCR**

PCR is a technique which allows rapid duplication of a segment of DNA [60, 61]. The two strands of a DNA template are first separated by heating them to 95°C. Then, the temperature is lowered to about 65°C to allow annealing of short single-stranded pieces of DNA called primers which are complementary to the template near its ends. Finally, the temperature is raised to 72°C, which activates a thermostable DNA polymerase [61]. This polymerase extends the primer DNAs by incorporating free nucleotides from the solution, until two copies of the template are produced. By repeating this cycle many times, the number of copies may be increased exponentially.

In addition to allowing simple and rapid amplification of DNA, PCR also can be used to incorporate modifications into DNA. The short primers are artificially synthesized, so it is relatively straightforward to incorporate a variety of chemical modifications into these primers, such as attachment of fluorophores or antibodies which can be bound to beads or surfaces. When such modified primers are used in PCR, the modifications are then incorporated into the resulting DNA product.

**Restriction enzyme digests**

Many bacteria possess enzymes capable of cutting DNA, designed to degrade foreign DNA [62]. These enzymes are extremely useful in biochemistry, as they cut DNA at
very specific sites based on sequence, and also frequently leave a specific single-stranded overhang. This overhang may then be used to attach the DNA to another piece of DNA with a complementary overhang with the aid of another enzyme called a ligase.

In addition to cutting DNA, these enzymes may also be used to test the accessibility of the DNA at the cut site. For example, by comparing the rate of cutting DNA both with and without a bound protein such as a nucleosome, we can determine the degree to which the protein blocks access to the DNA at specific positions [63].

**Cloning**

Cloning DNA into cells provides an excellent means to produce many copies of a DNA sequence with high fidelity [64]. If the DNA contains a gene capable of being expressed in the cells used, one can also create large quantities of proteins in cell cultures.

*E. coli* bacteria is the most commonly used organism for cloning, as it reproduces very rapidly (with cells dividing every 20 minutes on average), and it is very straightforward to clone DNA into *E. coli*. When treated with appropriate chemicals or electric fields, *E. coli* will readily uptake circular segments of DNA called plasmids and replicate them. *E. coli* naturally exchange such plasmids as a means of acquiring useful genes from their neighbors, and increasing their genetic diversity in the absence of true sexual reproduction [62, 64].

**1.3 Statistical physics may be used to interpret biological data**

All of the models described above are derived from statistical physics. Statistical mechanics provides a mechanism for predicting the behavior of systems where thermal fluctuations are important. The energy of thermal fluctuations at 37°C, human body temperature, is $4 \times 10^{-21} \text{J}$. Most biological interactions are stabilized primarily by hydrogen bonds, which have strengths of $8 \times 10^{-21} \text{J}$ to $50 \times 10^{-21} \text{J}$, only 2 to 12 times the thermal energy. Since the bonds which stabilize biomolecules are on the same energy scale as thermal fluctuations, the structure of biomolecules may be altered by such thermal fluctuations. Therefore, to properly model biological systems we must include thermal fluctuations.

**1.3.1 Boltzmann statistics**

We cannot predict the exact state of a system subject to random thermal fluctuations. However, the Boltzmann distribution gives the probability of the system exploring any state

$$P_i = e^{-E_i/k_B T}/Z,$$  \hspace{1cm} (1.13)
where $P_i$ is the probability of observing state $i$, which has energy $E_i$, $k_B$ is Boltzmann’s constant, and $T$ is the temperature. This probability is normalized by the partition function $Z$, which is defined as

$$Z = \sum_i e^{-E_i/k_B T},$$

(1.14)

where the sum over $i$ includes all possible states of the system.

Knowing the probability of observing any state of the system, it is then simple to calculate the statistical average, or expectation value, of any physical observable by just calculating its average value weighted by the Boltzmann distribution. If an observable $U$ has value $U(i)$ in state $i$, then its expectation value is

$$\langle U(i) \rangle = \sum_i U(i) e^{-E_i/k_B T}/Z.$$

(1.15)

Most biological systems are far too complicated to allow us to calculate an exact partition function including all possible states at the atomic scale. Instead, we must find an appropriate simplification which captures the physical behavior of interest. For example, single-stranded RNA or DNA can form complicated three-dimensional structures due to self-interactions. In many cases, one can get a good approximation of this structure by reducing the allowed self interactions to just pairing between the bases of the nucleic acid, and by modeling the structure in only two dimensions. While much information is lost, this method allows the nucleic acid structure to be calculated very rapidly, and in many cases the structure predicted is in very good agreement with experimentally determined structures [12, 13].

### 1.3.2 Extension of a freely-jointed chain

An example of a quantity that can be derived quite readily using using statistical mechanics is the extension of a FJC polymer. This model provides an accurate description for the flexibility of ssDNA (but not dsDNA) [3, 10]. As mentioned in Section 1.2.1, the FJC models a polymer as a series of connected rigid segments of length $b$, which are completely free to rotate at the joints between segments. There is no energy cost to bend the polymer, so if a force $F$ is applied to a FJC with extension $x$, the total energy of the system will just be the work done by the force [65]

$$E_{FJC}(x) = -F \cdot x = -\sum_{i=1}^{N} F \cdot b_i,$$

(1.16)

where $b_i$ is the displacement of segment $i$, and there are $N$ total segments in the chain. The partition function is then [65]
Figure 1.9: Diagram showing the extension of a FJC polymer. The applied force $F$ is shown in red. The polymer consists of $N$ segments of length $b$, each having displacement $b_i$, shown in black. The total extension $x$, shown in blue, is just the sum of the displacements of all the individual segments.

$$Z_{FJC} = \sum_{\text{all states}} e^{-E_{FJC}(x)/k_BT} = \sum_{\text{all states}} e^{F\cdot x/k_BT}$$  \hspace{1cm} (1.17)

$$= \prod_{i=1}^{N} \left[ \int d^2b_i e \cdot b_i \right] = \left[ \frac{4\pi k_BT}{Fb} \sinh(Fb/k_BT) \right]^N, \hspace{1cm} (1.18)$$

where the sum over all states is converted into a product of integrals over all possible orientations of each individual segment $b_i$. We next apply Eq. 1.15 to find the expectation value of $x$

$$\langle x \rangle = \sum_{\text{all states}} xe^{-E_{FJC}(x)/k_BT} / Z_{FJC} = \sum_{\text{all states}} xe^{F\cdot x/k_BT} / Z_{FJC}. \hspace{1cm} (1.19)$$

By comparing Eq. 1.19 to Eq. 1.17, we see that

$$\langle x \rangle = k_BT \frac{dZ_{FJC}}{dF} = Nb \left[ \coth \left( \frac{Fb}{k_BT} \right) - \frac{k_BT}{Fb} \right]. \hspace{1cm} (1.20)$$

Then, if we note that the contour length of the chain $L = Nb$, we see that Eq. 1.20 is equivalent to the result from [3] shown in Eq. 1.6.

### 1.4 Outline

This dissertation focuses on extracting mechanistic details about the behavior of nucleic acids from biophysical data. By applying and further refining the types of modeling described above, we are able to learn minute details about several interesting biological systems.

We first examine the flexibility of DNA at short length scales [66]. There have been
several studies, such as the one described in Section 1.2.4 [9], which suggest that the WLC underestimates the flexibility of DNA on short length scales. We attempt to develop a model which provides a mechanistic explanation for this flexibility, and then design and perform biochemical experiments which test this model.

Next we discuss a model for the structure of single-stranded nucleic acids (ssNAs) in the presence of other molecules which bind to the ssNAs [67]. We will focus on proteins which bind to regions of the ssNA which are not base paired. There are several important proteins which have this behavior, including the nucleocapsid protein from HIV [68, 69] the RecA DNA repair protein, which has analogues in almost all organisms including humans [70], and vascular endothelial growth factor, which is crucial to the development of cancerous tumors [6].

Finally, we will present a method for calculating the dynamics of DNA unwrapping from a nucleosome using a free energy landscape [71]. We calibrate this model using a variety of quantitative experimental data, and show how it can be applied to several different experimental systems.
Chapter 2
THE FLEXIBILITY OF LOCALLY MELTED DNA

Reproduced in part from Forties, R. A, Bundschuh, R, & Poirier, M. G.

Beyond the information content of a given DNA molecule, its physical properties as a polymer, such as its flexibility, are important for its biological function. The formation of sharp bends is critical to many biological processes such as gene regulation [72, 73], binding of transcription factors [29, 30], and DNA packaging [24]. An accurate quantitative understanding of the formation of such bends is important.

The flexibility of a polymer is characterized by its persistence length, the length over which orientations of the polymer become decorrelated. For DNA on scales longer than several hundred base pairs (bps) this flexibility has been investigated using many methods, including cyclization [48, 74, 75], direct mechanical measurements of force versus extension [1, 10], and atomic force microscopy [7, 8]. Data from these experiments all fit the worm-like chain (WLC) model, which treats DNA as a uniform rod with a given bending rigidity [76]. The numerical values of the persistence length recovered in these experiments are in the range of 40 to 55 nm.

Recent studies suggest that the WLC may not accurately describe sharp bends in DNA [9, 77–80]. For example, cyclization experiments for the biologically important regime of sequence lengths below 200 bp indicate that sharp bends occur significantly more frequently than predicted by the WLC model [77, 78]. While the results of these studies were recently challenged [75], the large sequence dependence in the flexibility reported remains unchallenged. Furthermore, fluorescence resonance energy transfer (FRET) measurements [79, 80] and atomic force microscopy (AFM) studies [9] also indicate that on
short length scales DNA is more flexible than predicted by the WLC model. Together, these studies suggest that a modification of the WLC model appears necessary to explain sequence dependent sharp bends in DNA.

One explanation of the increased flexibility of double-stranded DNA (dsDNA) on short length scales is that DNA may transition to kinked structures, which maintain base pairing [81, 82]. This notion is supported by recent investigations of DNA minicircles [83] and with AFM measurements, which show that the DNA bending energy is not harmonic for all deflections [9, 84]. An alternative but not mutually exclusive model for the measured flexibility of short DNA molecules is the formation of melted bubbles in DNA, where a number of contiguous nucleotides (nts) are unpaired. These bubbles may form spontaneously as thermal excitations. Indeed, Yan and Marko suggest that cyclization of short DNA sequences may be explained if small bubbles are significantly more flexible than dsDNA [47, 85], since for short molecules the gain in bending energy due to a flexible hinge can outweigh the loss in free energy of breaking the DNA base-pairing. In order to judge the relevance of this mechanism for the anomalous flexibility of short DNA molecules the gain in bending energy and the loss of base-pairing free energy must be quantified. The latter has been measured experimentally in a temperature and sequence dependent manner [15, 72, 86]. Thus, to accurately predict the flexibility of the molecule with this model, measurements of the effective flexibility of melted bubbles are required.

Here, we report measurement of the flexibility of small melted bubbles. We perform cyclization experiments for three different sequences at temperatures from 23°C to 42°C, and then fit the bubble flexibilities needed such that a sequence dependent Yan-Marko model makes predictions which agree with our cyclization data. We find that small bubbles have modest persistence lengths, of about 8 and 4 nm\(^1\) for 3 and 4 bp bubbles, respectively. This is consistent with the expectation that small bubbles should be significantly more flexible than dsDNA [79, 85] and is also consistent with measurements on DNA containing mismatches of similar length [79, 87, 88]. Our results imply that melted bubbles are important to DNA flexibility, but that other effects such as kinking [81, 82] are also needed to explain the very sharp bends observed in some AFM experiments [9] and in DNA minicircles [83].

We develop a model which predicts the impact of melted bubbles on the overall flexibility of dsDNA. We then make temperature dependent cyclization measurements on a 200 bp fragment of DNA from bacteriophage \(\lambda\), which has previously been studied at room temperature [74]. This allows us to make a rough estimate of the flexibilities for small melted bubbles, which we use to design a pair of DNA fragments with nearly identical sequences which our model predicts should have significantly different flexibilities. We then perform temperature dependent cyclization measurements [48] for these sequences to test

\(^1\)We report persistence lengths at a reference temperature of 294K. The persistence length decreases with temperature, so we must scale its value when comparing with experiments at different temperatures (Appendix A).
the predictions of our model.

To further test our model, we develop a novel cyclization procedure using a thermophillic ligase which allows us to probe DNA flexibility at temperatures between 42°C to 65°C. We find that our model continues to fit cyclization measurements up to 55°C using the same bubble flexibilities determined from low temperature measurements, but fails to reproduce DNA flexibility at higher temperatures. We also develop a technique for creating DNA constructs containing single base pair mismatches suitable for cyclization experiments, to allow us to compare the flexibility of these mismatches with melted bubbles of the same size. However, we have not yet been able to create a functional DNA construct containing mismatches due to multiple difficulties with its synthesis.

2.1 Extension of the worm-like chain to include melted bubbles within DNA

We describe the flexibility of dsDNA using a worm-like chain model reduced to discrete units (discretized) following the work of Yan, Kawamura, and Marko [47]. In this model the bending free energy of a discretized polymer of \( N \) segments is

\[
E_{WLC} = \sum_{n=1}^{N-1} \frac{A}{2b} \theta_n^2 \times k_B T.
\]  (2.1)

Here \( A \) is the persistence length, \( b \) is the segment length, \( \theta_n \) is the angle between two adjacent segments of the polymer, \( k_B \) is Boltzmann’s constant and \( T \) is temperature.

We apply this model to dsDNA with a segment length of one nucleotide, or \( b = 0.34 \) nm. In our calculations, we use \( A = A_{ds} \) between 40 and 55 nm, which is within the range of dsDNA persistence lengths reported in the literature [7, 10, 48, 74–76]. We then include the formation of melted bubbles. These are structures where a stretch of bases becomes unpaired resulting in a higher flexibility than base-paired dsDNA [47, 85]. The associated melting free energy for a bubble containing \( L \) open base pairs starting at base \( \alpha \) is \( \mu_{\alpha L} = \sum_{i=\alpha}^{\alpha+L} \mu_{si} + \mu_{L}^{\text{loop}} \), where \( \mu_{si} \) is the free energy cost to break the stacking interaction between base pairs \( i \) and \( i - 1 \), and \( \mu_{L}^{\text{loop}} \) is the entropic contribution that arises from the formation of a closed loop in a flexible polymer. Values have been explicitly measured for \( \mu_{si} \) [72, 86], and \( \mu_{L}^{\text{loop}} \) for \( L \leq 2 \) [15]. For larger values of \( L \), we use \( \mu_{L}^{\text{loop}} = \mu_{L}^{\text{avg}} + \mu_{d} + \mu_{a+L} \), where \( \mu_{L}^{\text{avg}} \) is the average loop entropy of a size \( L \) bubble, and \( \mu_{d} \) is a correction to this which depends on the last stacking interaction broken at each edge of the bubble. All needed values of \( \mu_{d} \) and \( \mu_{L}^{\text{avg}} \) have been measured [15]. We model bubbles whose flexibility is isotropic about the helical axis of the DNA. The free energy for a bubble is

\[
E_{\text{Bubble}} = \mu_{\alpha L} + \sum_{n=\alpha}^{\alpha+L-1} \frac{A_L}{2b_L} \theta_n^2 \times k_B T,
\]  (2.2)
where $A_L$ is the persistence length for a segment containing $L$ open base pairs, and $b_L$ is the segment length for an unpaired nucleotide, which we set equal to 0.7 nm, the segment length of single-stranded DNA (ssDNA). The total bending free energy $E$ for a configuration is calculated using Equation 2.1 for all closed base pairs and Equation 2.2 for base pairs inside melted bubbles.

In order to be cyclized, a DNA molecule must form a closed loop whose ends meet with a parallel orientation. In principle, the DNA molecule may explore many states due to thermal fluctuations most of which do not lead to closed loops. In order to quantitatively compare with the cyclization experiments we calculate the probability of loop closure, also called the $J$ factor, as the total probability of all the DNA configurations that lead to a closed loop with parallel ends. For details of the calculation of this probability, see Appendix A and [47].

The only free variables in our model are the persistence lengths of duplex DNA $A_{ds}$, and of internal bubbles $A_L$. We restrict the persistence length of duplex DNA to the reported values of 40 to 55 nm [7, 10, 48–76]. We fit the $J$ factor at 23°C by varying the duplex DNA persistence length. Variations in the $J$ factor at this temperature are likely due to sequence-dependent intrinsic bend and twist [82, 89, 90]. Once the duplex persistence length is determined for the DNA sequence, we compare the predictions of our model with experimentally measured $J$ factors, and vary $A_L$ to simultaneously give the best agreement at 30°C, 37°C and 42°C. While we must use different $A_{ds}$ for the different sequences, we fit all three sequences at all temperatures using a single set of $A_L$. 

Figure 2.1: Synthetic 116 bp DNA sequences. The sequences are identical except for 8 base pairs, which are highlighted in red. Both sequences have Cy5 fluorophores attached to thymine residues where shown.
Figure 2.2: Rate of ligation versus concentration of T4 ligase at 37°C. The open red circles show data from the 116o sequence, and the closed blue diamonds show data from the 116cl sequence. (a) The rate of formation of circular monomer increases linearly for all concentrations of ligase. (b) At high ligase concentration, the rate of dimer formation no longer increases linearly with ligase concentration. We find that ligation is linear up to 400 units/ml of ligase at 37°C, and therefore use 100 units/ml or less for our experiments at this temperature.

2.2 Cyclization Experiments

We begin by measuring a 200 bp λ DNA sequence which has already been well studied [74, 77]. The 200 bp λ sequence is PCR amplified from λ DNA, and then inserted into the pDrive plasmid (Qiagen). In addition, we study two synthetic 116 bp sequences, shown in Figure 2.1, which we design to have significantly different J factors using our model. The 116 bp sequences are synthesized by PCR amplification of custom DNA oligonucleotides (Operon). They are then digested with HindIII restriction endonuclease (20 units/µl; NEB) and inserted into the HindIII site of the pUC19 plasmid. All plasmid sequences are verified using a 3730 DNA Analyzer (Applied Biosystems).

In all cases, plasmids are replicated in DH5α E. coli and harvested with a QIAprep Spin Miniprep kit (Qiagen). Custom oligonucleotides containing Amino Modifier C6 dT (Operon) are labeled at the modified residues with Cy5 fluorophores (Amersham) and purified via reverse phase HPLC (Vidac). These oligonucleotides are then used to PCR amplify inserted sequences from plasmids, resulting in one Cy5 label per DNA molecule. The labeled DNA is digested with HindIII leaving AGCT overlapping ends, and then purified using ion exchange HPLC (Waters). We ligate with an excess of ligase to verify that ligatability of the digested ends is greater than 95%.

Cyclization experiments are performed using T4 DNA ligase (400 units/µl; NEB) in its standard buffer, with 0.1 mg/ml bovine serum albumin (BSA) (10 mg/ml; NEB) added.
Cyclization experiments require that ligase is the limiting reactant for ligation of both circles and dimers, such that the system samples the full equilibrium distribution of conformations, and only occasionally when the system is in a conformation which may be ligated, a ligase complex will bind and ligate the DNA. Because the rates of formation of monomer circles and dimers are not the same, ligase will cease to be the limiting reagent for dimer formation at lower concentration than for monomer circle formation. Therefore, if we use too much ligase, the ratio of monomer circle to dimer formed will be higher than at equilibrium, because the rate of monomer circle formation continues to increase with ligase concentration after the rate of dimer formation becomes saturated.

Ligation is found to be linearly proportional to the concentration of ligase for ligase concentrations less that 100 U/ml at 23°C, in agreement with previous studies [75]. Figure 2.2 shows that we also find that ligation is linear for ligase concentrations up to 400 units/ml at 37°C. Ligase activity is quenched by increasing the EDTA concentration to 0.05M and samples are digested with proteinase K (Invitrogen) for 20 minutes at 65°C to inactivate ligase. We repeat experiments at several T4 ligase concentrations to ensure reproducibility. For the 116 bp sequences samples are concentrated by precipitation with linear polyacrylamide (LPA) (see Appendix E.2 and [91]). Samples are visualized on 6% polyacrylamide gels. The gels are then imaged using a Typhoon Trio imager (GE Healthcare) set to detect the Cy5 labeled DNA in order to determine the concentration of the different ligation products.

The $J$ factor is calculated from cyclization experiments using [48]

$$J = 2M_0 \lim_{t \to 0} \frac{C(t)}{D(t)},$$ (2.3)

where $M_0$ is the initial concentration of the linear monomer of our DNA molecule, $C(t)$ is the concentration of cyclized monomer, which is one molecule ligated to itself forming a closed loop, and $D(t)$ is the combined concentration of linear dimers and dimer circles, which may form directly from linear dimers. Since the only difference between the reactions that generate circular monomer and linear dimer is that for the former case the DNA must bend into a closed loop while in the latter case it does not, the $J$ factor gives a direct measure of DNA flexibility. Figure 2.3 shows an example of how the $J$ factor is calculated from our experiments.
Figure 2.3: Example gel showing a ligation time course for the 116o sequence at a concentration of 0.33 nM ligated at 37°C with 50 units/ml T4 ligase. (a) Gel image showing both linearly and circularly ligated products. The bands in the gel are (from bottom to top) linear monomer (LM), linear dimer (LD), circular monomer (CM), linear trimer (LT), circular dimer (CD) and circular trimer (CT). The leftmost lane (labeled L) shows ligation with excess ligase, where almost all DNA is converted to circular products showing that it is at least 95% ligatable. The remaining lanes show a ligation time course, and are labeled with the reaction time in minutes. (b) A plot of the ratio $2M_0C(t)/D(t)$ created from the gel shown in (a). The intercept of this plot at $t = 0$ gives the $J$ factor measured in this experiment.
2.3 Extraction of bubble flexibilities from cyclization experiments

The results of cyclization experiments plotted alongside our sequence dependent Yan-Marko model predictions for the three DNA molecules are shown in Figure 2.4. Also shown is the result of Vologodskai et al. for the 200 bp λ sequence [74], which quantitatively agrees with our measurements at 23°C. We adjusted the dsDNA persistence lengths of the 200 bp λ, 116cl and 116o molecules to 51 nm, 44 nm and 48 nm, respectively, to fit the measured $J$ factors at 23°C. These persistence length values are within the range reported in the literature [7, 10, 48, 74–76]. At 23°C, the excitations of melted base pairs are so rare that they do not contribute significantly to the $J$ factor (Figure 2.4). These variations in $J$ factors are likely to be due to other effects such as sequence dependent permanent bend, twist and anisotropic bending and twisting fluctuations [81, 82, 89, 90, 92].

The measurements of Du et al. show that 116 bp sequences should have nearly an integer number of helical repeats [75], and therefore be relaxed with respect to twist when they are cyclized. However, the difference in $J$ factor between 116o and 116cl at 23°C suggests that 116o might have a different helical repeat, and therefore might not be relaxed when cyclized. This would make interpretation of our results more difficult as we would have to consider the impact of twist on the $J$ factor. Therefore we measure the cyclization of a 116o sequence truncated to 114 bp and one extended to 118 bp. We find little variation of the
\( J \) factor with length, indicating that the 116o sequence can be no more than 2 bp from an integer helical repeat, and that we are justified to neglect the effects of twist in our analysis (Appendix C).

While the \( J \) factors for 200 bp \( \lambda \) and 116cl may be fit at all temperatures by the WLC model, as indicated by the dashed lines in Figure 2.4, the 116o molecule cannot. It has a five-fold increase in the \( J \) factor as compared to the WLC model at 37\(^\circ\)C. This increase in \( J \) factor appears too large to be explained by known temperature dependent structural changes in DNA, such as changes in the helical repeat [93]. Since 116o forms melted bubbles more readily than the other sequences, this suggests that bubbles may impact the flexibility of short dsDNA molecules at 37\(^\circ\)C.

**Bubble flexibility**

The effect of local melting on DNA flexibility depends on the frequency with which local melting occurs, which we calculate based on known parameters, and on the (unknown) flexibility of those melted regions. Thus, by using our sequence dependent Yan-Marko model to simultaneously fit our \( J \) factor measurements for all three DNA molecules with a single set of persistence lengths, we can constrain the range of persistence lengths for different size bubbles. Clearly, the bubbles cannot be more stiff than dsDNA and cannot be less stiff than two parallel ssDNA molecules, i.e., 2 nm [94]. We also assume that the persistence length cannot increase as the bubble size increases. Furthermore, we find that we need only consider bubbles containing 4 bp or less, as larger bubbles occur so infrequently due to their large free energy cost that they cannot affect our model predictions.

We first fit the temperature dependence of all three DNA molecules with a constant persistence length for all melted bubbles. The observations provide a fit value of 13 nm as a lower limit for the persistence length of 1 bp bubbles. To determine the maximum persistence length for 1 bp bubbles, we determine how stiff we can make the 1 bp bubbles by reducing the persistence length of 2 bp and larger bubbles. We find that 1 bp bubbles can have persistence length equal to dsDNA if 2 bp and larger bubbles have a persistence length of 7 nm. This provides an upper limit of dsDNA persistence length for 1 bp bubbles and a lower limit for 2 bp bubbles of 7 nm. We then fit the data where 1 bp and 2 bp bubbles are as stiff as dsDNA and find that 3 bp and 4 bp bubbles must have a persistence length of 6 nm. This implies that 2 bp bubbles have a maximum persistence length equal to that of dsDNA and that 3 bp bubbles have a minimum persistence length of 6 nm. We finally attempt to fit the cyclization data assuming 1 bp, 2 bp and 3 bp bubbles are as stiff as dsDNA. However, we find no fit is possible even with a persistence length for 4 bp bubbles equal to that of 2 parallel single strand DNA molecules, 2 nm, unless we also include 3 bp bubbles with a persistence length of at most 11 nm.

The ranges of persistence lengths for different bubble sizes are summarized in Table 2.1. Plots of all fits compared to our experimental measurements are shown as supplemental
Table 2.1: Persistence lengths used to fit our data, given in nm at a reference temperature of 294 K. An entry of “DS” indicates a persistence length equal to that of dsDNA, which is 44, 48, and 51 nm for 116cl, 116o and the 200 bp λ fragment, respectively. The five “Fit” columns show values used in different fits to our data. The column labeled “Range” summarizes the minimum and maximum persistence lengths that may be used to fit our data for each bubble size, found by inspecting Fits 1-4.

<table>
<thead>
<tr>
<th>Bubble size</th>
<th>Fit 1</th>
<th>Fit 2</th>
<th>Fit 3</th>
<th>Fit 4</th>
<th>Fit 5</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 bp</td>
<td>13</td>
<td>DS</td>
<td>DS</td>
<td>DS</td>
<td>32</td>
<td>13 to DS</td>
</tr>
<tr>
<td>2 bp</td>
<td>13</td>
<td>7</td>
<td>DS</td>
<td>DS</td>
<td>16</td>
<td>7 to DS</td>
</tr>
<tr>
<td>3 bp</td>
<td>13</td>
<td>7</td>
<td>6</td>
<td>11</td>
<td>8</td>
<td>6 to 13</td>
</tr>
<tr>
<td>4 bp</td>
<td>13</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>2 to 13</td>
</tr>
</tbody>
</table>

data. We view Fits 1-4 as unlikely because we expect a smooth decrease in persistence length as the bubble size increases, which converges to the single strand persistence length limit at 4 to 5 bp. To obtain a more realistic fit we first assume that 3 bp bubbles have a persistence length of 8 nm, which is that measured for 3 bp mismatches [88]. We then arbitrarily choose a geometric scaling to provide a simple, smooth dependence. A scale factor of 2 results in the parameters shown in Table 2.1, Fit 5, which give a good fit to our data as shown in Figure 2.4. While there is a very large possible range of persistence lengths for 1 and 2 bp bubbles which fit our data, we view the values given by Fit 5 as the most physically reasonable approximate values for all persistence lengths.

Melted bubbles larger than 4 bp do not occur often enough to contribute to the cyclization of DNA, but we may still infer their flexibility indirectly from our experiments. We find that 4 bp bubbles display persistence length not much higher than that of two ssDNA molecules. Bubbles larger than 4 bp can neither display persistence length larger than the persistence length of 4 bp bubbles nor be less stiff than two single-stranded DNA molecules. Therefore, we conclude that bubbles larger than 4 base pairs have essentially the same persistence length as that of two ssDNA molecules.

### 2.4 Cyclization at high temperatures

Given that these melted bubbles form due to thermal fluctuations, they will naturally form more readily at higher temperatures. Therefore, if we perform cyclization experiments at higher temperatures, we expect larger deviations from the worm-like chain (WLC) due to the formation of melted bubbles, allowing us to more clearly determine the impact of melted bubbles on DNA flexibility, and more rigorously test our model.
Figure 2.5: Rate of ligation versus concentration of ligase for high temperature ligases acting on the 116To sequence. The experimentally measured rates of formation are show by red squares for dimers and blue diamonds for monomer circles. In all cases, the rate of formation of monomer circles is linear with ligase concentration at all concentrations measured, but the rate of formation of dimers saturates a high concentrations of ligase. (a) Rates of ligation using Taq ligase at 42°C. The rate of formation of dimers is not linear at any concentration of Taq ligase where we were able to measure ligation. Therefore, we cannot use this enzyme for cyclization experiments at this temperature. (b) Rates of ligation using Taq ligase at 65°C. The formation of dimers is linear up to 400 units/ml Taq ligase. (c) Rates of ligation using 9N ligase at 42°C. The formation of dimers is linear up to 40 units/ml 9N ligase. (d) Rates of ligation using 9N ligase at 65°C. The formation of dimers is linear at all concentrations measured, up to 80 units/ml 9N ligase.
Unfortunately, T4 ligase does not function above 42°C, so we must use a different ligase. We considered two high temperature ligases, Taq ligase (40 units/µl; NEB) and 9°N ligase (40 units/µl; NEB). To our knowledge, this is the first attempt to use such high temperature ligases for cyclization experiments. We find that these ligases do not function with ligatable ends having the 4 bp overhang produced by *Hind*III, but instead require an overhang of at least 9 bp, such as the one produced by TspRI (see Appendix D).

We first attempted to use Taq ligase (40 units/µl; NEB), but found that ligation was not linear with ligase concentration at 42°C at any concentration of Taq ligase we were able to achieve, as shown in Fig. 2.5a. We believe that at this temperature Taq ligase binds tightly to the ligatable ends of the DNA, such that even if we reduce its concentration it still remains bound and will ligate the DNA as soon as two ends come together. As described in Section 2.2, this does not allow us to measure the equilibrium distribution of ligation products, and therefore is unsuitable for cyclization experiments. We do find that Taq ligase is linear at 65°C, but we wish to perform cyclization experiments at temperatures at least as low as 42°C with our high temperature ligase so that we can confirm that it gives the same result as T4 ligase at these temperatures.

We therefore also explored the use of 9°N ligase. Figs. 2.5c and 2.5d show that ligation is linear at low concentrations of this ligase at both 42°C and 65°C. We conclude that this ligase is suitable for cyclization experiments across this temperature range at a concentration of 20 units/ml, and use this concentration for our experiments.

Cyclization experiments for the 116Tcl and 116To sequences were performed following the procedure outline in Section 2.2, except that 20 units/ml 9°N was used for ligation, DNA samples were at a concentration of 1 nM, and 0.01% Sodium dodecyl sulfate (SDS) was added to the samples before digestion with proteinase K for 20 minutes at 65°C to inactivate the ligase. The results of these measurements are shown in Fig. 2.6.

Measurements made with 116Tcl and 116To appear to agree with our model only at temperatures up to 50°C. One possible explanation for this deviation is that the two strands of a DNA may peel apart starting at the free ends at high temperature, and this unpeeling would likely block ligase activity. Such unpeeling would be even more favorable if a melted bubble was already present near the end of the DNA, and therefore could selectively block cyclization of constructs containing melted bubbles.

We therefore perform identical cyclization experiments with the 116Fcl and 116Fo sequences which have increased stability near the ends to discourage unpeeling (see Appendix D). These measurements are in slightly better agreement with our modeling at 55-60°C, but this difference from 116Tcl and 116To is within our experimental errors, so we cannot confirm that unpeeling is responsible for the differences between our model and experiments at high temperature.

It is also not clear whether there is truly agreement with our model between 42°C and
Figure 2.6: Results from cyclization experiments at high temperatures compared with model predictions and our low temperature data. In both plots, the predictions of the WLC model are shown by the dashed black line, and the predictions of the sequence dependent Yan-Marko model for Fit 5 from Table 2.1 are shown by a solid colored line. (a) Experimentally measured cyclization J factors for the 116Tcl sequence (blue filled squares) and the 116Fcl sequence (open blue circles), along with our low temperature results for 116cl (open blue diamonds) from Fig. 2.4b. (b) Experimentally measured cyclization J factors for the 116To sequence (red filled diamonds) and the 116Fo sequence (open red squares), along with our low temperature results for 116o (open red circles) from Fig. 2.4c. The J factors measured for 116Tcl and 116To appear to be in agreement with the Yan-Marko model up to 50°C, but then cease to have any temperature dependence. The J factors measured for 116Fcl and 116Fo agree with our modeling up to about 55°C, but then also level off.
55°C, as one could also argue that there might be an offset between measurements with T4 ligase and 9°N ligase. These enzymes could have different requirements for how straight the DNA must be at the site of ligation. If 9°N ligase accepts a substrate where the DNA is more bent at the ligation site than allowed by T4 ligase, this would mean that less bending would be required throughout the rest of the sequence, decreasing the energy cost of forming a ligatable circle, and increasing the rate of cyclization. Therefore, for equal comparison all of our J factors measured with 9°N ligase would need to be shifted downward, meaning that the entire temperature dependence between 23°C to 55°C could be fit about as well by the WLC model as by the Yan-Marko model.

If we make the assumption that T4 ligase and 9°N ligase do in fact require the same boundary conditions for ligation, then our measurements show good agreement with the Yan-Marko model over the temperature range from 23°C to 55°C. Our model’s prediction of temperature dependence in excess of the predictions of the WLC are also in agreement with measurements of topoisomer distributions in long circular DNAs which show a decrease in DNA persistence length at high temperature [95].

2.5 Cyclization of DNA containing single base pair mismatches

Our fit values for the persistence length of 3 bp melted bubbles agree with the flexibilities of 3 consecutive mismatches [88], suggesting that mismatches might behave like permanent melted bubbles. It would therefore be very interesting to directly measure the flexibilities of single base pair mismatches to compare with our fit values for the persistence length of single base pair bubbles. Furthermore, measurements of the binding of single base pair mismatches by the repair complex hMSH2-hMSH6 show a strong dependence on DNA sequence [96]. It has been proposed that hMSH2-hMSH6 recognizes mismatches based on their flexibility [23, 97, 98]. We could test this mechanism by verifying if the flexibility of mismatches within different DNA sequences correlates with their recognition by hMSH2-hMSH6.

We therefore attempt to create sequences containing single base pair mismatches suitable for cyclization experiments. By measuring cyclization J factors for such sequences and comparing with identical sequences that do not contain mismatches, we can determine the effective flexibility of the mismatches using our model. We place two single base pair mismatches within a 137 bp DNA sequence such that they will be opposite one another when the DNA is bent into a circle. If mismatches are in fact highly flexible, this should relax the majority of the bending strain in the DNA and promote rapid cyclization.

We also can determine if the mismatches can bend isotropically, like the melted bubbles described in our modeling, or if they instead form directional kinks. To do this, we vary
the position of the mismatches with respect to one another. If mismatches form directional bends, the cyclization J factor should be cyclic with respect to the distance between them, with a period equal to the DNA helical repeat (10.5 bp [75]). The J factor will be maximized when the mismatches are in phase and both bend in the same direction with respect to the helical axis, and significantly smaller when they are out of phase.

We have attempted several different methods to synthesize DNA substrates containing mismatches, but none of these so far has produced constructs which could then be readily ligated by T4 ligase. Cyclization experiments require a DNA substrate in which at least 95% of the DNA ends are ligatable [74, 75]. If this is not true, then there will be a significant number of DNA molecules which have only one ligatable end. Since these molecules can form linear dimers by ligating to another molecule with at least one ligatable end, but cannot be cyclized, this will cause a decrease the measured J factor which cannot be accurately corrected. All of the DNA constructs we have produced have significantly less than 95% ligatable ends, and are therefore not usable for cyclization experiments.

**Single-stranded plasmids**

We first attempted to create mismatches by annealing large single-stranded circular DNA fragments with shorter double-stranded fragments. We prepared two DNA sequences by polymerase chain reaction (PCR) which are identical except for two locations with single base pair differences, which will eventually become the mismatches. We cloned these sequences into the pDrive plasmid (Qiagen), and used M13KO7 bacteriophage (NEB) to create a single-stranded version of one of these sequences (see Appendix E.3) which we refer to as 137A because it contains adenines at the positions where we will create mismatches. The kanamycin resistance gene in the pDrive plasmid interferes with this protocol, so we deactivated it using site-directed mutagenesis. We then PCR amplified the desired portion of the other plasmid, called 137G as it contains guanines at the mismatch positions, using Cy5 labeled oligonucleotides to create a labeled dsDNA as described in Section 2.2.

We then annealed this dsDNA to the single-stranded plasmid by first denaturing it in 200 mM NaOH for 5 minutes, and then neutralizing the NaOH with Tris-HCl and allowing the DNA to reanneal. This creates G/T mismatches in two places where a guanine from the top strand of 137G is mispaired with a thymine from the bottom strand of 137A. We then digested the resulting DNA with *Hin*<sup>d</sup>III to form AGCT ligatable overhangs and cut off the excess ssDNA. Excess ssDNA was then extracted from solution using BND cellulose (see Appendix E.4).

We designed the labeled 137G sequence to contain a ClaI digest site overlapping with one of the mismatches. When 137G is annealed with 137A, a mismatch will be created within the ClaI site, suppressing digestion. We therefore digest the DNA with ClaI after annealing to cut any of the remaining 137G which does not contain mismatches. We then purified
this DNA by ion exchange HPLC (Waters). Unfortunately, the resulting DNA constructs had <50% ligatable ends, which is not nearly sufficient for cyclization experiments. It is not clear why the ligatability was so poor, but we suspect that either excess ssDNA or some of the reagents used in synthesis degraded the ligatable ends or were interfering with ligation.

\textbf{λ exonuclease}

We next tried another similar procedure, where the single-stranded DNA fragment was created by digesting a dsDNA with \textit{λ} exonuclease instead of using M13KO7 bacteriophage. We generated double-stranded versions of both 137G and 137A by PCR, making sure to kinase the primers on the strands of each which will \textit{not} be present in the final mismatched DNA construct using T4 polynucleotide kinase (10 units/µl; NEB). We then used \textit{λ} exonuclease (5 units/µl; NEB) to digest away the strands of each DNA which contain 5’ phosphates added by kinasing, leaving only the strands which we wish to be present in the final mismatched construct. We annealed the resulting ssDNAs by heating to 95°C for 5 minutes in 1X \textit{λ} exonuclease buffer and then slowly reducing their temperature to 25°C over 70 minutes. We digested the resulting construct with ClaI, and found that less than 40% of the sample could be digested, suggesting that 60% of the sample was mismatched DNA.

We then digested this sample with both \textit{Hind}III and ClaI, and HPLC purified it as described above. This resulted in a sample which was nearly 99.8% resistant to ClaI digestion, but which was only 73% ligatable, and therefore still was not suitable for cyclization experiments.

\textbf{Annealing dsDNA}

Given that our previous two approaches resulted in DNA constructs which contained mismatches, but were not ligatable, we decided to next attempt a much simpler approach which does not use any reagents not present in our procedure to create ligatable DNA constructs without mismatches. We create double-stranded versions of both 137G and 137A using PCR as described above. We then anneal these two sequences in a buffer containing 5 mM Tris pH 8.0, 0.5 mM EDTA and 50 mM NaCl by heating them to 95°C for 5 minutes and then reducing the temperature to 25°C over 210 minutes. Unfortunately, ClaI digestion showed that this only produced mismatches with only a few percent efficiency. We attempted to purify away the non-mismatched DNA via HPLC, but were unable to recover a significant amount of purified mismatched DNA.

To improve the efficiency of mismatch annealing, we increase the length of the DNA created by PCR from 337 bp to 420 bp by moving one of our primers. This extra DNA will be removed when we digest with \textit{Hind}III to create ligatable ends, so our final construct will remain the same. This will create large region of fully paired DNA at one end of the
mismatched construct. While this will not change the equilibrium between formation of mismatched DNA compared to non-mismatched DNA, it will increase the rate at which mismatched constructs can partially anneal at high temperature. Then, when we reduce the temperature, the DNA can be trapped in the mismatched state, allowing us to create a larger ratio of mismatched constructs. We anneal these DNAs in New England Biolabs restriction enzyme Buffer 3 by heating them to 95°C for 7.5 minutes and then reducing the temperature to 24°C over 2 hours, following the procedure from [99]. These constructs are approximately 40% resistant to ClaI digestion, suggesting that about this fraction contains mismatches.

Furthermore, we increase length of the 137A dsDNA by 3 bp on one end, adding in the sequence CTG. When this DNA is PCR amplified using Taq polymerase, the polymerase will add an additional adenine at the 3’ end. If the bottom strand of this 137A construct is then annealed to 137G, the result will have a 3’ overhang of CAGA. Following [99], we create an “anchor” segment of DNA which is 900 bp long, and contains a BstXI site which will leave the complementary overhang TCTG. We may then ligate our mismatched construct to this anchor DNA, and then easily purify it away from non-mismatched dsDNA via ion exchange HPLC (Waters) due to the large size difference between the combined mismatch plus anchor construct from the original PCR products. This project is ongoing, but we are confident that this approach will yield ligatable mismatched DNA.

2.6 Discussion

We provide a model which can accurately predict the impact of melted bubbles on DNA flexibility. To demonstrate the viability of this model, we perform cyclization measurements for nearly identical sequences, which our model predicts should have significantly different $J$ factors. The results of these experiments are in agreement with the predictions of our model. The worm-like chain model also correctly predicts $J$ factors for the 200 bp $\lambda$ and 116cl sequences, but not the 116o sequence, which we predict should readily form melted bubbles. Our model simultaneously predicts the behavior of all three sequences with a single set of persistence lengths for melted bubbles, indicating that the formation of melted bubbles may in fact explain the differences we measure between the 116o and 116cl sequences.

By comparing our model to experimentally determined $J$ factors at different temperatures, we constrain the persistence lengths of 3 bp and 4 bp bubbles to the ranges of 6 to 13 nm and 2 to 13 nm, respectively. We estimate that 1, 2, 3, and 4 bp bubbles have persistence lengths of roughly 32, 16, 8, and 4 nm. We also find that bubbles larger than 4 bp have a persistence lengths equal to that of two ssDNA molecules, or about 2 nm [94].

Our fit values for the persistence length of 3 bp melted bubbles are in very good agreement with measurements of the flexibility of 3 consecutive mismatches [88]. This implies
that mismatches may have similar physical properties to thermally excited melted bubbles, and that our values for persistence lengths of melted bubbles may be good estimates of mismatch flexibilities [96]. We have investigated several methods for creating constructs which will allow direct measurement of the flexibility of single base pair mismatches for comparison with our modeling, but as yet have not been able to prepare such a construct.

The bubble persistence lengths we find also suggest that the main impact of melted bubbles on DNA bending comes from those in the 2-4 bp range, and especially 3 bp bubbles. Larger bubbles occur too infrequently to make a significant impact on bending, while the large persistence length of 1 bp bubbles causes them to contribute very little to overall flexibility. Therefore a model such as ours with isotropic bending cannot explain experiments which measure large flexibilities for 1 bp mismatches [79, 87]. This would instead require a model including kinks or directional bends [81, 82].

Our results demonstrate that sequence dependent dsDNA melting can increase the flexibility of dsDNA such that the $J$ factor is increased to five fold more than predicted by the WLC model at 37°C. We show that such a large increase in $J$ factor cannot be due to a decrease in twist rigidity with temperature, reinforcing our assertion that this effect is due to the formation of melted bubbles. However, this increase in the $J$ factor is of the same order of magnitude as differences we observe between sequences at lower temperatures, where DNA melting does not contribute to increasing the $J$ factor. Moreover, we verified that DNA melting alone using the bubble flexibilities determined in our experiments cannot explain the sharp bends recently observed in AFM experiments [9]. We conclude that an accurate model of DNA flexibility at physiological temperatures must include both melted bubbles and other effects such as sequence dependent bends and twists [81, 82, 89, 90, 92].

We additionally develop a technique for performing cyclization experiments at high temperatures using the thermophillic 9°N ligase. While our results from these measurements are not conclusive because the measurements fail to follow the predictions of the Yan-Marko model above 55°C, they do lend further support to this model, and help to confirm that the bubble flexibilities we measure are relevant over a wide range of temperatures.
Chapter 3

MODELING THE INTERPLAY OF SINGLE-STRANDED BINDING PROTEINS AND NUCLEIC ACID SECONDARY STRUCTURE


Nucleic acids have been recognized to play a major role in many biological functions [100, 101]. Frequently they do not act alone but rather in conjunction with proteins. One important class of such proteins associating with nucleic acids are proteins that bind to single-stranded nucleic acids. E.g., the nucleocapsid protein from HIV binds and helps package its RNA genome [68, 69] and is involved in strand transfer during the initial steps of HIV’s replication cycle as well. Other examples include the RecA protein which binds single-stranded DNA and aids in strand exchange in bacteria [70] and the deactivation of vascular endothelial growth factor through binding with a small DNA aptamer [6].

An important aspect for the biological functions of nucleic acid molecules is the secondary structure of these molecules, i.e., the base pairings formed between their monomers. For an individual nucleic acid molecule by itself the secondary structure can be predicted from its specific sequence relatively reliably [12, 16]. However, the presence of proteins that directly interact with a nucleic acid molecule will change this secondary structure. In order to understand the biological functions of a nucleic acid molecule and its interactions with nucleic acid binding proteins it is thus important to fully understand the interplay
between the nucleic acid’s secondary structure and the proteins. Here, we develop a quantitative model that predicts secondary structures in the presence of proteins that bind to the single-stranded segments of a nucleic acid molecule, as well as the probability that protein is bound at any given position within the nucleic acid sequence.

Our model is implemented as an extension of the Vienna RNA package. The Vienna RNA package can predict the distribution of secondary structures for RNA in $O(n^3)$ time, if $n$ is the number of bases in the molecule. Through the use of a different energy parameter set the Vienna RNA package can also compute the secondary structure of single-stranded DNA [13, 15]. When a protein binds to a single-stranded nucleic acid, there is a free energy gain due to the favorable interaction between the nucleic acid and the protein. However, there is also a free energy cost because the region of the nucleic acid bound by the protein can no longer form base pairings. We incorporate the competition between these two interactions into the Vienna RNA package to determine the impact of the binding protein on nucleic acid secondary structure. Using this model, we can not only predict the secondary structure of nucleic acids in the presence of binding proteins, but also can compare with many experimental observables. These include end-to-end distance and melting of the nucleic acid, binding of the protein to the nucleic acid, and fluorescence resonance energy transfer (FRET) between dyes attached to the nucleic acid.

The binding of any single-stranded binding protein may be characterized by its affinity for the nucleic acid and its footprint, the number of nucleotides bound by one protein. For some proteins, the binding affinity is dependent on the nucleotide sequence, while for others sequence dependence may not be important. In some cases, proteins may bind cooperatively, or may form dimers. We therefore make our model as general as possible, such that it may accommodate any type of protein binding to single-stranded nucleic acids. By fitting experimental data with our model we can gain insight into the physical parameters of a specific binding protein, such as binding affinity, footprint, and cooperativity, that give rise to the experimental observations and are ultimately responsible for the biological consequences of the interaction between the protein and the nucleic acid. Once the interaction parameters have been determined by comparison with specific experiments, our model allows the prediction of the effect of the protein binding on the secondary structure formation and vice versa for arbitrary sequences of the nucleic acid and arbitrary concentrations of the protein.

The remainder of the paper is organized as follows. We first explain the methods we use to calculate nucleic acid secondary structures in the presence of binding proteins. Then we give an overview of some physical observables which may be calculated using the model presented here. Finally, we discuss the significance of such modeling to the study of nucleic acids in the presence of binding proteins.
3.1 Calculation of the partition function for nucleic acid secondary structure in the presence of single-stranded binding proteins

A single-stranded nucleic acid may form secondary structures by looping back on itself, allowing pairing between complementary bases. The probability of any such secondary structure forming can be predicted by calculating the partition function for a nucleic acid’s secondary structure. In general, the partition function is

\[ Z = \sum_s e^{-E_s/k_BT}, \]  

(3.1)

where the sum is taken over all possible states \( s \) of the system, \( E_s \) is the free energy of each state, \( k_B \) is Boltzmann’s constant and \( T \) is the temperature. In this formulation, the probability of state \( s \) is given by \( e^{-E_s/k_BT}/Z \). In our model, the possible states are all different nucleic acid secondary structures, and the free energy for a secondary structure is given by the sum of the free energies for all loops and base pairings formed in that structure according to the well established Turner energy model [14, 15, 17].

The partition function for nucleic acid secondary structure may be calculated efficiently by using recursive algorithms [12]. Several quantities related to the partition function are computed for successively larger segments of the nucleic acid until the entire partition function for the whole length of the nucleic acid is built up. We incorporate binding of single-stranded binding proteins in such a recursive framework. Here, we will illustrate our approach for a simplified model of nucleic acid secondary structure, in which the free energy to form a base pairing between nucleotides \( i \) and \( j \), \( E(i, j) \), only depends on the identities of \( i \) and \( j \), and not on the secondary structure formed by the rest of the nucleic acid. Our actual calculation is a modified version of the Vienna RNA package [13], which implements the full Turner energy model [14, 15, 17] in which the free energies are associated with base pair stacking and depend on the secondary structure of the region in which they form. However, since the method to add in binding proteins is essentially the same in the simplified model and in the full model, we explain it here in the more easily accessible case of the simplified model and give the full details for the Turner energy model in Appendix F.

We represent the partition function for a nucleic acid segment \( (i, j) \) with no base pairings and no proteins bound as \( O_{i,j} \). Since we define all free energies with respect to unpaired nucleic acids, \( O_{i,j} = 1 \) for any \( i \) and \( j \). Including \( O_{i,j} \) in our formulation allows us to explicitly keep track of unpaired regions, which will be useful later when we allow proteins to bind to these unpaired regions. We also represent the partition function for a segment \( (i, j) \) where \( i \) is paired with an (unspecified) base \( k \), \( k + 1 \) through \( j \) are unpaired, and the state of \( (i + 1, k - 1) \) is unconstrained as \( W_{i,j} \).

This allows us to calculate the unconstrained partition function \( Q_{i,j} \) for a segment \( (i, j) \)
starting at base $i$ and ending at base $j$ by the recursion relation

$$Q_{i,j} = O_{i,j} + W_{i,j} + \sum_{k=i}^{j-L_{\text{min}}-2} Q_{i,k}W_{k+1,j},$$  \hspace{1cm} (3.2)$$

where $L_{\text{min}}$ is the minimum number of nucleotides for the nucleic acid to bend back on itself and form a hairpin loop. No terms $Q_{i,k}$ where $k < i$ are included as the sum over $k$ is restricted to $k \geq i$. Therefore there is no ambiguity in the definition of $Q_{i,j}$. The restricted partition function $W_{i,j}$ is then given by

$$W_{i,j} = \sum_{k=i+L_{\text{min}}+1}^{j} Q_{i+1,k-1}e^{-E(i,k)/k_BT}O_{k+1,j}.$$  \hspace{1cm} (3.3)$$

Note that we use the convention $O_{i,j} = 1$ even for $i > j$, which corresponds to a substrand of negative length. Therefore there is no issue with the sum in Eq. 3.3 including the term $O_{j+1,j}$ when $k = j$. We show an arch diagram representation of the decomposition of the partition function given by Eqs. 3.2 and 3.3 in Fig. 3.1.

Using these recursion relations, we are able to calculate $W_{i,j}$ and $Q_{i,j}$ for any segment. We begin by setting $Q_{i,j} = 1$ and $W_{i,j} = 0$ for all segments where $j - i < L_{\text{min}} + 1$. These segments are too short to allow any base pairing to occur, so their unconstrained partition function $Q_{i,j}$ is equal to $O_{i,j} = 1$. We next calculate $W_{i,j}$ and then $Q_{i,j}$ for the segments with length $j - i = L_{\text{min}} + 1$ using Eqs. 3.3 and 3.2, respectively. We then continue to work our way up to larger and larger segments using these recursions. Finally, we obtain the total partition function for the whole sequence by calculating $Q_{1,n}$, where $n$ is the number of nucleotides in the sequence.

To include the impact of binding proteins, we make use of a chemical potential for
proteins binding to unpaired regions of the nucleic acid. A chemical potential gives the energy for a system to accept one particle from its environment. In this case, each protein from the surrounding solution which binds to the nucleic acid lowers the free energy of the system by the chemical potential $\mu^P(i)$, where $i$ indicates the position of the first nucleotide the proteins binds to. This helps to stabilize secondary structures of the nucleic acid which have unpaired regions where proteins may bind. The protein may bind more tightly to some nucleotide sequences than others, so the chemical potential can vary with position within the nucleotide sequence $i$. Additionally, as protein concentration is increased, it becomes more favorable for proteins to bind to the nucleic acid. For dilute solutions the chemical potential depends on protein concentration $c$ following $\mu^P(c,i) = \mu^P(c_0,i) + k_B T \ln (c/c_0)$, where $c_0$ is an arbitrary reference concentration [102].

As we now allow proteins to bind to unpaired regions of nucleic acid, we no longer represent these regions using $O_{i,j} = 1$. We instead use the sum over the contributions from all possible combinations of proteins which may bind to the nucleic acid segment

$$O_{i,j} = \sum_k e^{-\mu^P(c,k)/k_B T} + \sum_{k,l} e^{-[\mu^P(c,k)+\mu^P(c,l)]/k_B T} + \ldots,$$

(3.4)

where the sums over $k$ and $l$ must be restricted to only allow combinations of binding proteins which fit on the segment without overlapping. The sum over $k$ in the first term is taken over all positions in which a single protein may bind. The sums over $k$ and $l$ in the second term are taken over all combinations of two proteins which may simultaneously bind to the segment. Larger terms, which are not shown, would need to be included for each number of proteins which can simultaneously bind the segment. While cooperativity could easily be included in this formulation, the proteins are assumed here to bind to the nucleic acid independently, so the chemical potential for multiple proteins binding is just the sum of their individual chemical potentials.

This quantity can again be calculated recursively by

$$O_{i,j} = O_{i,j-1} + e^{-\mu^P(c,j-F+1)/k_B T} O_{i,j-F},$$

(3.5)

where $F$ is the footprint, or number of nucleotides bound by one protein. For $j - i < F - 1$, we still use $O_{i,j} = 1$, as such sequences are not long enough to accommodate even one protein binding. For longer sequences, the above recursion allows for all possible combinations of proteins which may bind to the sequence. Since $\mu^P(c,i)$ depends on both sequence and protein concentration, $O_{i,j}$ and thus $W_{i,j}$ and $Q_{i,j}$ do as well.

As indicated earlier, this yields the partition function for a simplified model of nucleic acid secondary structure, which we use to illustrate the method by which we model the impact of binding proteins on secondary structure. Our full model is identical to the calculation here except that more details of different types of base pairings are included when
calculating $W_{i,j}$, as required by the Turner energy model [14, 15, 17]. The full calculation of $W_{i,j}$ can be found in Appendix F. Source code for our modified version of the Vienna RNA package is freely available at http://bioserv.mps.ohio-state.edu/Vienna+P, implemented in C and running on Linux.

3.2 Observables

To compare this model with experiments we must be able to extract quantities that can be observed in experiments. By calculating restricted partition functions, we may predict melting probabilities, end-to-end distances, FRET values for fluorophores attached to the nucleic acids, and the probability that a protein is bound at any position within the nucleic acid sequence as a function of protein concentration. The dependence of the observables on protein concentration depends on the interaction parameters, i.e. the footprint size $F$ and the binding free energies $\mu^P(c_0,i)$. Thus, these parameters can be extracted from comparing model predictions with actual experiments.

3.2.1 Melting probability

Nucleic acid melting refers to breaking of base pairings between nucleotides. The probability that a given base is paired may be calculated directly using the partition function method outlined above, generalizing McCaskill’s approach [12] to our extended model.

We wish to calculate $Q_{\text{pair}}(p,q)$, the partition function for the full sequence with the constraint that base $p$ is paired to base $q$. In order to do this, we must first calculate the partition function for a sequence twice the length of the original. This will allow us to compute all possible base pairings between segments at the ends of the sequence while excluding those in the middle. We construct this sequence by setting bases $(1,n)$ equal to the original sequence, and then repeating this sequence again for bases $(n+1,2n)$. We next calculate the partition function for this doubled sequence, $Q_{i,j}'$, using Eq. 3.2. We may then calculate the constrained partition function using

$$Q_{\text{pair}}(p,q) = Q_{p+1,n-1}e^{-E(p,q)/k_BT}Q_{q+1,n+(p-1)}'$$

(3.6)

Since when we calculated $Q_{i,j}'$ we duplicated the sequence, $Q_{q+1,n+(p-1)}'$ actually gives the partition function for the union of segments $(1,p-1)$ and $(q+1,n)$.

Using this constrained partition function, we may then calculate the probability that base $p$ is paired with $q$ via

$$P_{\text{pair}}(p,q) = Q_{\text{pair}}(p,q)/Q_{1,n}.$$  

(3.7)

By calculating this pairing matrix as a function of protein concentration, we may then
examine the impact of binding proteins on DNA melting. (Note that for pedagogical reasons we ignore the fact that the minimum hairpin constraint $L_{\text{min}}$ should not apply to base pairs which bracket the connection between the two copies of the molecule. Again, our code contains the full calculation of the pairing probabilities.)

### 3.2.2 End-to-end distance

The end-to-end distance for a state of a nucleic acid is proportional to the number of exterior bases present, or the equivalent length of single-stranded nucleic acid which may be used to represent the state. Exterior bases are defined here as the number of unpaired bases which are not part of any loop, plus a factor $\Delta$ (the average width of a stem divided by the width of one nucleotide) times the number of stems not contained in any loop. To calculate end-to-end distance and FRET (see below), we need to first find the restricted partition function for the segment $(1, j)$ having exactly $m$ exterior bases.

$$Q_{r1,j}(m) = \sum_{k=L_{\text{min}}+3}^{j} W_{r1,k-1}(m - (j - k + 1))O_{k,j} + W_{r1,j}(m) + O_{1,j}\delta_{j,m},$$

(3.8)

where $\delta_{j,m} = 1$ for $j = m$ and zero otherwise. Therefore the last term in Eq. 3.8 is simply an unpaired segment of length $m$. $W_{r1,j}(m)$ is the restricted partition function for $(1, j)$ with $m$ exterior bases and the additional restriction that $j$ is paired to an unspecified base in $(1, j - 1)$. It is calculated from

$$W_{r1,j}(m) = \sum_{k=m-\Delta+1}^{j-L_{\text{min}}-1} Q_{r1,k-1}(m - \Delta)e^{-E(k,j)/k_BT}Q_{k+1,j-1}.$$

(3.9)

The probability that a nucleic acid of length $n$ nucleotides has exactly $m$ exterior bases is then simply

$$P(m) = Q_{r1,n}(m)/Q_{1,n}.$$  

(3.10)

If we then make use of a polymer model which predicts that an unpaired single-stranded nucleic acid of length $m$ nucleotides has a probability $P_p(x|m)$ of having end-to-end distance $x$, the predicted end-to-end distance distribution for a nucleic acid including base pairing and protein interactions is

$$P(x) = \sum_{m=\Delta}^{n} P(m)P_p(x|m).$$

(3.11)

One such polymer model is the worm-like chain model [104]. The end-to-end distance function for this model $P_p(x|m)$ is given in Thirumalai and Ha [105], and depends on the contour length of single-stranded DNA and its persistence length, the length over which orientations of the polymer become decorrelated. This model provides a good fit to experimental measurements of single-stranded DNA flexibility with a persistence length of 2.5 nm [106],
Figure 3.2: Predicted FRET distribution for an example sequence. (a) Our example sequence shown in its most likely conformation in the absence of any binding proteins, at 25°C. It forms two paired stems, with the stem near the ends of the sequence being less stable than the stem near the middle of the sequence. Cy5 and Cy3 fluorophores are attached at the 5’ and 3’ ends of the DNA, respectively. (b) The probability distribution as a function of FRET efficiency and protein concentration for the single-stranded DNA shown in (a). Darker shades of gray correspond to higher probability values, as indicated in the legend. In this example, the binding protein has a footprint of 6 nucleotides, and binds to the DNA with chemical potential of 10 kcal/mol at 1 M concentration. FRET values greater than 0.7 correspond primarily to states where all possible base pairings shown in (b) are formed. FRET values between 0.4 and 0.7 mainly correspond to the stem closest to the ends of the DNA being melted, while the stem at the middle is still intact. FRET values below 0.4 mainly correspond to DNA with no base pairings formed at all. (c) A line-out from (b) along the dashed line showing the probability distribution as a function of FRET efficiency for a protein concentration of 500 nM. At this protein concentration the states with both stems intact, one stem melted, and both stems melted all occur with modest probability.

and a contour length of 0.63 nm per nucleotide [107, 108].

3.2.3 FRET distribution

FRET experiments are a convenient way to experimentally measure the end-to-end distance of biological molecules [55, 56]. To this end, two different fluorophores are attached to the ends of a nucleic acid molecule. The donor fluorophore is excited by a laser tuned to its absorption maximum, and then may either decay or transmit its excitation to the acceptor fluorophore via FRET. The emission of the system is measured at the characteristic wavelengths of both the donor and the acceptor. Emission by the acceptor corresponds to an instance where FRET excitation transfer occurred, while emission by the donor corresponds to no FRET. The FRET efficiency is defined as $E_{FRET} = I_A/(I_D + I_A)$ [55], where $I_D$ is the
emission intensity of the donor and $I_A$ is the emission intensity of the acceptor. Therefore, the FRET efficiency is the fraction of the total emission which comes from the acceptor, after having undergone FRET transfer from the donor. The rate at which FRET occurs is highly dependent on distance, and is given by

$$k_{\text{FRET}} = k_0(R_0/R)^6,$$

where $k_0$ is the rate of fluorescent decay for an isolated donor fluorophore, $R_0$ is the FRET radius, which is a property of the fluorophores used and their environment, and $R$ is the distance between the donor and the acceptor, which in this case is the end-to-end distance for the nucleic acid. Therefore, FRET gives a measurement of the end-to-end distance of the nucleic acid.

Since $k_{\text{FRET}}$ depends nonlinearly on the end-to-end distance, it is nontrivial to convert the end-to-end distance distribution to a FRET distribution. Therefore we use a Monte Carlo simulation to sample the end-to-end distance distribution and predict when FRET will occur, as described in Murphy et al. [106].

For completeness we summarize this approach here. The Monte Carlo simulation proposes random steps in the distance between the two fluorophores, and accepts or rejects these steps based on the Metropolis algorithm using energies calculated from the the worm-like chain model. We use a step size $\Delta r = 0.55 \text{ nm}$ and a time step $\Delta t = 0.01 \text{ ns}$, which together correspond to a spatial diffusion constant of $3 \times 10^{-8} \text{ m}^2/\text{s}$ [106].

At each step, the donor may either decay with probability $k_0\Delta t$, or transmit its excitation to the acceptor with probability $k_{\text{FRET}}\Delta t$. These probability distributions are sampled using the GNU scientific library [109]. If either type of deexcitation occurs, we record whether a FRET transfer or decay was observed for this excitation, and then wait for the donor to become excited again. The end-to-end distance continues to fluctuate according to the worm-like chain model, and each step has a probability to become reexcited of $k_E\Delta t$. To be specific, we use for our illustration a rate for isolated donor fluorophore decay of $k_0 = 1 \text{ ns}^{-1}$ [106], and FRET radius of $R_0 = 6 \text{ nm}$ [56], which is appropriate for the Cy3/Cy5 donor/acceptor pair often used in such experiments. The rate of donor excitation depends on the irradiation intensity, and therefore can vary with the experimental apparatus. We choose a rate for donor excitation of $k_E = 1 \text{ ns}^{-1}$, which is a typical value that could be achieved in such experiments.

We run this Monte Carlo simulation for 1,000,000 donor excitations at each different number of exterior bases $m$. This allows us to accurately sample the probability distribution of the FRET intensity, $P_F(E_{\text{FRET}}|m)$, as a function of the number of exterior bases $m$ and the FRET efficiency $E_{\text{FRET}}$. We may then predict the observed FRET distribution for a
nucleic acid using
\[ P(E_{\text{FRET}}) = \sum_{m=\Delta}^{n} P(m)P_{F}(E_{\text{FRET}}|m). \]  \hspace{1cm} (3.13)

The distribution \( P_{F}(E_{\text{FRET}}|m) \) does not depend on the nucleic acid sequence used or any binding proteins present. Therefore, if we calculate \( P_{F}(E_{\text{FRET}}|m) \) once for a given set of experimental conditions and polymer model parameters, we then may calculate \( P(E_{\text{FRET}}) \) for any combination of nucleic acids and binding proteins.

As an example, we apply this method to simulate a single-stranded DNA containing 38 nucleotides, shown in Fig. 3.2a. We calculate the FRET distribution for this sequence at many different concentrations of single-stranded binding protein, for a protein with a chemical potential for binding of 10 kcal/mol at 1 M concentration and a footprint of 6 nucleotides. This distribution is shown in Fig. 3.2b. At low protein concentration, we only see FRET values very close to 1. This corresponds to the sequence forming almost all of the possible base pairings shown in Fig. 3.2a. As protein concentration is increased, we start to see a second FRET peak form between \( E = 0.4 \) and \( E = 0.7 \), which corresponds to melting of the 6 base pair stem closest to the ends of the DNA sequence. Finally, at very high protein concentration we see a third peak, which corresponds to all bases being unpaired, form between \( E = 0.1 \) and \( E = 0.4 \). This example shows that a relatively small single-stranded DNA in the presence of single-stranded binding proteins can result in a complicated FRET distribution, without any cooperativity or sequence specificity in protein binding. For instance, Fig. 3.2c shows that at a protein concentration of 500 nM all three FRET peaks coexist, which is a relatively complex distribution for such a small molecule. The details of the distribution arise purely from the promotion of DNA melting by adding increasing amounts of binding protein.

### 3.2.4 Protein binding

Using the methods described above, we may calculate the total partition function for the nucleic acid secondary structures in the presence of single-stranded binding proteins. The interaction between the nucleic acid and the binding proteins is governed by the chemical potential for protein binding \( \mu^{P}(c, i) \). Therefore, by varying the value of \( \mu^{P}(c, i) \) at a given nucleotide \( i \), we may examine the strength of the binding interaction between the nucleic acid and protein at nucleotide \( i \). We may then repeat this at every position within the nucleic acid sequence to find the probability that a protein is bound at each possible position. With this information we can compute experimentally measurable quantities such as the average number of proteins bound per nucleic acid, or the probability that at least one protein is bound to the nucleic acid.

Specifically, the probability that a protein is bound to the nucleic acid starting at nu-
Figure 3.3: Predicted protein binding probability for a protein with a chemical potential for binding of 10 kcal/mol at 1 M concentration and a footprint of 6 nucleotides binding to the sequence shown in Fig. 3.2a. (a) The binding probability $P_{\text{occupied}}(c, i)$ plotted as a function of both position within the nucleic acid sequence $i$ and protein concentration $c$. Darker shades of gray indicate higher binding probability, as indicated in the legend. At low protein concentrations, the protein first binds to the loop in the middle of the sequence. Then, as the concentration is increased the protein next binds near the ends of the sequence, breaking the relatively weak stem formed there. Finally, at very high concentrations the protein binds to the G and C nucleotides near the middle of the sequence, breaking the much stronger pairings between these bases. (b) The average number of proteins bound to each nucleic acid as a function of protein concentration calculated from Eq. 3.16. The number of proteins bound jumps up to 1.0 almost immediately at low concentration, as two proteins may bind to the loop in the middle of the DNA sequence without disturbing any base pairings. The number of proteins bound then increases approximately linearly with concentration from about 1.0 to 3.0 proteins bound. This corresponds to breaking the weak stem near the ends of the nucleic acid, allowing two more proteins to bind. At higher concentrations the number of proteins again increases linearly, but at a much slower rate, corresponding to breaking of the more stable stem near the middle of the nucleic acid and allowing two additional proteins to bind.
nucleotide $i$ is $[102]$

$$P_B(c, i) = -\frac{k_B T}{Q_{1,n}} \frac{\delta Q_{1,n}}{\delta \mu^P(c, i)}.$$  \hfill (3.14)

Since we can calculate $Q_{1,n}$ for any choice of $\mu^P(c, i)$ by Eq. 3.2, we can obtain the partial derivative $\delta Q_{1,n}/\delta \mu^P(c, i)$ and hence $P_B(c, i)$ numerically by calculating $Q_{1,n}$ for two nearby values of $\mu(c, i)$.

Each protein will bind $F$ consecutive nucleotides, where $F$ is the protein’s binding footprint. If the first nucleotide bound by a protein is $i$, then the protein will bind nucleotides $i$ through $i + F - 1$. Therefore, the probability that nucleotide $i$ is bound by a protein (regardless of whether it is the first nucleotide bound by this protein) is

$$P_{occupied}(c, i) = \sum_{j=i-F+1}^{i} P_B(c, j),$$  \hfill (3.15)

Where we have set $P_B(c, i) = 0$ for $i < 0$. Fig. 3.3a shows a plot of $P_{occupied}(c, i)$ as a function of protein concentration for the sequence shown in Fig. 3.2a. As in Fig. 3.2, we simulate a protein with a chemical potential for binding of $10 \text{ kcal/mol}$ at 1 M concentration and a footprint of 6 nucleotides. This plot illustrates the physical changes that occur to cause the changes in the FRET distribution shown in Fig. 3.2b. At very low concentration the protein binds to the loop at the middle of the nucleic acid, which does not require any of the base pairings which form in the absence of protein to be broken. The protein next invades the relatively weak stem formed by the ends of the nucleic acid, breaking this base pairing and reaching 50% binding to this region at a concentration of about 150 nM. Finally, at high protein concentration the much more stable stem formed near the middle of the nucleic acid is broken and the protein binds to this region, reaching 50% binding at about 700 nM.

Multiple proteins may bind to each nucleic acid, up to a limit of $n/F$, where $n$ is the number of nucleotides in the nucleic acid and $F$ is again the protein’s binding footprint. The average number of proteins bound per nucleic acid molecule is

$$P_{avg}(c) = \sum_{i=1}^{n} P_B(c, i).$$  \hfill (3.16)

Fig. 3.3b shows a plot of $P_{avg}(c)$ as a function of protein concentration. The number of proteins initially increases very rapidly, and reaches 1.0 at a concentration of about 30 nM, because up to two proteins my bind to the loop in the middle of the nucleic acid without disrupting any of the base pairings shown in Fig. 3.2a. While it is possible for two proteins to bind without breaking any base pairs, this is not the most likely state at low protein concentrations. This is because there are seven different ways a single protein may bind to the loop in the middle of the nucleic acid, while two proteins can only bind in one
configuration without breaking any base pairs.

The number of proteins bound then increases roughly linearly from 1.0 to 3.0 between concentrations of 50 and 200 nM. This corresponds to it becoming increasingly likely that the weak stem near the ends of the nucleic acid will be broken, allowing up to two more proteins to bind. The slope of the graph decreases between 200 and 400 nM, where there is a transition between breaking the weak stem formed by the ends of the nucleic acid and beginning to break the much stronger one near middle of the nucleic acid. Between 400 nM and 1000 nM the number of proteins bound increases slowly, as the likelihood of this strong stem being broken increases. Finally, at unrealistically high concentrations not shown in Fig. 3.3b the number of proteins bound will approach an asymptote at 6.0, as this is the maximum number of proteins with footprint 6 which can fit on a nucleic acid with a length of 38 nucleotides. Specifically, our numerics find that at 10,000 nM 5.8 proteins are bound per nucleic acid, while at 100,000 nM 6.0 proteins are bound.

We can also compute the fraction of nucleic acid molecules which are bound by at least one protein, which may be compared directly with experimental binding assays. First, we compute that the fraction of nucleic acid molecules not bound by any proteins at any position is

\[ P_{\text{unbound}}(c) = \prod_{i=1}^{n} [1 - P_B(c, i)], \]  

(3.17)

as \( 1 - P_B(c, i) \) gives the probability that no protein is bound starting at nucleotide \( i \). The probability that at least one protein is bound somewhere within the sequence is then

\[ P_{\text{bound}}(c) = 1 - \prod_{i=1}^{n} [1 - P_B(c, i)]. \]  

(3.18)

For the example protein and nucleic acid used in our illustrations, Eq. 3.18 predicts that 50% of the nucleic acid molecules will be bound by proteins when the protein concentration is 8.4 nM.

### 3.3 Discussion

We present a model for calculating nucleic acid secondary structure in the presence of proteins which bind to single-stranded nucleic acids. This model is very general, and may be applied to either DNA or RNA by using appropriate parameters for the Vienna RNA Package. The impact of protein is given by its free energy for interaction with a segment \( (i, j) \) of single-stranded nucleic acid. Therefore we may input any desired sequence dependence or footprint size for the protein.

This model allows us to exactly calculate the partition function for a nucleic acid in the presence of the binding protein. By computing restricted partition functions, we are able
to compare with a wide variety of experiments. We can predict FRET for dyes attached to the nucleic acid, the probability of the protein binding at every position in the nucleic acid sequence, as well as end-to-end distance distributions and base pairing probabilities for the nucleic acid. This allows interpretation of such experiments in terms of the physical binding parameters of the protein, such as its binding affinity and footprint.
Chapter 4

A QUANTITATIVE MODEL OF NUCLEOSOME DYNAMICS


Eukaryotic DNA is packaged into chromatin by wrapping around histone protein octamers to form nucleosomes [33, 110]. These nucleosomes sterically occlude replication, transcription and repair complexes from their DNA target sites. DNA processing complexes appear to access their target sites via nucleosome unwrapping [63, 111, 112] and disassembly [113]. However, the mechanism(s) associated with these nucleosome structural alterations are not well understood. An essential component in appreciating these nucleosome alterations is a quantitative and predictive model describing DNA unwrapping and disassembly from the histone octamer. Such a model will provide a critical tool in understanding in vivo DNA processing within chromatin.

The key concept in a model of nucleosomal DNA unwrapping is the unwrapping free energy landscape. This landscape quantitatively captures the free energy cost of unwrapping each individual base pair from the histone octamer. Previous studies determined qualitative features of the structure of this free energy landscape [63, 112]. Here, we develop a free energy landscape that not only accounts for these qualitative observations, but is calibrated from precise mechanical unzipping measurements of nucleosomal DNA [2]. These experiments apply a force to pull apart the DNA base pairing, which concurrently unwraps the DNA from the histone octamer. We extract detailed information about the energy landscape from these experiments. By carefully adjusting our energy landscape, we obtain an
excellent agreement with the highly complex measured unzipping profiles.

We then use this calibrated free energy landscape to model a completely separate experiment that investigates nucleosome disassembly by the human DNA mismatch recognition complex, hMSH2-hMSH6. This complex appears to function as a molecular switch that specifically binds DNA mismatches in the ADP bound form \cite{114, 115}. Upon the exchange of ADP for ATP, hMSH2-hMSH6 forms a sliding clamp that diffuses along duplex DNA. This allows for iterative loading of multiple hMSH2-hMSH6 complexes onto the duplex DNA surrounding a mismatch, which directs the excision reaction during DNA mismatch repair.

This mismatch recognition process must occur within chromatin. Recently, it was demonstrated that hMSH2-hMSH6 disassembles nucleosomes near a DNA mismatch and that histone modifications in the DNA-histone interface of the nucleosome dramatically enhance the rate of disassembly \cite{116, 117}. These results indicate that hMSH2-hMSH6 not only recognizes DNA mismatches but facilitates the DNA excision reaction by ensuring the DNA surrounding a mismatch is nucleosome and perhaps protein free. It was proposed that hMSH2-hMSH6 disassembles nucleosomes by trapping DNA unwrapping fluctuations through iterative loading of hMSH2-hMSH6 at the DNA mismatch \cite{116}.

We modeled this nucleosome disassembly process with our free energy landscape and found that disassembly of unmodified nucleosomes can be quantitatively understood using the same free energy landscape applied to unzipping experiments. We explain the influence of histone post-translational modifications (PTMs) and DNA sequence on the disassembly rates with modest changes to the free energy landscape based on the location of the modifications and the known influence of DNA sequence and histone modifications on DNA-histone binding free energies \cite{112, 117, 118}. Furthermore, our model provides detailed mechanistic insight into the process of nucleosome disassembly by hMSH2-hMSH6.

We also apply this model to experiments which directly measure nucleosome unwrapping fluctuations. LexA is a protein which tightly binds a specific DNA sequence. Placing its recognition site within a nucleosome allows DNA unwrapping fluctuations to be trapped by lexA binding. Our model is in excellent agreement with experimental measurements of such fluctuations monitored using fluorescence resonance energy transfer (FRET), showing it also accurately captures the kinetics of nucleosome unwrapping.

4.1 Modeling nucleosome unwrapping dynamics

4.1.1 Solution to the master equation for unwrapping dynamics

We treat the time evolution of the probability that $n$ base pairs of nucleosomal DNA are unwrapped as a continuous-time Markov process. This allows us to write the rates of
wrapping \( k_{\text{rewrap}}(n) \) and unwrapping \( k_{\text{unwrap}}(n) \) in a rate matrix \( T_{n,n'} \), which is defined by
\[
T_{n,n+1} = k_{\text{unwrap}}(n) \quad \quad \quad (4.1)
\]
\[
T_{n+1,n} = k_{\text{rewrap}}(n), \quad \quad \quad (4.2)
\]
with all other entries in \( T_{n,n'} \) equal to zero. This rate matrix satisfies the master equation
\[
\frac{dP_n(t)}{dt} = \sum_{n'} T_{n,n'} P_{n'}(t), \quad \quad \quad (4.3)
\]
where \( P_n(t) \) is the time dependent probability of observing the unwrapped state \( n \). The rate matrix \( T_{n,n'} \) contains all of the information about the influence of external constraints, such as an applied force or a protein complex, on the probability and time evolution of each partially unwrapped nucleosome state \( n \). Below, we describe how we determine the rate matrix that describes the influence of a mechanical unzipping force and a pressure from bound hMSH2-hMSH6 complexes on the nucleosome unwrapping probability and its time evolution. However, once this rate matrix is set up, the following solution to the master equation may be applied to any system where nucleosomes are unwrapped.

We solve the eigenvalue problem for this master equation to obtain \( P_n(t) \). We compute the eigenvalues \( \lambda_i \) and eigenvectors \( V_{i,n} \) of \( T_{n,n'} \) numerically [119]. We then use the inverse of the matrix of eigenvectors \( V_{i,n}^{-1} \) to express the initial state in the basis of the eigenvectors
\[
\tilde{P}_i(0) = \sum_n V_{i,n}^{-1} P_n(0), \quad \quad \quad (4.4)
\]
where \( P_n(0) \) is the initial state in the basis of the unwrapping states. If the nucleosome is fully wrapped at time zero, then \( P_0(0) = 1 \) and \( P_i(0) = 0 \) for \( i \neq 0 \). This then allows us to write the solution to Eq. 4.3 as
\[
P_n(t) = \sum_i V_{i,n} \tilde{P}_i(0) \exp(-|\lambda_i|t). \quad \quad \quad (4.5)
\]
The term \( \exp(-|\lambda_i|t) \) propagates each eigenvector with a decay rate given by its eigenvalue, and we multiply by \( V_{i,n} \) to transform back into the basis of unwrapping states. This solution allows us to calculate the complete dynamics of nucleosome unwrapping rapidly.

### 4.1.2 Rates of nucleosome unzipping

In unzipping experiments, an applied force \( F \) breaks apart the base pairing of DNA containing a nucleosome, which simultaneously unwraps the nucleosome. We compute \( k_{\text{unwrap}}(n) \) and \( k_{\text{rewrap}}(n) \), the rates of nucleosome unwrapping and rewrapping at \( n \) unwrapped base pairs, using the Gibbs free energy difference between the change in our energy landscape \( G_{\text{nuc}}(n) \) due to unwrapping and the work done on the nucleosome by the applied force. We
multiply the base rate for wrapping fluctuations \( k_0 \) by a Gibbs factor, resulting in

\[
k_{\text{unwrap}}(n) = k_0 e^{[(F - F_{\text{unzip}}(n)) \alpha \Delta x - \Delta G_{\text{unc}}(n)]/k_B T},
\]

(4.6)

\[
k_{\text{rewrap}}(n) = k_0 e^{[-(F - F_{\text{unzip}}(n))(1 - \alpha) \Delta x]/k_B T},
\]

(4.7)

where \( k_B \) is Boltzmann’s constant and \( T \) is the temperature. The applied force \( F \) acts over a distance \( \Delta x \) per opened base pair, and \( F_{\text{unzip}}(n) \) is the force required to break the DNA base pairing and stretch the DNA. Of the remaining force, a fraction \( \alpha \) increases the unwrapping rate and \((1 - \alpha)\) decreases the rewrapping rate. The parameter \( \alpha \) represents the relative position of the transition state between \( n \) and \( n + 1 \) unwrapped base pairs. We may then insert these rates of unwrapping and rewrapping into Eqs. 4.1 and 4.2, and follow the procedure outlined in Section 4.1.1 to calculate the dynamics of nucleosome unwrapping in unzipping experiments.

### 4.1.3 Calculation of dynamics in the presence of hMSH2-hMSH6

When hMSH2-hMSH6 is present we must model its binding to the DNA, the interactions between hMSH2-hMSH6 molecules, and the interactions between hMSH2-hMSH6 and the nucleosome. Given that the rate of hMSH2-hMSH6 diffusion along the DNA is faster than all other rates in the system, we treat hMSH2-hMSH6 diffusion as being in equilibrium with respect to all other processes in the system. We may then model the hMSH2-hMSH6 molecules as a one dimensional gas of hard spheres. This is known as a Tonks gas, and has a partition function given by [120]

\[
Z_{N,l} = (l - N\sigma)^N / N!,
\]

(4.8)

where \( N \) is the number of hMSH2-hMSH6 molecules bound to the DNA, \( \sigma \) is the footprint of hMSH2-hMSH6 in base pairs and \( l \) is the length of the DNA segment available to hMSH2-hMSH6 in base pairs (Fig. 4.1A). To confirm that use of the Tonks gas solution is valid, for select cases we also performed simulations of the dynamics including each hMSH2-hMSH6 particle individually. We find that approximating the hMSH2-hMSH6 behavior as a Tonks gas gives results nearly identical to these simulations, and is much more computationally efficient (Fig. 4.2).

We can then use the Tonks gas partition function to calculate the probability that the bound hMSH2-hMSH6 molecules will block nucleosome re-wrapping or the binding of more hMSH2-hMSH6. The probability that there are no hMSH2-hMSH6 particles within \( x \) base pairs of one end of the available DNA segment for \( x < (l - N\sigma) \) is

\[
\phi_{N,l}(x) = \frac{Z_{N,l-x}}{Z_{N,l}} = \left( \frac{l - N\sigma - x}{l - N\sigma} \right)^N.
\]

(4.9)
Figure 4.1: Model predictions compared with experimental measurements for nucleosome displacement by hMSH2-hMSH6. (A) Graphical depiction of the experimental system [116]. Streptavidin bound to the biotinylated left end of the DNA prevents hMSH2-hMSH6 (MSH) from sliding off this end, while a nucleosome inhibits hMSH2-hMSH6 from sliding off the right end of the DNA. There is a mismatch located \( n_M = 21 \) bp from the biotin bead, at which hMSH2-hMSH6 may bind. The total length of the DNA segment is \( L \). The hMSH2-hMSH6 is confined to length \( l \), the DNA between the biotin bead and the nucleosome. The footprint of one hMSH2-hMSH6 molecule is \( \sigma = 25 \) bp. Free energy landscapes for unwrapping nucleosomes on (B) the 601 positioning sequence and (C) the 5S positioning sequence. A comparison of the results of our modeling (lines) using these energy landscapes to experimental measurements (points) of nucleosome displacement by hMSH2-hMSH6 shows excellent agreement for both (D) the 601 positioning sequence and (E) the Xenopus 5S positioning sequence.
Figure 4.2: Full simulations of nucleosome displacement which explicitly include each hMSH2-hMSH6 particle compared to our approximate solution using a one dimensional Tonks gas to represent the hMSH2-hMSH6 particles. The solid lines all show the result of our approximate solution and the symbols show experimental data. These approximate curves and data are identical to those shown in Fig. 4.1. (A) Nucleosomes on the 601 sequence. The dashed light blue, orange and yellow dashed lines show full simulations for unmodified nucleosomes with a G/T mismatch, H3(T118ph) with no mismatch, and H3(T118ph) with a G/T mismatch, respectively. (B) Nucleosomes on the 5S sequence. The dashed light blue, orange and yellow dashed lines show full simulations for unmodified nucleosomes, H3(K56Q) and H3(K115ac,K122ac), respectively, all containing a G/T mismatch. In all cases the full simulations use identical parameters as in the corresponding approximate solution. The agreement is very good, but there is a slight systematic deviation. This is because in the full simulation it is possible for the nucleosome to undergo an unwrapping fluctuation, and then rapidly rewrap before the hMSH2-hMSH6 complexes have reached a new equilibrium, while in our approximation the hMSH2-hMSH6 complexes are assumed to reach equilibrium instantly. This causes our approximation to predict that displacement occurs slightly more rapidly than calculated by the full simulation.
To calculate the rates of unwrapping, \( k_{\text{unwrap}}(n) \), and rewrapping, \( k_{\text{rewrap}}(n) \), we multiply the base rate for wrapping fluctuations \( k_0 \) by a Gibbs factor determined from our free energy landscape, as in Eqs. 4.6 and 4.7

\[
\begin{align*}
    k_{\text{unwrap}}(n) &= k_0 e^{-\Delta G_{\text{nuc}}(n)/k_B T} \\
    k_{\text{rewrap}}(n) &= k_0 \varphi_{N,l}(1),
\end{align*}
\]

where \( \varphi_{N,l}(1) \) is the probability that an hMSH2-MSH6 particle does not block rewrapping of one base pair. Similarly, the rate of hMSH2-hMSH6 binding at the mismatch is

\[
k_{\text{bind}}^{N,l} = k_{\text{bind}}^0 \varphi_{N,l}(n_M + \sigma/2),
\]

where \( k_{\text{bind}}^0 \) is the rate of binding to bare DNA of length \( l \), \( n_M \) is the mismatch position, and \( \varphi_{N,l}(n_M + \sigma/2) \) is the probability that binding is not blocked by the presence of another hMSH2-hMSH6 molecule. This assumes that the mismatch is within 1.5\( \sigma \) base pairs of the end of the DNA segment, such that there cannot be any hMSH2-hMSH6 molecules between the end of the DNA and the mismatch when an hMSH2-hMSH6 molecule is bound at the mismatch. If this is not true, one must sum over all combinations for the number of hMSH2-hMSH6 complexes bound on either side of the mismatch.

We then calculate the dynamics of nucleosome unwrapping following the method described in Section 4.1.1. Instead of indexing the probability in Eq. 4.3 based only on \( n \), we use \( u = NL + n \), where \( N \) is the number hMSH2-hMSH6 molecules already bound to the DNA, and \( L \) is the maximum number of bases which can unwrap from the histone before it dissociates from the DNA. This allows us to write a rate matrix \( T_{u,u'} \) which includes not only rates of nucleosome wrapping and unwrapping, but also hMSH2-hMSH6 binding and dissociation.

### 4.2 Results

#### 4.2.1 Mechanical Unzipping of Nucleosomes

The most detailed method yet devised to probe nucleosome dynamics is mechanically unzipping the base pairing of DNA containing a nucleosome [2]. In experiments performed by Wang and coworkers, the nucleosome is positioned on the 601 high affinity sequence [121]. One strand of the DNA molecule is attached to a surface, and the other strand to a bead confined in an optical trap (Fig. 4.3A). By carefully controlling the movement of the optical trap, a constant force may be exerted on the DNA. This force sequentially breaks the DNA base pairing, increasing the contour length of the single-stranded DNA between the bead and the surface. Measuring the position of the bead as a function of time establishes the number of DNA bases that are unpaired. When DNA wrapped around the nucleosome
becomes critically unpaired, it must dissociate from the nucleosome. Thus, the number of broken base pairs equals the number of base pairs unwrapped from the nucleosome (solid line, Fig. 4.3B). This method probes the interaction between the nucleosome and the DNA as a function of the number of base pairs unwrapped from the nucleosome.

We calculate an energy landscape for unwrapping 601 positioning sequence DNA from a nucleosome, which is consistent with this unzipping data (Fig. 4.3A). To obtain this landscape, we first create a trial landscape that we use to compute rates of nucleosome unwrapping. We then use these rates to calculate unzipping dynamics, and compare the result with experiments. Finally, we iterate the trial landscape until we obtain good agreement with experiments (Fig. 4.3B) while respecting several qualitative features of the landscape known before our study.

In order to calculate experimental observables given an energy landscape, we compute $k_{\text{unwrap}}(n)$ and $k_{\text{rewrap}}(n)$, the rates of nucleosome unwrapping and rewrapping at $n$ unwrapped base pairs, using Eqs. 4.6 and 4.7. In these experiments, the total applied force
Figure 4.4: The force $F_{unzip}(n)$ needed to unzip the base pairing and stretch the DNA is shown as a function of position for the 601 sequence, calculated using [52]. Since the free energy cost to break one base pair is linearly proportional to $F_{unzip}(n)$, we show $F_{unzip}(n)$ in pN on the left hand axis, and the free energy cost of breaking one base pair at this force in $k_B T$ on the right hand axis. This free energy is just the force times the increase in length from breaking one base pair (1.12 nm) and then divided by 4.11 to convert from pN-nm to $k_B T$.

$F$ is held constant at 28 pN [2]. The force $F_{unzip}(n)$ needed to unzip the base pairing and stretch the DNA can be calculated as a function of the DNA sequence (Fig. 4.4) [52]. This results in an effective force responsible for nucleosome unwrapping averaging 10.9 pN, and ranging from 8.6 pN to 12.8 pN. Similarly, the increase in contour length from opening one base pair $\Delta x$ can be calculated to be 1.12 nm at this force [52]. Since the relative position $\alpha$ of the transition state between $n$ and $n+1$ unwrapped base pairs is not known under the geometry of the experiments, we treat it as a fit parameter.

We find that our results are not very sensitive to the value of $k_0$, the rate of nucleosome rewrapping in the absence of an applied force. We can fit the unzipping data with values of $k_0$ between $10^6$ to $10^7$ base pairs per minute by simultaneously making small adjustments to the nucleosome energy landscape and $\alpha$. We used the value of $k_0 = 6 \times 10^6$ base pairs per minute that is in the middle of this range for all subsequent calculations.

The trial landscape must conform to a number of experimental observations. First, the free energy landscape should be steepest in the dyad region (between about 63-84 bp), since the nucleosome contains the strongest contacts in this region [33, 112]. Second, the total height of the free energy landscape must be no more than 42 $k_B T$, the free energy cost to disassemble a nucleosome from a compacted chromatin fiber in Xenopus egg extracts [122]. The buffer conditions and presence of histone H1 in Xenopus egg extracts allow for strong interactions between nucleosomes, and promote chromatin compaction. We expect the
free energy cost to disassemble the recombinant histones examined here to be significantly smaller, as the measurements are performed under conditions where chromatin compaction does not occur.

Third, the steps in the landscape occur approximately every 5.25 bp, which is the average spacing between DNA contacts in the nucleosome [33] and between peaks in the unzipping measurements [2]. The positions of the peaks in the unzipping data [2] are extremely sensitive to the spacing between the steps in the unzipping landscape. E.g., if we force all steps to be separated by exactly 5.25 bp, we are not able to accurately fit the unzipping data (Fig. 4.5). We can thus directly read off the spacing of the steps from the positions of the peaks in the unzipping data [2]. This results in slight variations of the step spacing around the average of 5.25 bp consistent with the nucleosome crystal structure, which shows that DNA contacts in the nucleosome are not evenly spaced [33] either. We use discrete steps for simplicity, as we find no significant difference from using more gradual steps. Lastly, the landscape cuts off shortly after the dyad region, such that the histone octamer irreversibly dissociates from the DNA and cannot bind again once the system reaches this cut off. This corresponds with the observation that nucleosomes disassemble after the dyad is disrupted [2, 50].

The free energy landscape is varied until the calculated DNA unzipping dynamics (dashed line, Fig. 4.3B) agree with the experimental data. The free energy landscape shown (Fig. 4.3A) conforms to all the constraints outlined above and also accurately reproduces the measured unzipping dynamics. Thus, we are able to extract detailed information about the structure of the nucleosome free energy landscape from these studies.

Because we do not know the value of $\alpha$, the landscape is not completely constrained by the unzipping experiments. By making relatively small adjustments to the value of $\alpha$, we can obtain free energy landscapes that fit the unzipping data about as well as the one shown (Fig. 4.3A), but that have total height as small as $12k_BT$ or as large as $27k_BT$. We select a landscape with total height of $23.8k_BT$ as this allows us to simultaneously fit measurements of nucleosome disassembly by hMSH2-hMSH6 described in the next section.

### 4.2.2 Nucleosome Disassembly by hMSH2-hMSH6

To demonstrate the utility of our nucleosome free energy landscape, we apply it to measurements of the rate of nucleosome disassembly induced by the DNA mismatch recognition complex, hMSH2-hMSH6 [116, 117]. In these experiments, one end of a duplex DNA molecule is labeled with biotin, while the other end contains a nucleosome positioning sequence wrapped into a nucleosome (Fig. 4.1A). The DNA molecule contains 99 or 75 base pairs between the biotin and the 5S or 601 positioning sequence, respectively, and a mismatch 21 base pairs from the biotin labeled end. hMSH2-hMSH6 binds the DNA mismatch and following ATP binding forms a clamp that slides freely along the DNA [23, 115]. Strep-
Figure 4.5: A free energy landscape constrained to have steps separated by 5.25 bp, or half of one DNA helical repeat. (A) Dwell time as a function of the number of DNA base pairs (bp) unwrapped from the nucleosome. The solid blue line is determined from mechanical unzipping experiments, while the dashed red line is calculated from (B) a free energy landscape with exactly 5.25 bp separation between steps for DNA unwrapping from a nucleosome on the 601 positioning sequence. Because the separation between the peaks in the experimental unzipping data is not uniform, we are not able to accurately reproduce them with this landscape.
tavidin bound to the biotin labeled end of the DNA prevents hMSH2-hMSH6 from sliding off that DNA end. At the opposite end, hMSH2-hMSH6 disassembles the nucleosome, allowing for hMSH2-hMSH6 to slide off the DNA.

Nucleosome disassembly by hMSH2-hMSH6 was demonstrated to depend on ATP binding but not ATP hydrolysis [116]. We therefore consider a model where no ATP is hydrolyzed and hMSH2-hMSH6 cannot directly exert a force on the nucleosome. Rather, hMSH2-hMSH6 takes advantage of fluctuations in nucleosomal DNA unwrapping [63, 111, 123], which we calculate from our calibrated free energy landscape. hMSH2-hMSH6 binds at the DNA mismatch, forms a sliding clamp and then freely diffuses along the DNA. An additional hMSH2-hMSH6 complex can bind at the mismatch and form a sliding clamp once hMSH2-hMSH6 diffuses off the mismatch. As multiple hMSH2-hMSH6 clamps load onto the DNA, they can occupy unwrapped nucleosomal DNA, which prevents rewrapping thereby increasing the fraction of unwrapped nucleosomal DNA.

To accurately model hMSH2-hMSH6, we use known rates of hMSH2-hMSH6 sliding and binding. A rate of 5 per minute was used for hMSH2-hMSH6 binding and sliding clamp formation at a G/T mismatch, while 0.1 per minute was used for the rate of dissociation of the sliding clamp form of hMSH2-hMSH6 from DNA [23, 97, 114, 124]. A rate of $3.7 \times 10^7$ base pairs per minute was used for the step rate of hMSH2-hMSH6 clamps. This corresponds to the reported lower limit for the hMSH2-hMSH6 clamp diffusion constant on duplex DNA of 0.036 $\mu$m²/sec [124].

We calculated rates of hMSH2-hMSH6 induced nucleosome disassembly for comparison to rates measured by electrophoretic mobility shift assays [116, 117]. For computational efficiency, we used the fact that hMSH2-hMSH6 stepping along the DNA is the fastest process in our model, which allows us to model hMSH2-hMSH6 clamps as an equilibrated gas [120]. This gas consists of particles of size $\sigma = 25$ nm [114], which diffuse freely in one dimension along the DNA, but cannot pass through one another (Fig. 4.1A) (see Materials and Methods). We find the calculated nucleosome disassembly rates using this approximation to be nearly identical to those from a full simulation including every hMSH2-hMSH6 complex explicitly (Fig. 4.2).

### 4.2.3 DNA Sequence Dependence of Nucleosome Disassembly by hMSH2-hMSH6

We begin by examining nucleosome disassembly measurements with the 601 positioning sequence. This allows us to employ the same free energy landscape used to model unzipping experiments (solid line, Fig. 4.1B). We calculate the characteristic time for nucleosome disassembly to be 970 minutes, which agrees with the experimental observation that hMSH2-hMSH6 does not measurably disassemble nucleosomes in 60 minutes (diamonds, Fig. 4.1D) [117, 125].
Next, we consider the influence of the 5S nucleosome positioning sequence on nucleosome disassembly by hMSH2-hMSH6. We assume that the change in DNA sequence only influences the free energy landscape of nucleosome unwrapping. Therefore, we kept all model parameters fixed except for the nucleosome unwrapping free energy landscape. While detailed DNA unzipping measurements have not been reported for a 5S sequence, we expect this landscape to have the same general form as the 601 landscape, but a smaller total height. The sea urchin 5S sequence lowers the free energy of nucleosome formation by $4.7 \, k_BT$ relative to 601 [121] and thus we expect a reduction of several $k_BT$ for the Xenopus 5S sequence as well [116]. Furthermore, the DNA sequence near the dyad region appears to be responsible for the enhanced DNA-histone binding of the 601 sequence [112, 121].

We modified the 601 landscape by decreasing it by a total of $2.2 \, k_BT$ in the dyad region (solid line, Fig. 4.1C). Nucleosome disassembly by hMSH2-hMSH6 is not sensitive to the detailed structure of the free energy landscape, so our choice to modify the landscape at the dyad region is based on empirical evidence that the increased binding affinity of the 601 sequence relative to 5S is due mainly to contacts in the dyad region [112, 121]. This landscape gives a good fit to the experimental data for unmodified nucleosomes containing the Xenopus 5S sequence [116] (diamonds, Fig. 4.1E). This landscape allowed us to explain the behavior of nucleosome disassembly in the absence of a mismatch (squares, Fig. 4.1E) where hMSH2-hMSH6 is permitted to bind at any position within the DNA at a rate of $9 \times 10^{-4}$ per minute, the experimental rate of hMSH2-hMSH6 binding to duplex DNA without a mismatch [23, 97, 114]. These results suggest that modest changes in the free energy landscape due to alterations in the nucleosomal DNA sequence can lead to significant changes in the hMSH2-hMSH6 induced nucleosome disassembly rate.

**4.2.4 The influence of Histone Post-Translational Modifications of Nucleosome Disassembly by hMSH2-hMSH6**

Histone PTMs within the DNA-histone interface reduce the free energy of nucleosome formation [117, 118, 126, 127] and facilitate nucleosome disassembly by hMSH2-hMSH6 [116, 117]. We incorporated the influence of histone PTMs into our model by assuming that they specifically influence the DNA unwrapping free energy landscape close to the location of the PTM.

We first investigated the influence of histone PTMs near the nucleosome dyad: the phosphorylation of histone H3 at T118 [H3(T118ph)] and the double acetylation of H3 at K115 and K122 [H3(K115ac,K122ac)]. These modifications occur in vivo [128] and influence DNA repair [129]. Recently, it was reported that H3(T118ph) reduces the DNA-histone binding free energy by $3.5 \pm 0.3 \, k_BT$ and that it increases the rate of hMSH2-hMSH6 induced nucleosome disassembly by 25 times [117]. We found that a $3.8 \, k_BT$ reduction of our 601 free energy landscape in the dyad region results in excellent agreement between the
model predictions and the measured data (squares, Fig. 4.1D).

The dyad acetylations, H3(K115ac,K122ac), reduce the DNA-histone binding free energy by 0.8 ± 0.4 $k_B T$ [118] and increase the rate of nucleosome disassembly by hMSH2-hMSH6 [116]. We found that decreasing the free energy landscape for unmodified 5S by 1.2 $k_B T$ in the dyad region fits the experimental measurements for these modifications (circles, Fig. 4.1E).

Histone acetylation in the DNA entry-exit region of the nucleosome, H3(K56ac), reduces the free energy of the nucleosome binding to DNA [126] and increases DNA unwrapping [127]. From this increase in unwrapping, we estimate that H3(K56ac) reduces the binding free energy by 0.5 ± 0.3 $k_B T$ [127]. This modification occurs within newly replicated DNA and appears important for DNA repair [113]. Furthermore, the acetyllysine mimic at this location, H3(K56Q), enhances the rate of nucleosome disassembly by hMSH2-hMSH6 [116]. We incorporated H3(K56Q) into our model by reducing the energy landscape by 0.5 $k_B T$ in the entry-exit region compared to the unmodified 5S landscape (dashed line, Fig. 4.1C). This results in an excellent fit to the hMSH2-hMSH6 displacement data for H3(K56Q) (triangles, Fig. 4.1E).

Finally, using the energy landscapes described above, we reproduce the disassembly rates without a mismatch for H3(T118ph) (triangles, Fig. 4.1D) and H3(K115ac,K122ac) 5S nucleosomes (triangles, Fig. 4.1E). We use the same hMSH2-hMSH6 binding conditions used to describe unmodified nucleosomes with no mismatch on 601, i.e. without fitting any other parameters. As before, measurements without mismatches were modeled by allowing hMSH2-hMSH6 to bind at any position within the DNA.

4.2.5 Binding of LexA to the entry-exit region

LexA is a repressor protein from *E. coli* with a very specific 28 bp binding site and very high binding affinity (0.1 nM) [130]. By placing a binding site for LexA in the entry-exit region of a nucleosome, as shown in Fig. 4.6, Justin North was able to probe nucleosome unwrapping fluctuations [131]. In the absence of LexA, the DNA will be fully wrapped into the nucleosome, allowing efficient FRET to occur between the Cy3 fluorophore attached to the DNA and the Cy5 attached to the histone octamer. If the DNA unwraps from the nucleosome, FRET will decrease as the distance between the fluorophores is increased. By adding a large amount of LexA, Justin North is able to trap any nucleosome unwrapping fluctuations which occur, as the LexA remains bound throughout his experimental time frame due to its high affinity. Therefore, by measuring the rate at which FRET decays after the addition of LexA, he is able to determine the rate of nucleosome unwrapping fluctuations large enough to expose the LexA binding site. He estimates this rate to be 10-15 unwrapping events per second [131], while our model predicts that 11.7 such unwrapping events should occur each second.
Figure 4.6: Experimental setup for probing nucleosome unwrapping fluctuations with LexA. A binding site for LexA is placed in the entry-exit region of the nucleosome, such that 30 bp must unwrap from the nucleosome for this binding site to be exposed. The DNA is end-labeled with a Cy3 fluorophore, such that if the DNA is fully wrapped into the nucleosome FRET may occur between this dye and a Cy5 attached to the histone octamer.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Modification</th>
<th>$\Delta E_{\text{expt}}^{a} \quad (k_B T)$</th>
<th>$\Delta E_{\text{landscape}}^{b} \quad (k_B T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>601</td>
<td>H3(T118ph)</td>
<td>3.5 ± 0.3</td>
<td>3.8</td>
</tr>
<tr>
<td>5S</td>
<td>H3(K56Q)</td>
<td>0.5 ± 0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>5S</td>
<td>H3(K115ac,K122ac)</td>
<td>0.8 ± 0.4</td>
<td>1.2</td>
</tr>
</tbody>
</table>

$^{a}$Experimentally measured binding energies [117, 118, 127].

$^{b}$Binding energies determined from our free energy landscapes.

4.3 Discussion

To our knowledge, we propose the first quantitative free energy landscape for unwrapping DNA from a histone octamer. We develop this landscape from high resolution nucleosomal DNA unzipping experiments [2]. We then demonstrate that this landscape accounts for the different hMSH2-hMSH6 induced nucleosome disassembly rates for two separate DNA sequences and four separate histone modification states. All parameters are held constant except for the free energy landscape. We only introduce modest landscape alterations that are consistent with the locations of the modifications and their influence on DNA-histone binding free energies known from experiments (Table 4.1) [117, 118, 127], and DNA sequence differences [121]. These results demonstrate that the nucleosome unwrapping free energy landscape provides an important tool for understanding enzyme provoked nucleosome disassembly.
Figure 4.7: Force exerted on the nucleosome as a function of the number of hMSH2-hMSH6 complexes bound and the number of DNA base pairs unwrapped from the nucleosome. This force is calculated using our one dimensional Tonks gas approximation and shown by solid colored lines for between one and seven hMSH2-hMSH6 complexes bound to the DNA. For comparison, we also show the force required to unwrap unmodified and modified nucleosomes for (A) the 601 construct, which has 75 base pairs of free DNA outside the nucleosome, and (B) the 5S construct, with 99 base pairs of free DNA. Note that the difference in the forces exerted by hMSH2-hMSH6 on 601 and 5S is due to the different lengths of free DNA. The force required to unwrap the nucleosome is calculated from the derivative of our free energy landscape averaged over the distance between steps in the landscape (5 base pairs). The system can be at equilibrium at any position where the force exerted by the hMSH2-hMSH6 complexes intersects the force needed to unwrap the nucleosome. Therefore, with 3 or 4 hMSH2-hMSH6 complexes bound, it is possible for the system to reach an equilibrium with about 40 or 60 base pairs of DNA unwrapped from the nucleosome. From these states, it is then reasonably probable for a fluctuation to the completely unwrapped state to occur.

The model’s success in explaining numerous experimental results indicates that it captures the essential components of nucleosome disassembly by hMSH2-hMSH6 and that we can use the model to provide mechanistic insight into this disassembly process. The observation that multiple hMSH2-hMSH6 clamps loaded onto DNA behave as a one dimensional gas, also known as a Tonks gas [120], provides a mechanistic picture of hMSH2-hMSH6 induced nucleosome disassembly. The hMSH2-hMSH6 Tonks gas exerts a one dimensional pressure on the nucleosome (Fig. 4.7). The pressure builds as additional hMSH2-hMSH6 complexes load onto the DNA, which shifts the equilibrium amount of DNA wrapped into the nucleosome toward more unwrapped states. This enables the nucleosome to eventually reach an unwrapped state that allows for histone octamer disassociation.

The number of hMSH2-hMSH6 clamps required to disassemble the nucleosome is three.
Figure 4.8: Distribution of the number of bound hMSH2-hMSH6 complexes needed to displace nucleosomes. (A) hMSH2-hMSH6 complexes needed to displace a nucleosome from the 601 positioning sequence. The solid line shows the result for unmodified nucleosomes, and the dashed line for the dyad modification H3(T118ph). (B) hMSH2-hMSH6 complexes needed to displace nucleosomes from the 5S positioning sequence. The solid blue line shows the result for unmodified nucleosomes, the dashed red line for the modification H3(K56Q), and the dotted green line for the modifications H3(K115ac,K122ac). Note that more hMSH2-hMSH6 complexes are required to displace unmodified nucleosomes from the 5S DNA construct as compared to the 601 DNA construct because there is more free DNA outside the nucleosome (99 base pairs compared with 75 base pairs for the 601 construct). This allows up to three hMSH2-hMSH6 complexes to bind without unwrapping the nucleosome, and examination of Fig. 4.7 shows that four hMSH2-hMSH6 complexes are required to allow a partially unwrapped equilibrium with 40 or 60 base pairs of DNA unwrapped from the nucleosome. This is consistent with an average of four bound hMSH2-hMSH6 complexes required to displace nucleosomes from the 5S DNA construct.

to four (Fig. 4.8). However, this number of clamps is not sufficient to trap DNA unwrapping states out to the dyad. Instead, the plateau in the free energy landscape around 50 bp allows clamps to trap states with about 50 bp of unwrapped DNA often enough that an unwrapping fluctuation to the dyad occurs before the outermost clamp slides away from the nucleosome (Fig. 4.9).

The model indicates that the mechanism by which dyad modifications, H3(T118ph) and H3(K115ac,K122ac) influence nucleosome disassembly is distinct from the DNA entry-exit modification mimic, H3(K56Q). None of the modifications influence the distance the outermost hMSH2-hMSH6 clamp invades into the nucleosome. The dyad modifications increase the probability for an unwrapping fluctuation that allows for histone octamer release when the outermost hMSH2-hMSH6 clamp has trapped about 50 bp of DNA (Fig. 4.9). In contrast, H3(K56Q) increases the probability that hMSH2-hMSH6 has trapped an unwrapping fluctuation of about 50 bp. This increases the time period the nucleosome exists in a partially unwrapped state and in turn the probability that the nucleosome will unwrap to the dyad and release the histone octamer. These mechanistic insights elucidate how fractional changes in the nucleosome unwrapping free energy landscape result in an order of magnitude changes in the rate of disassembly.

The time for nucleosome disassembly varies from about 40 minutes to 16 hours. Yet, all
Figure 4.9: Distribution of the number of DNA base pairs unwrapped from the nucleosome in the presence of hMSH2-hMSH6 on (A) the 601 DNA construct and (B) the 5S DNA construct. We show all modified nucleosomes considered in this work with a single G/T mismatch in the DNA construct at which hMSH2-hMSH6 can bind. We also show the occupancy with no hMSH2-hMSH6 binding allowed for unmodified nucleosomes on each construct (unmodified -MSH). While it is always most probable for the nucleosome to have less than 30 base pairs of DNA unwrapped, in the presence hMSH2-hMSH6 there is also a significant probability that between 30 and 70 base pairs of DNA are unwrapped from the nucleosome. The modification H3(K56Q) further increases the probability of such unwrapping by lowering the energy barrier to unwrap past 30 base pairs. In contrast, the dyad modifications H3(T118ph) and H3(K115ac,K122ac) do not change the probability of such modest unwrapping, but rather increase the chance of a large fluctuation which totally unwraps the nucleosome.
Figure 4.10: Rate of binding three, four or five hMSH2-hMSH6 complexes to (A) the 601 DNA construct or (B) the 5S DNA construct containing an unmodified nucleosome. The distribution of hMSH2-hMSH6 complexes which can bind without displacing the nucleosome reaches equilibrium very rapidly, in about 1 minute. For both complexes three hMSH2-hMSH6 complexes can sometimes bind at equilibrium, with a larger probability for 5S because this construct contains more free DNA outside the nucleosome (99 base pairs compared to 75 base pairs for 601). For the 601 construct shown in (A), the probability of five hMSH2-hMSH6 complexes binding is insignificant, but the fraction of four hMSH2-hMSH6 bound increases very slowly with a characteristic time of 960 minutes. This is nearly equal to the characteristic time for displacement of the nucleosome from this complex. For the 5S construct shown in (B), the fractions of both four and five hMSH2-hMSH6 bound increase with a characteristic time of 170 minutes, which is again equal to the characteristic time for nucleosome displacement.

The processes involved with disassembly, including nucleosome unwrapping, hMSH2-hMSH6 sliding, and hMSH2-hMSH6 binding, are significantly faster. Our model explains this apparent discrepancy. The first and second hMSH2-hMSH6 clamps load on the scale of minutes. However, to load a third clamp, both loaded hMSH2-hMSH6 clamps must simultaneously move toward the nucleosome to provide enough space for the third hMSH2-hMSH6 clamp. This is strongly disfavored by entropy. Loading even larger numbers of hMSH2-hMSH6 clamps requires the coordination of nucleosome unwrapping with simultaneous sliding of multiple clamps, which is further disfavored by entropy. As a consequence, the number of hMSH2-hMSH6 clamps required for disassembly is three to four and the rates for loading the fourth clamp are similar to the rates for nucleosome displacement (Fig. 4.10). We conclude that the binding of multiple hMSH2-hMSH6 clamps is the rate-limiting step in nucleosome displacement.

Our model not only quantitatively explains recent experiments by Javaid et al. [116] and North et al. [117], but also explains the apparently contradictory studies that report nucleosomes are a barrier to hMSH2-hMSH6 sliding [125, 132]. The studies by Li et al. [125] observed no nucleosome disassembly by hMSH2-hMSH6. While their construct was similar
to studies by Javaid et al. and North et al., Li et al. only used the 601 positioning sequence and unmodified histone octamers. As discussed above, we find that the 601 sequence largely prevents disassembly by hMSH2-hMSH6 with unmodified histone octamer, which was confirmed by North et al. [117]. Gorman et al. [132] reported that nucleosomes create a boundary to yeast MSH2-MSH6 diffusion. Their experiments involve single molecule tracking of quantum dot labeled MSH2-MSH6 clamps on lambda DNA that contains nucleosomes but no mismatch. They track MSH2-MSH6 along the DNA after washing out free MSH2-MSH6. Because there is no mismatch and no additional MSH2-MSH6 to load onto the DNA, multiple MSH2-MSH6 clamps cannot be loaded onto the DNA between two nucleosomes. Thus, under the experimental conditions used by Gorman et al., the pressure of the one dimensional gas cannot reach the values required to disassemble a nucleosome and the nucleosome becomes a barrier to the sliding of a single MSH2-MSH6 clamp as they report.

The successful application of our nucleosome unwrapping free energy landscape indicates that it will prove useful for modeling a wide variety of experiments involving nucleosome unwrapping, and many different interactions between DNA binding proteins and nucleosomes in vivo. In particular, when a protein does not interact specifically with either the DNA or the nucleosome, as is the case with hMSH2-hMSH6, our free energy landscape completely describes the dynamics of the system. Even when this is not the case, the landscape should prove an invaluable tool. The model will facilitate the separation of the DNA-histone binding from additional interactions. This in turn will allow for determination of interactions between the nucleosome and other protein complexes, such as chromatin remodelers that disassemble, reposition and unwrap nucleosomes.
Bibliography


Appendix A

Calculation of cyclization $J$ factors

This section gives the details of the transfer matrix approach we use to calculate the thermal expectation values for $J$ factors. We closely follow the work of Yan, Kawamura and Marko as described in [47]. We expand their model for the flexibility of duplex DNA which may form melted bubbles to explicitly include the known free energy costs to form bubbles of different sizes as a function of DNA sequence and temperature. For a more general and complete formulation of the transfer matrix method discussed here, please see [47].

As described in the Eq. 2.2, the free energy for a melted bubble containing $L$ open DNA base pairs is

$$
E_{\text{Bubble}} = \mu_{\alpha L} + \sum_{n=\alpha}^{\alpha+L-1} \frac{A_L}{2b_L} \theta_n^2 \times k_B T
$$

$$
\approx \mu_{\alpha L} + \sum_{n=\alpha}^{\alpha+L-1} \frac{\beta_L T_0}{2} \left(\hat{t}_{n+1} - \hat{t}_n\right)^2,
$$

(A.1)

where $\mu_{\alpha L}$ is the melting free energy for a bubble containing $L$ open base pairs starting at base $\alpha$, $A_L$ is the persistence length for a bubble containing $L$ open base pairs, $b_L$ is the segment length for an unpaired nucleotide, $\theta_n$ is the angle between two adjacent DNA segments, $\beta_L$ is the bending rigidity for a bubble containing $L$ open base pairs, $T$ is the temperature, $T_0$ is the reference temperature for the bending rigidity, and $\hat{t}_n$ is a tangent vector describing the orientation of a DNA segment starting at base $n$ defined by $\hat{t}_{n+1} \cdot \hat{t}_n = \cos(\theta_n)$. For finite persistence length, $\theta_n$ will be small, so it is valid to make the substitution $\theta_n^2 = (\hat{t}_{n+1} - \hat{t}_n)^2$ as shown. The bending rigidity is given by $\beta_L = A_L/b_L \times k_B T$ at a reference temperature of $T_0 = 294K$. The persistence length explicitly decreases with temperature, while the bending rigidity has no explicit temperature dependence. In our calculations, we treat the bending rigidity as constant with respect to temperature, so the
persistence length scales with temperature as \( A_L = \beta_L \times b_L/k_BT \). Writing the bending free energy in terms of \( \beta_L \) makes it clear that the bending energy scales like \( 1/T \).

We may then write the total free energy for a base-pairing configuration \( S \) of DNA as

\[
E_S = \sum_{i=1}^{N-1} \left[ \mu_{\alpha_i} L_i + \frac{\beta \lambda_i T_0}{2} \left( \hat{t}_{i+1} - \hat{t}_i \right)^2 \right],
\]

(A.2)

where \( N \) is the total number of base pairs in the DNA sequence. If base \( i \) is contained in a melted bubble in state \( S \), \( L_i \) is the size of this bubble and \( \alpha_i \) is the first base pair contained in the bubble. If base \( i \) is in a double-stranded region, then \( L_i = 0 \) and \( \alpha_i = i \), and as \( \mu_i = 0 \) for all \( i \) and \( \beta_0 = \beta_{ds} \), there is no melting free energy associated with this base pair and it will have a persistence length equal to duplex DNA. We set \( L'_i = L_i \) for \( L_i > 0 \) and \( L'_i = 1 \) for \( L_i = 0 \), to avoid division by zero.

To calculate the probability of loop closure with no applied force, called the \( J \) factor, we evaluate the thermal expectation value of \( \delta^3(\sum_{n=1}^{N-1} b \hat{t}_n) \delta(\hat{t}_N - \hat{t}_1) \). The first delta function requires that the end-to-end distance is zero, while the second requires that the tangent vectors to the polymer are parallel at the two ends. In other words, we do not allow a kink at the point where the ends join together. We use a Fourier transform to decompose the first delta function into contributions from individual chain segments [47]

\[
\rho(0) = \int \frac{d^3k}{(2\pi)^3} \int d^2\hat{t}_1 \cdots d^2\hat{t}_N \delta(\hat{t}_N - \hat{t}_1) \sum_S \exp \left[ -E_S + i b \cdot \sum_{j=1}^{N-1} \hat{t}_j \right]
\]

(A.3)

where the sum over \( S \) denotes summing over all possible configurations of melted bubbles that may form for the DNA segment. The numerator of Equation A.3 gives the partition function for DNA bending subject to the constraint that it must form a closed loop with parallel ends at the boundary. The denominator gives the total partition function for the polymer with no constraints on the boundary conditions for its ends.

Using the transfer matrix

\[
T_{k,S}(\hat{t}_j, \hat{t}_{j+1}) = e^{i b \cdot \hat{t}_j} e^{-\frac{\mu_{\alpha_j} L_j}{L_j} - \frac{\beta \lambda_j T_0}{2} \left( \hat{t}_{j+1} - \hat{t}_j \right)^2}
\]

(A.4)

we may simplify Equation A.3 to

\[
\rho(0) = \int \frac{d^3k}{(2\pi)^3} \int d^2\hat{t}_1 d^2\hat{t}_N \delta(\hat{t}_N - \hat{t}_1) \sum_S \prod_{j=1}^{N-1} T_{k,S}(\hat{t}_j, \hat{t}_{j+1})
\]

(A.5)

The transfer matrix may then be expanded in terms of spherical harmonics as
\langle lm | T_{k,S} | l'm' \rangle = \int d^3\hat{t}_1d^3\hat{t}_2Y_{im}(\hat{t})T_{k,S}(|\hat{t}, \hat{t}'\rangle Y_{l'm'}(\hat{t}')
\begin{align*}
= 4\pi \sum_{l_3} i^{l_3}(-1)^m \delta_{mmm'}j_{l_3}(bk) \exp \left(-\beta_{L_i} \frac{T_0}{T}\right) i_{m'} \beta_{L_i} \frac{T_0}{T} \frac{\mu_{\alpha_{l_i} L_i}}{L_i} \left(\begin{array}{ccc}
l & l' & l_3 \\
m & -m & 0 \\
0 & 0 & 0
\end{array}\right)
\times (2l_3 + 1) \sqrt{(2l + 1)(2l' + 1)} \left(\begin{array}{ccc}
l & l' & l_3 \\
m & -m & 0 \\
0 & 0 & 0
\end{array}\right)
\end{align*}

where \(j_i\) are spherical Bessel functions of the first kind, and \(i_i\) are modified spherical Bessel functions of the first kind. The arrays in braces are Wigner-3\(J\) symbols, which represent integrals over products of spherical harmonics [47, 133].

When \(k = 0\), the transfer matrix reduces to

\begin{align}
\langle lm | T_{0,S} | l'm' \rangle = 4\pi \exp \left(-\beta_{L_i} \frac{T_0}{T}\right) i_{m'} \beta_{L_i} \frac{T_0}{T} \frac{\mu_{\alpha_{l_i} L_i}}{L_i} .
\end{align}

The unconstrained partition function (the denominator of Equation A.5) then becomes

\begin{align}
\int d^3\hat{t}_1d^3\hat{t}_N \sum_{S} \prod_{j=1}^{N-1} T_{0,S}(\hat{t}_j, \hat{t}_{j+1}) = 4\pi \left\langle 00 \right| \sum_{S} \prod_{j=1}^{N-1} T_{0,S}(\hat{t}_j, \hat{t}_{j+1}) \left| 00 \right\rangle
= 4\pi \sum_{S} \prod_{j=1}^{N-1} \exp \left(-\beta_{L_i} \frac{T_0}{T}\right) i_{0} \beta_{L_i} \frac{T_0}{T} \frac{\mu_{\alpha_{l_i} L_i}}{L_i} (A.8)
\end{align}

where the integration over tangent vectors with free boundary conditions is equivalent to \(4\pi\) times the \(l = 0\) state. The expression \(4\pi \exp \left(-\beta_{L_i} \frac{T_0}{T}\right) i_{0} \beta_{L_i} \frac{T_0}{T}\) in Equation A.8 gives the entropy of a base pair due to the flexibility of the DNA segment. However, the entropy contribution to the free energy of melted bubbles is already included in \(\mu_{\alpha_{l_i} L_i}\). Therefore, to avoid double counting we divide both \(\langle 00 | T_{0,S} | 00 \rangle\) and \(\langle lm | T_{k,S} | l'm' \rangle\) by \(4\pi \exp \left(-\beta_{L_i} \frac{T_0}{T}\right) i_{0} \beta_{L_i} \frac{T_0}{T}\), giving us

\begin{align}
\langle 00 | T_{0,S} | 00 \rangle &= e^{-\frac{\mu_{\alpha_{l_i} L_i}}{L_i}} .
\end{align}

\begin{align}
\langle lm | T_{k,S} | l'm' \rangle &= \sum_{l_3} i^{l_3}(-1)^m \delta_{mmm'}j_{l_3}(bk)(2l_3 + 1) \sqrt{(2l + 1)(2l' + 1)}
\times \frac{i_{l'} \beta_{L_i} \frac{T_k}{T}}{i_{0} \beta_{L_i} \frac{T_0}{T}} \left(\begin{array}{ccc}
l & l' & l_3 \\
m & -m & 0 \\
0 & 0 & 0
\end{array}\right)
\end{align}

The \(J\) factor is the probability of loop closure expressed as a concentration, meaning
that [47]

\[ J = \frac{4\pi}{N_A} \rho(0), \]  

(A.11)

where \( N_A \) is Avogadro’s number. Combining Equations. A.5, A.9, A.10 and A.11 gives

\[
\begin{align*}
J &= \frac{4\pi}{N_A} \int \frac{d^3 k}{(2\pi)^3} \sum S \sum_{lm} \langle lm | \prod_{j=1}^{N-1} T_{k,S}^{t_j} (t_j, t_{j+1}) | lm \rangle \\
&= \frac{1}{2\pi^2 N_A} \int \frac{k^2 dk}{S} \sum_{lm} \langle lm | \prod_{j=1}^{N-1} T_{k,S}^{t_j} (t_j, t_{j+1}) | lm \rangle \\
&= \frac{1}{2\pi^2 N_A} \int \frac{k^2 dk}{S} \sum_{lm} \frac{\langle lm | \prod_{j=1}^{N-1} T_{k,S}^{t_j} (t_j, t_{j+1}) | lm \rangle}{\langle 00 | \prod_{j=1}^{N-1} T_{0,S}^{t_j} (t_j, t_{j+1}) | 00 \rangle} \exp \left( -\frac{\mu_{n-1,L}}{L'} \right)
\end{align*}
\]

(A.12)

where we have replaced \( \delta(t_N - t_1) \) with \( \delta_{ll'} \delta_{mm'} \) in the numerator to maintain the parallel boundary condition for the two ends of the DNA molecule [47].

The sum over \( S \) may be performed using the recursions

\[
\begin{align*}
Z_{k,n} &= \sum_{L=0}^{L_{\text{max}}} Z_{k,n-L-1} \prod_{i=n-L}^{n-1} \langle lm | T_{k_{\{L,n-1\}}}(t_i, t_{i+1}) | lm \rangle \\
Z_{0,n} &= \sum_{L=0}^{L_{\text{max}}} Z_{0,n-L-1} \prod_{i=n-L}^{n-1} \exp \left( -\frac{\mu_{n-1,L}}{L'} \right),
\end{align*}
\]

(A.13)

where \( L_{\text{max}} \) is the maximum number of basepairs contained in a melted bubble (4 bp in our calculations), \( L' = L \) for \( L > 0 \) and \( L' = 1 \) for \( L = 0 \), and \( T_{k_{\{L,n-1\}}}(t_i, t_{i+1}) \) is evaluated in the state \( S \) where \( L_i = L \) and \( \alpha_i = n - 1 \). \( Z_{k,n} \) may be evaluated using the boundary conditions that \( Z_{k,0} = \langle lm | T_{k_{\{0,0\}}}(t_0, t_1) | lm \rangle \) and \( Z_{k,n} = 0 \) for \( n < 0 \). Similarly, \( Z_{0,0} = \exp (-\mu_{0,0}) \) and \( Z_{0,n} = 0 \) for \( n < 0 \). We may then rewrite Equation A.12 as

\[
J = \frac{1}{2\pi^2 N_A} \int \frac{k^2 dk}{S} \sum_{lm} \frac{Z_{k,n-1}}{Z_{0,n-1}},
\]

(A.14)

replacing the sums over \( S \) using the recursions given in Equation A.13.

The evaluation of spherical harmonics and Wigner-3\( J \) symbols, the multiplication of transfer matrices, and the integration over \( k \) are all performed numerically using the GNU Scientific Library [109]. The sums over \( l \) and \( m \) are truncated after reaching a large enough value that the calculated value of \( J \) converges. For the 200 bp \( \lambda \) sequence we find this is for \( l = 20 \), and for the 116 bp sequences this is for \( l = 24 \).

Finally, we must account for twisting of the DNA molecule about its axis. We choose DNA molecules having an integer number of turns in their DNA backbone so that cyclization may occur without any twist energy cost. Our bending model assumes that the molecule has no resistance to twisting, but that ends must be aligned with respect to twist for cyclization to occur [134]. Since the ends of our DNA are perfectly aligned with respect to twist, we have to multiply the calculated \( J \) factor by \( \gamma = \sqrt{2\pi C/(Nk_B T)} \) [134] where
\[ C = 2.4 \times 10^{-28} \text{ J m} \] \( [48, 74, 134] \) is the torsional rigidity of dsDNA.
Appendix B

FITTING BUBBLE FLEXIBILITIES

We model DNA as a semiflexible polymer which can locally increase its flexibility by forming melted bubbles. The frequency with which these bubbles form can be calculated using the known free energies to break DNA base-pairings. However, the flexibility of these bubbles is unknown, so we attempt to determine their flexibilities by comparing our model’s predictions with our experiments. We examine the range of persistence lengths for each different size of melted bubble with which our model can reproduce a good fit to our measured $J$ factors for all three DNA sequences. This allows us to place constraints on the values of each bubble persistence length, as shown in Table 2.1 of the main text.

As described in the main text, we require that all bubble persistence lengths be greater than or equal to that of two parallel strands of ssDNA (2 nm [94]) and less than or equal to that of dsDNA. In each of Fits 1-4 we then attempt to set one bubble persistence length to its maximum or minimum value which satisfies this requirement and still gives a reasonable fit to our experimentally measured $J$ factors, while constraining the persistence lengths of all larger bubbles to be the same. Figure B.1 shows that all five fits are in good agreement with our data, and that Fits 2-5 produce nearly indistinguishable results.

Fit 1, where all bubble persistence lengths are set equal, does not agree quite as well with our measured $J$ factors for the 200 bp λ fragment as the other fits. This supports our suggestion that bubble persistence lengths should decrease with bubble size. However, since the overall agreement of Fit 1 is still fairly good, we cannot rule out the possibility that all bubbles could have the same persistence length based on our experiments. We believe that Fit 5 represents the most physically reasonable set of bubble persistence lengths which fit our data because it uses bubble persistence lengths which increase smoothly with bubble size. However, as it does not give a significantly better fit to our data than Fits 1-4, our experiments do not provide any confirmation for this assertion.
Figure B.1: Fitting of measured $J$ factors as a function of temperature. In all plots, the predictions of the worm-like chain are shown by the dashed line, and the predictions of the sequence dependent Yan-Marko model for Fits 1-5 are shown by the solid green line, the dashed blue line, the dotted red line, the dot-dashed yellow line, and the double dot-dashed purple line, respectively. (a) A 200 bp fragment of $\lambda$ DNA, with our experimental data shown by the open triangles. (b) The 116cl sequence (open diamonds) is a 116 bp sequence designed to minimize the formation of local bubbles, while (c) the 116o sequence (open circles) is designed to readily form local bubbles in several locations.
Appendix C

Twist Measurements

Du et al. measure $J$ factors as a function of sequence length for sequences between 105 and 130 bp, and find a maximum at 116.5 bp [75]. This means that a 116 bp sequence should be nearly relaxed with respect to twist. However, one explanation for the difference in $J$ factor we measure between 116o and 116cl at 23°C is that 116o might not be relaxed with respect to twist. If this were the case, then the increase in $J$ factor we see as a function of temperature could be understood as a relaxation of twist rigidity, resulting in less twisting energy needed to cyclize 116o at higher temperatures.

We measure an increase in $J$ factor for the 116o sequence from 0.006 nM at 23°C to 0.063 nM at 37°C, or a ratio of 10. The worm-like chain also predicts an increase in $J$ factor with temperature because the persistence length decreases with temperature. The worm-like chain prediction at 37°C is 2 times its prediction at 23°C, so we measure an increase in $J$ factor with temperature 5 times that predicted by the worm-like chain model.

It is reasonable to expect that since the persistence length decreases with respect to temperature, the torsional rigidity might as well. Since we are unsure of the magnitude of this effect, we will consider the worst possible case - that the torsional rigidity relaxes effectively to zero at 37°C. The $J$ factor predicted by the worm-like chain when the torsional rigidity is relaxed, $J_{avg}$, is the linear average over the fluctuations in $J$ factor with length for a nonzero torsional rigidity. As discussed in Appendix A, the peak of the $J$ factor with respect to length is higher than what would be predicted for zero rigidity. At the peak the twist forms a perfect integer number of helical repeats, which is required for cyclization, while if there were no torsional rigidity the ends would fluctuate and explore all possible helical repeats. We would need this predicted $J$ factor with zero torsional rigidity to be 4.2 times our measured $J$ factor for 116o, or $J_{avg} = 0.025$ nM. To fit our measured $J$ factor for the 116o sequence, the sequence length $n*$ where the $J$ factor reaches a maximum because the sequence has an integer number of helical repeats would have to be shifted by 3 bp away from 116 bp, and this maximum value would be over an order of magnitude larger than the $J$ factor measured for 116o.
Figure C.1: The measured $J$ factors for 114o, 116o and 118o at 23°C are shown by black filled diamonds. We overlay the predictions of the worm-like chain model with a torsional rigidity of $C = 2.4 \times 10^{-28}$ J m and a helical repeat of 10.5 bp [48, 74, 134] for three different values of $n*$, the chain length which is torsionally relaxed when cyclized, and $J_{\text{avg}}$, the $J$ factor predicted for a 116 bp sequence if the torsional rigidity were zero. The solid blue line indicates the best fit to the data, and corresponds to $n* = 117.4$ bp and $J_{\text{avg}} = 0.0035$ nM. The green dashed line has $n* = 118.0$ bp and $J_0 = 0.0043$ nM and represents the largest deviation of $n*$ from 116 bp which can fit our data. The red dotted line has a $n* = 119.0$ bp and $J_{\text{avg}} = 0.025$ nM, and shows the deviation that would be required to produce the temperature dependence we measure for 116o from relaxation of twist alone.
To rule out this possibility, we measure the $J$ factors for 118o and 114o, variants of the 116o sequence with 2 more or less base pairs, respectively. Our results are shown in Figure C.1, and show very little change in $J$ factor with respect to sequence length. The large error in our measured $J$ factor for 114o is because it is difficult to detect $J$ factors smaller than about 0.003 nM in our cyclization experiments. However, we are able to place a good upper bound on the value for 114o and measure 118o accurately. Our results clearly show that the 116o sequence must be nearly relaxed with respect to twist. If 116o had a twist far away from an integer helical repeat, as shown by the dotted red line in Figure C.1, then we should measure a $J$ factor of nearly 0.1 for 118o. Since this is not the case, we conclude that we are in the regime where we may safely disregard any temperature dependent changes in twist rigidity when modeling our data.
Appendix D

Ligatable ends for high temperature ligases

Ligases which operate at high temperature, such as Taq ligase, cannot ligate DNA substrates containing the 4 base pair (bp) ligatable overhangs created by HindIII or other common restriction endonucleases. As shown in Fig. D.1a, Taq ligase requires ends which are at least 9 bp long, so we will not simply be able to use the same DNA constructs with HindIII ligatable ends we created for cyclization experiments at low temperatures.

The restriction enzyme TspRI (10 units/µl; NEB) cuts DNA at the site

\[
\begin{align*}
5' & \ldots NNCAGTGNV \ldots 3' \\
3' & \ldots NNGTSCGNN \ldots 5'
\end{align*}
\]

shown with the top strand of the DNA oriented from the 5' end to the 3' end and the bottom strand oriented from 3' to 5'. C, A, T, and G represent the bases cytosine, adenine, thymine and guanine, N can be any base, and S can be either C or G. The positions where TspRI cuts the DNA backbone are indicated by ▲ and ▼. Cutting with TspRI leaves a 9 bp overhang on the 3' side of the DNA, with considerable freedom in which specific bases may be included in this overhang.

We choose to use the site

\[
\begin{align*}
5' & \ldots CTCAGTGAGV \ldots 3' \\
3' & \ldots GAGTCACCTC \ldots 5'
\end{align*}
\]

as this site has a melting temperature of only 22°C, and therefore will not spontaneously anneal at room temperature. If we chose a site with more guanines and cytosines, we could form ends which would stick together relatively strongly at room temperature without being ligated. This would complicate our analysis, as we would not be able to distinguish ligated products from these spontaneously annealed ends.

We confirm that Taq ligase will function with the specific overhangs shown above. Fig. D.1b shows that Taq ligase will ligate TspRI ends having this sequence either formed
Figure D.1: Electrophoresis gels showing that Taq ligase will ligate overhangs of 9 bp or more, such as those made by TspRI, but not shorter overhangs. (a) Ligation of annealed primers designed to create ligatable ends of different lengths. Each lane is labeled to indicate the length of the ligatable overhang used. We used palindromic ends of length 4, 6, 8, and 10 bp, and also 9 bp TspRI ends labeled “T”. For each overhang, we performed controls with no ligase, ligations with Taq ligase, and ligations with T4 ligase. In the absence of any ligase, only unligated primers are visible at the bottom of the gel. When T4 ligase is added, all primers ligate with high efficiency to form a band near the middle of the gel labeled “ligated”. Some of the DNA remains bound to the T4 ligase, causing it to run slower and form the band near the top of the gel labeled “bound by ligase”. When Taq ligase is added, ligation only occurs for ends of length 9 (“T”) or 10 bp. Almost all of this ligated DNA remains bound to Taq ligase, likely because Taq has a very high affinity for DNA at room temperature. (b) Ligation of TspRI ends created both by annealing two sets of primers, or by cutting a solid piece of dsDNA with TspRI. Lane “S” shows a solid piece of dsDNA created by annealing two primers (second band is likely improperly annealed primers), Lane “S cut” shows “S” cut by TspRI, and lane “a/b” shows two pieces of dsDNA with 9 bp ligatable ends created by annealing two sets of primers. Each of these are shown with no ligase added, with T4 ligase, and with Taq ligase. Both “S cut” and “a/b” ligate to form the same products with T4 ligase and Taq ligase, confirming that Taq ligase can ligate TspRI ends whether created using TspRI or by annealing primers.
Figure D.2: DNA sequences for high temperature ligation. The sites where 116o is more likely to form melted bubbles are shown in red, with the corresponding stable sites in 116cl shown in blue. The TspRI ligatable overhangs are underlined and printed in bold. (a) and (b) show analogues of 116o and 116cl with the HindIII sites replaced with TspRI sites. (c) and (d) show these sequences further modified to decrease the probability of unpeeling from the ends. The sequence (CG)$^4$ (shown in green) is added at each end, and the remaining sequences is shifted by 7 bp toward the 3’ end to accommodate this addition without removing any of the melting sites.

by annealing together two sets of primers to leave the appropriate overhangs, or by cutting a solid piece of DNA with TspRI.

We modify our sequences for low temperature cyclization (Fig. 2.1) to include this TspRI site. As shown in Fig. D.2a and Fig. D.2b, we replace the HindIII with our TspRI site, and also make the region between the restriction sites 3 bp shorter, such that the total sequence including the restriction site remains the same length. Otherwise, the sequences 116To and 116Tcl are identical to 116o and 116cl.

In addition to changing the ligatable overhangs to allow ligation with a high temperature ligase, we are also concerned with the stability of the DNA near the ends of these constructs at high temperatures. If the DNA base pairings unpeel, breaking apart successively starting at one of the free ends, this would likely suppress ligation. Furthermore, such unpeeling would be more energetically favorable in any construct which already has a melted bubble near the end. Since our model predicts that such melted bubbles are required to promote rapid cyclization, unpeeling could suppress cyclization by destabilizing complexes containing melted bubbles. This could prevent us from accurately measuring the impact of melted bubbles on flexibility.

We therefore also create additional versions of the 116o and 116cl sequences shown in Fig. D.2c and Fig. D.2d where we increased the GC content near the ends of the sequences to increase their stability. We still keep the sequence the same except near the ends, but have to shift the sequence by 7 bp to fit in the extra G and C bases near the ends.
while maintaining all of the sites which promote formation of melted bubbles. We perform cyclization experiments with both the 116Fo and 116Fcl sequences and the 116To and 116Tcl sequences in order to test whether unpeeling impacts our measurements.
Appendix E

EXPERIMENTAL METHODS

E.1 Cyclization of DNA fragments with T4 ligase

1. Determine reaction volume
   - Each reaction should contain about 1 pmole DNA.
   - This means the volume will be 100 µl for 10 nM, and 3000 µl for 0.33nM.
   - The max volume without needing to concentrate aliquots is 200 µl (20 µl per aliquot).
   - The max volume using linear polyacrylamide (LPA) precip in 1.5 ml tubes is 3000 µl.

2. Prepare Heat Baths
   - Set one heat bath to the temperature at which ligations will be performed.
   - Set one heat bath to 65°C for heat killing ligase.

3. Prepare Aliquot tubes
   - Label one 1.5 ml centrifuge tube for each aliquot. Usually take 9 samples per experiment, at 0, 1, 2, 4, 6, 8, 12, 16 and 32 minutes.
   - Add 1% (e.g. 1 µl for every 100 µl) reaction volume 0.5M EDTA to each aliquot tube. The EDTA halts ligation.

4. Prepare Tubes for Ligase Dilution
   - Defrost 10X T4 ligase buffer and 100X BSA. Make sure the buffer does not contain any precipitates - if so it may be bad. Mix ligase buffer well with a vortex mixer, but do not vortex BSA (it is a protein).
• Dilute enough 10X ligase buffer and 100X BSA to make a stock containing 1X of both ligase buffer and BSA. (i.e., for 100 µl add 89 µl dH20, 10 µl 10X lig buffer, and 1 µl 100X BSA). Place the diluted buffer on ice.

• You will need enough 1X buffer to dilute the ligase (usually 200 µl to get 1:100 and plus 400 µl for each different concentration needed) plus the total reaction volume.

5. Prepare Reaction Mixture

• Add enough 1X buffer to give 120% of the reaction volume after DNA is added (the extra 20% volume will be used for ligation and quenching tests).

• Turn off the lights if desired. Light can photobleach the fluorophores used for labeling, but will not make a detectable difference over the timeframe of the experiment. Do avoid light exposure for stock solutions containing fluorophores.

• Add the desired amount of Cy5 labeled DNA, which has been HindIII digested and then purified on the Waters column.

• Mix well by pipetting up and down.

6. Prepare Tube for Excess Ligation

• Label one tube for excess ligation assay.

• Add 10% reaction volume of the mixture prepared in the previous step. Note - if the concentration of DNA used is very low, it may be necessary to prepare a separate reaction mixture with higher concentration of DNA to use for ligatability testing.

• Add 1 µl of pure T4 ligase. T4 ligase is stored in a -20°C freezer box in glycerol (so it remains liquid in the freezer). It must remain cold at all times - return to the freezer as soon as possible (after step 7).

• Note the time ligation is started. It should run for at least 30 minutes.

• This tube tests the ligatability of the HindIII digested ends. Ideally, no linear monomer should remain, and most DNA should be ligated into long linear polymers or circles. In these experiments, ligatability is the fraction of DNA not left in the bands for linear monomer or linear dimer.

7. Dilute Ligase

• Dilute 2 µl ligase into 198 µl cold 1X ligase buffer prepared in step 3 to make a "400 U/ml" dilution. Mix well by gently pipetting up and down or flicking.
• Ligase stock is 400,000 U/ml, and when adding ligase to the reaction we dilute it an additional 1:10, so to get the dilution factor needed divide 40,000 by the desired reaction concentration in U/ml.

• Prepare tubes containing enough 1X buffer to dilute the "400 U/ml" stock to the desired concentration for each reaction. Place each tube at the reaction temperature. DO NOT add the ligase until just before starting the reaction. For example, to get 100 u/ml prepare a tube with 300 µl 1X buffer, and then add 100 µl 400 U/ml just before starting the reaction.

8. Prepare the Quenched Tube

• Add 10% reaction volume of reaction mixture to aliquot tube 0.

• Pipette down on the EDTA already in the tube, and pipette up and down to mix.

9. Perform the Ligation Time Course

• Prepare a reaction tube containing 90% reaction volume of reaction mix.

• Place both the reaction tube and aliquot tube 0 in the heat bath so that they can warm up to the temperature at which ligations will be performed.

• Dilute the 400 U/ml ligase to the desired concentration in the preheated 1X buffer prepared in step 7. Do this approximately 1 min before starting the reaction (ligase begins to lose activity when heated).

• Add 10% volume of this diluted T4 ligase to the reaction. Start a stopwatch to measure the time elapsed since ligase was added. Mix well by pipetting up and down.

• Also add 1% volume diluted T4 ligase to aliquot tube 0.

• Take 10% reaction volume aliquots at desired times (usually 1, 2, 4, 6, 8, 12, 16 and 32 minutes). Pipette down on the 1% volume EDTA already in the tube and pipette up and down to mix.

10. Inactivate the ligase

• If necessary, quickly spin down the aliquot and ligation tubes to get all solution to the bottom of the tube.

• Add 1 µl of 1 ug/µl proteinase K to each aliquot tube.

• Place all aliquot tubes and the excess ligation tubes at 65°C for 20 minutes to heat kill.
• Spin down all tubes again half way through the heat killing to return all solution to the bottom of the tubes.

11. Concentrate the samples if needed.

• A maximum volume of 20 µl can be loaded in each lane of an acrylamide gel, so the samples must be concentrated to this volume using LPA precipitation or other means (see Appendix E.2).

12. Visualize on a 6% Acrylamide Gel

• For 116 bp sequences, the band for linear trimers runs on top of the monomer circle in a 5% acrylamide gel, so we must use 6% instead. For different size DNA fragments, a different percentage may be required to optimally resolve all bands.
• Add ficoll loading buffer to 3% final concentration (20% sample volume for 15% ficoll)
• Centrifuge briefly at max speed to make sure ficoll is mixed with sample and no sample is on the sides of the tubes.
• Load at maximum 20 µl of each sample
• Run at 300 V for 1.5 hours.

13. Image the Gel

• Image on a Typhoon imager.
• Best results are usually obtained with these settings:
  – Fluorescence: Cy5 dye (622nm)
  – PMT: 600V
  – normal sensitivity
  – 200 µm pixel size
• Image leaving the gel on the large glass plate covered with celophane, and with water between the gel and the glass platen of the imager.
• Quantify the bands using ImageQuant or ImageJ.
  – If using ImageJ, a version capable of properly rescaling Typhoon images must be used.
  – Each pixel of a Typhoon imager .gel data files contains the square root of the measured photon counts times 21025. Therefore, assuming this raw data is linearly proportional to the measured intensity will introduce a bias into the measurements.
E.2 DNA precipitation with linear polyacrylamide

Linear polyacrylamide (LPA) can be used to help ethanol precipitate DNA which is either too dilute or too short for an ordinary ethanol precipitation to be effective [91].

LPA Precip Mix preparation:

- 950 µl 0.25% LPA
- 3.5 ml saturated ammonium acetate
- 45.5 ml 95% EtOH (molecular biology grade)

Mix well in a 50 ml tube by inverting the tube. Store at -20°C.

Precipitation Procedure:

1. Mix the LPA Precip Mix by inverting.
2. Add 3 volumes LPA Precip Mix to sample and vortex well.
3. Freeze in liquid nitrogen, or place at -80°C overnight.
4. Spin 20 min at max speed (20,000g) at 4°C.
5. Gently dump out the contents of the tube (if being cautious, set aside until done).
6. Add 2-3 volumes 70% EtOH. Vortex extremely vigorously - make sure the pellet comes off the tube wall. **Note:** DO NOT vortex the pellet in an ordinary ethanol precipitation without LPA.
7. Spin 10 min at max speed (20,000g) at 4°C.
8. Gently dump out the contents of the tube (if being cautious, set aside until done).
9. Spin quickly at max speed (20,000g) at 4°C.
10. Pipette off remaining liquid with P200 (if being cautious, set aside until done).
11. Spin 1 minute in speedvac at low drying speed with heat off.
12. Add 20 µl (or more) of final buffer. If a volume of less than 20 µl is needed, then you must pipette up and down over the pellet about 50X now, and pipette again after heating in the next step.
13. Heat at 37°C for 30-60 min. The pellet usually disappears in this step.
14. Vortex thoroughly at modest speed (fluid should remain mostly in bottom half of tube). If using <20 µl pipette up and down instead.

15. If the pellet is still visible, pipette up and down over it until it disappears, or repeat steps 13 and 14.

**E.3 Procedure for generating single-stranded plasmid with M13 helper phage**

Adapted from the protocol provided by New England Biolabs “Use of M13KO7 Helper Phage for isolation of single-stranded phagemid DNA”.

**Materials:**

- M13K07 Helper phage (NEB N0315S)
- Kanamycin
- 2.5 M NaCl/20% PEG 8000 at 4°C (PEG 8000 Amresco 0159-500G)
- Phenol (NOT phenol chloroform) (Sigma P4557-400ML)
- Chloroform
- 2.5 M NaOAc (sodium acetate) pH=4.2 (pH IS IMPORTANT)
- 95% ethanol stored at 4°C (Sigma 459844-500ML)

**Wear** gloves for this procedure – everything is bad.

1. Transform plasmid into appropriate F’ strain of bacteria
   - plasmid must contain f1/m13 origin of replication.
   - plasmid must NOT contain kanamycin resistance.
   - XL1Blue cells are an F’ strain, DH5alpha are NOT.

2. Innoculate **50 ml** LB (no antibiotic) with a fresh colony
   - Can scale up volume for higher yield, all volumes which must be scaled together are marked in **red**.
   - first innoculate a 5 ml culture and grow overnight at 37°C on shaker.
   - withdraw 100 µl from LB flask before adding bacteria to use as a blank for the nanodrop.
• add 50 µl to a flask containing 50 ml LB.
• place on shaker at 37°C and grow until A600=0.05 at 1 cm pathlength
  – this means A600 must be 0.005 for the nanodrop’s 1mm path
  – every 15 minutes withdraw 100 µl
  – vortex well, then immediately spec 3 µl of sample

3. Add 50 µl M13K07 helper phage

  • Let sit at room temperature for 30-45 minutes.
  • Incubate for 1 hour at 37°C with shaking.

4. Add kanamycin to final concentration of 70 ug/ml

  • for 50 ml, add 350 µl of kanamycin 10 ug/µl.
  • grow at 37°C with shaking overnight (14-18 hours).

5. Spin culture at 3000 RPM for 10 minutes

  • divide equally into 50 ml conical tubes (add no more than 30 ml per tube).
  • spin in Avanti centrifuge at 3000 RPM for 10 minutes.
  • place supernatant in new tubes, and discard pellet (the pellet is bacterial cells which do not contain ssDNA).
  • repeat spins 2-3 times, until no visible pellet forms.

6. Precipitate ssDNA plasmids

  • Pipet supernatant into a beaker.
  • add 0.2 volumes 2.5M NaCl/20% PEG 8000.
  • Incubate at 4°C on a stir plate for at least 1 hour.
    – for better yield incubate overnight.

7. Centrifuge at 20,000g for 10 minutes to pellet ssDNA plasmids

  • divide evenly into round bottom polycarbonate tubes.
    – use a balance to make sure tubes are equal.
  • Spin at 20,000g for 10 minutes at 4°C in the Avanti centrifuge.
  • Decant supernatant and spin again briefly.
  • remove remaining supernatant.

8. Resuspend ssDNA plasmids in 1.6 ml 0.5X TE
• resuspend by pipetting up and down.
• place in 2 1.5ml centrifuge tubes.

9. Spin in a microcentrifuge at 3000g for 1 minute

• This should pellet any remaining cells.
• Transfer supernatant to new tubes, and discard the pellet.
• The pellet contains bacterial cells, and no ssDNA plasmids.
• Repeat this step until all cells have been removed and no pellet forms.

10. Add 200 µl of 2.5M NaCl/20% PEG 8000 to each tube

• Let sit at room temperature for 5 minutes.
• Spin max speed at 4°C for 10 minutes.

11. Decant the supernatant and spin again briefly

• remove any remaining supernatant with a pipette.

12. Resuspend each pellet in 300 µl 0.5X TE and phenol extract

• Add 300 µl phenol to each tube.
• Vortex very thoroughly.
• Let sit 15 minutes at room temperature.
• Spin max speed for 3 minutes.
• Extract top layer into a new tube.
  • If desired back extract in 100 µl 0.5X TE.
• Repeat this step 2 more times with 50/50 phenol/chloroform.
• Repeat 1 more time with chloroform.
• Add 1/10th volume 2.5M NaOAc pH=4.2, 2 volumes cold 95% ethanol.
• Vortex well.
• Freeze in liquid nitrogen.
• Place at -80°C for 15 minutes.
• Spin max speed at 4°C for 30 minutes.
• Pipette off supernatant, and wash with cold 70% ethanol.
• Pipette off 70% ethanol, and dry on speedvac for 1 minute.

13. Resuspend ssDNA plasmids in 25-50 µl 0.5X TE

• Yield should be >50 ug of single-stranded plasmid.
E.4 BND cellulose extraction of single-stranded DNA

Procedure provided by Chris Cook of Richard Fishel’s research group.

1. Wash 0.5 g BND cellulose three times with 1X TE + 1.7 M NaCl
   - Suspend BND in TE+NaCl.
   - Centrifuge 10 min at 3000 RPM.
   - Discard aqueous layer and repeat.

2. Resuspend BND Pellet in 2X volume TE+1.7M NaCl (BND Pellet should be half of total volume)

3. Annealed DNA should be in TE + 150mM NaCl (Or another buffer would work, but the salt is important)

4. Add: equal volume of well mixed BND cellulose slurry to annealed DNA
   (NOTE: New [NaCl] is 925mM - ideally should be 1M)

5. Mix well by pipetting
   - Cut the end off a P1000 tip with a razor blade to be able to pipette it.
   - Let sit for about 5 min after pipetting.

6. Spin in a centrifuge at max speed for 10 minutes

7. Remove supernatant to a new tube
   - Try not to take any BND particles.
   - To increase yield, back extract by adding additional TE + 150mM NaCl repeating steps 5 to 6.

8. Spin supernatant for 10 min at max speed to pellet any remaining particles, and remove supernatant to a new tube.

9. Run a small fraction of the sample on a 1% agarose gel to confirm that single-stranded DNA (ssDNA) has been removed.
   - Run at 150V for 1 hour.
   - ssDNA will run as a large smear.

10. If not all the ssDNA has been removed, repeat the extraction procedure using BND cellulose in 1X TE + 1M NaCl.
11. Filter using an Amicon .22 micron filter to remove any remaining BND particles

12. Concentrate and buffer exchange sample using a microcon 10
Appendix F

FULL PARTITION FUNCTION FOR
NUCLEIC ACID SECONDARY STRUCTURE

Here, we will describe in detail the calculation of the full restricted partition function $W'_{i,j}$ for a nucleic acid in the presence of a single-stranded binding protein. The algorithm presented below differs slightly from the implementation in the original Vienna RNA package (13) due to the need to explicitly keep track of all unpaired regions (which in the original Vienna RNA package acquire only an irrelevant factor of 1). However, the algorithm given here yields identical results to the standard Vienna RNA package in the absence of binding proteins.

The full restricted partition function where $i$ is paired with an (unspecified) base $k$, $k+1$ through $j$ are unpaired, and the state of $(i+1,k-1)$ is unconstrained is

$$W'_{i,j} = \sum_{k=i+L_{\text{min}}+1}^{j} B_{i,k}O_{k+1,j},$$

where $O_{k+1,j}$ is given by Eq. 3.5 and $B_{i,j}$ is the partition function for the segment $(i,j)$ with $i$ bound to $j$, and bases in the segment $(i+1,j-1)$ unconstrained. This is calculated using

$$B_{i,j} = Bh_{i,j} + Bl_{i,j} + Bm_{i,j},$$

where $Bh_{i,j}$ is the partition function for a hairpin loop, $Bl_{i,j}$ is the partition function for an internal loop, bulge or set of stacked base pairs, and $Bm_{i,j}$ is the partition function for a multiloop. A hairpin loop, described by the partition function

$$Bh_{i,j} = e^{-E_{\text{hairpin}}(j-i-1)/k_BT}O_{i+1,j-1},$$

has $i$ and $j$ paired, and all bases on $(i+1,j-1)$ unpaired, where $E_{\text{hairpin}}(n)$ is the free energy cost to form a hairpin loop. While we here use the same values of $O_{i,j}$ for unpaired segments both external and internal to loops, we could define four different values of the chemical potential $\mu^P(c,i)$ for unpaired segments outside loops, within hairpin loops, within internal
loops, and within multiloops. We could then use Eq. (3.5) to compute a different value of $O_{i,j}$ for each case. This would allow us to consider protein binding to different kinds of loops separately, in case the protein has different affinities for unpaired segments within different kinds of loops or external to loops.

The partition function for an internal loop closed by pairing $i$ with $j$ and $k$ with $l$, with $(i+1,k-1)$ and $(l+1,j-1)$ unpaired (where $i < k < l < j$), is

$$B_{i,j} = \sum_{k=i+1}^{j-L_{\min}+2} \sum_{l=k+L_{\min}+1}^{j-1} e^{-E_{i,j,k,l}^{\text{loop}}/k_{B}T} \times \times O_{i+1,k-1} B_{k,l} O_{l+1,j-1},$$

where $E_{i,j,k,l}^{\text{loop}}$ is the free energy cost to form this internal loop, bulge, or set of stacked base pairs. If $k > i + 1$ and $j > l + 1$, then the structure is an internal loop. If either $k = i + 1$ or $j = l + 1$ then the structure is a bulge. If both $k = i + 1$ and $j = l + 1$ then the structure is a set of stacked base pairs. All needed values of $E_{i,j,k,l}^{\text{loop}}$ for internal loops, bulges, and stacked base pairs have been extrapolated from experiments (15). We note that while use of Eq. (F.4) gives a partition function which scales with $O(n^4)$, the scaling is reduced to $O(n^3)$ by only allowing internal loops smaller than a cutoff size.

The partition function for a multiloop, in which $i$ and $j$ are paired, and there are at least two more stems on $(i+1,j-1)$ is $B_{m_{ij}}$. The free energy of a multiloop is given by the free energy of its components plus an additional cost to close the loop

$$C_{M} = C_{i,j}^{C} + n_{\text{unpaired}} C_{U} + n_{\text{stem}} C_{S}$$

where $C_{i,j}^{C}$ is the fixed cost for a multiloop closed by pairing bases $i$ and $j$, $n_{\text{unpaired}}$ is the number of unpaired nucleotides which form the multiloop, $C_{U}$ is a cost per unpaired base in the multiloop, $n_{\text{stem}}$ is the number of stems contained in the multiloop, and $C_{S}$ is the cost per stem. The partition function for a multiloop is then

$$B_{m_{ij}} = \sum_{k=i+L_{\min}+2}^{j-L_{\min}+2} Q_{m_{i+1,k}} W_{m_{k+1,j-1}} e^{-C_{i,j}^{C}/k_{B}T},$$

where $Q_{m_{i,j}}$ is the partition function for a segment $(i,j)$ containing at least one stem with multiloop closure energy included, given by

$$Q_{m_{i,j}} = W_{m_{i,j}} + \sum_{k=i+1}^{j-L_{\min}+1} W_{m_{k,j}} \times \times \left[ Q_{m_{i,k-1}} + O_{i,k-1} e^{-(k-i)C_{U}/k_{B}T} \right],$$

and $W_{m_{i,j}}$ is the partition function for $(i,j)$ where $i$ is paired with an (unspecified) base $k$, $k + 1$ through $j$ are unpaired, the state of $(i+1,k-1)$ is unconstrained, and multiloop energies are included. It is identical to $W'_{i,j}$ except for the addition of multiloop energies,
and is given by

\[ W_{m_{i,j}} = \sum_{k=i+L_{\text{min}}+1}^{j} B_{i,k} O_{k+1,j} e^{-(j-k)(C^{U}+C^{S})/k_{B}T}. \] (F.8)

The above gives the restricted partition function with all contributions for different types of loops found in the Vienna RNA package. We then need only to insert \( W'_{i,j} \) in place of \( W_{i,j} \) in Eq. 3.2 to calculate the full unconstrained partition function for nucleic acid secondary structure in the presence of binding proteins.