2013-12-12

Molecular Modeling of DNA for a Mechanistic Understanding of Hybridization

Terry Jacob Schmitt
Brigham Young University - Provo

Molecular Modeling of DNA for a Mechanistic Understanding of Hybridization

Terry Schmitt

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Thomas A. Knotts IV, Chair
William G. Pitt
Dean R. Wheeler
William C. Hecker
Morris D. Argyle

Department of Chemical Engineering

Brigham Young University

October 2013

ABSTRACT

Molecular Modeling of DNA for a Mechanistic Understanding of Hybridization

Terry Schmitt
Department of Chemical Engineering
Doctor of Philosophy

DNA microarrays are a potentially disruptive technology in the medical field, but their use in such settings is limited by poor reliability. Microarrays work on the principle of hybridization and can only be as reliable as this process is robust, yet little is known at the molecular level about how the surface affects the hybridization process. This work uses advanced molecular simulation techniques and an experimentally-parameterized coarse-grain model to determine the mechanism by which hybridization occurs on surfaces and to identify key factors that influence the accuracy of DNA microarrays. Comparing behavior in the bulk and on the surface showed, contrary to previous assumptions, that hybridization on surfaces is more energetically favorable than in the bulk. The results also show that hybridization proceeds through a mechanism where the untethered (target) strand often flips orientation. For evenly-lengthed strands, the surface stabilizes hybridization (compared to the bulk system) by reducing the barriers involved in the flipping event.

Additional factors were also investigated, including the effects of stretching or compressing the probe strand as a model system to test the hypothesis that improving surface hybridization will improve microarray performance. The results in this regard indicate that selectivity can be increased by reducing overall sensitivity by a small degree. Another factor that was investigated was the effect of unevenly-lengthed strands. It was found that, when unevenly-lengthed strands were hybridized on a surface, the surface may destabilize hybridization compared to the bulk, but the degree of destabilization is dependent on the location of the matching sequence. Taken as a whole, the results offer an unprecedented view into the hybridization process on surfaces and provide some insights as to the poor reproducibility exhibited by microarrays. Namely, the prediction methods that are currently used to design microarrays based on duplex stability in the bulk do a poor job of estimating the stability of those duplexes in a microarray environment.

Keywords: DNA, hybridization, molecular modeling, microarray

ACKNOWLEDGMENTS

I offer many thanks to my loving wife Alice, my understanding advisor Dr. Knotts, and all the other patient supporters who have helped me along.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

DNA microarrays–high throughput, parallel assays for determining in parallel the genes present in a sample–have been identified as a key technology in genomic sciences and emergent medical techniques. Despite their abundant use in laboratories, microarrays are not used in the clinical setting to their fullest potential. This is due to the fact that reproducible results are difficult to obtain. To date, efforts made to optimize microarrays have not lead to a robust device. The majority of these optimization efforts have focused on laboratory techniques and fabrication processes, but have not addressed the molecular level phenomena upon which microarrays function. The purpose of this research is to provide the needed understanding of these molecular-level phenomena so that better microarrays can be developed.

1.2 Microarrays

A DNA microarray is a flat surface, such as a glass slide, that has had thousands of single stranded DNA (ssDNA) attached to it for the purpose of determining gene expression in a sample. Fodor *et al.* first published an outline of the technology needed to develop both protein and DNA microarrays in the early 1990s [1, 2]. In these papers, Fodor *et al.* covered methods for building single stranded DNA onto the glass plate and analyzing where hybridization has occurred. While alternative methods have been developed to attach ssDNA to a solid substrate [3–5], microarrays still function the same way originally outlined in the Fodor *et al.* paper. Once the ssDNA has been attached to the surface, fluorescently tagged DNA samples are incubated over the chip. Hypothetically, these samples should hybridize with complementary sequences among the ssDNA tethered to the surface. Since the ssDNA

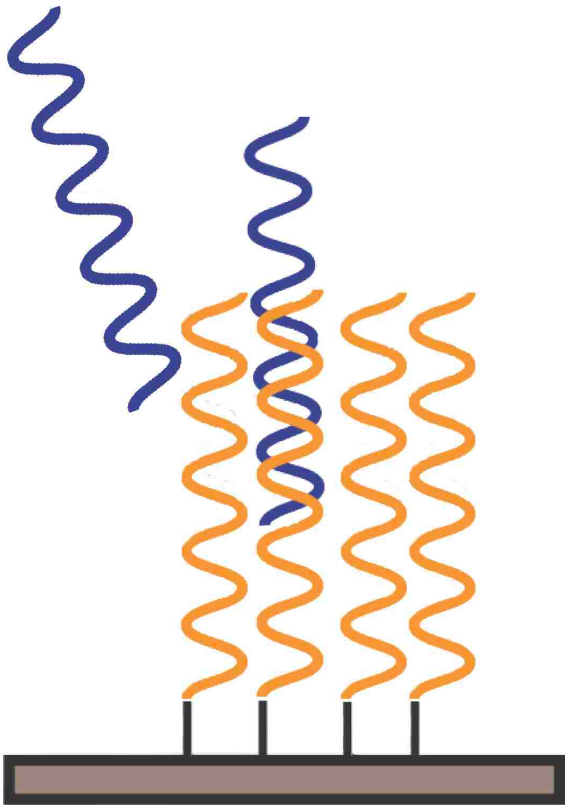

Figure 1.1: A simplified microarray at the molecular level. Target strands hybridize to probe strands attached to the surface.

sequences were attached to specific locations on the microarray, a computer can measure the fluorescent signals given off by the hybridized samples to determine the sequences present in the unknown samples. Figure 1.1 shows a cartoon representation of what a microarray surface looks like and Figure 1.2 gives a representation of how a microarray might appear to the naked eye. An example of how microarrays are used is given below.

Microarrays have revolutionized genomic science with uses in sequencing [6], analyzing DNA and RNA samples [7,8], drug discovery and delivery specialization [9], and monitoring gene expression [10–12]. The study performed by Hayashi *et al.* [13] is characteristic of the capabilities of microarrays. In this work, Hayashi *et al.* monitored the expression of genes in the white blood cells of diabetic and non-diabetic rats before and after feeding. The intent was to find out if diabetic traits could be observed and monitored from the white blood cells. They harvested RNA from the rats before and after feeding and analyzed the RNA to observe which genes were actively being transcribed under different conditions. The RNA

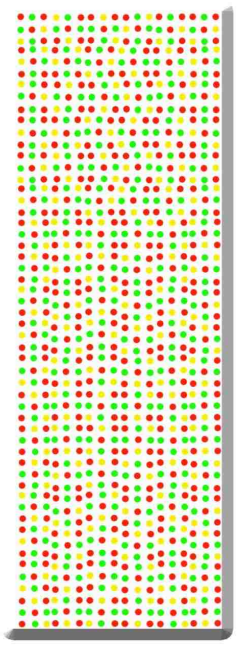

Figure 1.2: A used microarray platform. Computers analyze the fluorescence at varying locations to determine what genes were in the sample.

from the diabetic rats was tagged with a different fluorescent color than the RNA from the control rats. This allowed simultaneous comparative analysis of the two groups. Hayashi *et al.* compared the relative extent of expression of different genes in the diabetic and control rats. They showed that genes linked to diabetes may be monitored by testing just the white blood cells. This finding was important since obtaining and testing white blood cells is much easier than traditional diagnosis methods that require tissue testing samples of liver, adipose, or muscle cells.

Despite successes like the one above, microarrays have yet to be accepted for general use by the medical community [14]. To date, only one microarray is approved by the FDA for use in clinical settings [15]. The purpose of this array is to tailor drug dosage to the unique metabolism of the patient to maximize the effectiveness of the drug while minimizing risk of deleterious side effects. While many other drug treatments could also be tailored like this one, no others have been approved by the FDA. Drug tailoring is not the only possible clinical application of DNA microarrays. As shown above, they may also be used as a diagnostic

tool. By monitoring what genes are being expressed in a patient, doctors can determine which step in a complex biochemical system is leading to previously undiagnoseable health concerns. For example, insomnia patients could be tested to see if any of the hundreds of proteins involved in the circadian clock are not being produced in their body, causing the later steps in the response pathway to fail. Similarly, people worried about specific diseases may be tested to determine if they are expressing genes associated with those diseases.

1.2.1 Microarray Limitations

The problem that keeps microarrays from being widely used is that the results are not consistent [16–18]. Several published works demonstrate the inconsistency of microarrays [19–25]. One notable example is found in the works of three independent research groups studying embryonic, neural, and hematopoietic mouse cells [26–28]. The goal of the independent studies was to determine the genes that carried the "stem cell" traits, or which genes were "stemness genes." Each research group used nearly identical methods and materials, including microarrays made by the same company. Fortunel *et al.* found 385 genes that were common among the three types of stem cells, Ivanova *et al.* found 283, and Ramalho-Santos *et al.* found 230. When the three lists where compared, however, only one gene was found by all three groups [28]. To determine if methodology was the cause of the discrepancy, Fortunel *et al.* compared the results obtained for each type of stem cell and found considerable overlap. They determined that the aims of the experiment could not be achieved with current microarray technologies. Specifically, the variability in data was so high that specific conclusions could not be drawn.

1.2.2 Experimental Optimization

In an effort to eliminate possible inconsistencies originating from poor experimental skill and methods, and to determine the cause of conflicting results, the FDA performed an exhaustive study on microarray performance [29–39]. The study included researchers from government, industry, and academia in a group known as the MicroArray Quality Control (MAQC) consortium [33]. The MAQC consortium evaluated two high-quality RNA samples on seven microarray platforms with three methodologies. Each platform was used at three

independent sites with five replicates at each site. The experimental design included standardization of data reporting, common analysis tools, useful controls to provide confidence in the consistency and reliability, and stringent intersite protocol. In this manner the MAQC consortium was able to separate microarray performance variation from variations due to experimental procedures and reporting methods.

The MAQC consortium tested the microarrays on both the ability to detect genes that were present in the samples and also gene expression levels. The former is an easier test, while the latter is very sensitive to the amount of hybridization occurring. The results showed a concordance rate of 80-95% for multiple tests done at the same site with the same microarray brand. This means that when the scientists compared the gene lists from different replicates of the same experiment, 80-95% of the genes on the lists were the same. Inter-platform comparisons from different sites yielded 60-80% agreement in gene list overlap with only one site dropping below 60%. Comparison of the relative expression level between devices yielded a correlation of $R = 0.87 - 0.69$. The amount of overlap in the results show that genes are being identified by hybridization onto the microarrays at a statistically significant rate. In other words, DNA microarrays can give accurate results in some circumstances, but the cross-site variation shows that accuracy of microarray results can be improved. These improvements must be made if the devices are to be used in clinical settings or for medical diagnostics [37].

At first glance, agreement of 60-80% among different platforms at different sites seems a strong performance, but these results were somewhat biased in that they were obtained using only the strongly expressed genes. To eliminate noise, genes with low level expression were left out. Quantitatively, measurements of the relative amount of change in gene expression of genes expressed at low levels varied as much as 60% [30]. Moreover, since this is an idealized study with ideal samples, the concordance would likely drop if more realistic samples of biological relevance were used [37].

The MAQC consortium pointed out that "the expression patterns generated were reflective of biology." [37] In other words, the results show that hybridization is occurring on the surfaces of the microarrays even if it does not occur with high fidelity. Microarrays are designed based on melting points and relative stabilities of the possible probes in the

sequence of interest. These data, however, are from bulk systems and do not account for characteristics of a microarray system that can change the relative stability of a DNA duplex, such as the presence of a surface and the manner in which multiple probe strands are held in close proximity while binding to a target strand. Therefore, as demonstrated by the MAQC, optimizing only the laboratory procedures is not enough. Although the principle of identifying genes and gene expression levels using hybridization is sound, the engineering of the devices needs to address the hybridization process itself.

The mechanics of the hybridization process, however, are elusive. Experimental approaches have a limited resolution and molecular level data are sparse. This lack of data leaves much unknown about the actual mechanics of the hybridization process. In particular, very little is known about the effects of a surface on a hybridizing duplex or if bulk data can be used to approximate surface conditions. More data of higher resolution details of the hybridization process are needed in order to improve the engineering of DNA microarrays. Data with molecular level resolution is difficult to generate from experiments however. Therefore, it may be necessary to use data generated from computer simulations of molecular level interactions as a first step approximation for how to improve DNA microarray design [40].

Recent improvements in molecular modeling techniques and computational power have opened the doors to the possibility of generating the type of data that would be needed to improve microarray design. Until recently, computer simulations of DNA were too computationally demanding to obtain both the resolution and timescales that are necessary to characterize hybridization. Simulations of sufficient resolution could not reach adequate simulation time to characterize hybridization while simulations of long time frame events did not show the molecular level interactions that govern hybridization. Section 1.3 outlines the past work that has been done and the improvements that have been made to simulation techniques to make possible computer simulations that can generate the type of data that can be used to improve microarray design.

1.3 Increased Understanding via Simulation

1.3.1 Atomistic Simulations

Because hybridization is a single-molecule phenomenon which current experimental techniques cannot currently capture with high resolution, molecular simulations have been used in a limited way to understand microarray behavior. However, the computational demands of DNA simulations have limited the volume of work on the subject. In order to produce significant data about hybridization, the simulation must sample the phase space containing the hybridized duplex, the completely separated strands, and a continuous area of phase space connecting these points. To obtain correct physical properties, this must happen multiple times in a simulation. While atomistic simulations of DNA in bulk have received regular attention [41–48], they generally address stand orientation, persistence length, and hydration shells with little information on hybridization and other phenomena that are vital to improving microarrays. Some prominent all-atom studies that did explore hybridization have been done but the phenomena examined are limited. The works by Hagan *et al.* [49], Maiti *et al.* [50], and Perez *et al.* [42] are of particular interest. Hagan *et al.* studied the kinetic pathway for the binding and unbinding of a terminal base pair of a DNA duplex. Using transition path sampling, a computationally demanding technique, this group was able to simulate the binding events with enough efficiency to obtain statistically significant results. They mapped a detailed pathway of the mechanism for the initial stage of hybridization, but due to the simulation time required to observe this phenomenon, this is the only study to date that has produced these data. Even with the advanced simulation techniques this group was only able to generate the pathway for the *terminal base pair* of a three base pair duplex. Specifically, they held all the bases in the simulation fixed except the cytosine they endeavoured to observe making the flipping transition.

The other two studies are significant due to the size of the systems or the time scales explored. The work by Maiti *et al.* [50] investigated paranemic crossover DNA. Paranemic crossovers are complexes of DNA where four strands are intertwined. Their studies included super-complexes as long as 49 bases per strand for a contour length of $\sim$17 nm. However, as with other atomistic studies, the hybridization process itself was never observed or char-

acterized. The authors started the simulation with the four strands intertwined and simply let the system relax. The work by Perez *et al.* [42] is the first to simulate an atomistic DNA dodecamer on the μs time scale. At this long time scale, partial and total openings were observed, but base flipping and other phenomena needed for DNA melting were not sampled. Both these studies required advanced molecular dynamics simulation techniques, but were still limited by computational demands. The Maiti *et al.* study was only able to simulate 3 ns after equilibration and the Perez *et al.* study required 15 years equivalent of computation time.

In the aforementioned studies, the DNA was simulated to replicate bulk phase behavior. The microarray system is characterized by the presence of a surface and has received less treatment by researchers. Wong and Pettit performed one of the first simulations of DNA interacting with a surface [51]. In their simulation, the double stranded DNA dodecamer is attached to a surface and allowed to equilibrate to its native conformation. The surface was modeled as a glass substrate with epoxides grafted onto it. The DNA was attached to the surface using an amine linker. All interaction parameters were taken from the CHARMM forcefield [52–54]. In the 7 ns molecular dynamics simulation of this study, the DNA did not collapse to the surface. Due to salt induced colloid-like interactions, the DNA tilts towards its nearest neighbor, which was a periodic image of itself in this simulation.

Wong *et al.* completed a second atomistic study of DNA attached to a surface that was extended from 7 ns to 40 ns [55]. The longer time scale allowed two different conformations to be seen. The first was the same tilted confirmation seen in the first experiment. In the second conformation, the DNA remained upright, but the linker collapsed allowing the DNA to come in contact with the surface. In the 40 ns simulation the DNA only made one transition from its initial position to the tilted position and back to the upright position with the collapsed linker. This single transition did not make it possible to determine the most stable conformation.

These studies show that the detail gained from atomistic models is overshadowed by the computational demands. Moreover, only a few studies have been done for strands attached to a surface, and all simulated systems are limited to the amount of simulation time that can be generated computationally. In short, atomistic simulation reported in the

literature have not been able to determine the stability of the DNA duplex due to the fact that the hybridization/melting transitions can not occur on the time scales accessed with these models. A simpler model is needed to capture the hybridization process that is so fundamental to microarray perfomrance.

1.3.2 Meso-scale Simulations

The computational limitations of atomistic models have lead many researchers to use meso-scale models for DNA in both theory and simulation [56]. Simple mathematical and low-resolution models have been developed that can describe some DNA phenomena including the orientational dependence of successive bases and the elastic properties of the molecules [57–72]. Most of these simpler models, however, can not be used to simulate the microarray system because they do not describe melting/hybridization or are not directly applicable to molecular simulations. One example of this is the work by Bruant *et al.* where groups of atoms were represented as beads [71]. The model reproduced bending, torsional, and stretching rigidities, but did not address thermal denaturation or electrostatic interactions. Tepper and Voth published a model that represented DNA as a complex network of beads and springs in a coarse grain solvent [72]. Their model also failed to characterize hybridization.

Several coarse grain models have been proposed that allow melting and hybridization [56]. Drukker and Schatz developed a two site per nucleotide bead-spring model [73]. The model, with one site for the backbone and one for the base, allows for hybridization, but does not account for columbic interactions, describe the major and minor grooves of DNA, or address the mechanical properties of DNA. Another two-site model by Buyukdagli *et al.* accounts for stacking interactions, but still neglects columbic interactions and mechanical properties and does not have the correct geometry [74]. A model developed by Sales-Pardo *et al.*, is a bead-pin model with beads representing the sugar backbone and the pins representing the bases [75]. Despite its advantages, this model does not address the elastic properties of DNA or electrostatic interactions.

The most applicable meso-scale model was produced by Knotts, Rathore, Schwartz and de Pablo. Their model uses three sites to represent a nucleotide, one each for the

phosphate backbone, sugar, and base. Besides accurately portraying the geometry and intramolecular interactions of DNA, this model also closely reproduces melting curve data found in experiments [56]. As this model was originally parameterized, it predicted DNA melting, but the electrostatic repulsion of the backbones prevented the strands from hybridizing. A recent modification of this model by Sambriski *et al.* added solvation effects to make hybridization possible [76–78]. This model allows for the simulation of hybridization on a surface.

Before this study, only one meso-scale DNA model had been used to simulate hybridization of a target to a probe tethered to a surface. Jayaraman, Hall, and Genzer performed two studies attempting to optimize microarray construction [79, 80]. Their first study investigated the effect of probe length on hybridization [79]. It used a self avoiding polymer chain model on a lattice to represent the DNA. This means that each DNA strand was represented by a chain of beads. Each bead represented the length of DNA needed for one turn, or approximately 11 nucleotides. One of the draw backs of the lattice system used in this study is that it discretizes the space that the DNA strands may occupy. This means that the DNA must move through the simulation space like pieces on a peg board instead of strands in a fluid. Additionally, since each bead represented 11 nucleotides, each probe segment was restricted to interact only with its compliment on the target segment. One result of this approach is that the possibility of mismatches is eliminated. But despite these shortcomings in the model, this study was able to show that hybridization is most likely to start at the ends of the strands, and that the segments towards the stand centers would remain bound longer. Unfortunately, due to being restricted to a lattice, no mechanism for hybridization could be obtained. The second study used the same model to investigate the effect of probe density and length, but still suffered from the same weaknesses [80].

1.4 Goal and Outline

The goal of this project is to determine the characteristics of microarrays that need to be reengineered to improve the hybridization efficiency on the chip for improved microarray performance. *In vivo*, DNA is able to unbind, replicate, and bind again smoothly and efficiently with very high fidelity. The consistency with which DNA replication occurs is

likely due to a robust and efficient system that controls hybridization. The results of the MAQC consortium suggest that hybridization on the surface of a microarray is less robust than *in vivo*. Specifically, *this study uses molecular modeling to determine the effect of a surface and other microarray features on the hybridization of DNA*.

Possible characteristics of microarrays that affect DNA hybridization include the presence of the surface, similar target sequences competing to bind with the same probe strand, and hybridization between strands of varying length. Each of these are addressed in this research. The organization of the document is as follows. Chapter 2 contains information on the simulation techniques used and the model chosen to simulate hybridization both in the bulk and on the surface. Chapter 3 contains the work done to overcome the major limitation of previous simulation work–insufficient sampling–by determining the optimal simulation method that ensures the hybridization event occurs with sufficient fequency that reliable data are obtained. Chapter 4 describes the research performed to determine the effect of the surface on the hybridization of a tethered probe to a perfectly matching target. Chapter 5 expands the work performed in Chapter 4 by including a second reaction coordinate to further elucidate the effect of the surface on the hybridization mechanism. Chapter 6 explores the effects of manipulating the hybridizing strands by introducing non-complementary sequences on the stability of hybridized complexes and affecting the accessibility of the bases on the nucleotides. Chapter 7 covers the changes that occur in both the stability of the duplex and the mechanism of hybridization when the target strand is longer than the probe strand. Finally, Chapter 8 summarizes the results obtained and the conclusions reached from these results.

Chapter 2

Methodology

2.1 Simulation Techniques

Computer simulations were used to calculate thermodynamic stabilities and mechanistic pathways of DNA hybridization in the bulk and on a microarray surface. The differences in the thermodynamics and mechanisms between there two cases were then analyzed to determine the characteristics of microarray systems that lead to inconsistencies in microarray results. With an understanding of which characteristics cause changes in the thermodynamics and mechanisms of hybridization, steps can be made towards changing microarray platforms to be more reliable.

Thermodynamic data are obtained by sampling various areas of phase space with enough frequency to determine the probability that the system will reside at that state. To improve phase space sampling, advanced Monte Carlo techniques have been developed [81–84]. These include density of states [85–90], replica exchange [91–93], and umbrella sampling [85, 94, 95]. The weighted histogram analysis method (WHAM) is used with these methods to help calculate the density of states where it is not already explicitly calculated by the method itself [96]. The density of states is then applied to the partition function to find thermodynamic properties via the following equation:

$$X(T) = \langle X \rangle_T = \frac{1}{Q} \sum_i X(U_i)\Omega(U_i)e^{-\beta U_i}. \tag{2.1}$$

In Equation 2.1, X is an arbitrary property, Q is the canonical partition function, $\Omega(U_i)$ is the density of states for energy state i, $\beta = \frac{1}{k_B T}$, $< \ldots >_T$ denotes the ensemble average at temperature T, and the summation over i includes all populated energy states. Due to the

overwhelming degeneracy of an explicit solvent, advanced Monte Carlo methods are more powerful with coarse-grain models where solvent effects are measured implicitly.

Adequate phase space sampling is often impeded by energy boundaries. The system being simulated can become trapped in local energy minima due to a high energy state that blocks the transition path to the global energy minimum. Replica exchange and umbrella sampling are techniques designed to overcome this obstacle and are used in this research. Replica exchange uses multiple sets of the same system at different temperatures. At given intervals, a swap between adjacent systems is proposed. Swaps are accepted based on the Metropolis algorithm [92]. The theory is that systems at higher temperatures have more energy to overcome energy boundaries, so the energy landscape is effectively flattened. When a system at a low temperature is trapped in a local minimum, swapping with a higher temperature system will allow it to overcome the energy barriers.

Umbrella sampling is another advanced simulation method that forces the system to sample areas of phase space that might not be visited with regular molecular dynamics. The simulation includes an extra degree of freedom known as a reaction coordinate and the fluctuations in free energy along this reaction coordinate can be calculated to estimate the change in free energy between any two points along the reaction coordinate. The reaction coordinates need to be selected such that they can uniquely identify the possible states of the system. Important states in the hybridization process include the endpoints of the process–the canonical B-form of DNA and the melted, single-strand state–as well as the stable intermediates connecting the endpoints. The reaction coordinates must also distinguish between parallel and anti-parallel configurations.

Two order parameters that can serve as adequate reaction coordinates for the oligonucleotide lengths used in this study are the strand separation distance, ξ, and the angle between the two strands, θ. Both are depicted in Figure 2.1. ξ is defined as the distance between the central sugar on the probe strand and its corresponding sugar on the target strand. θ is found by first defining $\boldsymbol{u}$ as the vector from the sugar at the $5'$ end of the probe stand to the sugar at the $3'$ end of the probe stand and $\boldsymbol{v}$ as the vector from the sugar at the $3'$ end of the hybridizing sequence of the target strand to the sugar at the $5'$ end of the

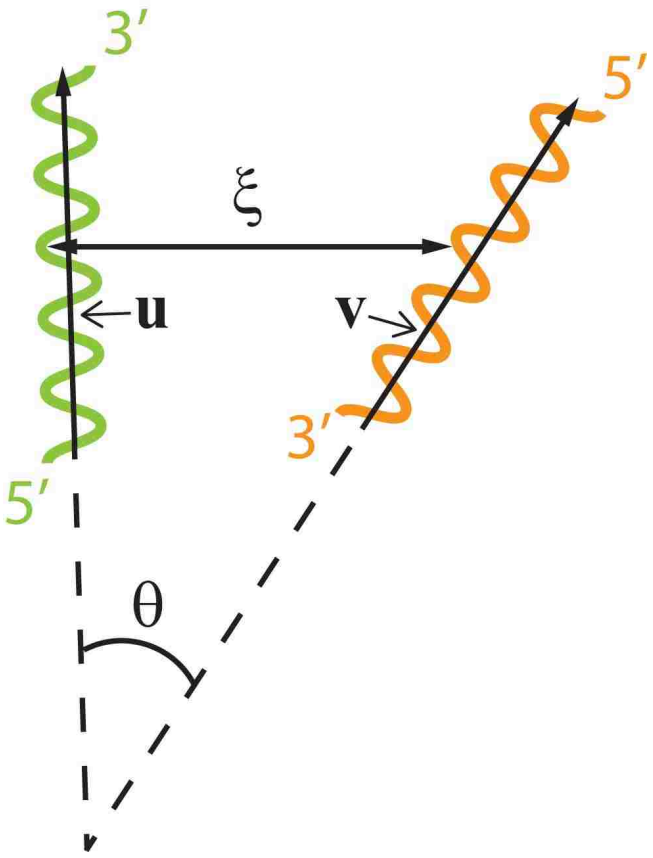

Figure 2.1: Order parameters used as reaction coordinates for projections of the free energy landscape of hybridization.

hybridizing sequence of the target strand. From this, θ is calculated according to

$$\theta = \cos^{-1}\left(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{||\boldsymbol{u}||\,||\boldsymbol{v}||}\right) \tag{2.2}$$

This definition yields $\theta = 0°$ for antiparallel configurations and $\theta = 180°$ for parallel configurations. More information on how these coordinates were applied to the various sequences is given in the individual chapters.

To force the system to sample an adequate portion of phase space along the reaction coordinates, a biasing potential is applied. The biasing potentials are of the form of Equations 2.3 and 2.4 where ξ_0 is the equilibrium distance between the strands for a particular umbrella, ξ is the instantaneous distance, k_ξ, θ_0 is the equilibrium angle between the strands for the particular umbrella, θ is the instantaneous value of the angle, and k_θ. Multiple independent

simulations with varying values of ξ_0 and θ_0 were performed to generate an energy landscape along these reaction coordinates. To avoid trapping the system in local minima along the rough potential landscape, care must be taken that the potential of each simulation overlaps the potential of all adjacent systems, or that the areas of phase space sampled in a simulation with a given value of ξ_0 and θ_0 include some of the same areas of phase space that were sampled in the simulations with a value of ξ_0 and θ_0 one step size greater and smaller. The reasoning for this is that sharp potential increases will trap the two strands within small areas of phase space along the reaction coordinate and discontinuities along the reaction coordinate prohibit the calculation of the energy landscape [85]. This leads to a delicate balancing act between the reaction coordinate step size, the amount ξ_0 and θ_0 are incremented from one simulation to the next, and the magnitude of the equilibrium constants k_ξ and k_θ. Step sizes must be small enough to give a high resolution along the reaction coordinate while remaining large enough that the number of simulations required to traverse the entirety of the reaction coordinate remains manageable. At the same time, equilibrium constants must hold the system around a specific point in the reaction coordinate while allowing it to periodically sample adjacent points. Groups of simulations were run with varying reaction coordinate step sizes and equilibrium constants to determine appropriate values for these factors. Histograms of the potential energy of the simulations were graphed to determine the overlap in phase space sampled by adjacent systems. Values were chosen as $\Delta\xi_0 = 0.25$ Å, $\Delta\theta_0 = 10°$, $\mathrm{k}_\xi = 10\frac{\mathrm{kJ}}{\mathrm{mol\ Å}^2}$, and $\mathrm{k}_\theta = 0.0382\frac{\mathrm{kJ}}{\mathrm{mol\ deg}^2}$ such that the histograms of adjacent systems overlapped $\approx 30-60\%$. These values produced high quality energy landscapes that were used to estimate properties and mechanisms that were not obtainable from regular molecular dynamic simulations.

$$\mathrm{U}_\xi = \mathrm{k}_\xi \left(\xi - \xi_0\right)^2 \tag{2.3}$$

$$\mathrm{U}_\theta = \mathrm{k}_\theta \left(\theta - \theta_0\right)^2 \tag{2.4}$$

Advanced simulation techniques, such as replica exchange and umbrella sampling, help to improve phase space sampling at a price. These techniques require the use of more processor time. In the case of replica exchange, the extra processor time is characterized by

the extra simulations at multiple temperatures run in parallel, while the cost of umbrella sampling is characterized by extra simulations for each point along the reaction coordinate. These prices are paid with the assumption that the amount of phase space explored from these extra simulations is greater than the amount that would be explored if a single regular molecular dynamics simulation was allowed to run for an equivalent amount of simulation time.

The strength of this assumption varies for different conditions. Because replica exchange techniques will effectively flatten the entire energy landscape, it is most effective when the energy barriers throughout the landscape are similar in size. Umbrella sampling forces a system through specific areas of phase space and is therefore useful when studying the phase space along a reaction coordinate particularly when those areas of phase space are blocked by energy barriers that are larger than ones found elsewhere in the system. While a system is held within a particular area of phase space by umbrella potentials, the roughness of the local energy landscape becomes more important than the global roughness. At the local level, it is probable that the magnitudes of the energy barriers are more uniform which is preferable for replica exchange techniques. Therefore, it was proposed to use these two techniques in tandem, but questions were raised as to the efficiency of this approach [97]. A small study (Reported in Chapter 3) was performed to determine if umbrella sampling alone was sufficient to investigate the systems of interest or if a combined umbrella sampling/replica exchange algorithm was needed.

It was determined that calculations of properties at a single temperature would not benefit from the improved sampling obtained from replica exchange. Therefore, only umbrella sampling was used to calculate the change in free energy and mechanistic pathways associated with hybridization. However, as will be described later, replica exchange dynamics were used to calculate the heat capacity of the DNA duplex over a range of temperatures and to validate the DNA model against experimentally-determined melting temperatures.

2.2 DNA Model

Recently, advanced coarse-grain models, which have been carefully parameterized to reproduce correct geometry and capture both the thermal and mechanical properties of

DNA, have been used to further study hybridization on surfaces. The one chosen for this study is the Three-Sites-Per-Nucleotide (3SPN) formalism of Knotts *et al.* [56] with recent improvements by Sambriski *et al.* [76–78] This model reduces a nucleotide to three interaction sites–one each for the phosphate, sugar, and base. All four bases–adenine, thymine, cytosine, and guanine–are represented. This model was carefully parameterized against experimental data. It possesses the correct double-helical geometry of the molecule and replicates both the thermal and the mechanical properties of DNA including salt-dependent effects [56, 76–78].

The force field of this model includes both bonded and non-bonded interactions described below. The total contribution to the potential energy is represented in Equation 2.5.

$$U_{\text{total}} = U_{\text{bond}} + U_{\text{bend}} + U_{\text{tors}} + U_{\text{stck}} + U_{\text{base}} + U_{\text{elec}} + U_{\text{solv}} + U_{\text{nnat}} \tag{2.5}$$

The first three terms represent the bonded interactions in the system. U_{bond} is the two-body term accounting for covalent bonding and is calculated as

$$U_{\text{bond}} = \sum_{i=1}^{n_{\text{bond}}} \left[k_1 \left(d_i - d_{0i} \right)^2 + k_2 \left(d_i - d_{0i} \right)^4 \right] \tag{2.6}$$

Where k_1 and k_2 are bond constants, d_i is the instantaneous bond distance, d_{0i} is the equilibrium bond distance, and i designates the bond in the set of n_{bond} bonds. U_{bend} is a three-body term for bend energy where

$$U_{\text{bend}} = \sum_{i=1}^{n_{\text{bend}}} \frac{k_\theta}{2} \left(\theta_i - \theta_{0i} \right)^2 \tag{2.7}$$

Here, k_θ is a bend constant, θ_i is the instantaneous bend angle for the ith three-body interaction, and θ_{0i} is the equilibrium angle for the ith angle in the set of n_{bend} angles. Four-body bonding interactions were calculated from

$$U_{\text{tors}} = \sum_{i=1}^{n_{\text{tors}}} k_\phi \left[1 - \cos \left(\phi_i - \phi_{0i} \right) \right] \tag{2.8}$$

where k_ϕ is a torsional constant and ϕ_i and ϕ_{0i} are instantaneous and equilibrium angles in the set of n_{tors} dihedral angles.

The rest of the terms in Equation 2.5 are non-bonded, pairwise interactions. U_{stck} is the energy associated with the sequence of the bases on the same strand and takes the form

$$U_{\text{stck}} = \sum_{i<j}^{n_{\text{stck}}} 4\varepsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.9}$$

This potential acts on each pair of bases in the same strand that are within a cutoff radius of 9Å. To expedite calculation, a list of n_{stck} native contacts within the cutoff radius is generated from the ideal reference structure of the DNA. ε_{ij} and σ_{ij} are the interaction specific potential strength and length scale between sites i and j that contribute to the stiffness of the backbone by controlling the instantaneous site-site separation distance r_{ij}.

Equation 2.10 gives the form of U_{base}, the potential that accounts for hydrogen bonding between complementary, or Watson-Crick, base pairs for both inter- and intrastrand interactions (complementary pairs that are not accounted for in the stacking interactions). The interstrand interactions account for the base pairing that drives hybridization and intrastrand interactions allow for hairpin formation. Bond strength ($\varepsilon_i \in \{\varepsilon_{\text{CG}}, \varepsilon_{\text{AT}}\}$) and length ($\sigma_i \in \{\sigma_{\text{CG}}, \sigma_{\text{AT}}\}$) are determined from the base type of sites i and j. Complementary base pairs are considered hydrogen bonded when the distance between the bases $r_{ij} < \sigma_i + 2.0\text{Å}$

$$U_{\text{base}} = \sum_{i=j}^{n_{\text{base}}} 4\varepsilon_{\text{b}i} \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] \tag{2.10}$$

$$U_{\text{elec}} = \sum_{i<j}^{n_{\text{elec}}} \frac{q_i q_j e^{-r_{ij}/\lambda_D}}{4\pi\epsilon_0\epsilon\left(T, I\right) r_{ij}} \tag{2.11}$$

$$U_{\text{solv}} = \sum_{i<j}^{n_{\text{solv}}} \varepsilon_s \left[1 - e^{-\alpha(r_{ij} - r_s)} \right]^2 - \varepsilon_s \tag{2.12}$$

The polyelectrolyte features of DNA are modeled with Equations 2.11 and 2.12. U_{elec} accounts for electrostatic contributions at the Debye-Hückel level of theory and is applied to all phosphate-phosphate interactions that are not accounted for in the bonded potential U_{bend}. λ_D is the Debye length, which defines the spatial extent of charge screening caused

by solvation conditions between sites i and j and separation distance r_{ij}. It is calculated as

$$\lambda_D = \left[\frac{\epsilon_0 \epsilon\left(T, I\right)}{2\beta N_A e_c^2 I}\right]^{1/2} \tag{2.13}$$

where ϵ_0 is the permittivity of free space, $\beta = \left(k_{\mathrm{B}} T\right)^{-1}$, N_A is Avogadro's number, e_c is the elementary charge, and I is the ionic strength of the implicit solution, and $\epsilon\left(T, I\right)$ is a function to calculate an effective dielectric constant.

$$\epsilon\left(T, I\right) = \epsilon\left(T\right) a\left(I\right) \tag{2.14}$$

The dielectric constant is calculated from the temperature and the salt concentration via product contributions of the form

$$\epsilon\left(T\right) = 249.4 - 0.788 T/\mathrm{K} + 7.20 \times 10^{-4} \left(T/\mathrm{K}\right)^2 \tag{2.15}$$

and

$$a\left(I\right) = 1.000 - 0.2551 \left(I/\mathrm{M}\right) + 5.151 \times 10^{-2} \left(I/\mathrm{M}\right)^2 - 6.889 \times 10^{-3} \left(I/\mathrm{M}\right)^3 \tag{2.16}$$

$\epsilon\left(T\right)$ is the static, or zero-frequency, dielectric constant at absolute temperature T and $a\left(I\right)$ is the salt correction for a solution with molarity I.

The second potential of the model used to reproduce the unique electrolytic properties of DNA systems is the solvation term U_{solv}. This term implicitly duplicates the effects associated with the ordering of water and ionic species around DNA to create hydration shells. Equation 2.12 shows the Morse-like potential with energy scale ε_s, particle separation r_{ij}, spatial range coefficient α^{-1}, and the minimum energy distance r_s. The values of these parameters were chosen to be compatible with the molecular geometry of DNA. Since the solvent-induced interactions reproduce the many-body effects seen experimentally, which depend on chain length and ionic conditions, the energy scale is approximated as $\varepsilon_s = A_I \varepsilon_N$.

A_I is the contribution of the salt dependence and is parameterized from empirical data as

$$A_I = 0.474876 \left(1 + \left\{0.148378 + 10.9553 \left[\mathrm{NA}^+\right]\right\}^{-1}\right) \tag{2.17}$$

while the effect of the chain length, ε_N, is parameterized as

$$\varepsilon_N = \varepsilon_\infty \left(1 - \left[1.40418 - 0.268231n\right]^{-1}\right) \tag{2.18}$$

with $\varepsilon_\infty = 0.504982$. These contributions had to be carefully parameterized such that both denaturation and renaturation could occur. Parameterization was performed by simulating multiple oligonucleotides with replica exchange molecular dynamics over a temperature range that yielded fully renatured and denatured duplexes. The melting temperatures of the oligonucleotides were calculated from the simulations and compared to experimental data. This term of the forcefield was parameterized to closely recreate the experimental melting temperatures [76].

The final term of the forcefield given in Equation 2.5 is the potential due to non-native contacts, shown in Equation 2.19. The n_{nnat} non-native contacts are all pairwise interactions not accounted for in other potentials including mismatched base pairs. This is a purely repulsive, excluded volume contribution based on a Weeks-Chandler-Anderson [98] interaction with energy scale ε. U_{nnat} does not contribute to the energy of the system until interparticle separation r_{ij} is less than cutoff length r_{coff}. This represents the energy cost of forcing particles close together. In the case of mismatched bases, $r_{\mathrm{coff}} = 1.00$Å, for all other cases, $r_{\mathrm{coff}} = 6.86$Å. This potential is designed such that U_{nnat} becomes zero for all $r_{ij} \geq 2^{-1/6} r_{\mathrm{coff}}$.

$$U_{\mathrm{nnat}} = \sum_{i<j}^{n_{\mathrm{nnat}}} \begin{cases} 4\varepsilon \left[\left(\frac{\sigma_0}{r_{ij}}\right)^{12} - \left(\frac{\sigma_0}{r_{ij}}\right)^{6}\right] & \text{if } r_{ij} < r_{\mathrm{coff}} \\ 0 & \text{if } r_{ij} \geq r_{\mathrm{coff}} \end{cases} \tag{2.19}$$

2.2.1 Validation of the 3SPN Model

To ensure that the model would provide data in agreement with experimental results, it was validated by calculating and measuring melting temperatures. Initially, salt-

dependant melting curves calculated from the model were compared to experimental results [99]. Owczarzy *et al.* measured the salt dependence on the melting point of hundreds of oligonucleotides by measuring the absorbance of light in samples of DNA. The relationship between the absorbance and the temperature produces a characteristic "s" shaped curve. The melting point of the oligonucleotide is determined to be the inflection point of the curve. To compare the 3SPN model to the experimental data, a 20-base-pair sequence was chosen from the sequences studied by Owczarzy *et al.* (**TAC TTC CAG TGC TCA GCG TA**) and was simulated over a range of temperatures using replica exchange molecular dynamics. The number of unpaired nucleotides was calculated as a function of temperature using Equation 2.1 since this is the physical property driving the absorbance of light.

Figure 2.2 shows the melting curves generated from simulation. The shape of the curves matches the shape seen in experimental absorbance data and the trend of increasing melting temperature with increasing sodium concentrations matches the results obtained by Owczarzy *et al.* The melting points predicted by calculating the inflection points in the curves in Figure 2.2 are 57°C, 67°C, and 80°C for 119mM Na^+, 220mM Na^+, and 621mM Na^+ respectively. The experimental values reported by Owczarzy *et al.* are 60.3°C, 64.4°C, and 67.7°C. While the melting points calculated from the 3SPN model deviated from the experimental values, they showed the same trends and offered a good basis of comparison between similar systems. This is important since the comparative stabilities of similar systems is a major metric in this study. Additionally, the simulation results were as accurate as another generally accepted prediction model, the salt-adjusted prediction method, which estimated melting points of 64.7°C, 69.1°C, and 76.6°C [100] for this same sequence at the salt concentrations given above.

To further validate the model, the melting point of the principle strand of interest for this study, the human topoisomerase II target discussed in Chapter 4, was determined from heat capacity data. The heat capacity was calculated from

$$C(T) = \frac{\langle U^2 \rangle_T - \langle U \rangle_T^2}{k_B T^2} \tag{2.20}$$

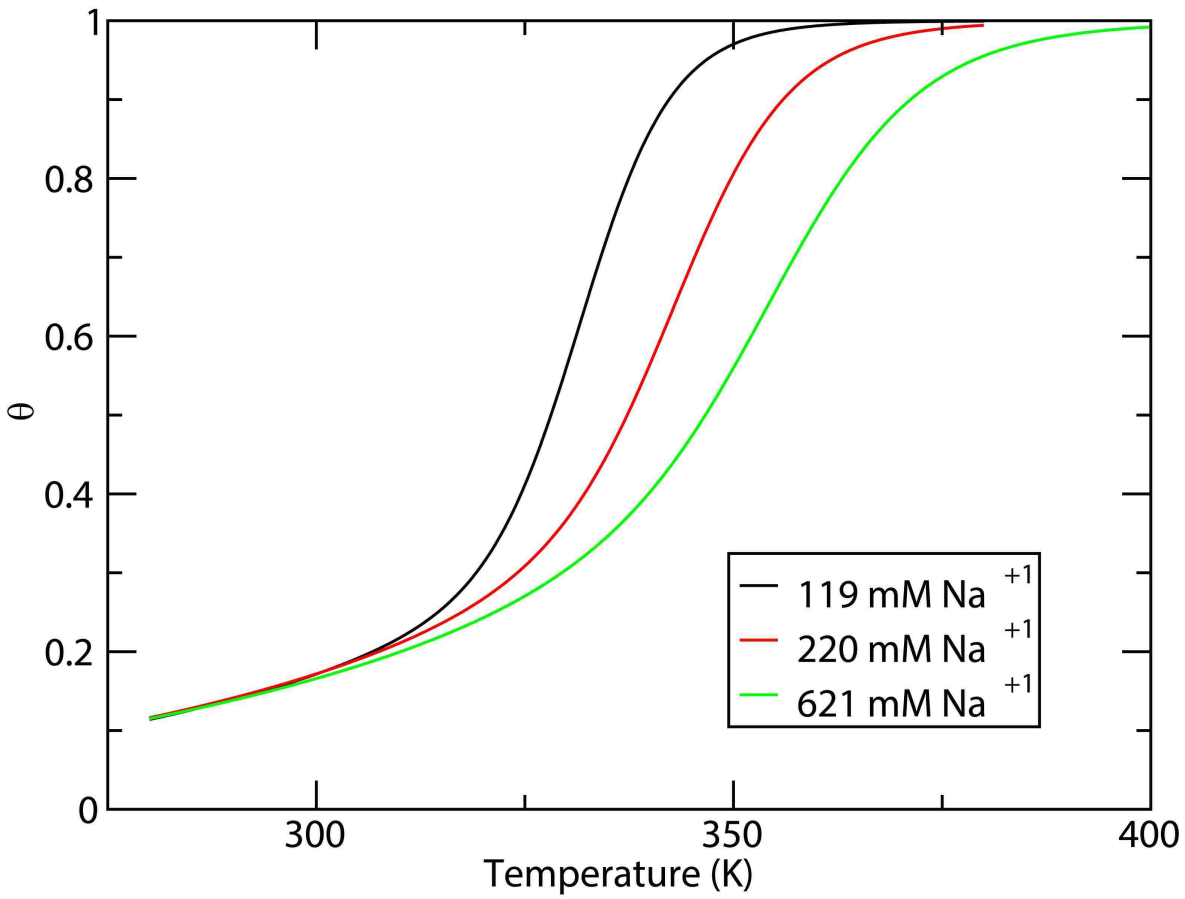

Figure 2.2: Salt-dependant hybridization curves follow the trends shown in experimental work

where the internal energy U was generated as a continuous function of T from WHAM data and Equation 2.1. This gives a continuous value of $C(T)$ as seen in Figure 2.3. Since heat capacity is a measurement of how much thermal energy the system requires to increase temperature and the thermal energy of a system is absorbed to break the hydrogen bonds between bases at the melting point, the melting point can be found as the maximum in the heat capacity curve. Replica exchange molecular dynamics was used with 20 boxes over the temperature range shown in Figure 2.3. Throughout this study, shaded regions around curves indicate the uncertainty of the results, or the statistical error in the numbers. Uncertainty was estimated as $\frac{\sigma}{\sqrt{N-1}}$, where σ is the standard deviation of the N values of the heat capacity for each value of temperature. The value obtained from simulation, 349.3 ± 0.7 K, was in good agreement with the nearest neighbor and salt-adjusted prediction methods, 350.15 K

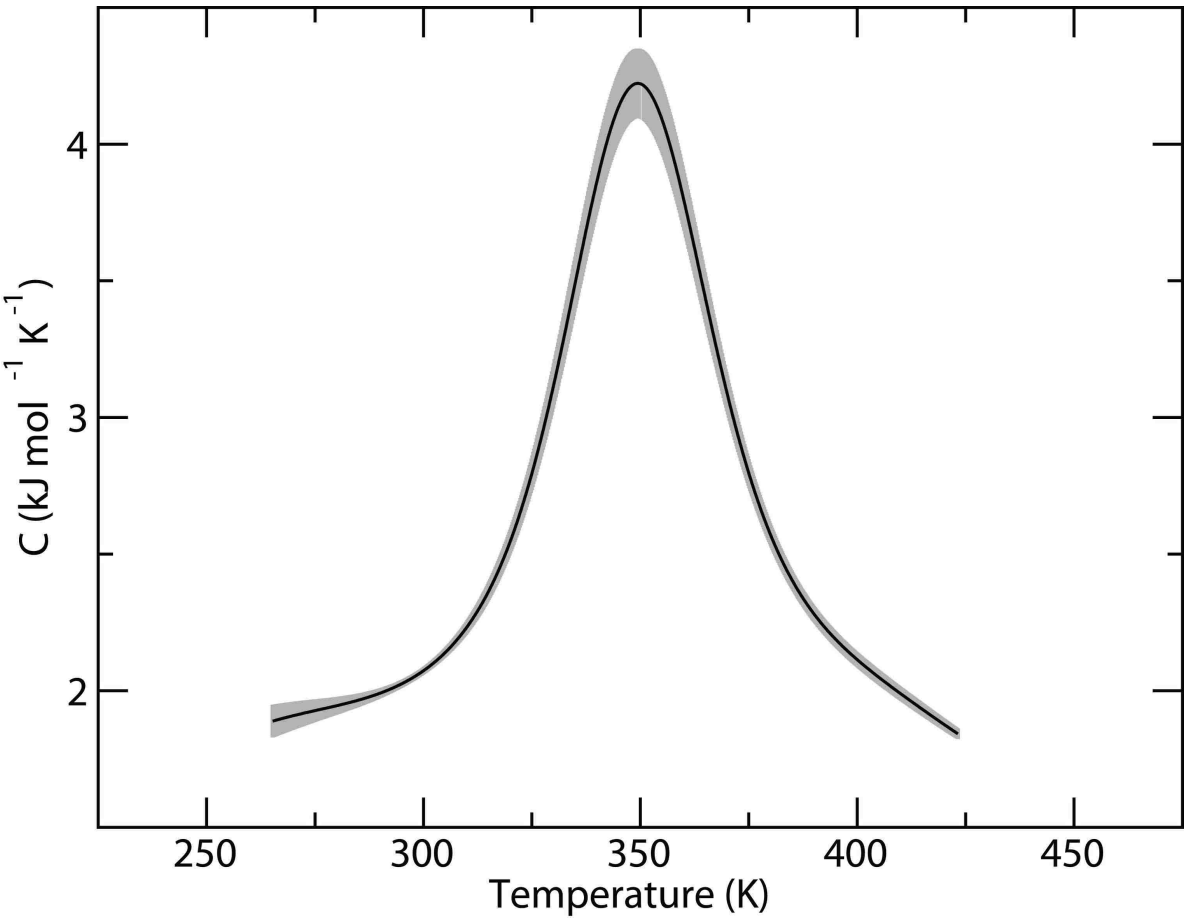

Figure 2.3: A heat capacity curve

and 345.15 K respectively [100]. Multiple temperature schemes were used in the replica exchange simulations and the calculated melting point remained the same for each scheme. These results indicate that results produced by the model are in satisfactory agreement with experiment.

2.3 Coding

In order to implement the methods and techniques outlined above, computer codes were adapted and scripts were written to produce and analyze hybridization data. This study built upon a rich suite of code that was developed for biological systems, [56, 101] but new code was added to implement enhanced models [102] and new techniques [103]. Additionally, the data generated by this research were large sets that required new methods

of post processing. These additions to the software suite were challenges that needed to be overcome before the research could be finished.

The first software challenge of this project was integrating the solvent-effect equations [102] into the existing code. Most of this implementation consisted of a straightforward process of adding new potentials into the already existing potentials and creating new switches so the program would apply the proper potentials to the correct multi-site interactions. Equation 2.18, however, presented an interesting challenge. In this equation, n is the number of base pairs in the duplex. This is easily calculated as the number of bases on one of the strands in the duplex when both strands are the same length. This research, however, included simulations of duplexes of strands of different lengths, creating the problem of determining what value should be used for n. It was determined that n represents the number of base pairs that may form and a short loop was written to find the shortest strand and use the number of bases in it as the value of n.

Once the model was running properly and calculating accurate representations of DNA hybridization, attention was turned to expanding the simulation protocols to include two dimensional umbrella sampling. This required creating a two dimensional matrix for every simulation within the reaction coordinates. Each element in the matrix gave the frequency in which a specific location along the two reaction coordinates was visited, with the first reaction coordinate binned along the rows and the second reaction coordinate binned along the columns. The element $[i, j]$, therefore, represents the number of times the simulation sampled a configuration in the ith bin of the first reaction coordinate and the jth bin in the second reaction coordinate. Memory was then allotted to store these histogram matrices according to the data given in one of the two new inputs required for this method.

Two-dimensional umbrella sampling required two new inputs to be read in by the program. The first contained the binning instructions the simulation would use to determine the size of the matrix required to store histogram data, the ranges and bin sizes the data would span, and the amount of memory that needed to be allocated so that the program could hold all of the histogram and consisted of a lower bound, an upper bound, and a bin size for each reaction coordinate. For example, this input might contain the information: LB1 10, UB1 125, BN1 0.25, LB2 0, UB2 180, BN 10. These numbers would allow the simulation

to create a $[464, 19]$ matrix with rows spanning the range from 10 to 125 in increments of 0.25 and columns spanning the range from 0 to 180 in increments of 10 with enough memory allocated to contain 8816 integers.

The second new input contained parameters for angle restraints. These parameters were the number of angle restraints, the sites used to define vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, the strength of the restraint k_θ, and the equilibrium angle θ_0. As outlined above, interstrand angle θ was calculated from the coordinates of the four sites given in this input using Equation 2.2 with

$$\boldsymbol{u} \cdot \boldsymbol{v} = x_{\boldsymbol{u}} x_{\boldsymbol{v}} + y_{\boldsymbol{u}} y_{\boldsymbol{v}} + z_{\boldsymbol{u}} z_{\boldsymbol{v}} \tag{2.21}$$

and

$$||\boldsymbol{u}|| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{2.22}$$

where $x_{\boldsymbol{u}} = x_2 - x_1$, $x_{\boldsymbol{v}} = x_4 - x_3$ and (x_1, y_1, z_1) are the coordinates of the first site *etc.* With the potential energy of the angle restraint, the simulation could iterate through its calculations using well established methods.

Once the simulations had completed, the histograms were analyzed using the weighted histogram analysis method [103] to generate potential-of-mean-force surfaces. Mathematically, this was done by iterating around the probability

$$P_{\{\lambda\}',\beta'}\left(\mathbf{V}_{\xi,\theta,R}, \xi, \theta\right) = \frac{\sum_{k=1}^{R} N_k \left(\mathbf{V}_{\xi,\theta,k}, \xi, \theta\right) \exp\left[-\beta' \sum_{j=1}^{L} \lambda_j' \mathbf{V}_{\xi,\theta,j}'\right]}{\sum_{m=1}^{R} n_m \exp\left[f_m - \beta_m \sum_{j=1}^{L} \lambda_{jm} \mathbf{V}_{\xi,\theta,j}\right]} \tag{2.23}$$

and the dimensionless free energy

$$\exp\left(-f_j\right) = \sum_{\mathbf{V}_{\xi,\theta,R},\xi,\theta} P_{\{\lambda\}_j,\beta_j}\left(\mathbf{V}_{\xi,\theta,R}, \xi, \theta\right). \tag{2.24}$$

Computationally, this was done using a Matlab script on the Fulton Supercomputing Laboratory BigMem processor. The first step of this analysis was to load all the two-dimensional histogram matrices into a three-dimensional matrix H. This matrix was used to sum up the number of samples taken in each simulation, n_m, for all R simulations over the entire

reaction coordinate, and the number of samples taken across all simulations within each bin combination of the two reaction coordinates, N_k. For example, if a distance reaction coordinate had a bin of distances from 10 Å to 10.24 Å and an angle reaction coordinate had a bin of angles from 0° to 9.99°, the previous step would count the number of times all simulations sampled a configuration that was within those ranges of the two reaction coordinates. Next, a second matrix $\mathbf{V}$ was generated containing the energy required for the system to exist in any given element in the matrix H. In other words, element $[1, 1, 1]$ contained a calculation of the amount of energy required for the first umbrella to contain a configuration in the first bin along both reaction coordinates. Finally, an initial value of f was estimated as a matrix of ones. With these inputs, a value for P was calculated from Equation 2.23. This value was plugged into Equation 2.24 to calculate a new f. The new f was compared to the old f and if the values were different the old f was replaced with the new f and another iteration was performed. For the systems studied in this project, these iterations continued for about 24 hours before converging. Although computationally expensive, these systems produced highly detailed thermodynamic data of complex biomolecular systems.

The systems outlined above were deceptively complex and time-consuming. In general, debugging the new code, once it was written, took twice the amount of time as writing it. Adding the solvation effects to the simulation suite, including new subroutines for integration, and rewriting a script that generated simulation inputs took about a month, but the new sections of code were adjusted and edited for an additional two months before it was confirmed that the suite was reproducing predicted values of a simulation standard. Two-dimensional umbrella sampling required about six weeks to figure out the pointers that would create dynamic memory allocation for efficient storage of the simulation histograms although the code for generating the histograms was originally written in just two weeks. Once the program would compile with the flags that allowed two-dimensional umbrella sampling, an additional month was required to track down the cause of a segmentation fault that occurred when the strands were perfectly aligned. Debugging the scripts for analyzing the two dimensional histograms presented an additional challenge since the analysis itself was so computationally demanding that script would often run entire days without giving any indication of whether it was running successfully or had stalled. Most of the computational

aspects associated with the two-dimensional umbrella sampling suffered from drawback. A single potential-of-mean-force curve required over 11000 independent simulations. Scripts were written that could generate and organize the unique inputs for each simulation, submit all the simulations to the super computer scheduling cue, and extract and order all the outputs from the simulations for post processing once they had finished. Each of these scripts required one to two hours to run in addition to the week required for all the simulations to proceed through the cue. While simulations and scripts ran, errors in the setup of the process could go unnoticed for up to a week before the they could be tracked down and corrected and the process could be restarted. This debugging process added to the total computational time required to obtain thermodynamic properties of the systems of interest.

Chapter 3

The Efficiency of Replica Exchange Techniques for Sampling Phase Space

3.1 Introduction

To obtain thermodynamic properties of the hybridization process, simulations must sample sufficient phase space. As mentioned in Chapter 2, various Monte Carlo techniques were considered. Replica exchange molecular dynamics was one of the techniques reviewed for this study since it is a robust tool that is widely accepted for computational studies of biological systems. It uses a biasing potential of temperature–a well understood thermodynamic quantity–to move systems in and out of local minima in rough energy landscapes. This shows the locations of such minima without allowing the system to get caught in them. Replica exchange molecular dynamics accomplishes this by simulating the same system at multiple temperatures, periodically proposing swaps between systems at adjacent temperatures, and accepting those swaps based on a Metropolis algorithm. This is particularly useful when the thermodynamic properties of a system are to be found over a range of temperatures since the system would be simulated at multiple temperatures regardless. By swapping with adjacent temperatures, a system stuck in a local energy minimum might swap to a temperature where it has enough energy to escape that minimum, or a system with a flat energy landscape might swap down to a temperature with a more characteristic energy landscape. Swapping is considered useful when the amount of simulation time between swaps is long enough that the system can relax between configuration changes and short enough to maximize the number of swaps in a simulation [104, 105]. As long as the energy profiles of the systems overlap, thermodynamic information from the individual systems can be combined to obtain the density of states over the entire temperature range. This allows the calculation of more detailed thermodynamic data.

The goal of this study, however, was to better understand the phenomena governing micoarray hybridization of DNA at a single temperature. Although the principle advanced Monte Carlo technique considered for this study was umbrella sampling, the viability of using replica exchange molecular dynamics techniques in conjunction with umbrella sampling techniques was explored. Replica exchange molecular dynamics is often used to study systems at a single temperature of interest despite the fact that this technique requires simulating a system at multiple temperatures. In these cases, the phase space sampled at the temperature of interest is analyzed while the phase space sampled at other temperatures is only used to enable the systems at the temperature of interest to move to areas of the energy landscape that they might not have otherwise explored. When this method is used while only the data at a single temperature of interest is desired, the computational cost of simulating the system at other temperatures is an additional cost of the method. This additional cost is often paid, by requisitioning the processor power required to simulate the systems at other temperatures, with the postulate that it will cause a net gain by decreasing the total amount of processor time to obtain the same results. In other words, it is believed that the amount of phase space explored by switching to and from higher temperatures is greater than the amount of phase space that would be explored by a single system that was allowed to run for an amount of processor time equivalent to the amount required for multiple systems. In 2006, Zuckerman *et al.* [97] proposed that this postulate is not well founded and could be false. Since the goal of the present study was to determine thermodynamic properties of DNA in a microarray setting, the Zuckerman claim was investigated. It was believed that evidence could be found validating the use of replica exchange molecular dynamics to obtain single temperature results. Specifically, it was hypothesized that *good thermodynamic data could be obtained with less processor time by using replica exchange than by using regular molecular dynamics alone.*

3.2 Methods

3.2.1 Model

Due to the complex nature of DNA hybridization and the large amounts of processor time needed to model it, it was decided that this initial investigation of simulation tech-

niques should be conducted with a system that is simpler and better understood. Two protein systems, modeled using a coarse-grain representation, were studied since they could be simulated rapidly. The chosen model has been shown to accurately capture *in vivo* folding mechanism [101, 106]. It is a bead spring model that uses a single bead for each amino acid in the protein. The potential energy of the system, U, is calculated as

$$U = U_{\text{bond}} + U_{\text{bend}} + U_{\text{tors}} + U_{\text{nat}} + U_{\text{nnat}} \tag{3.1}$$

where U_{bond} is the two body potential between bonded sites, U_{bend} is the three body potential between sites forming an angle, U_{tors} is the three body interaction between sites forming a dihedral, U_{nat} is the nonbonded energy between sites that are considered "native" contacts, and U_{nnat} is a two body potential between all pairs of sites not considered in one of the other potentials. The bonded potentials, U_{bond}, U_{bend}, and U_{tors}, are of the same form as the CHARMM [54] forcefield and are given below.

$$U_{\text{bond}} = \sum_{i=1}^{n_{\text{bond}}} k_d \left(d_i - d_{0i}\right)^2 \tag{3.2}$$

$$U_{\text{bend}} = \sum_{i=1}^{n_{\text{bend}}} k_\theta \left(\theta_i - \theta_{0i}\right)^2 \tag{3.3}$$

$$U_{\text{tors}} = \sum_{i=1}^{n_{\text{tors}}} k_\phi \left[1 + \cos\left(n\phi_i - \delta\right)\right] \tag{3.4}$$

A native-contact potential energy is applied to pairs of sites that do not form a bond or a bend but are considered to form a hydrogen-bond in the native state of the protein. These interactions form the secondary and tertiary structures of the protein and are determined from experimental structures. The native contact potential allows this model to accurately predict folding mechanisms [107, 108]. All nonbonded pairs that are not native contacts are considered non-native. The forms of the potentials for nonbonded interactions are as follows.

$$U_{\text{nat}} = \sum_{i<j}^{n_{\text{nat}}} \varepsilon_{ij} \left[13 \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 18 \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} + 4 \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{3.5}$$

$$U_{\text{nnat}} = \sum_{i<j}^{n_{\text{nat}}} \varepsilon_{ij} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} \tag{3.6}$$

Parameter and input files needed to simulate systems with this model were obtained from the Gō model builder on the Multiscale Modeling Tools for Structural Biology internet server [109]. Coordinates of the proteins were obtained from the Protein Data Bank. The first protein system was of residues 10-55 of the B domain of protein A (Protein Data Bank identification: 1BDD [110]). Simulations of this protein served as a metric for determining the efficiency of replica exchange molecular dynamics for obtaining thermodynamic properties. The second system was of ribonuclease H (PDB ID: 2RN2 [110]). RNase H has 155 residues. This represents a contour length of nearly 60 nm along the backbone when the protein is fully stretched. Umbrella sampling simulations along this length, with end-to-end separation as a reaction coordinate, were used to obtain mechanistic pathways. Since both mechanistic pathways and thermodynamic data are important goals of the DNA study, these systems showed the usefulness of this technique in the desired research.

3.2.2 Theory and Experimental Design

The optimal technique for this study was defined as the most efficient one. Efficiency was calculated by the amount of processer time required to accurately obtain data. The target data of the protein A system were three thermodynamic properties, the potential energy (U), heat capacity (C), and Gibbs energy of folding (ΔG_f) of a system. Heat capacity was calculated from Equation 2.20. For MD simulations, the ensemble average is simply the mean of the values at each time step throughout the simulation. For replica exchange simulations, the ensemble averages are calculated from the canonical partition function, Q. In general, the value of an arbitrary property, X, evaluated at temperature T is related to Q by Equation 2.1. Ω is estimated from the Weighted Histogram Analysis Method (WHAM) [111]. The partition function is related to the density of states through

$$Q = \sum_i \Omega(U_i) e^{-\beta U_i}. \tag{3.7}$$

The free energy of folding of protein A was calculated by classifying the configurations sampled during the simulation into "folded" and "unfolded" ensembles based upon the instantaneous fractional nativeness, or the fraction of the native contacts listed in the native contact list that are close enough to form hydrogen bonds. Since the native contacts represent the hydrogen bonds between amino acids that hold the protein in its folded structure, the protein is considered folded when more than half of the tertiary structure native contacts are formed and unfolded when less than half of those bonds are present, or contributing to the potential energy of the system. The free energy of folding at any temperature can then be calculated from

$$\Delta G = G_{\text{folded}} - G_{\text{unfolded}} = -kT \ln \left(\frac{P_f}{1 - P_f} \right), \tag{3.8}$$

where P_f is the probability of the folded state at temperature T.

With the relationships given above, the potential energy, heat capacity, and free energy of folding of the protein A systems were correlated with the amount of computational time needed to generate these values. The time required to obtain these properties was measured from the total amount of processor time needed for the simulation, or a total of the computational time of each processor used in the simulation. For example, a simulation that was run on one processor for two days was considered to have cost the same amount of processor time as a simulation that was run on two processors for one day. This is important since both replica exchange and umbrella sampling techniques require large numbers of processors and combining the two multiplicatively increases processors demand. If replica exchange can obtain good thermodynamic with less processor time, then the demand for more processors will be out weighed by the demand for less time. This would be a more efficient system. On the other hand, if replica exchange systems require the same amount of process time to calculate good thermodynamic data as regular molecular dynamics systems, than all that has been achieved is a more complex system with no net gain in efficiency.

The ribonuclease H systems provided an opportunity to test the efficiency of combining replica exchange techniques with umbrella sampling techniques. Previous studies have found that protein folding mechanisms differ when comparing thermal unfolding and

mechanical unfolding [101]. Thermal unfolding occurs when a protein unfolds due to an increase in energy in the system. The simplest method for inducing thermal unfolding is to raise the temperature of a system. Mechanical unfolding occurs when an external force acts on the ends of a protein, pulling them away from each other, and destroying the secondary structure. Experimentally, this is done with atomic force microscopy, or "optical tweezers." [112–115]

In simulations, mechanical unfolding is controlled by introducing a potential in the same form as Equation 2.3 between the two ends of the protein, sites 4 and 155 for RNase H, with spring constant $k_\xi = 10$ kcal/mol. ξ_0 is the equilibrium end-to-end distance to which the potential is pulling the protein and ξ is the instantaneous end-to-end distance. Mechanical unfolding is simulated by using the end-to-end distance as a reaction coordinate with umbrella sampling techniques. The system is moved along the reaction coordinate by simulating with ξ_0 being set to discrete lengths ranging from the distance between the two ends when the protein is in its native conformation to the contour length of the protein backbone. Specifics about how this was done for Ribonuclease H are given in Section 3.2.4.

From the umbrella sampling data, potentials of mean force were generated. These show the relative free energy of the system with respect to the reaction coordinate. Energy minima in the curve indicated stable folding states. Structural data such as native contacts and simulation snapshots from the simulations corresponding to strand separation distances with stable states were analyzed to determine the structures of the protein in these states and to hypothesize the mechanical folding pathway.

The efficiency of replica exchange molecular dynamics was tested by creating two sets of potentials of mean force of the system. The first contained only data that was generated from regular molecular dynamics simulations while the second set was only calculated from replica exchange molecular dynamics simulations. Potentials of mean force from the replica exchange systems were generated from only the first box of the replica exchange simulations; the other boxes simply provided higher energy environments where the protein could escape energy wells that might trap it. The number of production steps used for the replica exchange systems was equal to the number of production steps used in the regular molecular dynamics simulations divided by the number of boxes in the replica exchange simulations. This yielded

an equivalent amount of processor time for both sets so that efficiency could be compared as results per computation time. The PMFs for the replica exchange systems and the regular molecular dynamics simulations were compared to determine if the two schemes produced significantly different results. Specifically, the sizes and locations of energy wells were analyzed to determine which ones represented mechanical folding transition states and if any represented nonrealistic choke points where the simulations were trapped by energy barriers.

3.2.3 Simulation Protocols

Protein A

Before designing the simulations used to compare the amount of computation time needed for the two techniques to obtain good thermodynamic data, the model needed to be characterized with controls. Control simulations gave comparative values to help determine what constituted good thermodynamic data. They were obtained from conventional molecular dynamics and replica exchange systems that were larger than those used for the evaluation. In each simulation, the temperature was maintained using the Nosé-Hoover chain method [116] with five thermostats and the time step was 1 fs. The replica exchange control simulations were run with 75 ns of equilibration and 150 ns of production time. Each simulation consisted of 26 boxes spanning temperatures below and above the suspected melting point. Box density, the number of boxes spanning a given temperature interval, was set higher around the suspected melting point and lower on the ends of the temperature range.

Using the results of the large replica exchange runs, two temperatures of interest were chosen for the simulations designed to test the hypothesis that replica exchnage simulations are more efficient than regular molecular dynamic simulations. These temperatures were chosen far below and at the melting point for this model, 215 K and 257 K respectively. Control regular-molecular-dynamics simulations were then run at these temperatures with 100 ns of equilibration and 2 μs of production time. The results were compared with the RE results. Due to good agreement, the two sets of results were averaged and used for a basis of comparison for the heat capacity and potential energy. The basis of comparison for the free energy of folding was taken from only the replica exchange data.

Table 3.1: Simulation schemes used to determine the efficiency of the REMD technique. Temperatures are in units of kelvin.

Temperature of Interest	Simulation Scheme	Temperature 1	2	3
215	MD	215	-	-
	2 Box REMD	215	225	-
	3 Box REMD	215	225	235
257	MD	257	-	-
	2 Box REMD	257	267	-
	3 Box REMD	257	267	277

When comparing the efficiency of replica exchange and molecular dynamics, all simulations were run using the same code. Every simulation was started from the minimized average structure [110] with no equilibration. Equilibration was excluded to test phase space sampling of each method, to see which method more rapidly moved out of the unlikely area of phase space containing the ideal structure. Regular molecular dynamics simulations were run with only one box while replica exchange simulations had either two or three boxes with swaps between the boxes proposed every 2 ps. As discussed in Section 2.1, swaps were accepted based on the Metropolis algorithm. When a swap was proposed, a random number between zero and one was generated, if this number was smaller than the ratio of the energy of the current system and the proposed system, the swap was accepted. Table 3.1 shows how the schemes were organized.

Each scheme was set up to cause the simulation to run for a target amount of processor time. Target processor times were estimated by dividing the number of production steps by the number of boxes used in the simulation. Two box replica exchange systems had half the number of production steps as regular molecular dynamics systems and three box systems had one third. The actual amount of processor time used for each simulation was tracked and the values of the properties mentioned above are graphed below as a function of this actual processor time. Each simulation was run on a single node of a Dell 1955 Blade Cluster with 1260 Dual Core Intel EM64T processors 2.6 GHz known as Marylou4.

3.2.4 Ribonuclease H

Simulations of ribonuclease H were run on the same processors as the simulations of protein A. The canonical ensemble was generated with the Nosé-Hoover-Chain method, with four thermostats of mass 10^{-20} kg Å^2 and a time step of 3 fs. This time step was chosen as the highest time step possible that still maintains an accurate integration, as determined by monitoring the conservation of energy in the system. The conserved quantity of a canonical ensemble is the extended Hamiltonian which is calculated as

$$|\Delta \mathrm{E}| = \frac{1}{N} \sum_{k=1}^{N} \left| \frac{E_k - E_0}{E_0} \right| \tag{3.9}$$

where N is calculated as the total number of steps, E_k is the value of the conserved quantity at step k, and E_0 is the initial value of the conserved quantity. Quantitatively, integration is considered stable when $\log |\Delta E| \leq -2.5$. To optimize the time step, $\log |\Delta E|$ was calculated for $\Delta t = 1, 3, 5, 7, 9, 11$ fs with three independent simulations using each time step. From this, it was determined that 3 fs was the most aggressive yet stable time step. Choosing an aggressive time step was important to this inquiry as it served as a method to determine the extent of the phase space that may be explored by replica exchange and umbrella sampling techniques with a minimal amount of processor time.

Umbrellas of this system were performed with a reaction coordinate of ξ_0 ranging from 30 to 350.75 Å in 0.25 Å increments. Therefore, each pathway along the reaction coordinate required 1284 independent simulations. The data from these simulations were combined using WHAM [96] to create a potential of mean force curve. To test the efficiency of replica exchange molecular dynamics in increasing the phase space sampled, all of the 1284 simulations were repeated as small three box replica exchange simulations. Potentials of mean force were generated from only the first box of the replica exchange simulations, the box at the same temperature as the regular molecular dynamics simulations, 215 K, but the other boxes provided higher energy environments, 225 and 235 K, where the protein could escape energy wells that might trap it. Swaps were attempted every 6 ns (2000 steps). To keep the amount of simulation time constant between regular molecular dynamics and replica exchange, the number of production steps of each box in the replica exchange systems

was a third the number of production steps for the regular molecular dynamics simulations, or 30 ns of equilibration and 30 ns of production for a total of 180 ns for each state point. This gave a qualitative expression of the efficiency, given the same amount of processor time, for the exploration of phase space using the two methods.

This approach was much more qualitative than the quantitative approach used with the Protein A systems. There, numerical values of predicted thermodynamic properties are directly compared between one, two, and three box simulations; direct relationships can be drawn between computation time and thermodynamic results. Here, the general features of the curves from the two methods were compared to determine if, given equivalent amounts of processor time, the methods produced differing results. These two tests provided different vantage points into the quandary of whether replica exchange is more efficient at predicting thermodynamic data of biomolecular systems.

3.3 Results

3.3.1 Protein A

Figure 3.1 shows the results obtained from the simulations of Protein A far below the melting point of the protein; Panel A shows the calculation of potential energy from the simulations, Panel B shows that for heat capacity, and Panel C shows that of free energy of folding. All three schemes obtained the same value for the potential energy of the system after two days of processor time. In the calculations of heat capacity and free energy of folding, the lines for all three schemes follow the same trend; they decrease for ten days then level out and remain constant. The value obtained from the characterization simulations for heat capacity was 0.781 ± 0.001 kJ/mol K. The one box molecular dynamics simulation converged to 0.71 ± 0.01 kJ/mol K, the two-box replica exchange to 0.75 ± 0.01 kJ/mol K and the three-box replica exchange to 0.78 ± 0.01 kJ/mol K. The respective deviations are approximately 9%, 4%, and 0.1%. Characterization simulations calculated the Gibbs energy of folding as -12.7 ± 0.1 kJ/mol. The molecular dynamics, two-box replica exchange, and three-box replica exchange simulations obtained values of -14.2 ± 0.8 kJ/mol, -12.8 ± 0.4 kJ/mol, and -11.6 ± 0.3 kJ/mol respectively. These correspond to deviations of approximately 12%, 0.7%, and 8%.

The results near the melting temperature, in Figure 3.2, are similar. Panel A shows the large fluctuations in potential energy that would be expected for a system folding and unfolding around its melting point. These fluctuations all start to oscillate around the same average value, $\approx$ 118 kJ/mol, by the third day for all three schemes. Panel B shows an increase in the heat capacity for the first ten days of processor time and then all three schemes calculate similar values. Panel C shows the replica exchange schemes calculating similar values for free energy of folding after just two days of processor time, but the regular molecular dynamics scheme obtains the same value by the next data point. Discussion of these results follows the results of the Ribonuclease H simulations.

3.3.2 Ribonuclease H

Figure 3.3 shows the potentials of mean force generated by mechanically unfolding Ribonuclease H. The yellow line depicts the average energy for all of the regular molecular dynamics simulations at every point along the reaction coordinate and the blue line shows the same for the replica exchange simulations. The two lines are very similar and overlap through much of their range. The only area of divergence for the two methods was in the range between 175 Å and 250 Å where the general shape is the same, but the values predicted for the PMF are shifted up to 0.3 Å. Of note, the PMF predicted with regular molecular dynamics in this range contains a small local minimum that is not noticeable in the PMF predicted from replica exchange molecular dynamics.

3.4 Discussion and Conclusions

The original hypothesis on the efficiency of these techniques, that replica exchange is more efficient than molecular dynamics, was flawed. When efficiency was measured as the amount of processer time required to obtain accurate thermodynamic properties, replica exchange molecular dynamics techniques had only a slight advantage over regular molecular dynamics. From the protein A simulations, it was shown that the total amount of real time needed to obtain thermodynamic properties was reduced, but the amount of processor hours required required for the systems to converge to a constant value stayed about the same for all schemes. Although all three schemes converged after the same amount of time, the fact

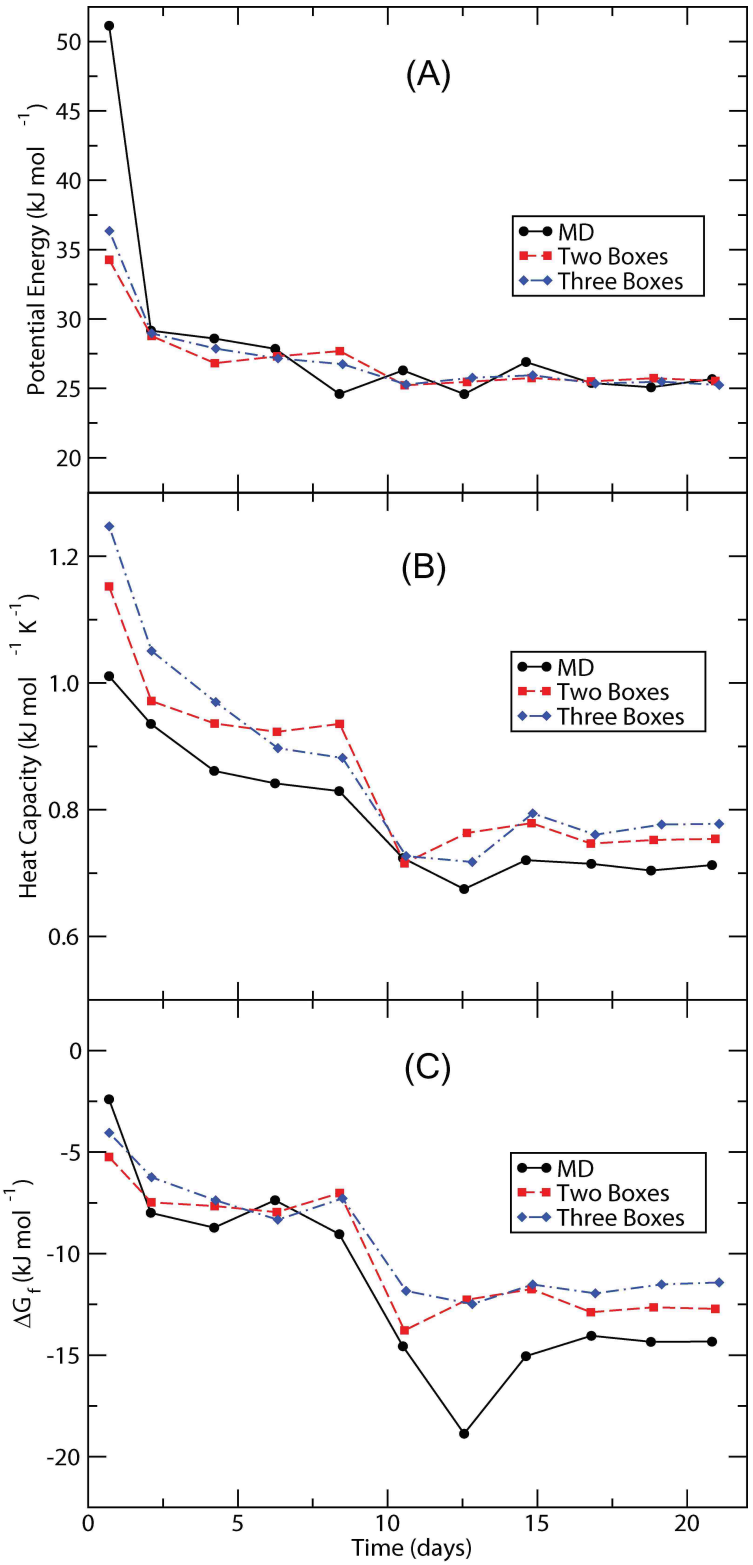

Figure 3.1: Below the melting point, all three schemes calculate similar values within the same amount of simulation time.

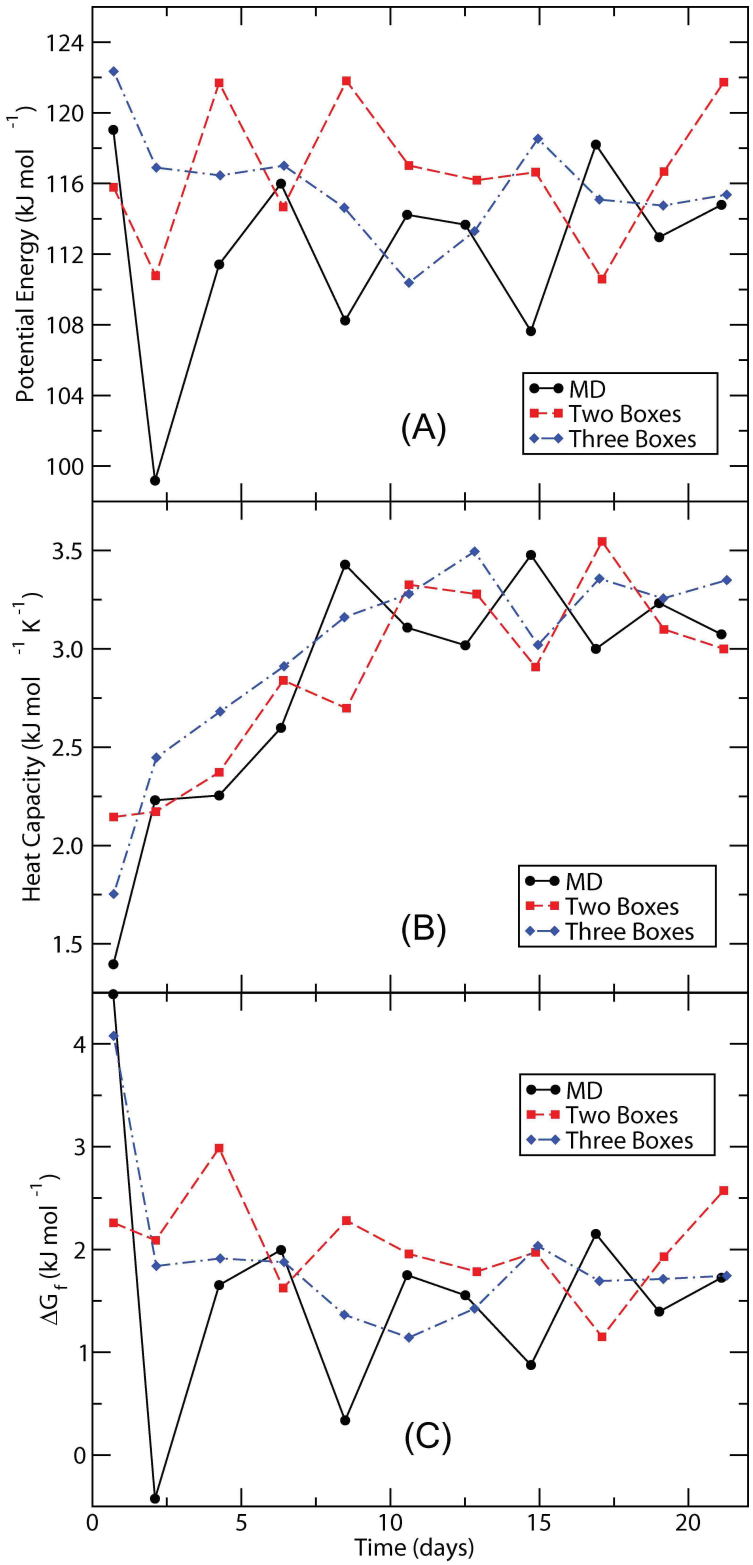

Figure 3.2: At the melting point, all three schemes calculate similar values for any given amount of simulation time.

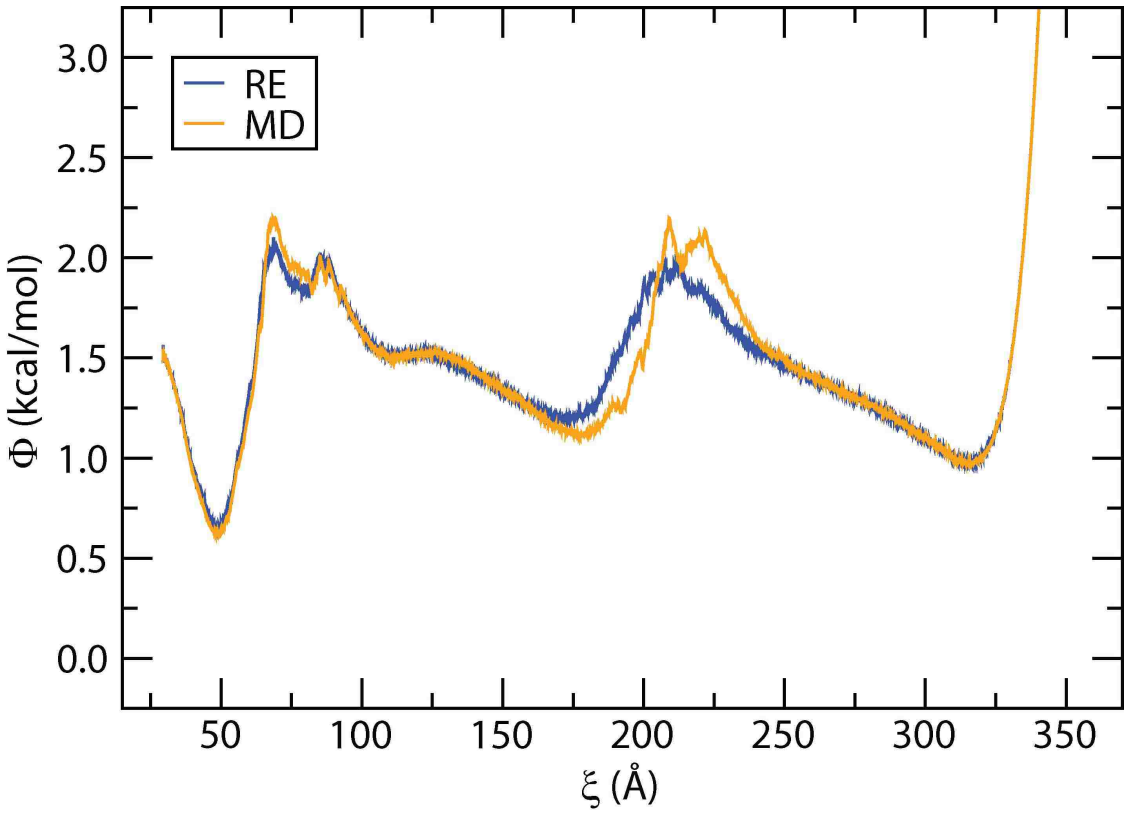

Figure 3.3: PMFs of the mechanical folding of RNase H as predicted from molecular dynamics umbrella sampling with and without replica exchange.

that they converged to different values for the heat capacity and the energy of folding below the melting point raised important concerns.

Initial theories on the reason for this difference in predicted properties were focused on the simulation protocols. These simulations were run without any equilibration steps and the heat capacity and folding energy are time averaged quantities; they are calculated from an average of all production steps. Since they were not given time to move from the idealized, physically unlikely, structure before data was collected, the thermodynamic properties of the crystal were averaged into thermodynamics of the protein at the simulation temperature. The problem with this supposition is that it ignores the relaxation time, the theory that states that, given a sufficient amount of time, a simulation will show no evidence of events that have already happened. It is more likely that this discrepancy is due to barriers in the rough energy landscape.

Heat capacity is calculated as the variance in potential energy. Although all three simulation schemes predicted the same average potential energy, it is possible that were visiting very different energy ranges while remaining around the same average. This would give different heat capacities while showing the same potential energy. The larger value seen in the three box system suggests a larger variance, or that this system was exploring more of the energy range of phase space. Additionally, energy of folding is calculated from the probability of breaking hydrogen bonds based on the captured transition events. The fact that these values are different for the different schemes suggests that some schemes are making more transition events than others, or that some schemes are crossing energy boundaries more than others. Again, the largest value for folding energy was calculated from the three box system. This suggests that it captured more transition events, or that it was crossing more energy barriers and exploring more phase space.

The replica exchange simulations were clearly exploring more phase space, but the hypothesis of the research was with respect to predicting accurate thermodynamic data. Increased phase space sampling is often considered synonymous with good thermodynamic data under the assumption that computer models cover so little simulation time that any real life observation would be made on a system that has spent ten orders of magnitude more time exploring phase space. Conversely, computer simulations are considered viable under the assumption that, at the molecular level, some atoms might be traveling through harder to reach areas of phase space, but the vast majority of atoms are staying in a common area of phase space. It is important, therefore, that efforts to increase phase space sampling do not cause rarely visited areas of phase space to be sampled more frequently than they are visited in real life. This was the reason for the emphasis on thermodynamic data. Although the replica exchange values were closer to the characterization values than the molecular dynamics ones were, the total difference in predicted heat capacities was only 0.07 kJ/mol K. The data for folding energy had a greater spread, but these values were only being compared to another replica exchange system. Fortunately, the ribonuclease H data offered more insight into the problem.

The biggest difference between the replica exchange data and the molecular dynamics data in the ribonuclease H results is the section from 175 to 250 Å. The molecular dynamics

results predict a small energy barrier here with a possible metastable folding state. This suggests that as the molecule is being pulled to unfolded, it snags into a short lived configuration before being pulled completely loose. Replica exchange, configured to facilitate traveling over larger energy barriers associated with other folding transitions of this molecule, effectively flattens out this section of the energy landscape losing these fine details. Although methods that can overcome energy barriers are desired, it is important to conserve the details of the energy landscape. DNA hybridization is characterized by a large energy barrier, associated with the alignment of complementary bases, that could obscure smaller transitions surrounding it. This could lead to the loss of valuable data about the hybridization mechanism; data that is desirable for understanding microarray design. Therefore, it was determined that although replica exchange can increase the sampling of hard-to-reach areas of phase space, this might not always be desirable. Additionally, since replica exchange carries a greater cost in the absolute number if processors required to simulate a given system, a limited resource, it was decided to use only regular molecular dynamics to explore hybridization of DNA on a microarray surface.

Chapter 4

Surface Effects

4.1 Introduction

The first component of the microarray environment that distinguishes it from the environment in which DNA hybridization occurs *in vivo* is the presence of a surface. The environments in which hybridization occurs in nature and most laboratory experiments have negligible interactions with surfaces due to the relative length scales of the hybridizing sequences and the reaction vessels. This is untrue in the case of microarray experiments, however, because half of the hybridizing duplex is tethered to the surface. Not only does this ensure that the hybridizing sequence must interact with the surface, it also restricts the range of motion of the tethered strand and reduces the phase space that the duplex can sample. The complete effect of this reduction in mobility and available phase space is unknown and was the first subject of interest in this study. It was hypothesized that *this reduction in mobility and available phase space would diminish the ability of the two strands to hybridize and reduce the stability of the duplex on surfaces.*

4.2 Methods

4.2.1 Experimental Design

The general approach used to test the hypothesis and determine the effects of the presence of the surface in microarray experiments involved comparing the thermodynamic stability of hybridization in the bulk (the control) with hybridization on the surface. For each system, the potential of mean force, Φ, was calculated as the two strands were brought together and allowed to hybridize. The stability of the duplex was quantified by defining the free energy of hybridization, ΔG_{hyb}, as the free energy of the hybridized duplex at the

minimum of the PMF, minus the free energy when the strands were separated by a long enough distance that they did not interact. In practice, the minimum in Φ occurred at approximately 13.5 Å and the interaction became zero for distances greater than 110 Å, so $\Delta G_{hyb} \simeq \Phi\left(13.5\text{Å}\right) - \Phi\left(110\text{Å}\right)$. To facilitate comparison, $\Delta\Delta G_{hyb}$ was defined as the free energy change which occurs upon hybridization *for the test system* minus the value *for the control*. The effect of the surface can then be quantified as $\Delta\Delta G_{hyb} = \Delta G_{hyb}^{surface} - \Delta G_{hyb}^{bulk}$. If $\Delta\Delta G_{hyb} > 0$, the tested system has inhibited hybridization; if $\Delta\Delta G_{hyb} < 0$, the tested system has enhanced hybridization.

4.2.2 Simulation Protocols

Umbrella sampling was used to simulate the hybridization process. The temperature of each system was maintained at 300 K using the Nosé-Hoover chain method [116]. This temperature is far enough below the melting point of the oligonucleotides to ensure stable duplexes. Each umbrella was simulated with a time step of 1 fs for 4 ns of equilibration and 100 ns of production. Systems usually equilibrated, in terms of the potential energy reaching a steady-state value, in approximately 20 ps of simulation time. Only data from the production time steps were analyzed. In total, each potential of mean force curve consisted of 48.4 μs of simulation time.

Following the methods outlined in Chapter 2, the reaction coordinate for this set of umbrella simulations, ξ, was defined as the distance between the central sugar in each strand. Each point along the reaction coordinate was simulated as an independent system with a biasing potential of the form

$$U_{restraint} = k\left(\xi - \xi_0\right)^2 \tag{4.1}$$

where ξ_0 is the equilibrium value of the reaction coordinate and k is the spring constant of the potential. The distance between the molecules ranged from 10 Å to 130.75 Å in 0.25 Å increments and $k = 10\frac{\text{kJ}}{\text{mol Å}^2}$.

The surface was modeled as an infinite slab of Lennard-Jones particles placed on the z = 0 plane. The radial and angular dependencies were integrated across the respective

domains to create a potential dependent only on the distance of a particle from the plane. The resulting potential was truncated at the minimum and shifted by the well depth to create a purely repulsive, short-range interaction potential is of the form [98]

$$V_{\text{surface}} = \sum_{i}^{N} \left\{ \epsilon_{sur} \left[\left(\frac{\sigma_{sur}}{z_{is}} \right)^{9} - 7.5 \left(\frac{\sigma_{sur}}{z_{is}} \right)^{3} + c \right] \right\} \tag{4.2}$$

where z_{is} is the distance between site i and the surface, $\epsilon_{sur} = 0.0363$ kcal/mol, and the value of σ_{sur} is site-specific. Previous work has shown that the exact value of ϵ_{sur} has little effect on the behavior of the molecule attached to the surface [101]. The parameter c is chosen such that the potential falls smoothly to zero at $z_{is} = \left(\frac{2}{5}\right)^{\frac{1}{6}} \sigma_{sur}$. This model creates a short-ranged, purely repulsive surface which captures the most important features of the surface and has been used previously for similar systems [101]. Additionally, since the dielectric constant is calculated from Equation 2.14 as a function only of temperature and ionic strength, the surface has no effect on the dielectric constant of the system. The probe strand was attached to the surface via a bead-spring, coarse-grain linker based upon the atomistic model of Wong and Pettit [51].

4.2.3 DNA Model

Simulation of DNA hybridization was made possible using the carefully-parameterized coarse-grain model of Knotts that was previously discussed in Chapter 2. Three DNA sequences of varying length were simulated with this model to study the effects of the surface. The first was a target for human topoisomerase II (ACA GCT TAT CAT CGA TCA CGT; PDB ID: 2JYK). This sequence was chosen since it is a biologically important sequence of optimal length for maximum microarray specificity [117]. The other sequences are also biologically significant but are longer and shorter than the first sequence to test the effect of the surface on multiple sequence lengths. The second sequence was an operator for a Restriction-Modification Controller Protein (ATG TGA CTT ATA GTC CGT GTG ATT ATA; PDB ID: 3CLC chain E). The final sequence was a target sequence for a site specific recombinase, Gamma-Delta Resolvase (GCA GTG TCC GAT AAT; PDB ID: 1GDT chain C). Initial coordinates for all sequences were generated following the scheme outlined by

Knotts *et al.* [56] Umbrella sampling was used to induce melting/hybridization in an implicit solvent with a 750 mM salt concentration which was chosen based on experimental precedence [118] and model constraints [76].

4.2.4 Statistics

Due to the large amounts of computer time required for each potential of mean force in this study, replications of results were limited. This limitation impacted the statistical analysis given herein. To obtain an estimate of the statistical significance of the results, several independent umbrella simulations were repeated in their entirety. The results presented in the figures below show the mean of the N independant simulations as a solid line. The uncertainties are represented by the shaded areas, where there was sufficient data to calculate uncertainties, and were estimated as $\frac{\sigma}{\sqrt{N-1}}$, where σ is the standard deviation of the N values of the PMF for each value of ξ. Defined in this manner, the largest error in the ΔG_{hyb} was 0.7 while the smallest was 0.1. As such, $\|\Delta\Delta G_{hyb}\| > 1.4$ identifies treatments which are significantly different from the control.

4.3 Results

4.3.1 Human Topoisomerase II Target Sequence

Figure 4.1 shows the comparison of the potentials of mean force for hybridization in the bulk (the control) and on the surface for the human topoisomerase II target sequence (2JYK). These PMFs are the reversible work done as the target strand approaches the probe strand and hybridizes. The solid line in each PMF represents the average of four sets of simulations and the gray shading represents the error across those simulations at each point on the PMF, calculated as outlined in Section 2.2.1. A low energy minimum is present in both cases at $\xi \approx 13.5$ Å and represents the perfectly hybridized state. At large values of ξ, the PMF of both systems is flat indicating the two strands do not interact. As the two strands move closer together, the PMF increases for two reasons. First, the charged backbones repel each other. Second, an entropic penalty is paid as the interstrand distance is reduced. The degree of increase in the PMF, as the two strands move closer together, is higher on the surface as the excluded volume interaction of the surface reduces

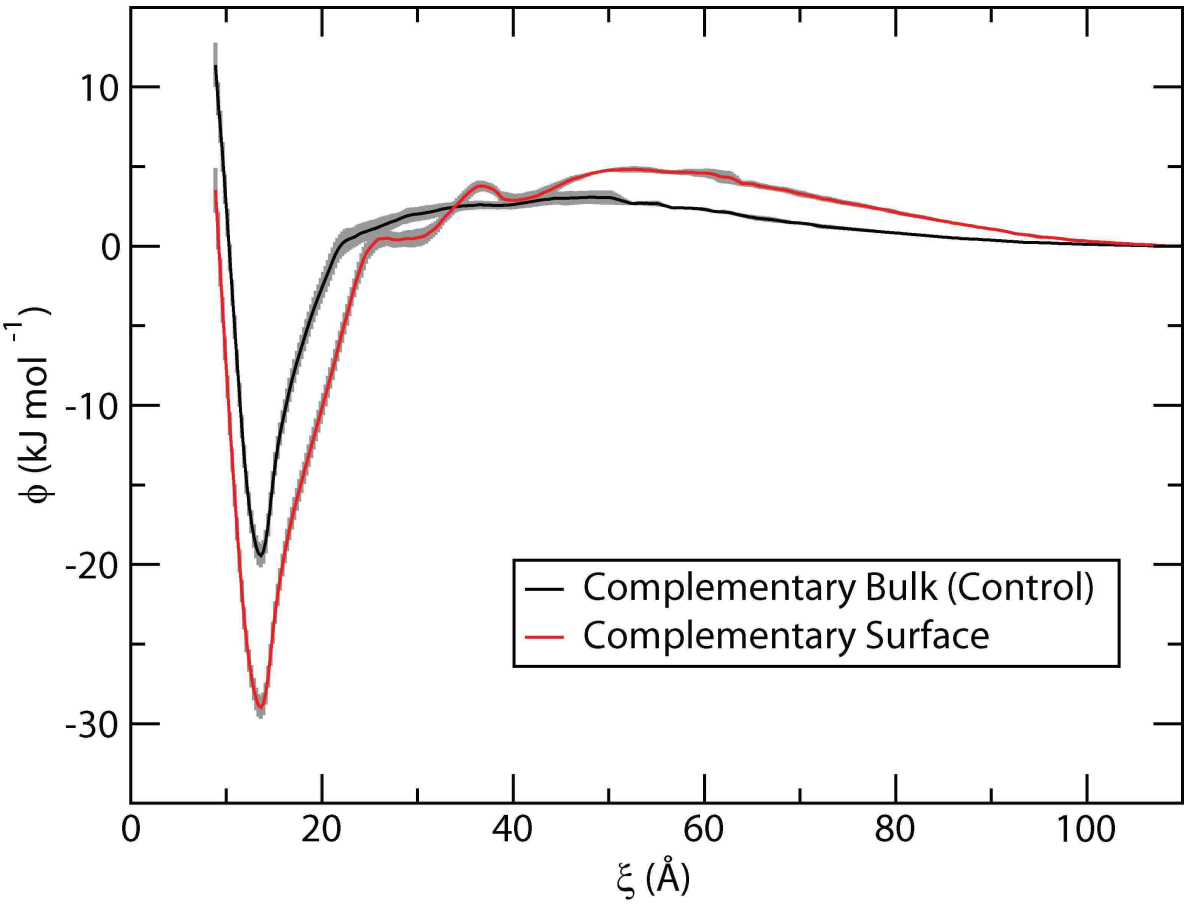

Figure 4.1: Hybridization of the human topoisomerase II target sequence on the surface is thermodynamically more favorable than in the bulk. The shaded regions indicate the standard error of the calculations.

the phase space in which the two strands may move to align with each other. Besides having a larger entropic barrier, the surface cases also have more defined local minima and maxima throughout the entropic barrier region. These features are discussed in detail in Chapter 5.

As described above, the stability of the duplex is quantified by defining the free energy of hybridization as the free energy of the duplex minus that of the separated strands. Defined in this way, the bulk system gives $\Delta G_{\text{hyb}}^{\text{bulk}} = -19.4 \pm 0.726$ kJ/mol and the surface system gives $\Delta G_{\text{hyb}}^{\text{surface}} = -28.8 \pm 0.721$ kJ/mol. Accordingly, $\Delta\Delta G_{\text{hyb}}^{\text{surface}} = -9.4 \pm 1.45$ kJ/mol indicating that, contrary to preconceived expectations, the surface stabilized the hybridizing duplex.

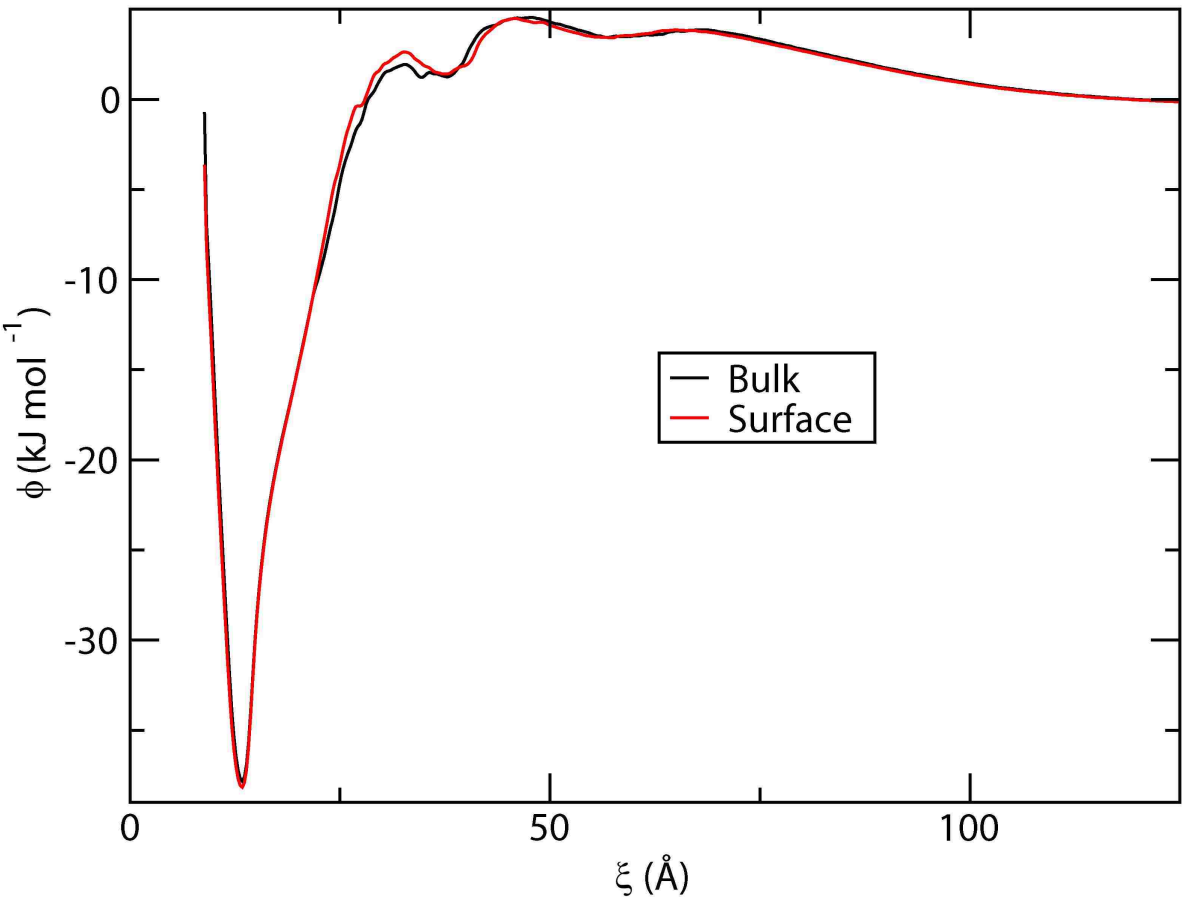

Figure 4.2: Hybridization of the restriction-modification controller protein sequence on the surface is thermodynamically more favorable than in the bulk.

4.3.2 Other Sequences

Similar results were obtained from simulations of the other sequences. Figure 4.2 shows the difference in the potentials of mean force for hybridization of the restriction-modification controller protein sequence both in the bulk and on a surface. In the bulk, the change in free energy for hybridization is $\Delta G_{hyb}^{bulk} = -37.819$ kJ/mol. On the surface, $\Delta G_{hyb}^{surface} = -38.158$ kJ/mol. This translates into $\Delta\Delta G_{hyb}^{surface} = -0.339$ kJ/mol, an insignificantly small stabilization on the surface. Figure 4.3 shows stronger results for the Gamma-Delta Resolvase target sequence. For this sequence, $\Delta G_{hyb}^{bulk} = -15.442$ kJ/mol, $\Delta G_{hyb}^{surface} = -17.678$ kJ/mol, and $\Delta\Delta G_{hyb}^{surface} = -2.236$ kJ/mol. All three sequences were stabilized when hybridization occurred on the surface.

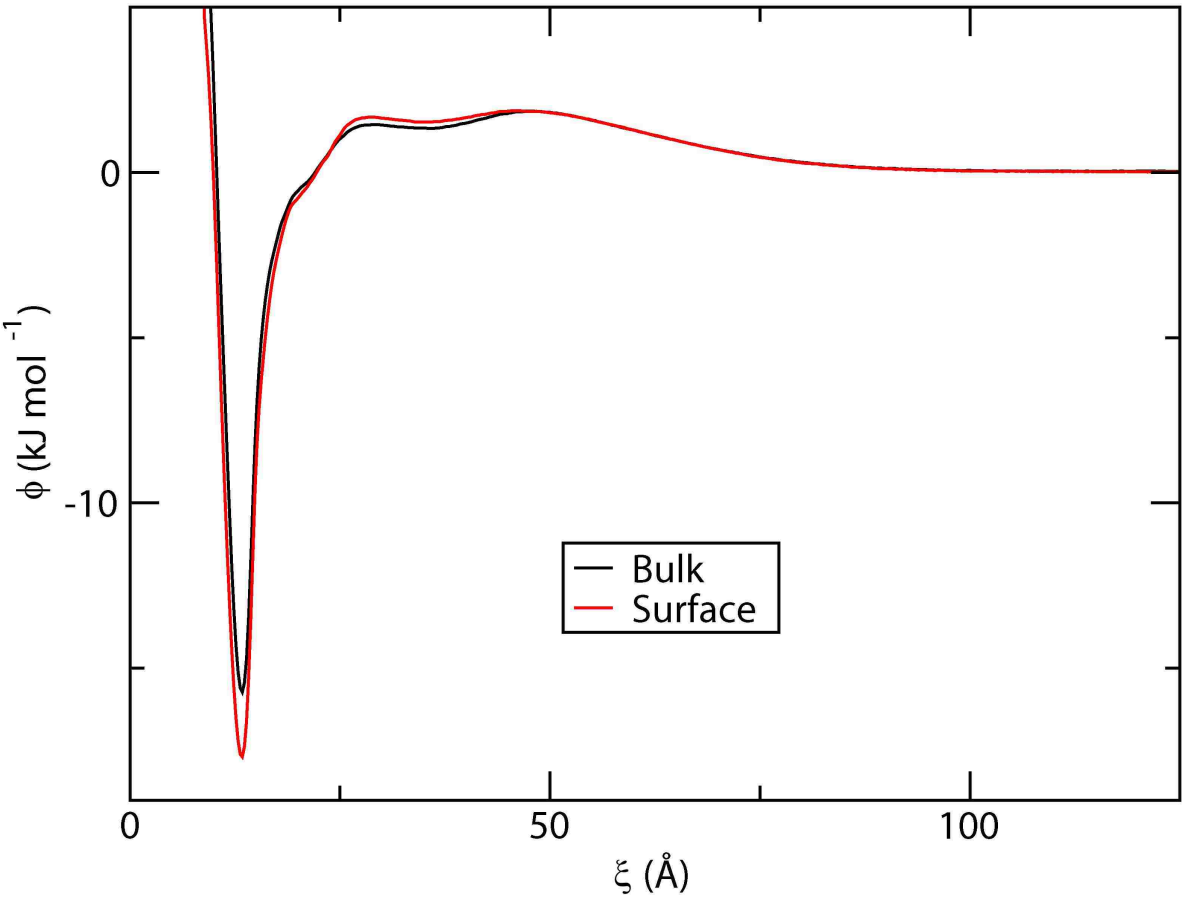

Figure 4.3: Hybridization of the Gamma-Delta Resolvase target sequence on the surface is thermodynamically more favorable than in the bulk.

Table 4.1: Thermodynamics of hybridization on a Surface.

System	ΔG_{hyb} (kJ/mol)	$\Delta\Delta G_{\text{hyb}}$ (kJ/mol)	Base Pairs
Topoisomerase Target in Bulk	-19.4	0	21
Topoisomerase Target on a Surface	-28.8	-9.4	21
Restriction-Modification Controller in Bulk	-37.8	0	27
Restriction-Modification Controller on a Surface	-38.2	-0.339	27
Resolvase Target in Bulk	-15.4	0	15
Resolvase Target on a Surface	-17.7	-2.24	15

4.4 Discussion and Conclusions

Table 4.1 summarizes the data given above. All the surface duplexes were more stable than their bulk counterparts. This stabilization effect varied from not significantly different to six standard deviations of stabilization. Although the stabilization of the restriction-modification controller protein sequence is not significant, it seems to maintained the trend seen in the other sequences, that the surface stabilized DNA duplexes of ideal microarray length. The sequence that was stabilized the most was the human topoisomerase target, which is the sequence that is the closest to the ideal microarray probe length. Chapter 7 shows that this magnitude of stabilization is consistent for other sequences of similar length to the topoisomerase target.

Another noteworthy result seen in the data above is the presence of a local energy minimum at about 40 Å seen in all three sequences. These minima are evidence of a metastable transition state that could arise from the hybridization mechanism. Figure 4.1 shows that when the results from multiple replicates of the entire reaction coordinate are averaged together, the minimum in the bulk case is smoothed out while the minimum in the surface case is enhanced. This phenomenon is explored in Chapter 5. Figure 4.2 shows that the number and size of the minima increase as the strand length increases. This phenomenon is explored in Chapter 7.

4.4.1 Analysis of Hypotheses

The first hypothesis of this research was that *hybridization on surfaces is less favorable than in the bulk*. From a *thermodynamic* perspective, the data indicate that there is little support for this hypothesis. All three tested *surface* duplexes were more stable than their counterparts in the bulk. The fact that the surface duplex is more stable than the bulk duplex confirms recent experimental findings by Hurst *et al.* [119]. In that work, it was found that when DNA was attached to gold nanoparticles, the particles stabilized DNA duplexes with non-complementary sequences. The stabilization effect increased with particle size to such an extent that, for 150 nm particles, duplexes would form with just one matching base pair. It should be noted that as the particle size increases, the particle surface behaves more like a flat microarray surface on the molecular level.

One reason for examining this hypothesis is that previous theories about the poor performance of microarrays were based on the belief that surfaces destabilize DNA duplexes by restricting the ability of complementary ends to correctly align and that microarray accuracy would increase if the destabilization was counteracted. Since the duplexes are already stabilized on surfaces, there must be other factors affecting microarray consistency. These factors could be linked to the hybridization process, or the mechanism of hybridization.

The shape of the PMF curves shown previously offers useful insights into the hybridization process. One would intuitively expect that as the two strands are brought together, they overcome an energy barrier due to Coulombic interactions and the energy needed to align bases. Once the bases are aligned, the strands form their double helix and settle into an energy minimum. As seen in Figure 4.1, the PMFs for the bulk cases seem to support this logic. The surface cases, however, show features not seen in the bulk, such as, the bump between 20 and 50 Å of Figure 4.1. Equilibrium configurations for each point along the PMF were analyzed to determine the cause of the features, and the behavior was linked to the surface's ability to restict the reorientation of probe/target complexes. Representative snapshots of these equilibrium configurations are shown in Figure 4.4. These structures were chosen from the umbrella simulations for being indicative of the hybridization process on a surface. The process begins as the two strands move along each other until they meet in a way that the major and minor grooves are aligned. They then start to wind around each other to start a helical formation. Once this occurs, the two strands slide into the hybridized state by aligning complementary bases. The fact that the bases do not align until after the initial partial helix is formed is seen in steps three and four where the oligonucleotides are clearly offset. Chapter 5 explores these data in greater detail.

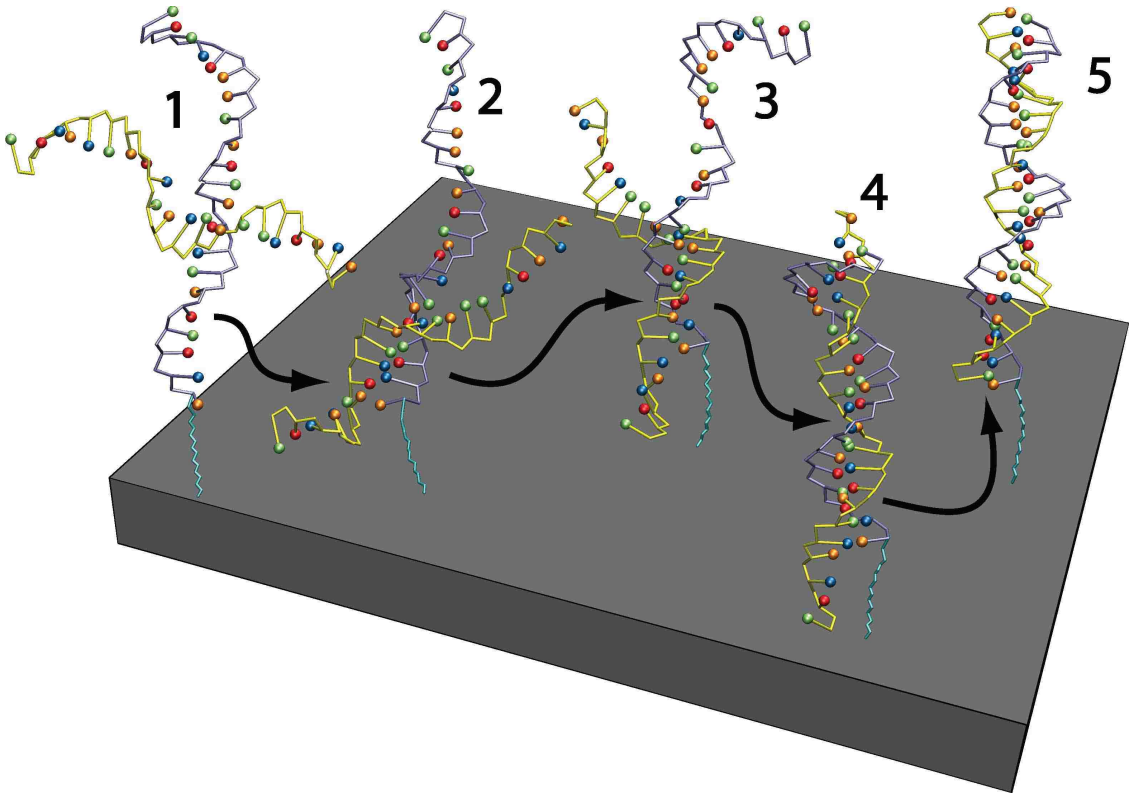

Figure 4.4: Snapshots of the hybridization process on the surface. The two strands of DNA match grooves, wind around each other, and shift into position.

Chapter 5

Analyzing the Hybridization Mechanism with Two-Dimensional Umbrella Sampling

5.1 Introduction

It was shown in Chapter 4 that surfaces improve the stability of DNA hybridization. This unexpected result was calculated from potential of mean force (PMF) curves along a reaction coordinate of strand separation distance by comparing the free energy of the non-interacting ssDNA with the free energy of the hybridized duplex of DNA of the same strands. The PMFs for the human topoisomerase II target also showed that the surface systems, in addition to being more thermodynamically stable, contained an unexpected feature that appears to be a meta-stable intermediate state and was not present in the PMF of the bulk case. The purpose of this chapter was to explain these features by quantifying the mechanisms of hybridization on a surface and in the bulk and better understand why surface duplexes were stabilized. It was hypothesized that these features were evidence of the formation of intermediates in the hybridization mechanism due to a "flipping" transition in which the duplex would switch from a parallel to an anti-parallel orientation. It was proposed that the hybridization mechanism in both surface and bulk cases was the same, but that the greater stability of intermediates and the greater entropic barriers to move into and out of these intermediates on the surface caused these features to only be visible for surface cases.

The hybridization of two strands of DNA occurs in two regimes. The first is a regime governed by colloidal interactions and occurs when the strands are separated by large distances. In this stage, colloidal interactions cause attractive forces which draw the strands together without regard to sequence effects. This means that as two strands approach each other there is roughly a 50% chance that the strands will be properly aligned (*i.e.* the 5'

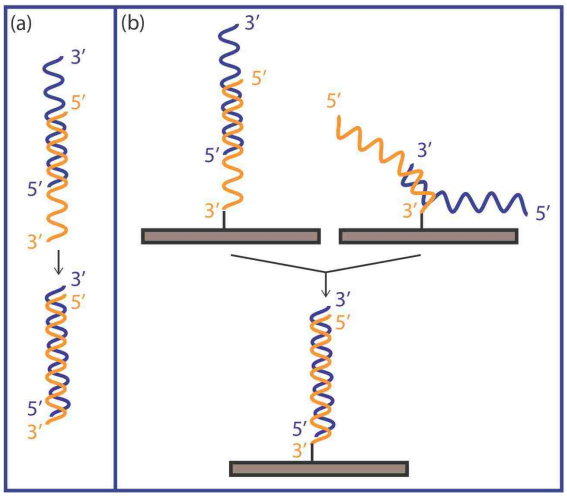

Figure 5.1: The process of hybridization in the bulk, Panel (a), and on the surface, Panel (b), when the approach produces an anti-parallel configuration. No reorientation is needed in these cases and the strands slide into place.

end of one strand with the 3' end of the other) and a 50% chance for improper alignment. In the second regime, which occurs at short interstrand distances, strand-strand interactions are governed by specific base pairing interactions. In this regime, strands slide up and down each other seeking proper sequence alignment and must often "flip" to complete the hybridization process. One important feature of the hybridization process is the presence of an energy barrier separating the two regimes. The previous research suggested that this energy barrier was entropic in nature and originated in the reorientation process. As the two strands approach each other, the strands must overcome this barrier before falling into the low-energy, hybridized state.

If the approach produced a parallel configuration, complete hybridization cannot occur until the strands reorient to be antiparallel. This situation is shown in Figure 5.2. The top panel is possible pathways a bulk duplex may follow to move from a parallel to an antiparallel orientation. In this case the process can occur via movement of only one strand or by both strands simultaneously. Once the strands have moved into an antiparallel orientation, hybridization can proceed as depicted in Figure 5.1. The bottom panel of Figure 5.2 shows how the reorientation process is different for a surface attached duplex. The first distinction is that the probe strand can not participate in the movement due to its connection to the surface. Second, the target strand is limited in the ways it may pivot due

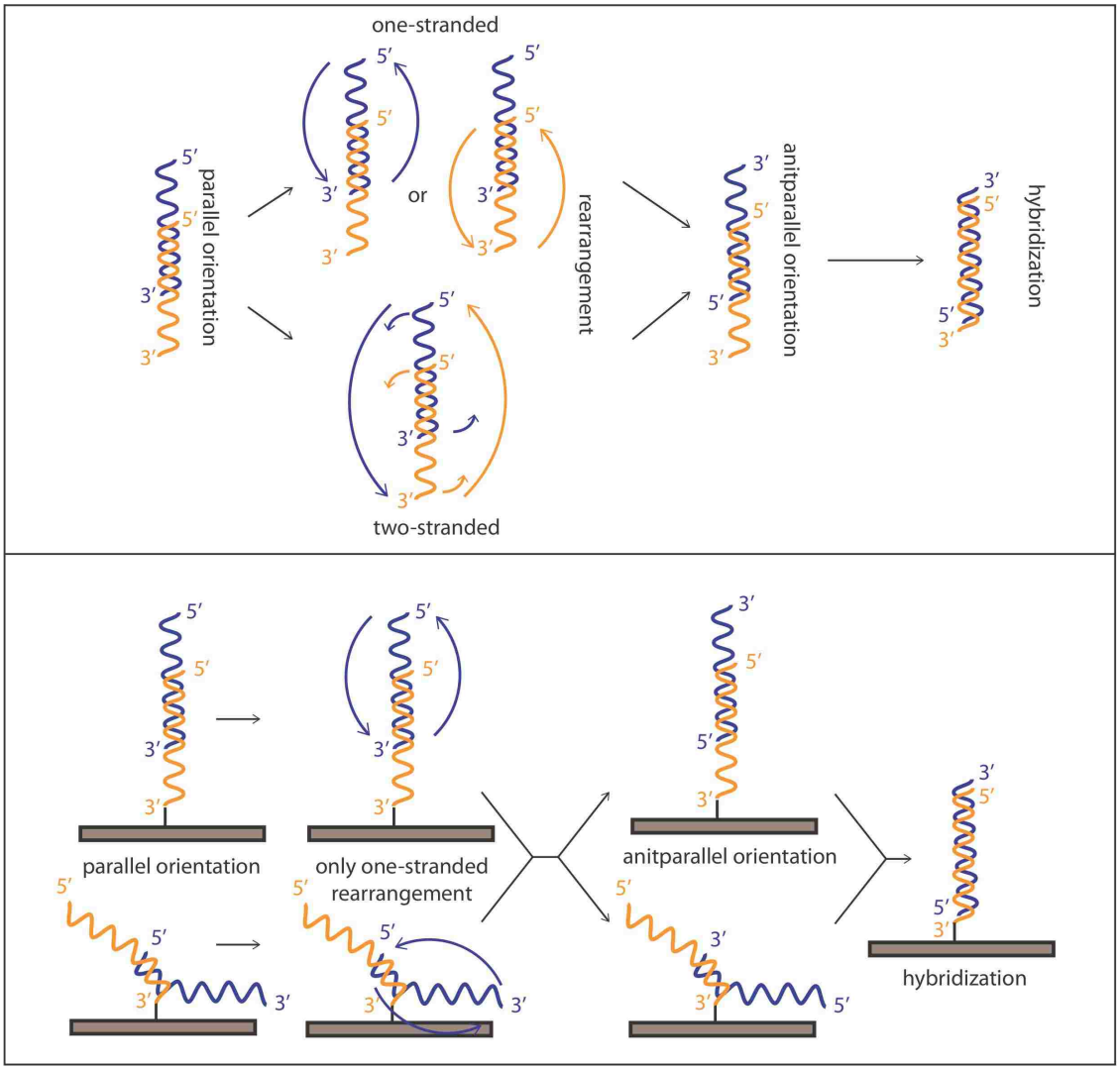

Figure 5.2: The process of hybridization in the bulk, Panel (a), and on the surface, Panel (b), when the approach produces a parallel configuration. Here the strands must reorient into an anti-parallel orientation before completing hybridization.

to the presence of the surface, particularly when the target strand initiates hybridization with the end of the probe strand nearest to the surface. The dual facts that *only* the target strand can reorient, and that reorientation can result in the target approaching the probe from the surface-bound end, could cause the "bumps" seen only in the PMFs for the surface cases, see Figure 4.1.

5.2 Methods

5.2.1 Experimental Design

Determining the mechanism to test the hypothesized method given above required projecting the free energy landscape onto measurable reaction coordinates. As outlined in Chapter 2, the reaction coordinates were chosen as strand separation distance ξ and strand interaction angle θ. These coordinates allowed the calculation of the free energy landscape by moving the two strands through all important states in the hybridization process with two-dimensional umbrella sampling.

Two-dimensional umbrella sampling– using the two reaction coordinates of strand separation distance and the angle formed by the two strands–was used to test if this rough energy landscape was caused by the surface inhibiting transitions between antiparallel and parallel orientations. This technique produced a three-dimensional energy landscape as a function of both of the reaction coordinates. The energy landscape was then used to determine the relative stability of all possible transition states by using the same method outlined in Section 4.2.1.

5.2.2 Simulation Protocols

The simulations of this chapter followed many of the conventions of the previous chapter. The simulated systems were all held at a temperature of 300 K using the Nosé-Hoover chain method [116]. The salt concentration was set equal to 750 mM. Surfaces were modeled with a Weeks-Chandler-Anderson potential [98], to create a short-ranged, purely repulsive surface. The probe strand was attached to the surface via a bead-spring, coarse-grain linker based upon the atomistic model of Wong and Pettit [51]. Umbrellas were used to generate potentials of mean force (PMFs), Φ, of the hybridization process along the reaction coordinates using WHAM [120, 121]. The reaction coordinates spanned from 10 to 50 Å in 0.25 Å increments for ξ and from 0° to 180° in 10° increments for θ. These reaction coordinates required 3059 simulations for each PMF compared to the 484 simulations required per 1D PMF in Chapter 4. Due to this increase in computer time, only the sequence for a target of

human topoisomerase II (ACA GCT TAT CAT CGA TCA CGT; PDB ID: 2JYK) was used for this part of the study.

The PMF was used to quantify the stability of the various duplexes by defining the free energy of a transition, ΔG_{trans}, which was calculated in the same manner as was previously discussed with the following specifications. Several changes of free energy were calculated. The most important was the change in free energy of hybridization, or ΔG_{hyb}, defined as the energy value at a location where the PMF becomes flat minus the value of the PMF for the perfectly hybridized structure, an absolute minimum. The former occurs at long distances of ξ and physically means that the strands do not interact. In practice, the absolute minimum in Φ occurred at approximately $\xi = 13.5$ Å and $\theta = 0°$. Another important change in free energy was that for flipping events where the strands change from a parallel to an antiparallel configuration. As will be shown below, such a transition does indeed occur. To facilitate comparison of the stabilities of the intermediates, the change in free energy for this *target flip transition* is defined as $\Delta G_{flip} \simeq \Phi\left(13.5Å, 0°\right) - \Phi\left(13.5Å, 180°\right)$. To facilitate comparisons, $\Delta\Delta G_{flip}$ was defined as the free energy change of flipping for a certain treatment minus the free change energy of flipping for the control case ($\Delta\Delta G_{flip} = \Delta G_{flip}^{treatment} - \Delta G_{flip}^{control}$). As mentioned above, the control case was bulk hybridization of the probe with the target sequence.

All systems were simulated with a 1 fs time step for 4 ns of equilibration and 100 ns of production time. Equilibration, in terms of the potential energy reaching a steady-state value, usually occurred within 100 ps of simulation time. Only the data from the production steps were analyzed. The total amount of simulation time needed to generate the 2D PMF, calculated as the sum of the simulation times for all umbrellas used to generate the curve, was 319 μs.

5.3 Results

5.3.1 DNA Hybridization Mechanism in Bulk

Figure 5.3 shows the free energy for hybridization of the human topoisomerase II target sequence duplex in bulk as a function of the distance and angle between the two strands. The arrows on the figure indicate representative pathways that the two strands can

take as they proceed from large separation distances to the hybridized state and are only present to help guide the following discussion. The orientations of the strands along each pathway are also found in schematic form in the figure, and the changes in free energy as the system travels along these pathways are summarized in Table 5.1 for convenience. It is important to note that ϕ raises at the edges of the graph due to end effects associated with the cutoff of a reaction coordinate in umbrella sampling techniques.

When the distance separating the stands is greater than $\xi \approx 37$ Å, the strands show little preference with respect to orientation over most θ values. The exceptions occur when θ approaches 0° (antiparallel) and 180° (parallel). The increase in free energy for these two cases is entropic in nature in that a penalty is paid to order the system into such aligned configurations.

As the strands come closer together, they are no longer free to orient themselves at any value of θ. Specifically, an energy barrier begins to appear between the parallel and antiparallel orientations at $\xi \approx 36.5$ Å. This barrier is largest at $\theta = 102.5°$ and separates the two energy minima found on the landscape. Both minima occur at $\xi = 13.5$ Å. One is found at $\theta = 0°$ while the other is found at $\theta = 180°$. The former is the global minimum and is the antiparallel, hybridized state. The latter is a state where the strands are duplexed in the wrong orientation where the complementary bases do not align.

The strands can essentially take two paths as they come closer together. These are indicated as Path A and B on the figure and the resulting configurations are parallel and antiparallel respectively. If the strands follow Path A, producing a parallel state, a reorientation event must take place for correct hybridization to occur. The free energy associated with this flipping event is $\Delta G^{\text{bulk}}_{\text{flip}} = -19.7$ kJ/mol which favors the correctly folded state, but various paths can be followed for this to happen. Path D shows one way that this transition can occur where the strands remain close together. Specifically, if the strand separation distance remains constant at $\xi = 13.5$ Å, the energy barrier that must be overcome for the flip to occur (measured as the difference in free energy between the transition state separating the two minima and the parallel state) is $\Delta G^{13.5\text{ Å, bulk}}_{\text{barrier}} = G^{\text{bulk}}_{\text{transition state}} - G^{\text{bulk}}_{\text{parallel}} = 16.5$ kJ/mol. However, if the strands separate before flipping,

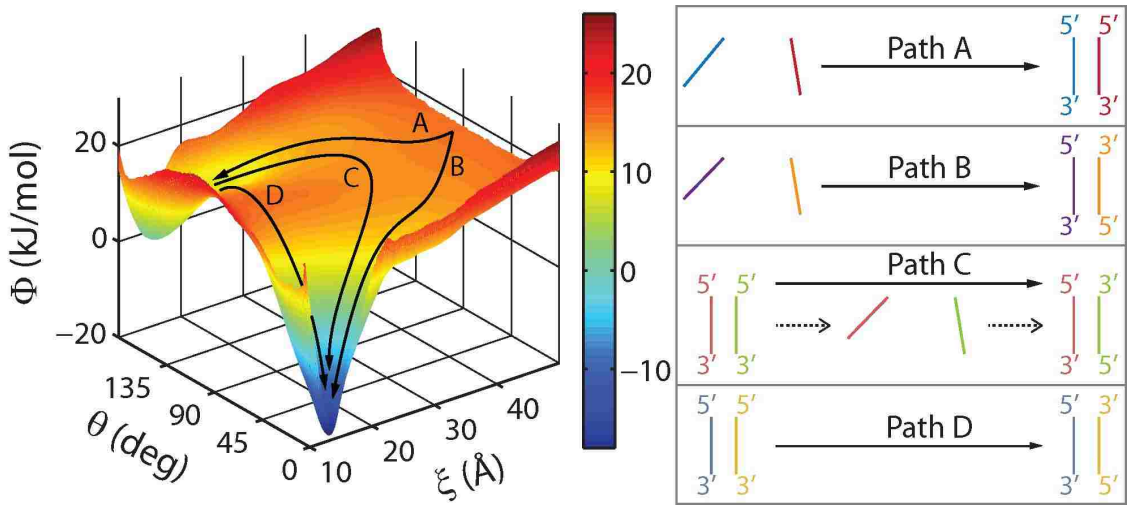

Figure 5.3: Free energy of hybridization of bulk hybridization as a function of strand separation distance (ξ) and angle (θ).

as illustrated in Path C, this barrier is reduced. For example, if the strands separate to approximately $\xi = 40$ Å before reorienting, $\Delta G_{\text{barrier}}^{40\text{ Å, bulk}} = 10.0$ kJ/mol.

5.3.2 DNA Hybridization Mechanism on the Surface

Figure 5.4 shows the free energy of hybridization on the surface as a function of ξ and θ. As before, the arrows are representative pathways that the system can follow, the orientations of the strands along each pathway are shown in schematic form, and the changes in free energy are summarized in Table 5.1. Similar to the bulk situation, two minima exist at close distances corresponding to the parallel ($\theta = 180°$) and antiparallel ($\theta = 0°$) configurations. However, the bulk and surface cases are different in multiple ways, including the depths of the energy wells, the heights of the barriers between the wells, and the number of wells, as is now described.

Unlike the bulk case, a third minimum is found at $\xi \approx 41$ Å and $\theta \approx 160°$ for hybridization on the surface. The presence of this additional intermediate helps the system arrive at the correct, hybridized state more easily than in the bulk. The two strands can follow Path A and produce the antiparallel duplex without a flip occurring. If the strands follow Path B, the system has two options. It can proceed along Path C to form the parallel duplex or it can proceed through Path D to form the correct final structure. Following

Table 5.1: Changes in free energy for the hybridization process of 2JYK. Data for duplexes in the bulk and on the surface.

ΔG	Value (kJ/mol)	Description
$\Delta G_{\text{flip}}^{\text{bulk}}$	-19.7	Change in free energy to flip from parallel to antiparallel in the bulk. (See Paths C and D of Figure 5.3.)
$\Delta G_{\text{barrier}}^{\text{13.5 Å, bulk}}$	16.5	The barrier that must be overcome to flip in the bulk if the strands remain at close distances (ξ = 13.5 Å). (See Path D of Figure 5.3.)
$\Delta G_{\text{barrier}}^{\text{40 Å, bulk}}$	10.0	The barrier that must be overcome to flip in the bulk if the strands first separate. (See Path C of Figure 5.3.)
$\Delta G_{\text{flip}}^{\text{surface}}$	-43.5	Change in free energy to flip from parallel to antiparallel on the surface. (See Path E and the reverse of Path C followed by Path D of Figure 5.4.)
$\Delta G_{\text{Path C}}^{\text{surface}}$	-0.2	Change in free energy to move from the intermediate to the parallel state on the surface. (See Path C of Figure 5.4.)
$\Delta G_{\text{Path D}}^{\text{surface}}$	-43.7	Change in free energy to move from the intermediate to the antiparallel state on the surface. (See Path D of Figure 5.4.)
$\Delta G_{\text{barrier}}^{\text{13.5 Å, surface}}$	3	The barrier that must be overcome to flip on the surface if the strands remain at close distances (ξ = 13.5 Å). (See Path E of Figure 5.4.)

Path C, while energetically favorable with $\Delta G_{\text{Path C}}^{\text{surface}} = -0.2$ kJ/mol, requires the system to overcome an energy barrier on the order of k_bT which will reduce the kinetic likelihood that this transition occurs. However, Path D follows a valley through the free energy landscape where the barriers are all much less than k_bT. This transition is not only thermodynamically favorable, with $\Delta G_{\text{Path D}}^{\text{surface}} = -43.7$ kJ/mol, it is also kinetically more likely as the barriers are very small.

If the strands arrive in the parallel duplex, the system can proceed to the correct final state by either separating and flipping (*e.g.* traveling backwards along Path C and then taking Path D as was discussed above) or by flipping at close distances. The latter is energetically favorable with $\Delta G_{\text{flip}}^{\text{surface}} = -43.5$ kJ/mol compared to a value of -19.7 kJ/mol in the bulk. Also different from the bulk case is the energy barrier that must be overcome

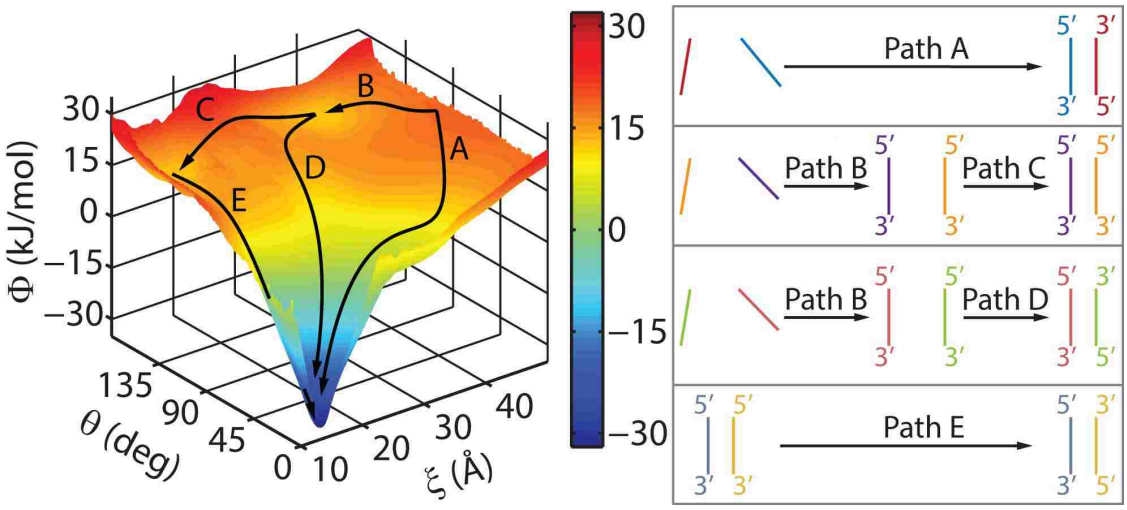

Figure 5.4: Free energy of hybridization on the surface as a function of strand separation distance (ξ) and angle (θ).

to flip at close distances. Specifically, $\Delta G_{\text{barrier}}^{13.5\text{ Å, surface}} = 3$ kJ/mol on the surface compared to the bulk value of 16.5 kJ/mol.

5.4 Discussion and Conclusions

5.4.1 Analysis of Results

Figures 5.3 and 5.4 show that flips occur in both the bulk and surface cases which supports the theory that DNA hybridization occurs via a mechanism with metastable states that are due to the strands moving from parallel to antiparallel orientations. However, no intermediate state is found in the bulk. Specifically, in the bulk, two strands separated at long distances will be attracted to each other, but the attraction can lead to either a parallel or an antiparallel stable state at short distances. Until ξ approaches the value found at hybridization there is no preference for parallel vs. antiparallel configurations and no low-energy minimum in the energy landscape. On the surface, the situation changes such that a stable intermediate forms at longer distances. This intermediate, located at $\xi \approx 41$ Å and $\theta \approx 160^\circ$ is such that the completion of the hybridization preferentially follows a path that leads to the antiparallel hybridized state rather than a parallel state (Path D rather than Path C of Figure 5.4). This intermediate state is a parallel configuration which requires a flip to occur to complete hybridization. Thus, there is evidence for the theoretical hybridization

mechanism with metastable intermediate states on the surface. In the bulk, however, no intermediate is found but flips are still required to complete hybridization if the system arrives at the parallel state at close distances.

The results presented above give rise to two important questions. First, if surfaces stabilize duplexes, why is the parallel duplex less stable on the surface than it is in the bulk? Surface duplexes are stabilized by entropic effects. The surface acts as a barrier that blocks the duplex from exploring certain areas of phase space. This reduction in available phase space lowers the entropy of the system, which increases the thermodynamic stability. The enthalpy of the system also contributes to the thermodynamic stability of the duplex. It is likely that the decreased entropy of the parallel duplex on the surface is accompanied by a decrease in enthalpy that is greater in magnitude than the change in entropy and therefore represents a net decrease in thermodynamic stability.

The second important question raised by Figure 5.4 is why is there an energy minimum at $\xi \approx 41$ Å and $\theta \approx 160^\circ$ on the surface but not in the bulk? The reduction in free energy is most likely due to interactions between the two strands since the formation of hydrogen bonds between complementary bases will lower free energy. On the surface, the degeneracy of ways in which the two strands can arrange themselves at a separation distance of 41 Å and an angle of 160° is much less than in the bulk. This reduction in possible configurations forces the strands on the surface over any energy barriers blocking the formation of these hydrogen bonds while strands in the bulk may continue through configurations of low energy that do not require crossing those barriers.

5.4.2 The Thermodynamic Origins of Surface Stability

Chapter 4 showed that surfaces stabilize DNA hybridization of tethered probe/target complexes over their bulk counterparts, but the reason for the stabilization was not known [122]. The results presented above demonstrate that there are at least two reasons for the stabilization. The first, which was discussed in Section 5.3.2, is that the surface creates a stable intermediate at long distances that alters (compared to the bulk case) the free energy landscape in such a way that the system can follow a low-energy valley from long distances to the properly-hybridized state. There is no such valley connecting the intermediate state to the

parallel state, so the system preferentially follows a path to arrive at the correct final state. In the bulk, no path to either the parallel or antiparallel state follows a low-energy valley. The presence of such a path on the surface results in a more stable hybridization process.

A second factor, the ease with which a "flip" can occur at short distances, also contributes to the stabilization of hybridization on the surface. In both the bulk and on the surface, it is possible for the two strands to arrive at short intermolecular distances in both antiparallel and parallel configurations. The "flip" is more easily done on the surface because it 1) decreases the free energy of the system by a greater amount than in the bulk and 2) has to overcome a smaller energy barrier to do so. Specifically, $\Delta G_{\text{flip}}^{\text{surf}} = -43.5$ kJ/mol compared to a value of -19.7 kJ/mol in the bulk and $\Delta G_{\text{barrier}}^{13.5\text{ Å, surf}} = 3$ kJ/mol compared to the bulk value of 16.5 kJ/mol. This means that, when compared to the bulk case, both the overall move from parallel to antiparallel *and* the first step of overcoming the energy barrier are more thermodynamically favorable. Specifically, the difference between the surface and the bulk cases for the flip gives $\Delta\Delta G_{\text{flip}} = -23.8$ kJ/mol and $\Delta\Delta G_{\text{trans}} = -13.5$ kJ/mol.

In summary, the stabilizing influence of the surface arises from at least two factors. First, the presence of the intermediate on the surface drives any duplexes that are parallel at long distances towards the antiparallel configuration along the energy valley (Path D of Figure 5.4). Second, any surface system that finds itself in the parallel duplex can more easily make the change to the antiparallel configuration compared to the bulk case. This is true whether the system first separates to the intermediate and follows Path D or follows Path E which has a lower energy barrier than in the bulk (See Figure 5.4). Chapter 7 expands on the universality of these results by applying a similar treatment to a completely different sequence.

Chapter 6

The Effects of Strand Manipulation

6.1 Introduction

In a microarray system, not only does hybridization occur on a surface, it must also occur with high fidelity. Probe strands should only hybridize with perfectly complementary target sequences. Hybridization with non-complementary sequences would result in false positives. It is important, therefore, to determine how preferentially a DNA microarray probe strand hybridizes to the target it is designed to find over a target strand that is similar but not completely complementary. In microarray design, this preference is called specificity. This chapter explores the relative stabilities of perfectly complementary DNA duplexes and DNA duplexes with a single nucleotide mismatch, or a single nucleotide polymorphism (SNP), both in the bulk and on a surface. The first hypothesis of this chapter was that *the differences in stability of perfectly complementary DNA duplexes and SNP duplexes would be significant and measurable both in the bulk and on a surface.*

The second goal of this chapter was to find ways to enhance hybridization and increase selectivity. It was proposed that stretching one strand of DNA along its backbone would increase the accessible surface area of the bases and thereby allow the bases on the strands to interact more freely. The increased interactions would then lead to higher fidelity in the hybridization process both in the bulk and on a surface. These hypotheses were motivated by work by Southern *et al.* [123] who proposed that stretching facilitates molecular recognition between the two molecules thereby enhancing hybridization. The second hypothesis of this chapter was that *stretching the probe strand would preferentially stabilize complementary sequences.*

6.2 Methods

6.2.1 Experimental Design

The simulations and analysis used for this chapter follow the same pattern as Chapter 4. Umbrella sampling techniques were used to simulate hybridization in the bulk and on a surface along a single reaction coordinate, ξ. PMFs were then generated from these simulations and ΔG_{hyb} was calculated as a measurement of duplex stability. $\Delta\Delta G_{hyb}$ was then calculated to determine the relative stabilities of varying duplexes and to quantify the specificity of microarray systems. The first hypothesis of this chapter was tested by calculating $\Delta\Delta G_{hyb}$ of SNP duplexes in the bulk and on a surface and the second hypothesis was tested by calculating $\Delta\Delta G_{hyb}$ of complementary and SNP duplexes with backbone restraints in the bulk and on the surface. For a better contrast among the simulations with backbone restraints, one strand was also compressed along its backbone to decrease the accessibility of its bases. In all cases, the control system used to determine the $\Delta\Delta G_{hyb}$ of a system was the completely complementary duplex in the bulk with no artificial backbone restraints. See Chapter 4 for more details.

6.2.2 DNA Model

This chapter uses the same DNA model and sequences that were used in previous chapters. Single nucleotide polymorphisms were modeled by changing the central nucleotide in the target strand to a non-complementary pyrimidine while keeping the probe strand the same. For human topoisomerase II, the probe strand remained the same (ACA GCT TAT **C**AT CGA TCA CGT) and the target strand was ACG TGA TCG AT**T** ATA AGC TGT. This sequence represents a single nucleotide polymorphism at the tenth base pair (shown in boldface type). Mismatched strands were simulated under the same conditions as the perfectly complementary strands with the single reaction coordinate ξ to generate one dimensional PMFs. The stability of the duplexes was quantified by calculating $\Delta\Delta G_{hyb}$ in an analogous manner as described previously.

6.2.3 Simulation Protocols

Simulations were performed in the same manner as has been previously discussed with the following changes. To assess how efforts to improve hybridization on the surface affects microarray performance, simulations were performed where the probe strand was stretched or compressed by an external force. Stretching and compressing does not correspond to any process currently used in microarray experiments but could serve in the future as a way to thermodynamically stabilize DNA hybridization [118]. The idea behind stretching is to improve molecular recognition between the hybridizing strands. In nature, oligonucleotides form coils and hairpins in solution. To hybridize, the molecule must elongate enough for the bases on both strands to interact. Stretching by external means is a way to drive this process and is expected to improve the hybridization process. In contrast, compressing will increase the propensity of the ssDNA molecule to fold back on itself which inhibits the base/base interactions.

In simulation, one of the DNA strands was stretched or compressed with a harmonic potential of the same form as Equation 4.1 with $k = 42\frac{\text{kJ}}{\text{mol Å}^2}$, tuned to maintain the backbone length of the probe molecule. The "stretching/compressing" potential acted between the first and last sugars of the probe strand and the distance was maintained at values of 80%, 85%, 90%, 95%, 105%, 110%, and 115% of the equilibrium distance found in dsDNA when no external force was applied (66.3Å for the human topoisomerase II target).

6.3 Results

6.3.1 Single Nucleotide Polymorphisms

The effect of a SNP on hybridization in the bulk and on the surface of the human topoisomerase II target sequence is depicted in Figure 6.1. The complementary bulk and surface cases are shown along with the SNP cases in the bulk and on the surface. The least stable process is hybridization of the SNP in the bulk with $\Delta G_{\text{hyb}}^{\text{bulk SNP}} = -18.7$ kJ/mol. This gives $\Delta\Delta G_{\text{hyb}} = 0.7$ kJ/mol which is not significantly different from the bulk case. On the surface, the SNP duplex is more stable than the control with $\Delta G_{\text{hyb}}^{\text{surface SNP}} = -29.4\frac{\text{kJ}}{\text{mol}}$ and $\Delta\Delta G_{\text{hyb}} = -10.0$ kJ/mol. These data indicate that, on a surface, mismatched duplexes

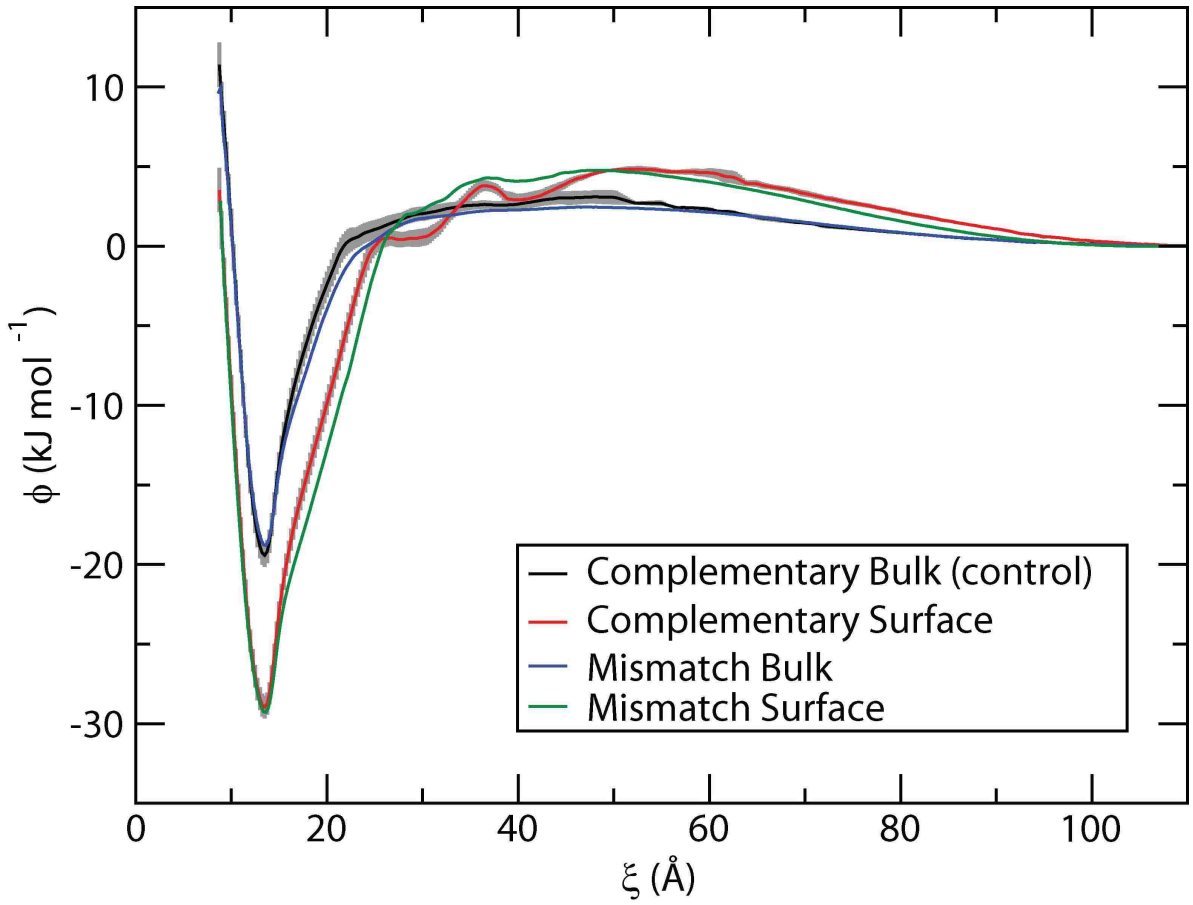

Figure 6.1: The surface stabilizes both the complementary and the mismatched sequences compared to the bulk case.

were more stable than perfectly matched duplexes in bulk. The most stable systems occured on the surface for both the complementary and the mismatched strands. Similar results were found for the other two sequences.

6.3.2 The Effects of Manipulating Base Accessability in the Bulk

Figure 6.2 shows the effect of stretching/compressing the probe strand on the hybridization of perfectly complementary strands. As discussed above, multiple degrees of stretch and compression (80%, 85%, 90%, 95%, 105%, 110%, and 115%) were simulated; however, only two cases are shown (along with the control) for clarity. Those depicted are stretching to 110% of the native length and compressing to 95% of the native length since

these were the most stable degrees of stretching and compressing. The results for all cases, those shown and not shown, are consistent with the discussion that follows.

The free energy barrier for hybridization, the entropic barrier that must be overcome for hybridization to occur, is similar in height for all cases peaking at $\xi \approx 45$Å but the depth of the free energy minimum for the duplex changes as the probe strand is stretched or compressed. For all compressed strands, the minimum free energy of the duplex is greater (less negative) than that of the duplex without any stretch/compression. The 95% PMF is shown as it results in the least amount of destabilization with $\Delta\Delta G_{hyb}^{95\%} = 0.6$ kJ/mol, a value that is not significantly different from the control. The implications of this will be discussed later. As the probe is stretched, the minimum free energy for the duplex decreases up to a certain point (110%). After this, further stretching causes the free energy of the duplex to rapidly increase (not shown). At 110%, the most stable case, $\Delta\Delta G_{hyb}^{110\%} = -1.9 \pm 0.9$ kJ/mol.

6.3.3 Effects of Compressing Mismatched Strands on the Surface

Figure 6.3 emphasizes the effect of compressing the probe strand on the hybridization of complementary and mismatched strands on the surface. Depicted are four curves: the control, non-compressed complinentary strands on the surface, 95% compressed complementary strands on the surface, and 95% compressed mismatched strands on the surface. As has been seen in other cases, the barrier to hybridization is similar in each instance, but the depth of the well in the hybridized state is different. Each of the surfaces cases is more stable than the control. The most stable states are the uncompressed, complementary, surface case (the same system seen in Figure 4.1) and the 95% compressed surface case, which are not significantly different. These results are a combination of two competing forces. Compression acts to hinder hybridization and destabilize the duplex compared to the control (see Figure 6.2) while the surface acts to stabilize hybridization (see Figure 4.1). However, the surface effects dominate the behavior yielding $\Delta\Delta G_{hyb}^{surface,\ 95\%} = -9.6 \pm 0.4$ kJ/mol which is not significantly different than the $\Delta\Delta G_{hyb}^{surface} = -9.4 \pm 1.5$ kJ/mol for the uncompressed state. In contrast to the complementary strands, compression affects mismatched strands

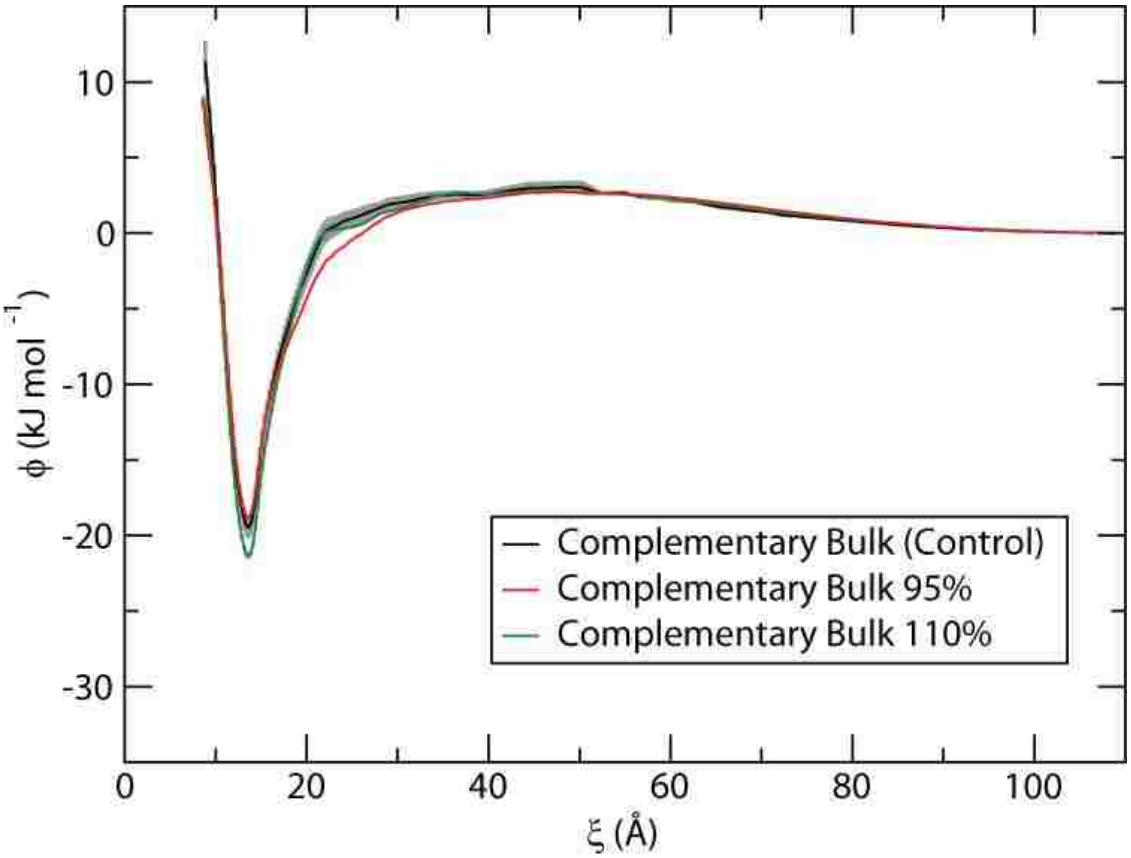

Figure 6.2: Stretching slightly enhances the stability of the duplex while compressing slightly destabilizes the duplex compared to the bulk.

to a greater degree. For the 95% mismatched surface case, $\Delta\Delta G_{hyb}^{\text{surface, SNP, 95\%}} = -7.6 \pm 0.1$ kJ/mol.

6.3.4 Effects of Stretching Mismatched Strands on the Surface

Figure 6.4 illustrates the effect of stretching both complementary and mismatched strands tethered to a surface. Four cases are depicted: the control, unstretched complementary strands on the surface, 110% stretched complementary strands on the surface, 110% stretched mismatched strands on the surface. As has been seen before, the barrier to hybridization is greatest on the surface, but the least stable state is the complementary strands in the bulk (the control). The most stable state is stretching of the complementary strand on the surface with $\Delta\Delta G_{hyb}^{\text{surface, 110\%}} = -14.3$ kJ/mol. Strikingly, the next-most stable situation is stretching the mismatched strand on the surface with $\Delta\Delta G_{hyb}^{\text{surface, SNP, 110\%}} = -11.9$

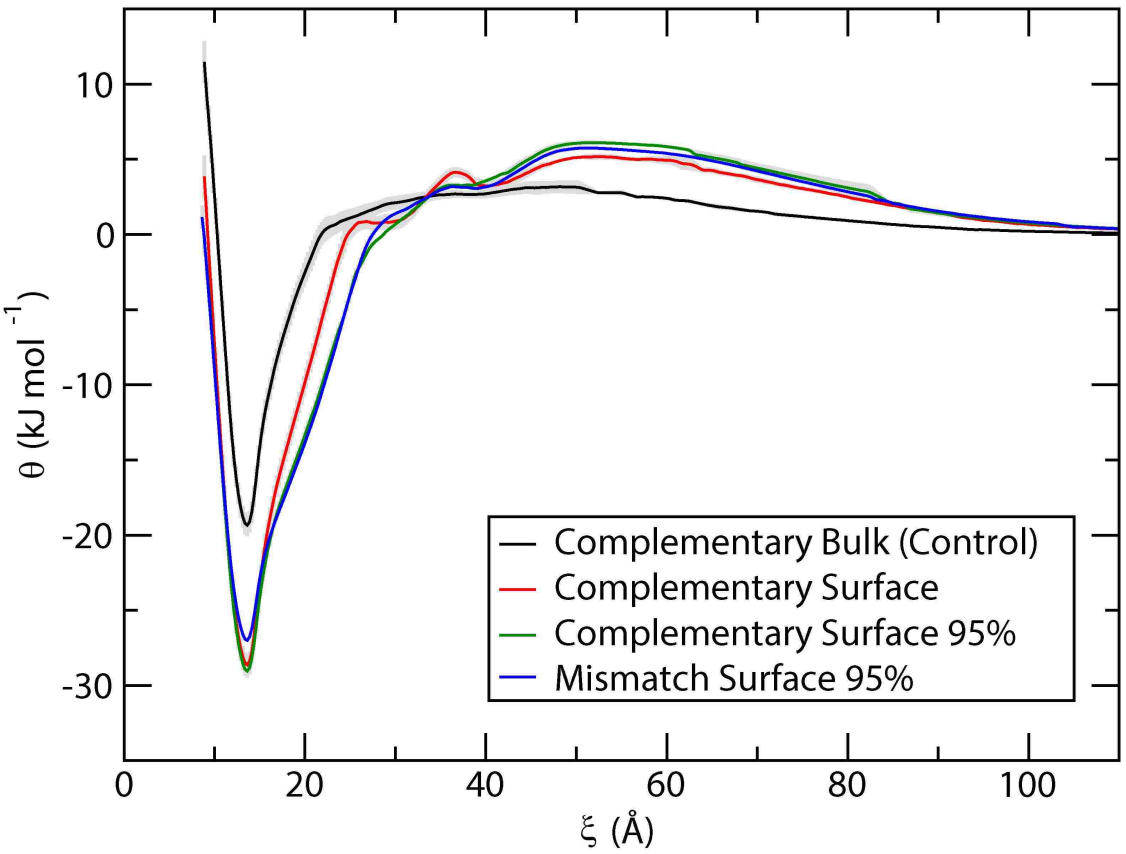

Figure 6.3: Compressing the probe strand on the surface has no significant effect on the complementary duplex but significantly destabilizes the duplex with a SNP.

kJ/mol. This represents a stabilization of ≈ 1 kJ/mol over the complementary surface case with no stretch (see Figure 4.1).

6.3.5 Summary of Results

The data presented in the figures found in this section were selected to make the most useful comparisons as placing all of the data on one figure would have made identification of trends difficult. To summarize all of the results, Table 6.1 contains ΔG_{hyb} and $\Delta\Delta G_{\text{hyb}}$ for every situation discussed above. The table is ordered with the most stable duplex at the top and the least stable duplex at the bottom.

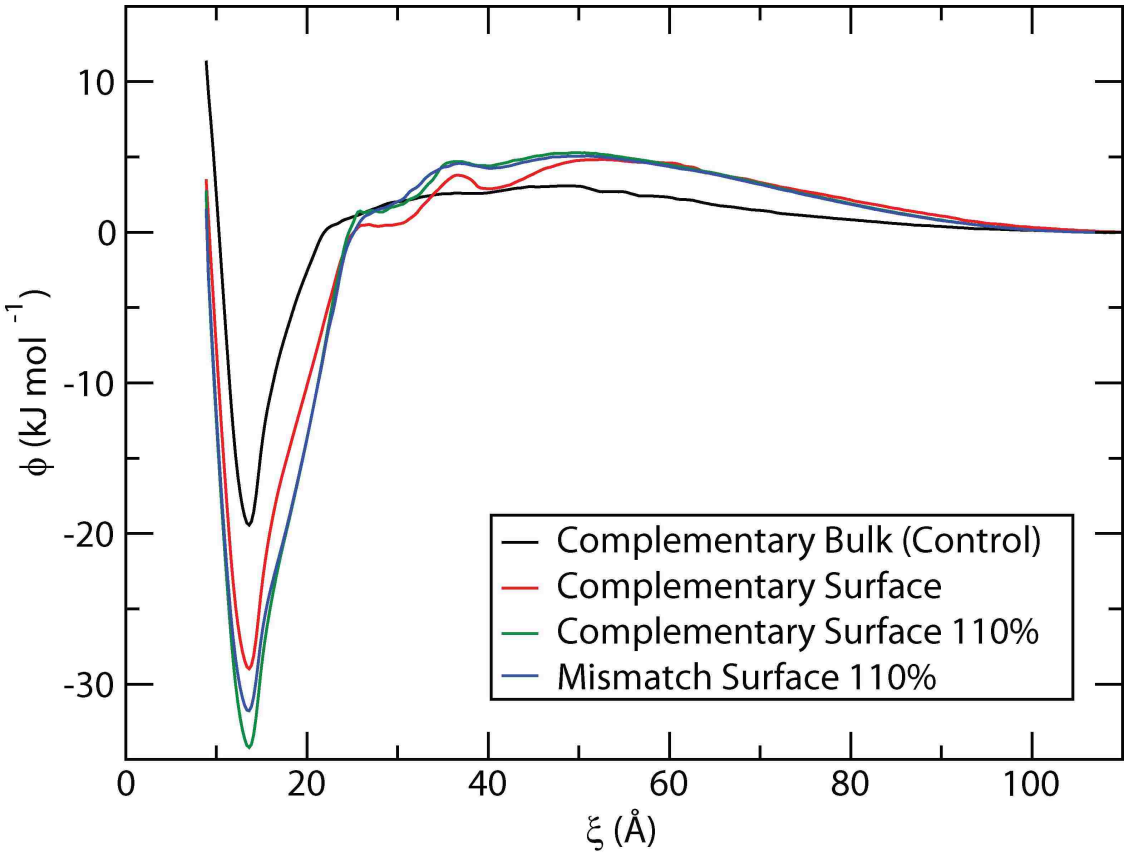

Figure 6.4: Stretching stabilizes both the complementary and the mismatched strands.

Table 6.1: Thermodynamics of hybridization for complementary and mismatched strands. Data for duplexes in multiple environments.

System	ΔG_{hyb} (kJ/mol)	$\Delta\Delta G_{\text{hyb}}$ (kJ/mol)
complementary, Surface, 110%	-33.7	-14.3
SNP, Surface, 110%	-31.3	-11.9
SNP, Surface	-29.4	-10.0
complementary, Surface, 95%	-29.0	-9.6
complementary, Surface	-28.8	-9.4
SNP, Surface, 95%	-27.0	-7.6
complementary, Bulk, 110%	-21.3	-1.9
complementary, Bulk (Control)	-19.4	0.0
complementary, Bulk, 95%	-18.8	0.6
SNP, Bulk	-18.7	0.7

6.4 Discussion and Conclusions

The first hypothesis of this chapter was that *a SNP would significantly decrease the stability of a DNA duplex*. This was found to be incorrect as neither of the SNPs that were simulated without backbone restraints were significantly less stable than their complementary counterparts. The implications of this finding are of greater import for the surface case. As in previous chapters, this set of simulations found that the surface stabilizes duplexes. While this might increase the sensitivity of a microarray device, it does not improve its specificity, or the ability to preferentially hybridize with completely complementary sequences over mismatching sequences, given that the stability of a SNP duplex on a surface is not significantly different than the stability of a complementary duplex on a surface. Further more, since hybridization of complementary strands on a surface is more stable than in the bulk, SNP hybridization on a surface is also more stable than the hybridization of the perfectly complementary duplex in the bulk.

One reason this is important is that microarray probes are designed based upon oligonucleotide behavior in the bulk. The expected melting temperatures of both complimentary and possible mismatched probe/target duplexes are important parameters taken into account when designing microarrays. When microarrays are designed, probes are chosen such that all surface duplexes will have approximately the same melting temperature. This garners confidence that when samples are incubated over the platform at a temperature below but near the melting point, complementary sequences (the ones the microarray was designed to detect) will be stable enough to form but mismatched sequences will not. Current prediction techniques, such as the nearest-neighbor and salt-adjusted methods, were developed using melting temperature data obtained in the bulk. The data presented above indicate that such an approach under predicts the melting temperature for a given probe/target duplex on the surface. Thus mismatched duplexes occur more easily than accounted for in the microarray design. This is likely part of the reason for the high variability seen in microarray data.

The hypothesis that *stretching probe strands improves hybridization on surfaces*, was investigated as a model system to test the idea that improving base accessibility on the probe strand would improve microarray selectivity. These simulations found that stretching indeed

improved the stability of the hybridized duplex for the DNA sequences tested; however, this stabilizing effect occurred for both the complementary and the mismatched sequences. When the perfectly-matched sequences were compressed on the surface, there was little change in the stability of the duplex ($\Delta\Delta G_{\text{hyb}} = -10.5$) compared to the uncompressed perfectly-matched sequence ($\Delta\Delta G_{\text{hyb}} = -9.4$). Compressing the mismatched sequence on the surface destabilized the duplex ($\Delta\Delta G_{\text{hyb}} = -8.7$). This might suggest that rather than designing strategies to *enhance* hybridization on a surface, efforts should be made to hinder the process. The latter could result in a small decrease in overall sensitivity while increasing specificity. Since the variability in microarray results is likely due to false positives, increasing specificity would increase the reproducibility of microarray results.

Chapter 7

Dangling Ends

7.1 Introduction

Chapter 5 explored the mechanism of hybridization for dsDNA with strands of the same length. It was found that the surface affected the ability of the target to change its orientation, or to flip. Because flipping is likely length dependent, investigation was made into systems where the probe and target strands differ in both length and the location of the complementary sequence on the target. This last point is important in advancing the knowledge of the biophysics of DNA microarrays as target strands are usually much larger than the probe strands which increases the opportunity of the former to interact with the surface and change the hybridization mechanism compared to the bulk system. This study uses a more detailed model compared to previous efforts examining such effects [79], and the results offer unprecedented insight into the hybridization process.

The purpose of this chapter is to examine how the hybridization mechanism changes when the two hybridizing strands are of different length. In practice, probe strands on a microarray surface are carefully designed with specific composition characteristics [124] and lengths [117], but the target strands in a sample will vary in length and composition. Usually, the target strands are much longer than probe strands which likely has a significant effect on the hybridization mechanism. For example, if a long target strand approaches the probe from the bulk with parallel orientation, the reorientation can cause a large portion of the target to interact with the surface. This will likely destabilize the molecule compared to an approach where the target does not interact with the surface. Other scenarios are possible, and the hypothesis of this section seeks to understand these phenomena. Stated specifically, the hypothesis is *duplexes of uneven length will be destabilized on a surface only if they must*

reorient with the dangling end on the 5' end of the target, such that it extends towards the surface when it is properly hybridized with the probe.

7.2 Methods

7.2.1 DNA Model

This chapter required the use of new sequences of DNA. Simulations were performed using *probes* of 20 base pairs and *targets* of 60 base pairs. The sequences were selected from an exon for human insulin [125]. A well designed probe sequence [124] of 20 bases [117] was chosen from the exon and 60 base strands around the complementary sequence were selected for target sequences. The specific sequences are found in Table 7.1. The first line contains the 20 base probe sequence. The next three lines have the sequence for the three targets, each of which were simulated with the probe sequence found in the first line of the table. The portion of the target that is complementary to the probe is indicated in bold font.

The probe strand was attached to the surface at the $3'$ end. The targets were chosen to satisfy three distinct bonding motifs [79] as indicated in Figure 7.1. The first target was designed so that hybridization would occur at its $3'$ end. In this case, the location of the complementary sequence forces the rest of the strand down towards the surface. This case is termed "top" because the top of the target, relative to the surface, contains the complementary sequence. The second target, termed "middle," was designed to hybridize with the probe from bases 21 to 40 leaving a segment of the target extending towards the surface and a segment extending away from the surface. The last target was designed to hybridize with the probe at the $5'$ end of the target. This causes the target to hybridize in a way that leaves all the bases not involved in hybridization extending away from the surface. This case is termed "bottom" because the complementary sequence is found at the bottom of the target.

The control for each insulin target was simulation in the bulk with the probe strand listed in the first line of Table 7.1 and only the equally-lengthed complementary portion of the targets (found in the bold font.) This control was chosen because in practice, microarray probes are designed based upon bulk melting temperatures of evenly-lengthed probe/target duplexes. This choice also helps isolate the effects of "extra," nonhybridizing bases, or

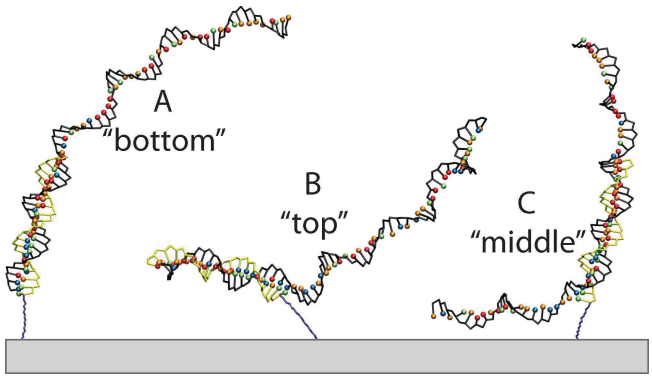

Figure 7.1: Different bonding motifs for probe-target complexes of unequal length: A) matching sequence on target strand is at the 5′ ("bottom") end of the molecule, B) matching sequence on the target strand is at the 3′ end ("top") of the molecule, and C) matching sequence on the target strand is in the "middle" of the molecule.

Table 7.1: Sequences used to study the effect of dangling ends.
Taken from the sequence for human insulin.

Probe		CCA GCC GCA GCC TTT GTG AA
Targets	Matching Sequence Location	
Target 1	3′ End (Top)	GTA GAG AGC TTC CAC CAG GTG TGA GCC GCA CAG GTG TTG GTT CAC AAA GGC TGC GGC TGG
Target 2	Center (Middle)	GTG AGC CGC ACA GGT GTT GGT TCA CAA AGG CTG CGG CTG GGT CAG GTC CCC AGA GGG CCA
Target 3	5′ End (Bottom)	TTC ACA AAG GCT GCG GCT GGG TCA GGT CCC CAG AGG GCC AGC AGC GCC AGC AGG GGC AGG

dangling ends. Since the matching sequence is the same in all three cases, the hydrogen bonds that form to build the final duplex will be the same. If the majority of the change in free energy that occurs upon hybridization is due to the formation of hydrogen bonds, then the free energy of hybridization should be the same for all treatments and the control. Any changes that do appear in the thermodynamic stability of the hybridized duplex are due to effects of the sections that do not directly participate in base pairing.

7.2.2 Simulation Protocols and Experimental Design

Simulations were organized and run according to the methods outlined in Chapters 4 and 5. For this query, two types of PMFs were calculated: one-dimensional PMFs using ξ as the reaction coordinate (following the method of Chapter 4) and two-dimensional PMFs using both ξ and θ as reaction coordinates (following the method of Chapter 5). One-dimensional PMFs were used where appropriate due to the large computational demands required to calculate the two-dimensional variety. In both cases, each point along the reaction coordinate was simulated as an independent system with appropriate biasing potentials.

For the one-dimensional PMFs of the unevenly-lengthed insulin pairs, ξ_0 ranged from 10 to 325.75 Å in increments of 0.25 Å requiring a total of 1264 simulations. For the one-dimensional PMF of the evenly-lengthed control for the insulin pairs, ξ_0 ranged from 10 to 150 Å in increments of 0.25 Å requiring necessitating 561 simulations. For the two-dimensional PMF's, θ_0 ranged from 0° to 180° in 10° increments and ξ_0 ranged from 10 to 160 Å in 0.25 Å increments. A two-dimensional PMF of the system, therefore, required 11476 simulations across both ξ and θ and represented a total of 1.19 ms of simulation time.

7.3 Results

Figure 7.2 which shows the potential of mean force for hybridization of uneven strands in both the bulk and on the surface as well as the control case of hybridization in the bulk for two strands of the same length. Panel A is the 1D PMF when hybridization occurs at the $3'$ end of the longer strand (Target 1, "Top"), Panel B in the center of the longer strand (Target 2, "middle"), and Panel C at the $5'$ end of the longer strand (Target 3, "bottom"). The solid line on each graph is the control while the dashed and dotted lines correspond to uneven hybridization in the bulk and on the surface respectively.

Below, the bulk data are discussed first followed by the surface data. In these discussions, comparisons are made between the free energies of hybridization for the control and each treatment. As mentioned above, ΔG_{hyb} is defined as the free energy of the hybridized state minus the free energy of the state where the strands do not interact. As shown in Figure 7.2, the hybridized state (located at the minimum of the free energy curve) occurs at approximately $\xi = 13.5$ Å while the non-interacting state ($\Phi = 0$) occurs for

Table 7.2: Changes in free energy for the hybridization of human insulin. Data of various probe/target pairs in the bulk and on the surface.

ΔG	Value (kJ/mol)	Description
$\Delta G_{hyb}^{control}$	-34.0	Change in free energy upon hybridization of the evenly-length, insulin probe/target pair in the bulk. (See solid lines of all panels in Figure 7.2.)
$\Delta G_{hyb}^{top,\ bulk}$	-17.6	Change in free energy upon hybridization of the insulin probe with Target 1 ("top") in the bulk. (See dashed line of Panel A of Figure 7.2.)
$\Delta G_{hyb}^{middle,\ bulk}$	-22.1	Change in free energy upon hybridization of the insulin probe with Target 2 ("middle") in the bulk. (See dashed line of Panel B of Figure 7.2.)
$\Delta G_{hyb}^{bottom,\ bulk}$	-18.6	Change in free energy upon hybridization of the insulin probe with Target 3 ("bottom") in the bulk. (See dashed line of Panel C of Figure 7.2.)
$\Delta G_{hyb}^{top,\ surface}$	-21.4	Change in free energy upon hybridization of the insulin probe with Target 1 ("top") on the surface. (See dotted line of Panel A of Figure 7.2.)
$\Delta G_{hyb}^{middle,\ surface}$	-24.2	Change in free energy upon hybridization of the insulin probe with Target 2 ("middle") on the surface. (See dotted line of Panel B of Figure 7.2.)
$\Delta G_{hyb}^{bottom,\ surface}$	-24.5	Change in free energy upon hybridization of the insulin probe with Target 3 ("bottom") on the surface. (See dotted line of Panel C of Figure 7.2.)

ξ greater than approximately 235 Å. This leads to the following functional definition of $\Delta G_{hyb} = G(\xi = 13.5\ Å) - G(\xi = 250\ Å) = G(\xi = 13.5\ Å)$. The errors associated with these ΔG_{hyb} values ranged from ± 0.1 to 0.6 so $\Delta\Delta G$ values greater than 1.2 are considered significant. For convenience, each of the ΔG_{hyb} values are summarized in Table 7.2.

Figure 7.2 shows that hybridization of two strands of different length *in the bulk* is destabilized compared to bulk hybridization of the same sequence without extra nucleotides on one of the strands. This destabilization occurs regardless of the location of the complementary sequences on the longer strand. Specifically, the free energy of hybridization for the control (Panels A, B, and C, solid line) is $\Delta G_{hyb}^{control} = -34.0$ kJ/mol. When the

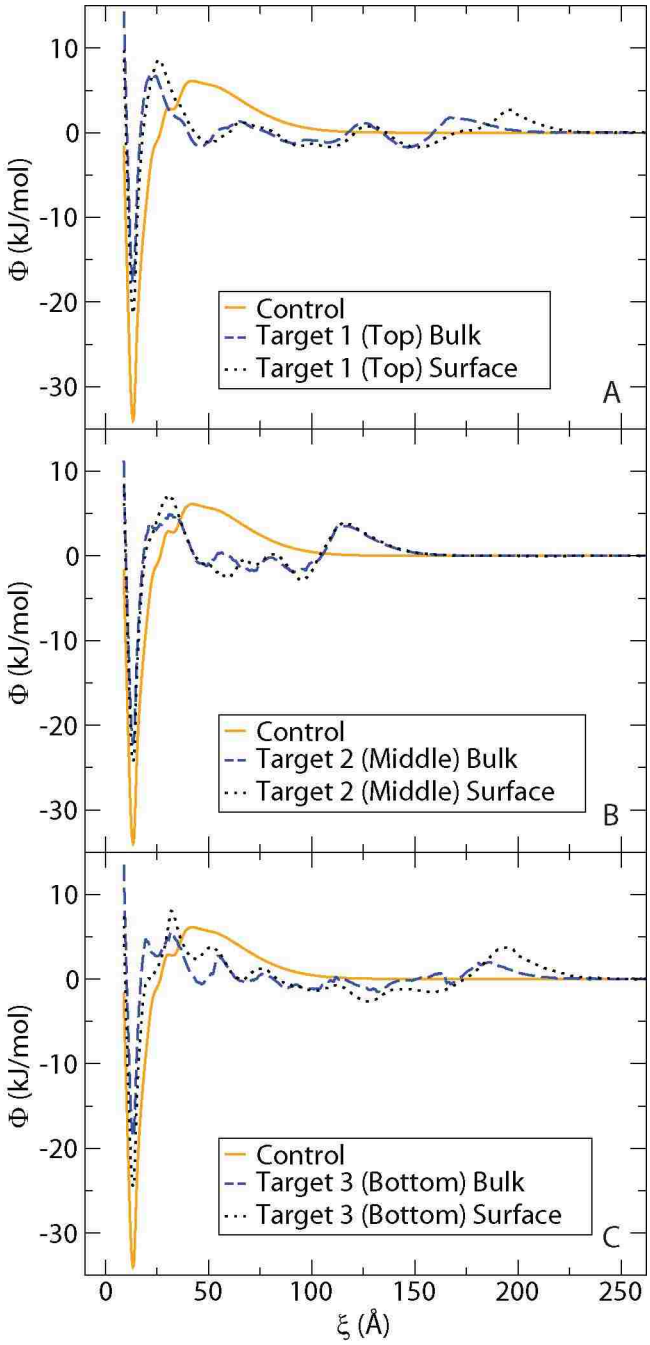

Figure 7.2: Potential of mean force (Φ) of DNA hybridization of unevenly-lengthed strands in the bulk and on the surface as a function of the distance between the strands (ξ). Panel A) "top", Panel B) "middle", Panel C) "bottom."

matching sequence is found at the "top" of the longer target strand (Panel A, dashed line), $\Delta G_{hyb}^{top,\ bulk} = -17.6$ kJ/mol. This represents $\Delta\Delta G_{hyb}^{top,\ bulk} = 16.4$ kJ/mol, or a destabilization of the duplex compared to the evenly-lengthed strands. When hybridization occurs in the "middle" of the longer strand (Panel B, dashed line), the duplex is also destabilized as $\Delta G_{hyb}^{middle,\ bulk} = -22.1$ kJ/mol and $\Delta\Delta G_{hyb}^{middle,\ bulk} = 11.9$ kJ/mol. Hybridization at the $5'$ end of the longer strand (Panel C, dashed line) continues this same pattern with $\Delta G_{hyb}^{bottom,\ bulk} = -18.6$ kJ/mol and $\Delta\Delta G_{hyb}^{bottom,\ bulk} = 15.4$ kJ/mol.

The PMFs for hybridization of unevenly-lengthed strands *on the surface* are shown as dotted lines in Figure 7.2. Panel A is the "top" case where the non-hybridizing nucleotides of the target strand must extend down towards the surface. In this case, the $\Delta G_{hyb}^{top,\ surface} = -21.4$ kJ/mol which translates to $\Delta\Delta G_{hyb}^{top,\ surface} = 12.6$ kJ/mol. This is more stable than when these two molecules hybridize in the bulk, but it is still a destabilization from the bulk duplex with strands of the same length (the control). Panel B is the "middle" case where hybridization occurs in the center of the target strand. Here, $\Delta G_{hyb}^{middle,\ surface} = -24.2$ kJ/mol and $\Delta\Delta G_{hyb}^{middle,\ surface} = 9.8$ kJ/mol. As with the "top" configuration, hybridization of these two molecules is more stable on the surface than in the bulk, but it is still less favorable than hybridizing two evenly-lengthed strands in either environment. This same pattern is also found in the "bottom" case (Panel C) where $\Delta G_{hyb}^{bottom,\ surface} = -24.5$ kJ/mol and $\Delta\Delta G_{hyb}^{bottom,\ surface} = 9.5$ kJ/mol.

One of the features of hybridization of unevenly-lengthed strands, seen in both the bulk (dashed lines of Figure 7.2) and the surface (dotted lines) is the presence of rough landscapes along the hybridization free energy curves that are not present for hybridization of evenly-lengthed strands (solid lines). To determine the cause of these minima, whether they are local energy minima or noise, 2D PMF's were generated for each unevenly-lengthed system. Panel A of Figure 7.3 shows the "top" case, Panel B the "middle" case, and Panel C the "bottom" case. To aide in the discussion that follows, Figure 7.4 shows representative snapshots of configurations of the "top" and "bottom" cases at various minima on the free energy landscapes. The dashed lines indicate transitions between neighboring low-energy minima. Transitions that move the system between antiparallel and parallel orientations with respect to sequence are labeled "flip." Those that move a system initially in either

a parallel or antiparallel configuration to a perpendicular configuration are labeled "partial flip," and a subsequent partial flip from perpendicular to either parallel or antiparallel will produce a complete flip. Transitions that move the system away from an ability to finish hybridization by having the target *slide* into place are colored in gray.

Complete hybridization in the "top" situation forces unmatched nucleotides into the surface. Panel A of Figure 7.3 and the top panel of Figure 7.4 suggest that the system seeks to minimize this surface interaction while maximizing the ability of the strands to contact each other without respect to sequence. At long distances ($\xi \approx 144$ Å) the system prefers an orientation where the matching portion of the target adopts a more parallel orientation ($\theta \approx 138°$). This allows the strands to contact each other and create both canonical and non-canonical base pairing motifs even though the sequence alignment is incorrect. A stable intermediate progressing to the final hybridized state is found at $\xi \approx 88$ Å and $\theta \approx 109°$ Here, the two molecules are intertwined with the matching portion of the target farthest away from the surface and a small number of the non-hybridizing bases contacting the surface. Hybridization finishes as the target slides down the probe. The sliding process occurs in an energy "valley" (with only small energy barriers) at low values of θ and forces the remaining non-hybridizing bases into the surface. This configuration is accommodated by the duplexed portion of the molecule leaning down toward the surface. Also note that there are no local minima at short distances for the parallel state. This contrasts with what is seen for strands of equal length (See Figure 5.4).

When the hybridizing sequence is found in the center of the probe strand, the "middle" case, flipping between parallel and antiparallel configurations seems to occur more easily at shorter distances compared to the "top" case. As shown in Panel B of Figure 7.3, at long distances, the most favorable configuration occurs when the strands are approximately perpendicular to each other ($\theta \approx 90°$). As the strands move closer together, they prefer an antiparallel orientation at $\xi \approx 107$ Å but very quickly adopt a more perpendicular to parallel orientation from $\xi \approx 75 - 95$ Å. At $\xi \approx 58$ Å, the strands prefer an antiparallel orientation. From here, the system can follow two paths. The first is they can remain in the antiparallel configuration and hybridize into the low-energy, duplexed state crossing a few small energy barriers. The second is that the strands can adopt a parallel orientation

centered at $\xi \approx 45$ Å. To complete the hybridization process from this state requires a "flip" event. If the strands do not separate, a small energy barrier must be overcome for this to occur. However, if the strands separate, the "flip" can follow a low-energy path to go back to the antiparallel state at $\xi \approx 58$ Å and follow the first path mentioned.

Hybridization for the "bottom" case starts similar to that of the "top" case in that at long distances, the system prefers a parallel orientation. As the bottom panel of Figure 7.4 depicts, in this configuration ($\xi \approx 145$ Å and $\theta \approx 155°$) the non-matching portion of the target can interact with the probe without significant surface effects. As the distance between the two strands decreases, the system drifts toward a perpendicular arrangement where a small number of the bases from the non-matching portion of the molecule are attempting to hybridize with the probe ($\xi \approx 85$ Å and $\theta \approx 89°$). From here, the system can proceed one of two ways depending on the movement the target takes. If the target moves in a counterclockwise direction (relative to the position depicted in the bottom panel of Figure 7.4) the system will arrive at a stable intermediate at 69 Å and 146° where the system appears hybridized, but the strands are in the incorrect orientation and the sequences do not match. The second is where the target moves clockwise so that it can be in the correct orientation to slide along the probe and complete hybridization. The sliding process occurs in an energy "valley" (with only small energy barriers separating multiple energy minima) at low values of θ. Sliding initially forces the matching bases into the surface ($\xi \approx 59$ Å and $\theta \approx 61°$), but the final state is the properly-hybridized duplex with the extra bases extending away from the surface.

Of note in the "bottom" case is that no parallel state is favorable less than 69 Å. The intermediate found at 69 Å and 146° is the last time a true parallel state is seen. The low energy state at 47 Å and 158° (snapshot not depicted but is similar to the 51 Å, 41° state for the "top" target) is not a true parallel state but a J-like structure produced from the sliding process where the portion of the matching bases that are not yet hybridized are seeking to minimize their contact with the surface. Moreover, a significant energy barrier to flipping forms at about 30 Å from $90 - 120°$. This particular barrier is not found in the other cases. Another distinct feature of the "bottom" case is that the well for the properly-hybridized state is deeper over a wider range of ξ values than both the "top" and "middle" cases.

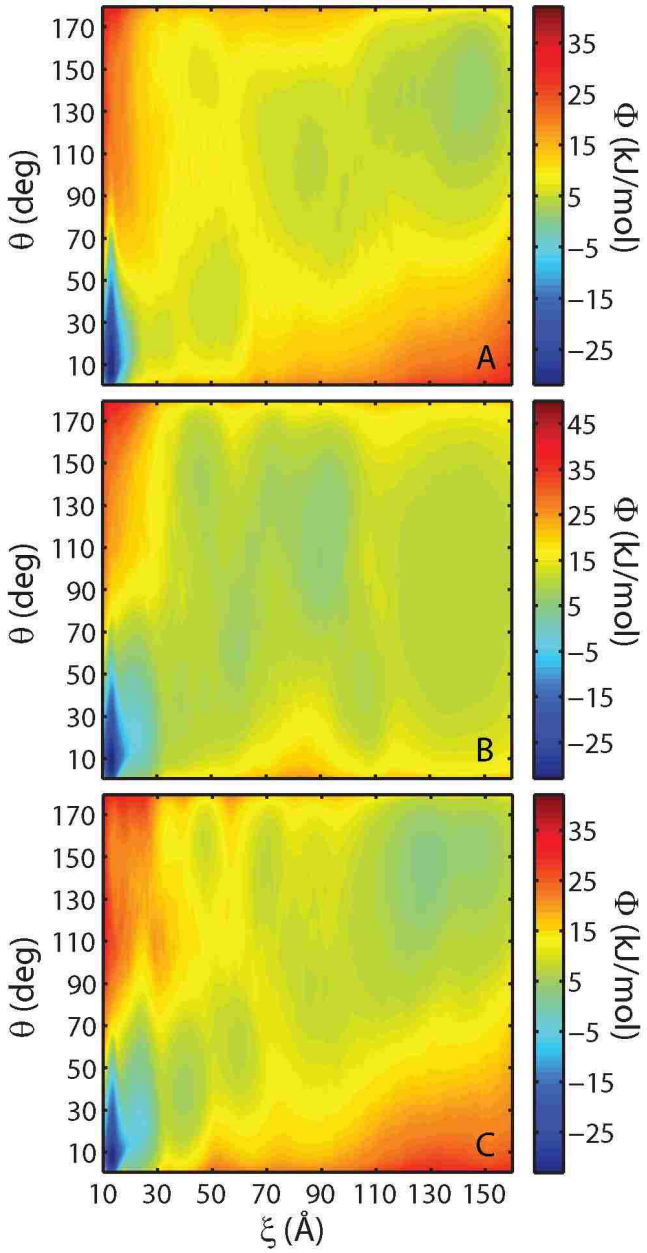

Figure 7.3: Potential of mean force (Φ) of DNA hybridization of unevenly-lengthed strands on the surface as a function of distance between the strands (ξ) and angle made by the strands (θ). Panel A) "top", Panel B) "middle", Panel C) "bottom."

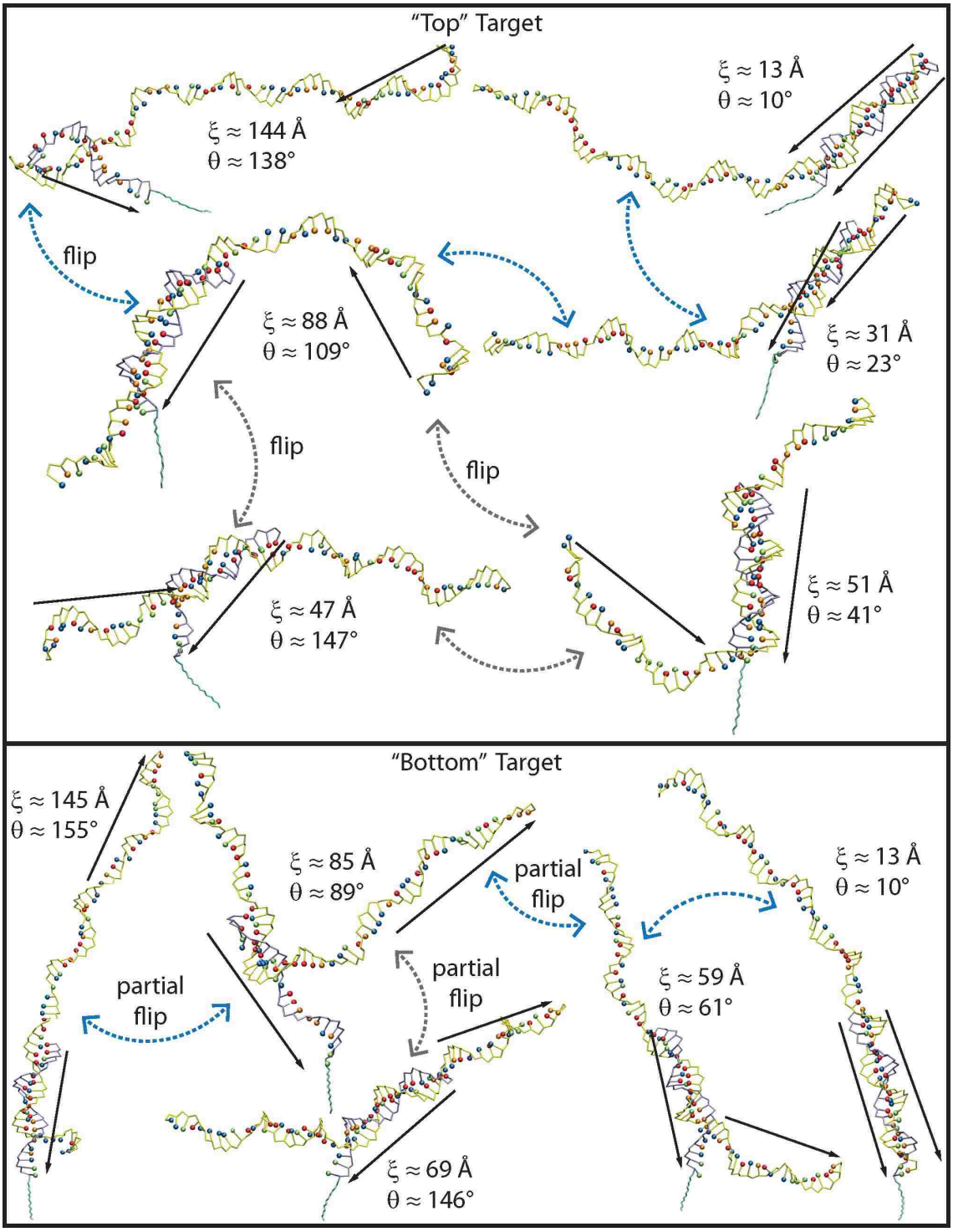

Figure 7.4: Representative snapshots for configurations of the system for the "top" and "bottom" unevenly-lengthed strands corresponding to low-energy minima in Figure 7.3.

7.4 Discussion and Conclusions

This chapter explored how the flip mechanism explained in Chapter 5 changes with strands of uneven length and how those changes affect the stability of the duplex. Figures 7.3 and 7.4 show that flipping occurs even for unevenly-length pairs and involves a metastable state at intermediate distances. For the "top" case, flipping occurs at long distances ($\xi \approx 144$ Å and $\theta \approx 138°$) to properly orient the target strand for sliding into place and produces a metastable state at 88 Å and 155°. From here, a flip leads to stable states at shorter distances ($\xi \approx 47$ Å and $\theta \approx 147°$ and $\xi \approx 41$ Å and $\theta \approx 41°$) that cannot result in proper hybridization without a subsequent flip that takes the system back to 88 Å and 155°. In the "middle" case, the flipping occurs at both long and short distances with relative ease as an "extra" amount of bases are found on both sides of the molecule so that the surface does not affect either end preferentially. However, a metastable state is also present at $\xi \approx 90$ Å and centered around $\theta \approx 110°$. In the "bottom" case, flipping occurs gradually to move the system at long distances and parallel orientations ($\xi \approx 145$ Å and $\theta \approx 155°$) through a metastable state with a perpendicular orientation ($\xi \approx 85$ Å and $\theta \approx 89°$) to finally end up at short distances and an antiparallel orientation ($\xi \approx 59$ Å and $\theta \approx 61°$) from which hybridization can finish by the target molecule sliding into the proper duplex. Complete flipping is not found at long distances since both sequence and the extra bases work together to keep the molecule in the correct orientation. In summary, the data for unevenly-lengthed strands on the surface support the findings in Chapter 5 with respect to flipping mechanisms and metastable states.

The hypothesis of this chapter was *duplexes of uneven length will be destabilized on a surface only if the reorientation process to correctly align complimentary bases for hybridization increases the interaction of the target molecule with the surface.* The results do not support this hypothesis. The data in Figure 7.2 and Table 7.2 demonstrate that targets of longer length than the probe, regardless of the location of the matching sequence on the target, result in less favorable hybridization both on the surface and in the bulk compared to evenly-lengthed strands (the control). Moreover, even if the comparison is restricted to only uneven strands in the bulk and on the surface, the data show that the surface stabilizes duplexes even if the unmatching bases must extend into the surface (the "top" case). These

results are in agreement with previous work on evenly-lengthed strands that showed that the surface stabilizes hybridization (see Chapter 5). In short, the data do not support the hypothesis of this chapter whether the comparison is made to the control (hybridization of evenly-lengthed strands in the bulk) or between the same unevenly-lengthed strands in the bulk and on the surface. Extra bases that extend toward the surface do have a destabilizing effect, as indicated by the fact that the "bottom" case is more stable than the "top" case, but the degree of this destabilization is less than the overall stabilizing effect of the surface.

The data presented above also help place the role of unevenly-lengthed probe/target complexes in perspective. In the bulk, the most stable duplex was the one that hybridized at the center of the longer strand. On the surface, this duplex was just as stable as the duplex that hybridized at the $5'$ end of the target strand (the "bottom" case). This indicates that the thermodynamic stability gained by hybridizing at the center of the strand is lost to the unfavorable interactions between the surface and the extra length of the target that extends towards the surface. Further evidence of the negative effect of nonhybridizing bases interacting with the surface is seen in the case of the probe hybridizing at the $3'$ end of the target (the "top" case). This duplex, which was just as stable as the duplex at the $5'$ end of the longer strand in the bulk, is less stable than the other duplexes on the surface. Although this destabalization is small, ≈ 3 kJ/mol, it does mark a significant change of more than three standard deviations in the $\Delta\Delta G_{hyb}^{surface}$. This is in agreement with previous results [79].

Finally, the data also indicate that there are two competing forces that affect the stability of unevenly-lengthed strands: the destabilizing effect of the extra bases and the stabilizing effect of the surface. Extra bases, regardless of the location of the hybridizing sequence, destabilize DNA duplexes both in the bulk and on the surface. Surfaces, for both evenly- and unevenly-lengthed strands act to stabilize the duplex. The latter appears to be the dominant force in this interplay as seen in the fact that all surface cases for the unevenly-lengthed duplexes are more stable than their bulk counterparts.

Because the stabilizing surface effect seems to overcome at least part of the destabilizing effect of the extra bases, the location of the hybridizing sequence on the target is likely to be of secondary importance. Figure 7.2 shows that the difference in the stability between the "top," "middle," and "bottom" cases on the surface is only 4 kJ/mol at the largest. This

means that while there are differences in the hybridization pathway in each case, the effect of the location of the hybridizing sequence on the performance of real microarrays is likely to be small if the targets do not become too long.

Chapter 8

Conclusions

8.1 Summary of Results

8.1.1 Surfaces and Hybridization Mechanisms

This study found that the presence of a surface thermodynamically stabilizes hybridization. Specifically, for each duplex system simulated, $\Delta G_{hyb}^{surf} < \Delta G_{hyb}^{bulk}$. This can be attributed to the way the hybridization mechanism changes on the surface compared to the bulk. As shown in Chapter 5, strand reorientation, or flips, occur both in the bulk and on the surface. In the bulk, however, no stable intermediate state is present. Specifically, in the bulk, two strands separated at long distances will be attracted to each other, but the attraction can lead to either a parallel or an antiparallel stable state at short distances. Until ξ approaches the value found at hybridization there is no preference for parallel vs. antiparallel configurations and no low-energy minimum in the energy landscape. On the surface, the situation changes such that stable intermediates form at longer distances. These stable intermediates help to guide the system along the hybridization process and into proper orientation.

8.1.2 Mismatches and External Forces

It was found in Chapter 6 that surfaces stabilize duplexes with single nucleotide polymorphisms to the same extent that they stabilize perfectly matched duplexes. Chapter 6 showed that efforts to counteract that stabilization via external forces would be more effective at improving microarray specificity than trying to further stabilize the correctly hybridized duplexes. The overall unexpected result of these chapters was that the stabilities of mismatched and perfectly matched duplexes were too similar on a microarray surface.

8.1.3 Unevenly-Length Strands

Chapter 7 showed that long dangling ends destabilize duplexes. This destabilization occurs in the bulk and on a surface. The presence of a surface counteracts some of the destabilization seen in the bulk and some binding motifs are more stable than others, but the overall effect is still a destabilization.

8.2 Implications of Results

These results have implications in designing microarray probes. One of the parameters used to design probe sequences is the melting temperature of the probe/target complex. These melting temperatures are generated assuming the pair is composed of two single-stranded DNA molecules of equal length and neglects the fact that targets are usually much longer than probes. This approach, according to the unevenly-lengthed PMF data presented in Chapter 7, results in duplexes that are less stable than expected. Though melting temperature is not directly analogous to stability, it has been shown [101, 126–128] that increases in stability result in higher melting temperatures and vice versa. It is therefore reasonable to suspect that the melting temperatures with which an array is designed are higher than those that actually occur on the chip. Additionally, according to the results of Chapter 6, the melting points calculated for possible competitive duplexes on a microarray surface might not be as different as those that actually occur on the chip. These discrepancies are likely part of the cause for the limited reproducibility seen in microarrays. As such, changing the prediction methods used to design microarrays might improve accuracy and reproducibility on microarrays.

Despite the results presented above, further work is needed to fully understand how microarray environments affect duplex stability. Experimental studies on the subject give varied results that both agree and disagree with those presented above. Primarily, the work by Hurst *et al.* confirms the findings that surfaces stabilize both perfectly matched and mismatched DNA duplexes [119]. Chapter 7, however, showed that the destabilizing effects of dangling ends was greater than the stabilizing effect of the surface and the experimental results on this topic are more varied. For example, Guckian *et al.* [129] found that adding a single nucleotide dangling end stabilizes duplexes, but Bommarito *et al.* [130] showed that

such can produce both stabilization and destabilization. In more recent work, Isaksson and Chattopadhyaya [131] confirmed that both stabilization and destabilization are possible; however, Moreira *et al.* [132] found that dangling ends have little effect on duplex stability.

Future simulation work should seek to interpret the experimental findings and help determine the cause for the apparent discrepancies. Before doing so, one difference between the experiments and simulations needs to be addressed. Specifically, the simulations presented above were done in the single molecule regime while the experiments were done in the thermodynamic limit. It has been shown that this difference can have a pronounced effect on thermodynamic transitions [133]. Moreover, this difference may be crucially important to the problem at hand as it has been hypothesized that stabilization is due to the dangling ends of multiple molecules interacting in non-canonical base pairing motifs [134]. Future work needs to be done to determine how the presence of multiple probe and target molecules affects the hybridization mechanism and the duplex stability of unevenly-lengthed strands.

8.3 *Summa Summarum*

The purpose of this study was to develop a more detailed understanding of the DNA hybridization in microarray environments. The results show that the surface changes the hybridization process compared to the bulk situation by creating a stable intermediate on the free energy landscape. This intermediate makes it easier for strands to "flip" orientations if such is required to complete hybridization. The surface stabilizes all duplexes compared to the same duplex in the bulk, including mismatched duplexes, but if one strand is shorter than the other the duplex becomes less stable. Due to the fact that microarrays are designed based on the thermodynamic stability of evenly-length duplexes in the bulk, microarray design may be improved by considering the effects of strands of varying lengths in the design calculations. Finally, it was found that the location of the hybridizing sequence on the longer strand was of secondary importance for surface hybridization.

Bibliography

[1] S. Fodor, J. Read, M. Pirrung, L. Stryer, A. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *Science*, vol. 251, no. 4995, pp. 767–773, FEB 15 1991. 1

[2] S. Fodor, R. Rava, X. Huang, A. Pease, C. Holmes, and C. Adams, "Multiplexed biochemical assays with biological chips," *Nature*, vol. 364, no. 6437, pp. 555–556, AUG 5 1993. 1

[3] A. Roda, M. Guardigli, C. Russo, P. Pasini, and M. Baraldini, "Protein microdeposition using a conventional ink-jet printer," *Biotechniques*, vol. 28, no. 3, pp. 492–496, MAR 2000. 1

[4] D. Graves, H. Su, S. McKenzie, S. Surrey, and P. Fortina, "System for preparing microhybridization arrays on glass slides," *Analytical Chemistry*, vol. 70, no. 23, pp. 5085–5092, DEC 1 1998. 1

[5] G. Yershov, V. Barsky, A. Belgovskiy, E. Kirillov, E. Kreindlin, I. Ivanov, S. Parinov, D. Guschin, A. Drobishev, S. Dubiley, and A. Mirzabekov, "Dna analysis and diagnostics on oligonucleotide microchips," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 10, pp. 4913–4918, MAY 14 1996. 1

[6] A. Mirzabekov, "Dna-sequencing by hybridization - a megasequencing method and a diagnostic-tool," *Trends in Biotechnology*, vol. 12, no. 1, pp. 27–32, JAN 1994. 2

[7] J. Vijg, "Two-dimensional dna typing - a cost-effective way of analyzing complex mixtures of dna fragments for sequence variations," *Molecular Biotechnology*, vol. 4, no. 3, pp. 275–295, DEC 1995. 2

[8] D. Shalon, S. Smith, and P. Brown, "A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization," *Genome Research*, vol. 6, no. 7, pp. 639–645, JUL 1996. 2

[9] A. Butte, "The use and analysis of microarray data," *Nature Reviews Drug Discovery*, vol. 1, no. 12, pp. 951–960, DEC 2002. 2

[10] J. Hall, C. LeDuc, A. Watson, and A. Roter, "An approach to high-throughput genotyping," *Genome Research*, vol. 6, no. 9, pp. 781–790, SEP 1996. 2

[11] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnology*, vol. 14, no. 13, pp. 1675–1680, DEC 1996. 2

[12] M. Schena, D. Shalon, R. Davis, and P. Brown, "Quantitative monitoring of gene-expressioin patterns with a complementary-dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, OCT 20 1995. 2

[13] Y. Hayashi, S. Iida, Y. Sato, A. Nacaya, A. Sawada, N. Kaji, H. Kamiya, Y. Baba, and H. Harashima, "Dna microarray analysis of type 2 diabetes-related genes co-regulated between white blood cells and livers of diabetic otsuka long-evans tokushima fatty (oletf) rats," *Biological & Pharmaceutical Bulletin*, vol. 30, no. 4, pp. 763–771, APR 2007. 2

[14] G. D. Vladutiu, "The fda announces new drug labeling for pharmacogenetic testing: Is personalized medicine becoming a reality?" *Molecular Genetics and Metabolism*, vol. 93, no. 1, pp. 1–4, JAN 2008. 3

[15] T. U. S. Food and D. Association, "Guidance for industry: Pharmacogenomic data submissions," March 2005. [Online]. Available: http://www.fda.gov/CbER/gdlns/pharmdtasub.htm 3

[16] R. Simon, M. Radmacher, and K. Dobbin, "Design of studies using dna microarrays," *Genetic Epidemiology*, vol. 23, no. 1, pp. 21–36, jun 2002. 4

[17] M. Branca, "'omic diagnostics trip up on way to clinic," *Nature Biotechnology*, vol. 23, no. 7, p. 769, JUL 2005. 4

[18] A. Abdullah-Sayani, J. M. Bueno-de Mesquita, and M. J. van de Vijver, "Technology insight: tuning into the genetic orchestra using microarrays - limitations of dna microarrays in clinical practice," *Nature Clinical Practice Oncology*, vol. 3, no. 9, pp. 501–516, SEP 2006. 4

[19] M. Carter, T. Hamatani, A. Sharov, C. Carmack, Y. Qian, K. Aiba, N. Ko, D. Dudekula, P. Brzoska, S. Hwang, and M. Ko, "In situ-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling," *Genome Research*, vol. 13, no. 5, pp. 1011–1021, MAY 2003. 4

[20] R. Kothapalli, S. Yoder, S. Mane, and T. Loughran, "Microarray results: how accurate are they?" *BMC Bioinformatics*, vol. 3, 2002. 4

[21] W. Kuo, T. Jenssen, A. Butte, L. Ohno-Machado, and I. Kohane, "Analysis of matched mrna measurements from two different microarray technologies," *Bioinformatics*, vol. 18, no. 3, pp. 405–412, MAR 2002. 4

[22] M. Lenburg, L. Liou, N. Gerry, G. Frampton, H. Cohen, and M. Christman, "Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data," *BMC Cancer*, vol. 3, NOV 27 2003. 4

[23] J. Li, M. Pankratz, and J. Johnson, "Differential gene expression patterns revealed by oligonucleotide versus long cdna arrays," *Toxicological Sciences*, vol. 69, no. 2, pp. 383–390, OCT 2002. 4

[24] N. Mah, A. Thelin, T. Lu, S. Nikolaus, T. Kuhbacher, Y. Gurbuz, H. Eickhoff, G. Kloppel, H. Lehrach, B. Mellgard, C. Costello, and S. Schreiber, "A comparison of oligonucleotide and cdna-based microarray systems," *Physiological Genomics*, vol. 16, no. 3, pp. 361–370, FEB 13 2004. 4

[25] P. Tan, T. Downey, E. Spitznagel, P. Xu, D. Fu, D. Dimitrov, R. Lempicki, B. Raaka, and M. Cam, "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5676–5684, OCT 1 2003. 4

[26] M. Ramalho-Santos, S. Yoon, Y. Matsuzaki, R. Mulligan, and D. Melton, ""stemness": Transcriptional profiling of embryonic and adult stem cells," *Science*, vol. 298, no. 5593, pp. 597–600, OCT 18 2002. 4

[27] N. Ivanova, J. Dimos, C. Schaniel, J. Hackney, K. Moore, and I. Lemischka, "A stem cell molecular signature," *Science*, vol. 298, no. 5593, pp. 601–604, OCT 18 2002. 4

[28] N. Fortunel, H. Otu, H. Ng, J. Chen, X. Mu, T. Chevassut, X. Li, M. Joseph, C. Bailey, J. Hatzfeld, A. Hatzfeld, F. Usta, V. Vega, P. Long, T. Libermann, and B. Lim, "Comment on " 'stemness': Transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature" (i)," *Science*, vol. 302, no. 5644, OCT 17 2003. 4

[29] Anonymous, "Making the most of microarrays," *Nature Biotechnology*, vol. 24, no. 9, p. 1039, SEP 2006. 4

[30] R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid, "Evaluation of dna microarray results with quantitative gene expression platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1115–1122, SEP 2006. 4, 5

[31] D. A. Casciano and J. Woodcock, "Empowering microarrays in the regulatory setting," *Nature Biotechnology*, vol. 24, no. 9, p. 1103, SEP 2006. 4

[32] D. J. Dix, K. Gallagher, W. H. Benson, B. L. Groskinsky, J. T. McClintock, K. L. Dearfield, and W. H. Farland, "A framework for the use of genomics data at the epa," *Nature Biotechnology*, vol. 24, no. 9, pp. 1108–1111, SEP 2006. 4

[33] F. W. Frueh, "Impact of microarray data quality on genomic data submissions to the fda," *Nature Biotechnology*, vol. 24, no. 9, pp. 1105–1107, SEP 2006. 4

[34] L. Guo, E. K. Lobenhofer, C. Wang, R. Shippy, S. C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F. M. Goodsaid, P. Hurban, K. L. Phillips, J. Xu, X. Deng, Y. A. Sun, W. Tong, Y. P. Dragan, and L. Shi, "Rat toxicogenomic study reveals analytical consistency across microarray platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1162–1169, SEP 2006. 4

[35] H. Ji and R. W. Davis, "Data quality in genomics and microarrays," *Nature Biotechnology*, vol. 24, no. 9, pp. 1112–1113, SEP 2006. 4

[36] T. A. Patterson, E. K. Lobenhofer, S. B. Fulmer-Smentek, P. J. Collins, T.-M. Chu, W. Bao, H. Fang, E. S. Kawasaki, J. Hager, I. R. Tikhonova, S. J. Walker, L. Zhang, P. Hurban, F. de Longueville, J. C. Fuscoe, W. Tong, L. Shi, and R. D. Wolfinger, "Performance comparison of one-color and two-color platforms within the microarray quality control (maqc) project," *Nature Biotechnology*, vol. 24, no. 9, pp. 1140–1150, SEP 2006. 4

[37] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Scherf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-h. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, W. Slikker, Jr., and M. Consortium, "The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, SEP 2006. 4, 5

[38] R. Shippy, S. Fulmer-Smentek, R. V. Jensen, W. D. Jones, P. K. Wolber, C. D. Johnson, P. S. Pine, C. Boysen, X. Guo, E. Chudin, Y. A. Sun, J. C. Willey, J. Thierry-Mieg, D. Thierry-Mieg, R. A. Setterquist, M. Wilson, A. B. Lucas, N. Novoradovskaya, A. Papallo, Y. Turpaz, S. C. Baker, J. A. Warrington, L. Shi, and D. Herman, "Using rna sample titrations to assess microarray platform performance and normalization techniques," *Nature Biotechnology*, vol. 24, no. 9, pp. 1123–1131, SEP 2006. 4

[39] W. Tong, A. B. Lucas, R. Shippy, X. Fan, H. Fang, H. Hong, M. S. Orr, T.-M. Chu, X. Guo, P. J. Collins, Y. A. Sun, S.-J. Wang, W. Bao, R. D. Wolfinger, S. Shchegrova, L. Guo, J. A. Warrington, and L. Shi, "Evaluation of external rna controls for the assessment of microarray performance," *Nature Biotechnology*, vol. 24, no. 9, pp. 1132–1139, SEP 2006. 4

[40] D. Coman and I. Russu, "A nuclear magnetic resonance investigation of the energetics of basepair opening pathways in dna," *Biophysical Journal*, vol. 89, no. 5, pp. 3285–3292, NOV 2005. 6

[41] D. Beveridge and K. McConnell, "Nucleic acids: theory and computer simulation, y2k," *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 182–196, APR 2000. 7

[42] A. Perez, F. J. Luque, and M. Orozco, "Dynamics of b-dna on the microsecond time scale," *Journal of the American Chemical Society*, vol. 129, no. 47, pp. 14 739–14 745, NOV 28 2007. 7, 8

[43] T. Cheatham, "Simulation and modeling of nucleic acid structure, dynamics and interactions," *Current Opinion in Structural Biology*, vol. 14, no. 3, pp. 360–367, JUN 2004. 7

[44] M. Orozco, A. Noy, and A. Perez, "Recent advances in the study of nucleic acid flexibility by molecular dynamics," *Current Opinion in Structural Biology*, vol. 18, no. 2, pp. 185–193, APR 2008. 7

[45] T. Cheatham and P. Kollman, "Molecular dynamics simulation of nucleic acids," *Annual Review of Physical Chemisrty*, vol. 51, pp. 435–471, 2000. 7

[46] E. Giudice and R. Lavery, "Simulations of nucleic acids and their complexes," *Accounts of Chemical Research*, vol. 35, no. 6, pp. 350–357, JUN 2002. 7

[47] J. Norberg and L. Nilsson, "Molecular dynamics applied to nucleic acids," *Accounts of Chemical Research*, vol. 35, no. 6, pp. 465–472, JUN 2002. 7

[48] M. Orozco, A. Perez, A. Noy, and F. Luque, "Theoretical methods for the simulation of nucleic acids," *Chemical Society Reviews*, vol. 32, no. 6, pp. 350–364, NOV 2003. 7

[49] M. Hagan and A. Chakraborty, "Hybridization dynamics of surface immobilized dna," *Journal of Chemical Physics*, vol. 120, no. 10, pp. 4958–4968, MAR 8 2004. 7

[50] P. Maiti, T. Pascal, N. Vaidehi, J. Heo, and W. Goddard, "Atomic-level simulations of seeman dna nanostructures: The paranemic crossover in salt solution," *Biophysical Journal*, vol. 90, no. 5, pp. 1463–1479, MAR 2006. 7

[51] K. Wong and B. Pettitt, "A study of dna tethered to surface by an all-atom molecular dynamics simulation," *Theoretical Chemistry Accounts*, vol. 106, no. 3, pp. 233–235, JUL 2001. 8, 46, 57

[52] A. MacKerell, D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3586–3616, APR 30 1998. 8

[53] A. MacKerell, N. Banavali, and N. Foloppe, "Development and current status of the charmm force field for nucleic acids," *Biopolymers*, vol. 56, no. 4, pp. 257–265, 2000. 8

[54] N. Foloppe and A. MacKerell, "All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data," *Journal of Computational Chemistry*, vol. 21, no. 2, pp. 86–104, JAN 30 2000. 8, 30

[55] K. Wong and B. Pettitt, "Orientation of dna on a surface from simulation," *Biopolymers*, vol. 73, no. 5, pp. 570–578, APR 5 2004. 8

[56] T. A. Knotts, N. Rathore, D. C. Schwartz, and J. J. de Pablo, "A coarse grain model for dna," *Journal of Chemical Physics*, vol. 126, no. 8, FEB 28 2007. 9, 10, 17, 23, 47

[57] B. Coleman, W. Olson, and D. Swigon, "Theory of sequence-dependent dna elasticity," *Journal of Chemical Physics*, vol. 118, no. 15, pp. 7127–7140, APR 15 2003. 9

[58] A. Flammini, A. Maritan, and A. Stasiak, "Simulations of action of dna topoisomerases to investigate boundaries and shapes of spaces of knots," *Biophysical Journal*, vol. 87, no. 5, pp. 2968–2975, NOV 2004. 9

[59] M. Hao and W. Olson, "Modeling dna supercoils and knots with b-spline functions," *Biopolymers*, vol. 28, no. 4, pp. 873–900, APR 1989. 9

[60] J. LaMarque, T. Le, and S. Harvey, "Packaging double-helical dna into viral capsids," *Biopolymers*, vol. 73, no. 3, pp. 348–355, FEB 15 2004. 9

[61] A. Srinivasan, R. Torres, W. Clark, and W. Olson, "Base sequence effects in double helical dna .1. potential-energy estimates of local base morphology," *Journal of Biomolecular Structure & Dynamics*, vol. 5, no. 3, pp. 459–&, DEC 1987. 9

[62] R. Maroun and W. Olson, "Base sequence effenct in double-helical dna .2. configurational statistics of rodlike chains," *Biopolymers*, vol. 27, no. 4, pp. 561–584, APR 1988. 9

[63] ——, "Base sequence effenct in double-helical dna .3. average properties of curved dna," *Biopolymers*, vol. 27, no. 4, pp. 585–603, APR 1988. 9

[64] A. Matsumoto and W. Olson, "Sequence-dependent motions of dna: A normal mode analysis at the base-pair level," *Biophysical Journal*, vol. 83, no. 1, pp. 22–41, JUL 2002. 9

[65] M. Peyrard, "Nonlinear dynamics and statistical physics of dna," *Nonlinearity*, vol. 17, no. 2, pp. R1–R40, MAR 2004. 9

[66] D. Sprous and S. Harvey, "Action at a distance in supercoiled dna: Effects of sequence on slither, branching, and intramolecular concentration," *Biophysical Journal*, vol. 70, no. 4, pp. 1893–1908, APR 1996. 9

[67] D. Sprous, R. Tan, and S. Harvey, "Molecular modeling of closed circular dna thermodynamic ensembles," *Biopolymers*, vol. 39, no. 2, pp. 243–258, AUG 1996. 9

[68] R. Tan and S. Harvey, "Molecular mechanics model of supercoiled dna," *Journal of Molecular Biology*, vol. 205, no. 3, pp. 573–591, FEB 5 1989. 9

[69] R. Tan, D. Sprous, and S. Harvey, "Molecular dynamics simulations of small dna plasmids: Effects of sequence and supercoiling on intramolecular motions," *Biopolymers*, vol. 39, no. 2, pp. 259–278, AUG 1996. 9

[70] A. Vologodskii, "Brownian dynamics simulation of knot diffusion along a stretched dna molecule," *Biophysical Journal*, vol. 90, no. 5, pp. 1594–1597, MAR 2006. 9

[71] N. Bruant, D. Flatters, R. Lavery, and D. Genest, "From atomic to mesoscopic descriptions of the internal dynamics of dna," *Biophysical Journal*, vol. 77, no. 5, pp. 2366–2376, NOV 1999. 9

[72] H. Tepper and G. Voth, "A coarse-grained model for double-helix molecules in solution: Spontaneous helix formation and equilibrium properties," *Journal of Chemical Physics*, vol. 122, no. 12, MAR 22 2005. 9

[73] K. Drukker and G. Schatz, "A model for simulating dynamics of dna denaturation," *Journal of Physical Chemistry B*, vol. 104, no. 26, pp. 6108–6111, JUL 6 2000. 9

[74] S. Buyukdagli, M. Sanrey, and M. Joyeux, "Towards more realistic dynamical models for dna secondary structure," *Chemical Physics Letters*, vol. 419, no. 4-6, pp. 434–438, FEB 26 2006. 9

[75] M. Sales-Pardo, R. Guimera, A. Moreira, J. Widom, and L. Amaral, "Mesoscopic modeling for nucleic acid chain dynamics," *Physical Review E*, vol. 71, no. 5, Part 1, MAY 2005. 9

[76] E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo, "A mesoscale model of DNA and its renaturation," *Biophys. J.*, vol. 96, no. 5, pp. 1675–1690, 2009. 10, 17, 20, 47

[77] ——, "Uncovering pathways in DNA oligonucleotide hybridization via transition state analysis," *Proc. Natl. Acad. Sci.*, vol. 106, no. 49, p. 21007, 2009. 10, 17

[78] E. J. Sambriski, V. Ortiz, and J. J. de Pablo, "Sequence effects in the melting and renaturation of short DNA oligonucleotides: structure and mechanistic pathways," *J. Phys.-Condes. Mat.*, vol. 21, no. 3, p. 034105, 2009. 10, 17

[79] A. Jayaraman, C. K. Hall, and J. Genzer, "Computer simulation study of molecular recognition in model dna microarrays," *Biophysical Journal*, vol. 91, no. 6, pp. 2227–2236, SEP 2006. 10, 75, 76, 87

[80] ——, "Computer simulation study of probe-target hybridization in model dna microarrays: Effect of probe surface density and target concentration," *Journal of Chemical Physics*, vol. 127, no. 14, OCT 14 2007. 10

[81] B. Berg and T. Neuhaus, "Multicanonical algorithms for 1st order phase-transitions," *Physics Letters B*, vol. 267, no. 2, pp. 249–253, SEP 12 1991. 12

[82] F. Escobedo and J. dePablo, "Expanded grand canonical and Gibbs ensemble Monte Carlo simulation of polymers," *Journal of Chemical Physics*, vol. 105, no. 10, pp. 4391–4394, SEP 8 1996. 12

[83] D. Gront, A. Kolinski, and J. Skolnick, "Comparison of three Monte Carlo conformational search strategies for a proteinlike homopolymer model: Folding thermodynamics and identification of low-energy structures," *Journal of Chemical Physics*, vol. 113, no. 12, pp. 5065–5071, SEP 22 2000. 12

[84] F. Yasar, T. Celik, B. Berg, and H. Meirovitch, "Multicanonical procedure for continuum peptide models," *Journal of Computational Chemistry*, vol. 21, no. 14, pp. 1251–1261, NOV 15 2000. 12

[85] E. Kim, R. Faller, Q. Yan, N. Abbott, and J. de Pablo, "Potential of mean force between a spherical particle suspended in a nematic liquid crystal and a substrate," *Journal of Chemical Physics*, vol. 117, no. 16, pp. 7781–7787, OCT 22 2002. 12, 15

[86] N. Rathore, T. Knotts, and J. de Pablo, "Density of states simulations of proteins," *Journal of Chemical Physics*, vol. 118, no. 9, pp. 4285–4290, MAR 1 2003. 12

[87] ——, "Configurational temperature density of states simulations of proteins," *Biophysical Journal*, vol. 85, no. 6, pp. 3963–3968, DEC 2003. 12

[88] F. Wang and D. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states," *Physical Review Letters*, vol. 86, no. 10, pp. 2050–2053, MAR 5 2001. 12

[89] ——, "Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram," *Physical Review E*, vol. 64, no. 5, Part 2, NOV 2001. 12

[90] Q. Yan and J. de Pablo, "Fast calculation of the density of states of a fluid by Monte Carlo simulations," *Physical Review Letters*, vol. 90, no. 3, JAN 24 2003. 12

[91] Y. Sugita and Y. Okamoto, "Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape," *Chemical Physics Letters*, vol. 329, no. 3-4, pp. 261–270, OCT 20 2000. 12

[92] D. Earl and M. Deem, "Parallel tempering: Theory, applications, and new perspectives," *Physical Chemistry Chemical Physics*, vol. 7, no. 23, pp. 3910–3916, 2005. 12, 13

[93] U. Hansmann and Y. Okamoto, "Monte Carlo simulations in generalized ensemble: Multicanonical algorithm versus simulated tempering," *Physical Review E*, vol. 54, no. 5, pp. 5863–5865, NOV 1996. 12

[94] C. Bartels and M. Karplus, "Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations," *Journal of Computational Chemistry*, vol. 18, no. 12, pp. 1450–1462, SEP 1997. 12

[95] ——, "Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy," *Journal of Physical Chemistry B*, vol. 102, no. 5, pp. 865–880, JAN 29 1998. 12

[96] S. Kumar, D. Bouzida, R. Swendsen, P. Kollman, and J. Rosenberg, "The weighted histogram analysis method for free-energy calculations on biomolecules .1. The method," *Journal of Computational Chemistry*, vol. 13, no. 8, pp. 1011–1021, OCT 1992. 12, 36

[97] D. M. Zuckerman and E. Lyman, "A second look at canonical sampling of biomolecules using replica exchange simulation," *Journal of Chemical Theory and Computation*, vol. 2, no. 4, pp. 1200–1202, JUL 11 2006. 16, 29

[98] J. D. Weeks, D. Chandler, and H. C. Andersen, "Role of repulsive forces in determining the equilibrium structure of simple liquids," *J. Chem. Phys.*, vol. 54, no. 12, pp. 5237–5247, 1971. 20, 46, 57

[99] R. Owczarzy, Y. You, B. Moreira, J. Manthey, L. Huang, M. Behlke, and J. Walder, "Effects of sodium ions on dna duplex oligomers: Improved predictions of melting temperatures," *BIOCHEMISTRY*, vol. 43, no. 12, pp. 3537–3554, MAR 30 2004. 21

[100] W. A. Kibbe, "OligoCalc: an online oligonucleotide properties calculator," *Nucleic Acids Res.*, vol. 35, pp. 43–46, 2007. 21, 23

[101] T. A. Knotts IV, N. Rathore, and J. J. de Pablo, "Structure and stability of a model three-helix-bundle protein on tailored surfaces," *Proteins*, vol. 61, pp. 385–397, 2005. 23, 30, 33, 46, 90

[102] E. J. Sambriski, V. Ortiz, and J. J. de Pablo, "Sequence effects in the melting and renaturation of short dna oligonucleotides: structure and mechanistic pathways," *Journal of Physics - Condensed Matter*, vol. 21, no. 3, JAN 21 2009. 23, 24

[103] S. Kumar, J. M. Rosenberg, D. Bouzida, R. Swendsen, and P. A. Kollman, "Multidimensional free-energy calculations using the weighted histogram analysis method," *Journal of Computational Chemistry*, vol. 16, no. 11, pp. 1339–1350, 1995. 23, 25

[104] X. Periole and A. E. Mark, "Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent," *J. Chem. Phys.*, vol. 126, p. 014903, 2007. 28

[105] D. Sindhikara, Y. Meng, and A. E. Roitberg, "Exchange frequency in replic exchange molecular dynamics," *J. Chem. Phys.*, vol. 128, p. 024103, 2008. 28

[106] N. Rathore, M. Chopra, and J. J. de Pablo, "Optimal allocation of replicas in parallel temperating simulations," *J. Chem. Phys.*, vol. 122, p. 024111, 2005. 30

[107] J. Karanicolas and C. Brooks, "The origins of asymmetry in the folding transition states of protein l and protein g," *Protein Science*, vol. 11, no. 10, pp. 2351–2361, OCT 2002. 30

[108] ——, "Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions," *J. Mol. Biol.*, vol. 334, pp. 309–325, 2003. 30

[109] C. L. Brooks, "http://mmtsb.org." 31

[110] H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, and I. Shimada, "Three-dimensional solution structure of the b domain of staphylococcal protein a: comparisons of the solution and crystal structures," *Biochemistry*, vol. 31, pp. 9665–9672, 1992. 31, 35

[111] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," *J. Comp. Chem.*, vol. 13, pp. 1011–1021, 1992. 31

[112] T. Fisher, A. Oberhauser, M. Carrion-Vazquez, P. Marszalek, and J. Fernandez, "The study of protein mechanics with the atomic force microscope," *TRENDS IN BIOCHEMICAL SCIENCES*, vol. 24, no. 10, pp. 379–384, 1999. 33

[113] D. Allison, P. Hinterdorfer, and W. Han, "Biomolecular force measurements and the atomic force microscope," *Current Opinion In Biotechnology*, vol. 13, no. 1, pp. 47–51, 2002. 33

[114] J. R. Forman and J. Clarke, "Mechanical unfolding of proteins: insights into biology, structure and folding," *Current Opinion In Structural Biology*, vol. 17, no. 1, pp. 58–66, 2007. 33

[115] R. Best, D. Brockwell, J. Toca-Herrera, A. Blake, D. Smith, S. Radford, and J. Clarke, "Force mode atomic force microscopy as a tool for protein folding studies," *Analytica Chimica Acta*, vol. 479, no. 1, pp. 87–105, 2003. 33

[116] G. Martyna, M. Klein, and M. Tuckerman, "Nosé-Hoover chains - the canonical ensemble via continuous dynamics," *J. Chem. Phys.*, vol. 97, no. 4, pp. 2635–2643, 1992. 34, 45, 57

[117] S. Suzuki, N. Ono, C. Furusawa, A. Kashiwagi, and T. Yomo, "Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays," *BMC Genomics*, vol. 8, no. 373, 2007. 46, 75, 76

[118] A. Ozel, "Theoretical and experimental investigation of the impact of surfaces on DNA melting temperature," Ph.D. dissertation, The University of Michigan, 2009. 47, 67

[119] S. J. Hurst, H. D. Hill, and C. Mirkin, ""three-dimensional hybridization" with polyvalent DNA-gold nanoparticle conjugates," *J. Am. Chem. Soc.*, vol. 130, pp. 12 192–12 200, 2008. 51, 90

[120] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, "The weighted histogram analysis method for free-energy calculations on biomolecules. 1. the method," *J. Comput. Chem.*, vol. 13, pp. 1011–1021, 1992. 57

[121] S. Kumar, J. Rosenberg, D. Bouzida, R. Swendsen, and P. Kollman, "Multidimensional free-energy calculations using the weighted histogram analysis method," *Journal of Computational Chemistry*, vol. 16, no. 11, pp. 1339–1350, 1995. 57

[122] T. J. Schmitt and T. A. Knotts, "Thermodynamics of dna hybridization on surfaces," *J. Chem. Phys.*, vol. 134, no. 20, MAY 28 2011. 63

[123] E. Southern, K. Mir, and M. Shchepinov, "Molecular interactions on microarrays," *Nat. Genet.*, vol. 21, no. Suppl. S, pp. 5–9, 1999. 65

[124] J. G. Mulle, V. C. Patel, S. T. Warren, M. R. Hegde, D. J. Cutler, and M. E. Zwick, "Empirical evaluation of oligonucleotide probe selection for dna microarrays," *PLOS ONE*, vol. 5, no. 3, MAR 29 2010. 75, 76

[125] G. Bell, R. Pictet, W. Rutter, B. Cordell, E. Tischer, and H. Goodman, "Sequence of the human insulin gene," *Nature*, vol. 284, no. 5751, pp. 26–32, 1980. 76

[126] N. Rathore, T. A. Knotts IV, and J. J. de Pablo, "Confinement effects on the thermodynamics of protein folding: Monte Carlo simulations," *Biophys. J.*, vol. 90, pp. 1767–1773, 2006. 90

[127] T. A. Knotts IV, N. Rathore, and J. J. de Pablo, "An entropic perspective of protein stability on surfaces," *Biophys. J.*, vol. 94, pp. 4473–4483, 2008. 90

[128] S. Wei and T. A. Knotts IV, "Predicting stability of alpha-helical, orthogonal-bundle proteins on surfaces," *J. Chem. Phys.*, vol. 133, p. 115102, 2010. 90

[129] K. M. Guckian, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, D. C. Tahmassebi, and E. T. Kool, "Factors contributing to aromatic stacking in water: Evaluation in the context of DNA," *J. Am. Chem. Soc.*, vol. 122, pp. 2213–2222, 2000. 90

[130] S. Bommarito, N. Peyret, and J. SantaLucia Jr., "Thermodynamic parameters for dna sequences with dangling ends," *Nucleic Acids Res.*, vol. 28, pp. 1929–1934, 2000. 90

[131] J. Isaksson and J. Chattopadhyaya, "A uniform mechanism correlating dangling-end stabilization and stacking geometry," *Biochemistry*, vol. 44, pp. 5390–5401, 2005. 91

[132] B. G. Moreira, Y. You, M. A. Behlke, and R. Owczarzy, "Effects of fluorescent dyes, quenchers, and dangling ends on dna duplex stability," *Biochem. Biophys. Res. Comm.*, vol. 327, pp. 473 – 484, 2005. 91

[133] J. I. Lewis, D. J. Moss, and T. A. Knotts IV, "Multiple molecule effects on the cooperativity of protein folding transitions in simulations," *J. Chem. Phys.*, vol. 136, no. 24, p. 245101, 2012. 91

[134] O.-S. Lee and G. C. Schatz, "Interaction between DNAs on a gold surface," *J. Phys. Chem. C*, vol. 113, pp. 15 941–15 947, 2009. 91