THE UNIVERSITY OF CHICAGO

DNA HYBRIDIZATION MECHANISM ELUCIDATED BY TEMPERATURE-JUMP INFRARED SPECTROSCOPY AND COMPUTATIONAL MODELING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
RYAN JAMES MENSSEN

CHICAGO, ILLINOIS
AUGUST 2020

I dedicate this thesis to my family, friends, teachers, professors, and the many others who supported and encouraged me along this academic journey. This work would not have been possible without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This thesis is the culmination of a long journey that would not have been possible without a great number of people that I have been very fortunate to know. I owe a great deal of gratitude to my advisor Professor Andrei Tokmakoff. His unwavering support and encouragement always meant a great deal to me. He has always been around to discuss the latest results or even the craziest of ideas of which there were many. When the decision was made to go in an entirely new direction and build a computational model he projected more confidence in the project than I honestly had in it myself at the time. I truly believe that I could not have ended up with a better advisor.

The majority of my time at The University of Chicago was spent in the GCIS basement with my fellow Tokmakoff Group members, who all deserve my thanks. When I first joined Luigi De Marco was one of the first people I met and is one of the most welcoming and friendly people I have ever met. My early years in the group would have been very different without his incredible patience when explaining topics in the lab and his backyard BBQs and evenings at the Cove outside of it. When I began in the Tokmakoff Group ”Big Lab” I joined with the Pauls, Stevenson and Sanstead. Paul Stevenson knew every detail about the spectrometers and I greatly appreciated every bit of knowledge that he shared with me. Paul Sanstead worked on DNA with me and in addition to showing me the ropes he was always willing to lend a hand in lab, help interpret data, or talk through and flush out ideas no matter how vague they were initially. Brennan Ashwood joined the DNA project after me and brought with him a lot of new and fresh ideas. I am excited to see the work that he will do with the new temperature-jump spectrometer. I, like many others in the Tokmakoff Group before me, spent time on a side project to build our own flow cell without making significant progress, but Ram Itani picked up the project and has made great strides. I look forward to seeing how what these new capabilities will bring. Sam Penwell was always willing to lend an ear regardless of if the topic was current projects in the lab or talking about life in general and he always provided a welcome perspective and good advice.

Joe Fournier was an excellent lab mate and friend and I enjoyed our time together both in the office and golfing at the driving range. I joined the lab with two classmates, Memo Carpenter and Chi-Jui Feng who have been fantastic support over the past six years. Memo's upbeat personality and inquisitive nature provided useful suggestions at many points throughout my work. When I started my computational work Chi-Jui's knowledge of all things computers was a significant boost to getting the project off to a fast start. Prior to joining the Tokmakoff Group I spent my first year sharing an office with Yining Han, Kade Head-Marsden, Jaehyeok Jin, and Ziwei He. I could not have asked for a better group of people to spend my first year with who made everything from grading to problem sets way more bearable.

I also owe a great deal of gratitude to Professor Jeff Schwinefus, my undergraduate research advisor at St. Olaf. I can honestly say I wouldn't have gotten to this point if he hadn't believed in my and given me my first research opportunity. I learned a lot about academic research from him and after only a short time working in his lab I made the decision to pursue a graduate degree in chemistry.

This thesis would also not have been possible without my strong support system of friends. I need to deeply thank my friends Nick Lee, Paul Carroll, and Maggie Flint. Over the last six years you have helped to keep me grounded while also providing much needed levity and distraction. Matt Boltz and Julia Zinkus-Boltz have been both fantastic friends and amazing neighbors. I have greatly enjoyed our culinary adventures and appreciated our time spent together. A few years ago I joined the West Side Slammers soccer team. I could not have asked for a better group of guys to have as teammates and your friendship and league matches gave me something to look forward to outside of lab each week.

Of course none of this would have been possible without the love and support of my family. My mom and dad, Juli and Mike Menssen, who have always supported me and encouraged me to pursue my passions no matter how difficult things got. My siblings Andrew, Katie, and Ellie who have always been at my side every step of the way. Finally,

my wife Rebecca for the most extraordinary companion I could hope for always providing never ending love and support throughout all of the highs and the lows.

Funding

The research presented in this thesis has been made possible by funding from the National Science Foundation and the National Institutes of Health.

ABSTRACT

The work presented in this thesis utilized experimental and computational methods to investigate the association and dissociation of small DNA oligonucleotides. Fourier transform infrared spectroscopy (FTIR) and temperature-jump (T-jump) infrared (IR) spectroscopy were used to investigate the thermodynamics, mechanism, dynamics, and kinetics of DNA oligos with the sequence 5'-C(AT)$_n$G-3' where n = 2-6. To compliment the experiments a Markov state Monte Carlo kinetic model, intended to be accessible to experimentally focused researchers with regards to the model's complexity and computational expense, was built to simulate association and dissociation trajectories of these sequences plus 5'-ATATGCATAT-3' (GC-core) and 5'-ATATATATAT-3'. These sequences were selected to make a first attempt at separating the different factors that impact DNA dynamics and kinetics focusing initially on sequence length and composition.

IR spectroscopy is ideal for studying DNA due to its ability to resolve adenine-thymine (A:T) and guanine-cytosine (G:C) base pairs. Additionally, the kinetics of the sequences studied here fall within the nanosecond to millisecond time window the T-jump instrument can resolve. The Markov state Monte Carlo model provides improved base pair resolution by independently tracking each base pair providing new insights into the mechanism and dynamics of association and dissociation. The experimental results of the 5'-C(AT)$_n$G-3' series were analyzed using an Eyring analysis of a two-state model providing a clearer interpretation of the reaction energetics by extracting the activation entropy, activation enthalpy, and activation free energy. Global analysis links the thermodynamic and kinetic parameters utilizing a linear dependence on oligo length of the entropic and enthalpic activation barriers. Analysis incorporating the thermodynamic nearest neighbor parameters and the experimentally determined activation enthalpy found that the critical nucleus, the minimum number of base pairs such that the partially formed duplex is stable and will proceed downhill to the fully formed dimer, increases in size with increasing temperature and sequence length.

Association and dissociation trajectories from the kinetic model were analyzed directly and utilizing transition pathway theory (TPT). The dominant association pathways, isolated by TPT, showed two primary motifs: initiating at or next to a G:C base pair, which is enthalpically driven, and initiating in the center of the sequence, which is entropically driven. For GC-core these motifs overlap resulting in a strong preference for initiating association at the central G:C base pairs. For 5'-C(AT)$_n$G-3' sequences the paths compete resulting in a preference for initiating association events either at or next to a terminal G:C base pair or in the center. Configurations in the transition state ensemble were found to increase in size with increasing sequence length and temperature, in good agreement with the literature and the experimentally determined critical nucleus size. Finally, terminal end fraying experimentally observed in GC-core was replicated by the model and shown to be driven by increased thermodynamic accessibility of the frayed states after the T-jump. This was compared to fast dynamics observed for longer 5'-C(AT)$_n$G-3' sequences, the physical origins of which were not previously clear, and suggests that this fast response is also due to thermodynamically driven end fraying.

CHAPTER 1

INTRODUCTION

1.1 DNA Hybridization and Dehybridization

The hybridization of DNA single strands to their complement and the dehybridization of a DNA duplex are fundamental to biological function. There are also a number of interesting applications of DNA hybridization and dehybridization outside of natural biological functions. DNA biosensors use a DNA sequence, commonly a short oligonucleotide, which upon hybridizing with a specific target of interest generates a detectable signal. These biosensors can be used to detect a large number of targets including small molecules, such as drugs, proteins, or complimentary DNA sequences in addition to a wide range of other applications.[1] DNA origami involves designing DNA sequences that fold into complex three dimensional structures with a wide range of applications.[2] Numerous studies have investigated the thermodynamics and kinetics of the folding process, and the energetic driving forces behind it, to better understand and predict how these complex structures fold.[3–5] Similarly, DNA is a common model system for the study of self-assembling polymers and is widely used as a building block for nanomachines.[6] Another fascinating application is DNA computing.[7] DNA computing takes inspiration from the idea of DNA as a molecule for information storage and processing. It has been demonstrated that DNA computers can carry out logical operations,[7,8] simple mathematical operations such as multiplication,[9] and could be used as a practical and cost-effective solution for archiving data.[10] DNA computers have even solved a basic version of the Hamiltonian path problem, a special case of the traveling salesman problem,[11] and simple chess puzzles.[12] All of these applications require the consistent, repeatable, and highly selective hybridization of DNA to function necessitating a complete understanding of the hybridization process.

DNA thermodynamics, kinetics, and dynamics all play an important role in a large vari-

ety of biological processes. Processes varying from DNA replication to protein expression all involve the hybridization of nucleic acid sequences ranging from long sequences to short RNA primers. While it has long been known that the thermodynamic stability of DNA plays an important role in its biological function, more recent findings have demonstrated that DNA dynamics and kinetics also play a significant role. DNA undergoes a variety of dynamical changes including structural changes within an intact duplex and fluctuations involving the localized loss of base pairing. A large number of distortions to the "ideal" double helix have been identified and extensively categorized. It has been recognized that these significantly impact the physical properties of DNA which likely plays a role in biological processes, particularly protein-DNA interactions.[13,14] In addition to these internal fluctuations it is well known that DNA undergoes dynamical breathing modes, which range from the localized loss of a single base pair, referred to as base flipping,[15] to continuous stretches of broken base pairs that can grow quite large, which are referred to as bubbles.[16] Base flipping is the process by which a single base ends up in an extrahelical position, a configuration that is known to be adopted in a variety of instances where a base is in the active site of a protein. This occurs in the context of a variety of processes, an example of which is the removal of a mismatched base pair or a modified base pair, such as a methylated cytosine.[17] It has been proposed that the dynamics through which the base adopts an extrahelical configuration play a significant role in such processes, an example of which is that these dynamics aid in the recognition of the target base by the protein.[15,17] DNA breathing modes, where base pairs dynamically open and close along a stretch of the double helix, have been thoroughly studied utilizing a number of different techniques and have been shown to play a role in a large number of biological processes. Two such processes, among many others, are the recognition of thymine dimers formed due to UV damage, which if unrepaired may lead to skin cancer, and the initiation of DNA transcription, one of the most fundamental processes in biology.[16,18–22]

The thermodynamics of nucleic acid hybridization and dehybridization have been ex-

tensively studied by experimental and computational methods. While the thermodynamics of DNA are quite well understood, there are still many details regarding the kinetics and dynamics that are not. In particular, even though initial kinetic studies were conducted over half a century ago and continue to this day, questions about the description of the energy landscape and underlying mechanism, particularly the form of the transition state, remain unanswered. Despite interest in the folding of nucleic acid hairpins[23–28], the diffusion-limited association of small nucleic acid oligomers has received less attention. Recent work on DNA duplex dynamics and kinetics has focused more on longer lengths, often ranging from 20 to over 100 base pairs, often looking at more complex dynamics, such as bubble formation, that do not necessarily involve the complete hybridization or dehybridization of the duplex.[16,29–32]. Additional studies have been conducted that look at how factors such as salt concentration affect the formation of DNA duplexes.[31,33]

1.2 Hybridization Dynamics

Before jumping into specific models for association and dissociation it is worth taking a broader look at the overall processes and the various dynamics that occur. In this section we will discuss common dynamics and the terminology that is used to describe them. The purpose of this is to provide a foundation for understanding both the physical processes that occur and the language that is used to describe them.

DNA association and dissociation are commonly discussed utilizing the conceptual picture of the nucleation-zipper mechanism. This mechanism has been incorporated into a number of models which will be discussed in more detail in the next section. Here our focus is broader and we will use the mechanism to qualitatively discuss different aspects of DNA association and dissociation. A general overview of the nucleation-zipper mechanism is shown in Figure 1.1. In the nucleation-zipper picture there are three distinct phases of the association process. The first step is two monomers diffusing together to form the first base

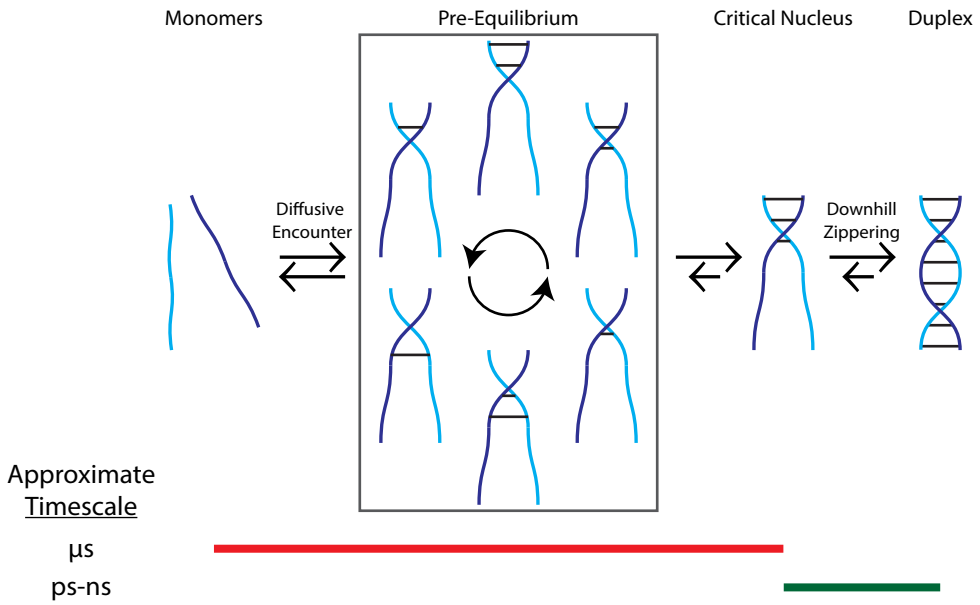

Figure 1.1: Overview of the different stages, and their approximate time scales, of DNA association according to the nucleation-zipper mechanism.

pair. This process is largely considered to be diffusion controlled.[34,35] In a simple case the rate constant for the reaction of two molecules that occurs when the molecules come within a distance *R* of one another is given by $k_D = 4\pi RDN_A$ where k_D is the diffusion controlled reaction rate, D is the sum of the diffusion coefficients for the two molecules and N_A is Avogadro's number.[36] The first base pair can form anywhere along the sequence either as an in-register or out-of-register base pair, with internal rearrangement required in the case of an out-of-register base pair.

Upon forming the first base pair the process enters what is commonly known as the pre-equilibrium.[34,35,37] During this portion of the reaction the partially formed duplex rapidly interconverts between a variety of configurations all of which are not thermodynamically stable. The partially formed duplex remains in the pre-equilibrium until it either returns to the monomer state, after which the monomers may reenter the pre-equilibrium or diffuse apart, or the partially formed duplex forms a structure known as the critical nucleus. The critical nucleus is a structure that contains the minimum number of intact base pairs such that the partially formed duplex is stable and has a significantly greater probability of

rapidly zipping up the remaining bases in a downhill fashion relative to returning to the pre-equilibrium. The pre-equilibrium portion of the process involves not just the formation and breaking of base pairs, but in situations where out-of-register base pairs are present it also includes internal rearrangements that result in the formation of in-register base pairs.[38]

The non-fundamental nature of the pre-equilibrium kinetic step in the association has been linked to a negative activation energy[35,37–39] that has been observed by a number of different studies.[35,37–41] However, there is some disagreement in the literature as to what the sign of the association activation energy should be with some studies finding a positive activation energy.[42,43] Some potential explanations for the differing signs of the activation energy have been proposed including a sequence effect due to the increased stability of G:C base pairs potentially limiting the extent of the pre-equilibrium step.[42] However, the fact that sequences with G:C base pairs have been found with both positive[42,43] and negative[39,40] activation energies suggests that the presence of G:C base pairs does not necessarily dictate the sign of the activation energy. Temperature has also been proposed as a contributing factor due to the association rate taking the form of a bell shaped curve as a function of temperature with a maximum rate below T_{m},[34] which has been experimentally observed.[41] This would result in a differently signed activation energy on each side of the maximum. As a result the sign of the association activation energy may prove to be a useful indicator of the underlying mechanism due to its potential connection to the existence of a pre-equilibrium step. However, additional research is necessary to provide a unified explanation for this connection.

The downhill zipping portion of the reaction involves the sequential formation of the remaining base pairs from the critical nucleus to the ends of the sequence, a process that occurs orders of magnitude faster than the formation of the critical nucleus. In Figure 1.1 the critical nucleus is shown including a terminal base pair, with the resulting zipping occurring towards the other end. However, the critical nucleus may form anywhere along the sequence and if it does not include a terminal base pair zipping will proceed out in

both directions.

It is worth pausing here to make an important note about the terminology used with respect to the critical nucleus and a similar structure, the transition state. While we will more rigorously, and quantitatively, define the transition state later on, for the time being we will consider it to be a structure that sits at the peak of a standard reaction free energy diagram. This implies that the probability of the transition state going to the monomer state is equal to the probability of going to the fully formed dimer state. Since the critical nucleus is stable and proceeds in a downhill fashion to the fully formed dimer state we can think of it as the first configuration found on the dimer side of the free energy peak where the transition state is found.

Finally we will discuss common dissociation dynamics that occur. Dissociation follows the same general mechanism provided in Figure 1.1, but in the opposite direction. A fully formed duplex will begin to dissociate and will rapidly form and break base pairs until enough are broken such that the structure is no longer stable, which by definition is one base pair smaller than the corresponding critical nucleus since that structure is stable, at which point the remaining bases will break apart resulting in the strands entering the monomer state.

There are two ways in which the bases begin to dissociate, fraying or bubble formation. Fraying is the sequential loss of base pairing that initiates at the end of the sequence and the base pairs break sequentially towards the center.[40] Configurations where dissociated base pairs exist but both terminal base pairs are intact are known as bubbles.[16,29,30] It is possible to see fraying occurring from both termini simultaneously. Additionally, in long enough sequences it is possible for multiple bubbles to occur, or for bubbles and fraying to both be present at the same time. There is also the potential for interplay between the two, if a frayed end becomes long enough the base pairs can reattach resulting in the formation of a bubble, or create both a bubble and a frayed state. Bubble states that expand far enough will also eventually create frayed states, especially in the context of

shorter sequences.

Moving forward we will explore how the different dynamics and configurations provide insight into mechanistic and dynamical questions about DNA association and dissociation. We start by taking this broad view of the mechanism and exploring specific models, methods, and techniques that have been utilized to explore these questions. We start with a closer look at specific models that are built off of the canonical nucleation-zipper mechanism.

1.3 Proposed Mechanisms for DNA Association and Dissociation

One of the earliest and most commonly employed models describing nucleic acid association and dissociation is the zipper model which was first proposed in the 1950s.[44,45] In some contexts the zipper model is also sometimes referred to as the nucleation-zipper model, since it is closely aligned with the nucleation-zipper mechanism. The version discussed here follows the formalism of Craig, Crothers, and Doty which is one of the more extensive versions of the model.[37] The model describes the association and dissociation of DNA as a series of sequential steps each involving forming or breaking a single base pair. The reaction scheme for this version of the zipper model is presented in Figure 1.2. The k_f and k_b parameters are rate constants for forming and breaking base pairs at the end of a long helical segment. The d parameter is a degeneracy factor that accounts for the number of different ways the system can move between states based on the different possible configurations for a given number of intact base pairs. For example, there are N

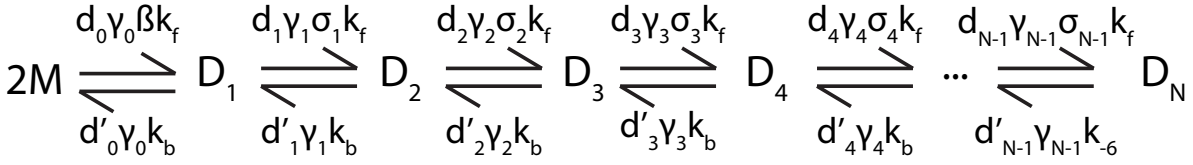

Figure 1.2: DNA zipper model mechanism using the formalism of Craig, Crothers, and Doty.[37]

different ways to form the first intact base pair, where N denotes the total number of base pairs in the sequence. For the reverse steps the degeneracy is two, with the exception of breaking the very last base pair which has a degeneracy of one since the zipper model requires that all intact base pairs must be sequential and thus a base pair can only be broken at one of the two ends of the stretch of intact base pairs. Another way to put this is that bubble states are not allowed. The γ parameter is a kinetic parameter that provides additional flexibility for reducing the equilibrium constant either through decreasing the forward rate or increasing the backwards rate, in comparison to σ, another attenuation parameter, that only impacts the forward rate constant. The σ parameter attenuates the forward rate to account for the fact that k_f is the rate of formation for a base pair at the end of a long helical stretch, where base pair formation is expected to be the fastest, and base pairs will form slower earlier in the process. Finally, β serves a similar purpose to σ, though it only attenuates the formation of the first base pair from two monomer strands. The zipper model has been utilized extensively in the literature for studying the association and dissociation of DNA oligos. The model consistently follows this general form, however small changes in the symbols used for each parameter and their definitions do exist.[35,37,42,44–48]

The zipper model is often referred to as the nucleation-zipper model to reflect an important aspect of the mechanism that is not necessarily apparent from simply looking at the reaction scheme shown in Figure 1.2, which is the two components of the association mechanism discussed previously and shown in Figure 1.1. The reaction scheme in no way identifies the formation of the critical nucleus nor where along the scheme its formation occurs. Currently, there is no clear consensus in the literature with respects to the size of the critical nucleus, though most estimates put it somewhere in the realm of one to four base pairs and suggest it is impacted by factors including base pair composition, temperature, and sequence length.[35,37,38,42]

More recent work on the original theory of the zipper model has focused on evaluating

the scaling of the association rate with oligomer length, originally proposed by Wetmur and Davidson[34] and studied by many others,[31,32,35,37,42] and attempting to determine its underlying causes.[31,32] This scaling behavior predicts that for sequences under 100 base pairs the association rate should be proportional to length (L) and for sequences over 100 base pairs it should be proportional to $L^{0.5}$.[34]

More recent computational studies have seen evidence of critical nucleus formation while also proposing new mechanisms such as "slithering" or the "inch-worm" mechanism that may also play a role in the formation of DNA duplexes.[31,38] Both of these mechanisms differ significantly from the zipper model in that they involve configurations with out-of-register base pairs as intermediates on the way to a fully hybridized duplex. While sequences do not need to be perfectly repetitive for these mechanisms to play a significant role, they will be more relevant for repetitive sequences. Additionally, the probability for out-of-register binding is higher for longer sequences suggesting these mechanisms would be expected to be more relevant for longer sequences.[38] However, these mechanisms have not yet been experimentally observed.

1.4 Infrared Spectroscopy of DNA

Infrared (IR) spectroscopy is a powerful tool for studying DNA hybridization and dehybridization. The most prominent benefit of IR spectroscopy is that each of the four DNA bases has a distinct IR spectra meaning changes in adenine-thymine (A:T) base pairs can be observed independently from changes in guanine-cytosine (G:C) base pairs. This can be observed in Figure 1.3 which shows both the two-dimensional infrared (2DIR) spectrum and the linear Fourier transform infrared spectroscopy (FTIR) spectrum for each of the four DNA bases, the collection and processing of these spectra and a more detailed discussion of their analysis is provided in later chapters. The peak assignments for each base have been determined and their frequencies are given above each of the FTIR spectra in Figure

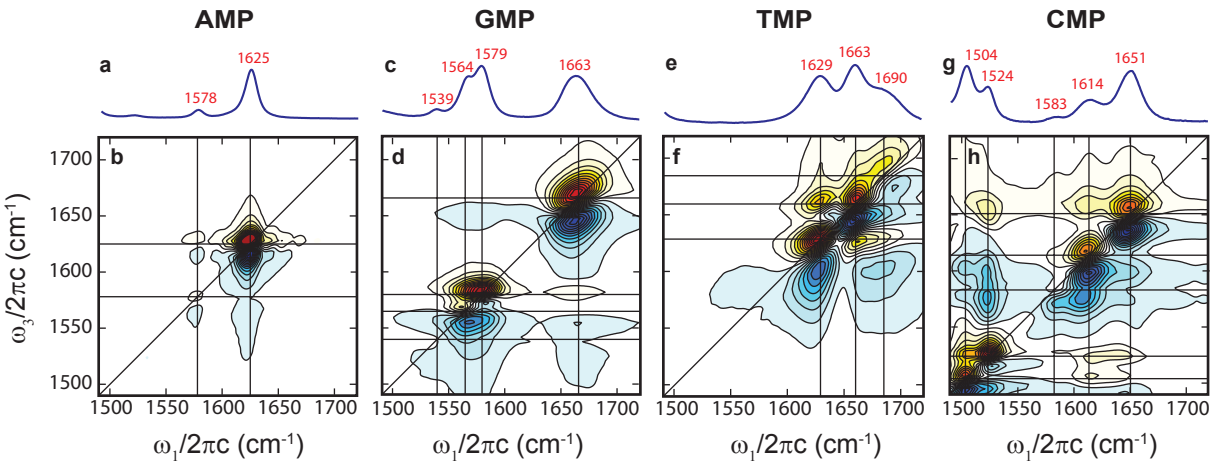

Figure 1.3: 2DIR and FTIR spectra of adenine monophosphate (a-b), guanine monophosphate (c-d), thymine monophosphate (e-f), and cytosine monophosphate (g-h), adapted from Ref 49.

1.3. Each of these peaks is made up of a convolution of individual vibrations that have been previously outlined by DFT calculations.[49] Generally, these peaks are broken down into two categories. Peaks below 1650 cm^{-1} are commonly referred to as ring modes and contain contributions from in-plane ring vibrations. Peaks above 1650 cm^{-1} are referred to as carbonyl modes because they contain strong contributions from symmetric and asymmetric carbonyl stretching modes. These terms will be used to generally describe the two sets of peaks throughout this work.

We can see that both guanine and cytosine have regions where they are the only base pair with a peak, at around 1560-1580 cm^{-1} and 1500-1525 cm^{-1}, respectively. Adenine and thymine don't have individual peaks that are quite as clean, though thymine does have a small shoulder that sits relatively isolated at around 1690 cm^{-1}, but this can be hard to observe in sequences without a large percentage of thymine bases. However, around 1625 cm^{-1} both adenine and thymine have a very strong absorption, on top of a weak cytosine absorption, which means this peak can be used as a relatively clean marker for A:T base pairs. Since the research described here is primarily focused on hybridization and dehybridization, the ability to isolate adenine from thymine is of less concern compared to the ability to separate A:T base pairs from G:C base pairs.

All of the peaks, with the exception of the thymine shoulder found at approximately 1690 cm^{-1}, are suppressed by duplex hybridization. The experiments conducted here perturb the system through heating which means that they induce dehybridization. As a result our experiments primarily track the increasing IR signal that occurs as the result of a loss of base pairing. Additional signals in the nonlinear experiments occur due to the existence of cross peaks that primarily appear, in the context of this work, due to coupling between different vibrational modes. Intramolecular coupling appears in Figure 1.3 as cross peaks between different modes, easily observed in Figure 1.3d between the guanine ring modes and carbonyl modes. Intermolecular cross peaks also appear between vibrational modes of hydrogen bonded base pairs, such as cross peaks between adenine ring modes and thymine carbonyl modes, which disappear as a result of the loss of hydrogen bonding. This provides additional signal changes that can be tracked, and is a clear marker of dehybridization in the 2DIR experiments since these cross peaks only appear when intact hydrogen bonds between the complimentary base pairs exist.

Now that we have introduced the ability of IR to differentiate the signal from different base pairs we will highlight why IR is useful for observing hybridization and dehybridization

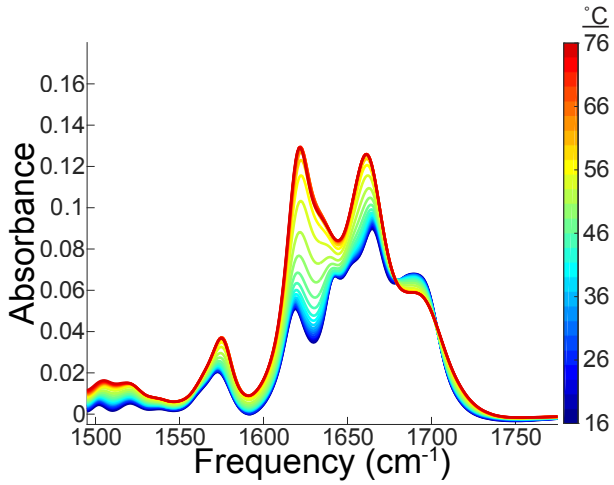

Figure 1.4: Temperature ramp series of FTIR spectra taken every 3 °C between 16 °C and 76 °C for the sequence 5'-CATATATATATATATG-3' showing the change in absorbance as a result of the loss of base pairing.

processes. Figure 1.4 contains a series of FTIR spectra for 5'-CATATATATATATATG-3' taken approximately every 3 °C from 16 °C to 76 °C tracking the duplex as it dehybridizes with increasing temperature. We see that, with the exception of the thymine shoulder at 1690 cm^{-1}, the signal increases with increasing temperature because the duplex structure suppresses these vibrational modes. This provides the ability to track the loss of A:T base pairs, looking primarily at the growth of the 1625 cm^{-1} peak, and the loss of G:C base pairs, looking primarily at the growth of the 1580 cm^{-1} peak. With the ultimate goal of gaining insight into the mechanism of DNA hybridization, especially when considering the effect of base pair composition, the ability to independently monitor A:T and G:C base pairs that IR provides makes it particularly well suited for these types of studies.

Beyond the benefits of IR spectroscopy, the kinetics of the DNA sequences studied here are an ideal fit to the time range our temperature-jump experiment is able to resolve. The spectrometer is able to resolve kinetics that fall in the nanosecond to millisecond timescale. For the sequences studied here the full dissociation occurs on a timescale of tens to hundreds of microseconds with fast dynamics occurring on the order of nanoseconds which places these processes directly in our resolvable time window. This makes these systems ideal candidates for study utilizing these techniques.[39,40,50]

1.5 Computational Methods

One of the most well-known and utilized DNA thermodynamic models is the nearest neighbor (NN) model.[51–54] The NN model is utilized in the construction of the lattice model[55] that the kinetic model presented here is built off of. The NN model, and resulting parameters, are also used as a point of comparison during analysis of the thermodynamic and kinetic results presented here. The model is based on the assumption that the stability of a given base pair depends on its identity and the identify of the neighboring base pair. The dinucleotides are often represented with a slash denoting the two strands in

antiparallel orientation such that AC/TG refers to a 5'-AC-3' that is paired with 3'-TG-5'.[51] As an example, the sequence 5'-CATG-3', which is self-complimentary, contains three dinucleotide subunits (CA/GT, AT/TA, and TG/AC).

The model can accurately predict DNA secondary structure in a variety of salt conditions. The NN model breaks down thermodynamic parameters for sequences of several different DNA motifs into individual parameters containing the contribution of each dinucleotide, of which there are ten, to each particular thermodynamic quantity. In addition to the dinucleotide parameters, additional parameters exist to account for the impact other factors have on the thermodynamic parameters. These include accounting for the initiation, a penalty applied to sequences with terminal A:T base pairs, and a symmetry correction for self-complimentary sequences.[51–54] Parameters are provided for ΔH^0 and ΔS^0. In some cases additional parameters are provided for ΔG^0_{37} which is the value of ΔG^0 at 37 °C.

The ΔH^0 and ΔS^0 are determined by identifying each dinucleotide in a given sequence and summing the parameters provided by the model in addition to the values for any relevant additional parameters. The value of ΔG^0 at any temperature, and the T_m, can then be determined from ΔH^0 and ΔS^0. It is worth noting that the individual NN parameters have been shown to not carry any dependence on sequence length, however the salt correction does.[51] Parameters are often determined and provided for a particular salt concentration, however equations to correct for salt concentration have been determined.[52,56] The specific NN parameter set used in the thermodynamic lattice model, and the analysis presented here, was determined by SantaLucia utilizing a set of 108 oligonucleotide duplexes.[51] The most common method for obtaining thermodynamics from experiment to develop NN parameters is melting curves monitored by UV spectroscopy, though DSC and other techniques have also been used.[51]

The NN parameters have been shown to be highly accurate for predicting thermodynamic values. A study examining 264 duplexes of length 4 to 16 base pairs found

an average absolute deviation of 1.6 °C between the experimental T_m and calculated T_m, which is particularly good agreement since the model was optimized to predict ΔG^0, ΔH^0, and ΔS^0 rather than the T_m.[52] While next nearest neighbor models, that break down the thermodynamics into parameters for each possible trinucleotide, exist; evidence shows that they do not provide any significant improvement over NN models.[57] Additionally, NN parameter sets have been determined by a number of different research groups that have been shown to all be in agreement.[51]

Lattice models are another common method for modeling DNA thermodynamics. Lattice models are a class of models where the physical space occupied by the system is a discrete lattice rather than a continuous space. This simplification greatly reduces the computational cost. It has been shown that lattice models can accurately reproduce melting thermodynamics and study the energy landscape of nucleic acid duplex oligomers and hairpins.[24,55,58]

Advances in coarse-grained molecular dynamics (MD) simulations have provided new insights into DNA association mechanisms and the resulting kinetics.[31,38,59,60] Two commonly used coarse-grained MD models are the 3SPN.2 model[31,59–62] and the OxDNA model.[38,63] Coarse-grained MD differs from all atom MD in that it groups atoms together into single entities, commonly referred to as interaction sites, each of which represents a portion of the DNA base, which greatly reduces computing costs. For example, the 3SPN.2 model uses three interaction sites to model the nucleotide that represent the sugar, the phosphate, and the nucleobase itself. These models have been used extensively to investigate the kinetics and mechanism of DNA oligo hybridization and examine the effect of a number of different parameters including length, sequence, and salt concentration on the association process.[31,38,60] These methods have made numerous contributions to the study of DNA kinetics including, but not limited to, examining the scaling relationship proposed by Wetmur and Davidson,[34] proposing new mechanisms by which DNA associates including the previously mentioned “inch-worm” and ”pseudoknot”

mechanisms,[31,38] and examining the configurational states, and their size, that make up the transition state for particular sequences.[31,38,59,60] Coarse-grained MD simulations are usually carried out under equilibrium conditions where DNA hybridization is a rare event making sampling the kinetics and dynamics difficult. To overcome this, a few different sampling methods are utilized including umbrella sampling, transition path sampling, or forward flux sampling.[31] Another method for studying the transitions is to generate a set of key kinetic states from the trajectories which can be analyzed as a Markov state model.[64–66] While these techniques are very powerful they have a relatively high computational cost which, combined with the complexity of running and analyzing them, puts them out of reach for many researchers.

We now look to another method used to study the kinetics of biomolecular systems, the use of Monte Carlo methods. Monte Carlo methods are used to simulate trajectories of a system evolving through a given state space. One option for generating the state space is to utilize the states generated from a thermodynamic lattice model.[4] Other approaches for generating states include using the NN model[5] and building Markov state models from MD simulations.[64–66]

Here we broadly discuss a few methods, and their applications, from the literature. Monte Carlo simulations are commonly run in two different ways. With discrete and constant time steps or in continuous time. In the continuous time case the model moves forward in discrete time steps, but each discrete time step is randomly selected from a continuous distribution of potential time steps. One commonly utilized algorithm for discrete time step models is the Metropolis-Hastings algorithm,[67,68] which has been used to study both proteins[69] and DNA.[4,20] In brief, at each discrete time step the algorithm randomly selects a potential move based on the probability density for the system. Then the model decides whether to accept that move and go to the new state, or to reject that move and remain in the current state. The probability of accepting or rejecting is proportional to the probabilities of the two states and is commonly referred to as an acceptance ratio. The

algorithm decides whether or not a move is accepted by comparing the acceptance ratio against a randomly generated number. After making the move if accepted, or remaining in the same state if not, the model steps forward one time step. The Metropolis-Hastings method is particularly useful for resolving mechanisms and pathways, however it is difficult to extract meaningful kinetic information without additional complexity.

To extract kinetic information it is beneficial to utilize a Monte Carlo algorithm that operates in continuous time. One commonly used algorithm is the Gillespie algorithm.[70,71] This algorithm is used to generate the trajectories for the model presented in this work and as such the theory and methodology required for carrying it out will be described in detail later. The Gillespie algorithm has been used to study DNA in a variety of contexts including breathing dynamics[19,72] and the hybridization of a variety of DNA motifs and structures.[3,5,73,74]

Now we will briefly introduce two analysis techniques commonly utilized in conjunction with these computational methods that are also used in the work presented here. Transition path theory (TPT) was developed for the purpose of analyzing the statistical properties of the pathways between any two subsets in the state space of continuous-time Markov chains on discrete state spaces.[75–78] A common application of TPT is determining the dominant reactive pathways between two states in a Markov state model through calculating the reactive flux between all the intermediate states that make up the possible pathways. This is often utilized to determine and study the dominant folding pathways of proteins.[64–66] A particularly useful aspect of TPT analysis is that it does not require that trajectories are run to disseminate pathway information; it simply requires the transition rate matrix from a Markov state model to determine the pathways and relevant statistical information.

While TPT is very useful for isolating the dominant pathways between two states, usually the initial and final states of a folding process, it does not easily provide significant insight into the intermediate stages of the process beyond what states the pathway moves

through. This misses out on a significant aspect of the mechanism which is the identity of the transition state and critical nucleus. One way to isolate the identity of the transition state is through defining the transition state ensemble (TSE). The TSE is a collection of states that represent the transition state, which is considered to be an ensemble since processes involving large complex biomolecules are unlikely to have a single configuration that makes up the transition state. One method for determining the configurations in the TSE for the hybridization of DNA oligos is based on the probability of a given intermediate configuration first reaching the fully formed dimer state versus the monomer state. A configuration is considered to be in the TSE if the probability of going to the fully formed dimer state is roughly equal to that of going to the monomer.[59,60] Isolating the TSE in conjunction with the identification of the dominant reactive pathways provides a comprehensive picture of the association and dissociation processes and provides a useful framework for understanding multiple aspects of the mechanism.

1.6 Research Question and Goals

Even with the significant interest in the thermodynamics, kinetics, and dynamics of DNA hybridization and dehybridization there are a number of remaining open questions. The transition state, which is correlated with the critical nucleus, has not been conclusively characterized and inconsistencies in the literature exist with regards to whether the association mechanism of certain sequences is described by Arrhenius or anti-Arrhenius behavior.[40,42] This carries mechanistic importance since the anti-Arrhenius kinetics have been proposed to be connected to the pre-equilibrium step and Arrhenius kinetics might imply that this step is not present.

Additionally, little exists with regards to accessible predictive models for DNA kinetics based purely on sequence, such as a kinetic analog to the NN thermodynamic parameters. NN parameters can be used to predict barriers to dissociation within a two-state model,

allowing the dissociation rate to be predicted from the association rate.[54] However, conclusive evidence for the validity of the two-state model is lacking. Additionally, more work is needed to determine the robustness of the predictive power and how sequence, secondary structure, and mechanism all impact the effectiveness of predictions.

The research presented here is motivated by multiple factors. Recent research has demonstrated that IR methods are capable of observing mechanistic changes as a result of base pair composition.[40,49] Additionally, developing accessible computational models to pair with these experiments greatly improves our ability to study the mechanism, kinetics, and dynamics of DNA oligo hybridization. This motivated the development of an accessible and computationally inexpensive kinetic model that could be used in conjunction with our experimental techniques and thermodynamic lattice model. We were motivated to revisit the length dependent trends studied decades ago to reexamine these systems utilizing modern experimental techniques. It is also our hope that examining the length dependence of DNA oligos will help drive forward the study of longer and more biologically relevant oligos utilizing our IR spectroscopic methods in combination with our thermodynamic and kinetic models. Utilizing modern label-free sequence specific spectroscopies on the length dependent samples, and a new stochastic model on both the length dependent samples and the sequence dependent samples, provides new insights into the process by which DNA associates and dissociates while also providing new insights into equilibrium fluctuations and dynamics.

The research conducted here focused on a few overarching goals. The first was to build an accessible and computationally inexpensive kinetic model for use in conjunction with experimental results to obtain more specific mechanistic insight than can be provided with experiment alone. Beyond the goal of developing new tools, the ultimate goal is focused on understanding and providing a more robust description of the association mechanism, transition state, and their underlying energetics. While sequence specificity and length are just two of the many variables that impact the hybridization and dehybridization

of DNA oligos we believe that this work provides a significant step forward by clarifying the role of these two significant variables. Additionally, we believe that the methods proposed here can provide a framework for investigating other important variables to continue to provide greater clarity with regards to this long discussed problem.

1.7 Acknowledgements

I would like to thank Greg Kimmel and Ram Itani for their careful reading and thoughtful comments on this chapter.

1.8 References

1. Zhao, W. W.; Xu, J. J.; Chen, H. Y. Photoelectrochemical DNA Biosensors. *Chem. Rev.* **2014**, *114*, 7421–7441.
2. Wang, P.; Meyer, T. A.; Pan, V.; Dutta, P. K.; Ke, Y. The Beauty and Utility of DNA Origami. *Chem* **2017**, *2*, 359–382.
3. Dunn, K. E.; Dannenberg, F.; Ouldridge, T. E.; Kwiatkowska, M.; Turberfield, A. J.; Bath, J. Guiding the Folding Pathway of DNA Origami. *Nature* **2015**, *525*, 82–86.
4. Cumberworth, A.; Reinhardt, A.; Frenkel, D. Lattice Models and Monte Carlo Methods for Simulating DNA Origami Self-Assembly. *J. Chem. Phys.* **2018**, *149*, 234905.
5. Dannenberg, F.; Dunn, K. E.; Bath, J.; Kwiatkowska, M.; Turberfield, A. J.; Ouldridge, T. E. Modelling DNA Origami Self-Assembly at the Domain Level. *J. Chem. Phys.* **2015**, *143*, 165102.
6. Bath, J.; Turberfield, A. J. DNA Nanomachines. *Nat. Nanotechnol.* **2007**, *2*, 275–284.
7. Ezziane, Z. DNA Computing: Applications and Challenges. *Nanotechnology* **2005**, *17*, R27–R39.
8. Genot, A. J.; Bath, J.; Turberfield, A. J. Reversible Logic Circuits Made of DNA. *J. Am. Chem. Soc.* **2011**, *133*, 20080–20083.
9. Orbach, R.; Lilienthal, S.; Klein, M.; Levine, R. D.; Remacle, F.; Willner, I. Ternary DNA Computing Using 3×3 Multiplication Matrices. *Chem. Sci.* **2015**, *6*, 1288–1292.

10. Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; LeProust, E. M.; Sipos, B.; Birney, E. Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized DNA. *Nature* **2013**, *494*, 77–80.

11. Adleman, L. M. Molecular Computation of Solutions to Combinatorial Problems. *Science* **1994**, *266*, 1021–1024.

12. Faulhammer, D.; Cukras, A. R.; Lipton, R. J.; Landweber, L. F. Molecular Computation: RNA Solutions to Chess Problems. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 1385–1389.

13. Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, I., Thomas; Dixit, S. et al. A Systematic Molecular Dynamics Study of Nearest-Neighbor Effects on Base Pair and Base Pair Step Conformations and Fluctuations in B-DNA. *Nucleic Acids Res.* **2010**, *38*, 299–313.

14. Olson, W. K.; Bansal, M.; Burley, S. K.; Dickerson, R. E.; Gerstein, M.; Harvey, S. C.; Heinemann, U. et al. A Standard Reference Frame for the Description of Nucleic Acid Base-Pair Geometry. *J. Mol. Biol.* **2001**, *313*, 229–237.

15. Roberts, R. J.; Cheng, X. Base Flipping. *Annu. Rev. Biochem.* **1998**, *67*, 181–198.

16. Altan-Bonnet, G.; Libchaber, A.; Krichevsky, O. Bubble Dynamics in Double-Stranded DNA. *Phys. Rev. Lett.* **2003**, *90*, 138101.

17. Dai, Q.; Sanstead, P. J.; Peng, C. S.; Han, D.; He, C.; Tokmakoff, A. Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability. *ACS Chem. Biol.* **2016**, *11*, 470–477.

18. von Hippel, P. H.; Johnson, N. P.; Marcus, A. H. Fifty Years of DNA “Breathing”: Reflections on Old and New Approaches. *Biopolymers* **2013**, *99*, 923–954.

19. Ambjörnsson, T.; Banik, S. K.; Krichevsky, O.; Metzler, R. Breathing Dynamics in Heteropolymer DNA. *Biophys. J.* **2007**, *92*, 2674–2684.

20. Blagoev, K.; Alexandrov, B.; Goodwin, E.; Bishop, A. Ultra-Violet Light Induced Changes in DNA Dynamics May Enhance TT-Dimer Recognition. *DNA Repair* **2006**, *5*, 863–867.

21. Kalosakas, G.; Rasmussen, K. Ø.; Bishop, A. R.; Choi, C. H.; Usheva, A. Sequence-Specific Thermal Fluctuations Identify Start Sites for DNA Transcription. *Europhys. Lett.* **2004**, *68*, 127–133.

22. Alexandrov, B. S.; Gelev, V.; Yoo, S. W.; Alexandrov, L. B.; Fukuyo, Y.; Bishop, A. R.; Rasmussen, K. Ø. et al. DNA Dynamics Play a Role as a Basal Transcription Factor in the Positioning and Regulation of Gene Transcription Initiation. *Nucleic Acids Res.* **2010**, *38*, 1790–1795.

23. Melnykov, A. V.; Nayak, R. K.; Hall, K. B.; Van Orden, A. Effect of Loop Composition on the Stability and Folding Kinetics of RNA Hairpins with Large Loops. *Biochemistry* **2015**, *54*, 1886–1896.

24. Ma, H.; Proctor, D. J.; Kierzek, E.; Kierzek, R.; Bevilacqua, P. C.; Gruebele, M. Exploring the Energy Landscape of a Small RNA Hairpin. *J. Am. Chem. Soc.* **2006**, *128*, 1523–1530.

25. Sorin, E. J.; Engelhardt, M. A.; Herschlag, D.; Pande, V. S. RNA Simulations: Probing Hairpin Unfolding and the Dynamics of a GNRA Tetraloop. *J. Mol. Biol.* **2002**, *317*, 493–506.

26. Bonnet, G.; Krichevsky, O.; Libchaber, A. Kinetics of Conformational Fluctuations in DNA Hairpin-Loops. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 8602–8606.

27. Ansari, A.; Kuznetsov, S. V.; Shen, Y. Configurational Diffusion Down a Folding Funnel Describes the Dynamics of DNA Hairpins. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 7771–7776.

28. Ma, H.; Wan, C.; Wu, A.; Zewail, A. H. DNA Folding and Melting Observed in Real Time Redefine the Energy Landscape. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 712–716.

29. Rapti, Z.; Smerzi, A.; Rasmussen, K. Ø.; Bishop, A. R.; Choi, C. H.; Usheva, A. Lengthscales and Cooperativity in DNA Bubble Formation. *Europhys. Lett.* **2006**, *74*, 540–546.

30. Dasanna, A. K.; Destainville, N.; Palmeri, J.; Manghi, M. Strand Diffusion-Limited Closure of Denaturation Bubbles in DNA. *Europhys. Lett.* **2012**, *98*, 38002.

31. Hinckley, D. M.; Lequieu, J. P.; de Pablo, J. J. Coarse-Grained Modeling of DNA Oligomer Hybridization: Length, Sequence, and Salt Effects. *J. Chem. Phys.* **2014**, *141*, 035102.

32. Sikorav, J.-L.; Orland, H.; Braslau, A. Mechanism of Thermal Renaturation and Hybridization of Nucleic Acids: Kramers' Process and Universality in Watson−Crick Base Pairing. *J. Phys. Chem. B* **2009**, *113*, 3715–3725.

33. Dupuis, N. F.; Holmstrom, E. D.; Nesbitt, D. J. Single-Molecule Kinetics Reveal Cation-Promoted DNA Duplex Formation Through Ordering of Single-Stranded Helices. *Biophys. J.* **2013**, *105*, 756–766.

34. Wetmur, J. G.; Davidson, N. Kinetics of Renaturation of DNA. *J. Mol. Biol.* **1968**, *31*, 349–370.

35. Pörschke, D.; Eigen, M. Co-operative Non-enzymatic Base Recognition III. Kinetics of the Helix-Coil Transition of the Oligoribouridylic • Oligoriboadenylic Acid System and of Oligoriboadenylic Acid Alone at Acidic pH. *J. Mol. Biol.* **1971**, *62*, 361–381.

36. Atkins, P. W.; De Paula, J. *Atkins' physical chemistry*; W. H. Freeman and Company: New York, 2006.

37. Craig, M. E.; Crothers, D. M.; Doty, P. Relaxation Kinetics of Dimer Formation by Self Complementary Oligonucleotides. *J. Mol. Biol.* **1971**, *62*, 383–401.

38. Ouldridge, T. E.; Šulc, P.; Romano, F.; Doye, J. P. K.; Louis, A. A. DNA Hybridization Kinetics: Zippering, Internal Displacement and Sequence Dependence. *Nucleic Acids Res.* **2013**, *41*, 8886–8895.

39. Menssen, R. J.; Tokmakoff, A. Length-Dependent Melting Kinetics of Short DNA Oligonucleotides Using Temperature-Jump IR Spectroscopy. *J. Phys. Chem. B* **2019**, *123*, 756–767.

40. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved Through Infrared Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138*, 11792–11801.

41. Chen, C.; Wang, W.; Wang, Z.; Wei, F.; Zhao, X. S. Influence of Secondary Structure on Kinetics and Reaction Mechanism of DNA Hybridization. *Nucleic Acids Res.* **2007**, *35*, 2875–2884.

42. Pörschke, D.; Uhlenbeck, O. C.; Martin, F. H. Thermodynamics and Kinetics of the Helix-Coil Transition of Oligomers Containing GC Base Pairs. *Biopolymers* **1973**, *12*, 1313–1335.

43. Rauzan, B.; McMichael, E.; Cave, R.; Sevcik, L. R.; Ostrosky, K.; Whitman, E.; Stegemann, R. et al. Kinetics and Thermodynamics of DNA, RNA, and Hybrid Duplex Formation. *Biochemistry* **2013**, *52*, 765–772.

44. Applequist, J.; Damle, V. Theory of the Effects of Concentration and Chain Length on Helix—Coil Equilibria in Two-Stranded Nucleic Acids. *J. Chem. Phys.* **1963**, *39*, 2719–2721.

45. Gibbs, J. H.; DiMarzio, E. A. Statistical Mechanics of Helix-Coil Transitions in Biological Macromolecules. *J. Chem. Phys.* **1959**, *30*, 271–282.

46. Applequist, J.; Damle, V. Thermodynamics of the Helix-Coil Equilibrium in Oligoadenylic Acid from Hypochromicity Studies. *J. Am. Chem. Soc.* **1965**, *87*, 1450–1458.

47. Eigen, M.; Pörschke, D. Co-operative Non-enzymic Base Recognition: I. Thermodynamics of the Helix-Coil Transition of Oligoriboadenylic Acids at Acidic pH. *J. Mol. Biol.* **1970**, *53*, 123–141.

48. Zimm, B. H. Theory of "Melting" of the Helical Form in Double Chains of the DNA Type. *J. Chem. Phys.* **1960**, *33*, 1349–1356.

49. Peng, C. S.; Jones, K. C.; Tokmakoff, A. Anharmonic Vibrational Modes of Nucleic Acid Bases Revealed by 2D IR Spectroscopy. *J. Am. Chem. Soc.* **2011**, *133*, 15650–15660.

50. Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A. Transient Two-Dimensional IR Spectrometer for Probing Nanosecond Temperature-Jump Kinetics. *Rev. Sci. Instrum.* **2007**, *78*, 063101.

51. SantaLucia, J. A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 1460–1465.

52. SantaLucia Jr, J.; Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 415–440.

53. Sugimoto, N.; Nakano, S.-i.; Yoneyama, M.; Honda, K.-i. Improved Thermodynamic Parameters and Helix Initiation Factor to Predict Stability of DNA Duplexes. *Nucleic Acids Res.* **1996**, *24*, 4501–4505.

54. Ohmichi, T.; Nakamuta, H.; Yasuda, K.; Sugimoto, N. Kinetic Property of Bulged Helix Formation: Analysis of Kinetic Behavior Using Nearest-Neighbor Parameters. *J. Am. Chem. Soc.* **2000**, *122*, 11286–11294.

55. Sanstead, P. J.; Tokmakoff, A. A Lattice Model for the Interpretation of Oligonucleotide Hybridization Experiments. *J. Chem. Phys.* **2019**, *150*, 185104.

56. Owczarzy, R.; Moreira, B. G.; You, Y.; Behlke, M. A.; Walder, J. A. Predicting Stability of DNA Duplexes in Solutions Containing Magnesium and Monovalent Cations. *Biochemistry* **2008**, *47*, 5336–5353.

57. Owczarzy, R.; Vallone, P. M.; Goldstein, R. F.; Benight, A. S. Studies of DNA Dumbbells VII: Evaluation of the Next-Nearest-Neighbor Sequence-Dependent Interactions in Duplex DNA. *Biopolymers* **1999**, *52*, 29–56.

58. Everaers, R.; Kumar, S.; Simm, C. Unified Description of Poly-and Oligonucleotide DNA Melting: Nearest-Neighbor, Poland-Sheraga, and Lattice Models. *Phys. Rev. E* **2007**, *75*, 041918.

59. Sambriski, E. J.; Schwartz, D. C.; de Pablo, J. J. Uncovering Pathways in DNA Oligonucleotide Hybridization via Transition State Analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 18125–18130.

60. Hoefert, M. J.; Sambriski, E. J.; de Pablo, J. J. Molecular Pathways in DNA-DNA Hybridization of Surface-Bound Oligonucleotides. *Soft Matter* **2011**, *7*, 560–566.

61. Sambriski, E. J.; Schwartz, D. C.; de Pablo, J. J. A Mesoscale Model of DNA and Its Renaturation. *Biophys. J.* **2009**, *96*, 1675–1690.

62. Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; de Pablo, J. J. An Experimentally-Informed Coarse-Grained 3-Site-Per-Nucleotide Model of DNA: Structure, Thermodynamics, and Dynamics of Hybridization. *J. Chem. Phys.* **2013**, *139*, 144903.

63. Šulc, P.; Romano, F.; Ouldridge, T. E.; Rovigatti, L.; Doye, J. P. K.; Louis, A. A. Sequence-Dependent Thermodynamics of a Coarse-Grained DNA Model. *J. Chem. Phys.* **2012**, *137*, 135101.

64. Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of *ab Initio* Protein Folding for a Millisecond Folder NTL9(1-39). *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.

65. Noé, F.; Fischer, S. Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struc. Biol.* **2008**, *18*, 154–162.

66. Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011–19016.

67. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

68. Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109.

69. Przytycka, T. Significance of Conformational Biases in Monte Carlo Simulations of Protein Folding: Lessons From Metropolis–Hastings Approach. *Proteins* **2004**, *57*, 338–344.

70. Gillespie, D. T. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comput. Phys.* **1976**, *22*, 403–434.

71. Gillespie, D. T. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **1977**, *81*, 2340–2361.

72. Banik, S. K.; Ambjörnsson, T.; Metzler, R. Stochastic Approach to DNA Breathing Dynamics. *Europhys. Lett.* **2005**, *71*, 852–858.

73. Schaeffer, J. M.; Thachuk, C.; Winfree, E. Stochastic Simulation of the Kinetics of Multiple Interacting Nucleic Acid Strands. In *DNA Computing and Molecular Programming*, Boston and Cambridge, MA, August 17-21, 2015; Phillips, A., Yin, P., Eds.; Springer: Cham, 2015; pp 194–211.

74. Zolaktaf, S.; Dannenberg, F.; Winfree, E.; Bouchard-Côté, A.; Schmidt, M.; Condon, A. Efficient Parameter Estimation for DNA Kinetics Modeled as Continuous-Time Markov Chains. In *DNA Computing and Molecular Programming*, Seattle, WA, August 5-9, 2019; Thachuk, C., Liu, Y., Eds.; Springer: Cham, 2019; pp 80–99.

75. Weinan, E.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503.

76. Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Illustration of Transition Path Theory on a Collection of Simple Examples. *J. Chem. Phys.* **2006**, *125*, 084110.

77. Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7*, 1192–1219.

78. Weinan, E.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Ann. Rev. Phys. Chem.* **2010**, *61*, 391–420.

CHAPTER 2

THEORY AND FORMALISM OF IR

2.1 Introduction

In this chapter the basics necessary for understanding the physical origin of the signals measured with nonlinear IR techniques are discussed along with how these signals are processed into the final spectrum. 2DIR spectroscopy is a technique within the broader family of ultrafast spectroscopies. Numerous references exist for ultrafast spectroscopy in general[1–4] and 2DIR[5–8] itself that comprehensively discuss the theory and formalism of the technique. As a result this discussion will be limited to what is necessary to understand the experiment as it is utilized here and the spectra that are analyzed. Readers interested in learning more about ultrafast spectroscopy or 2DIR are encouraged to seek out the previously mentioned resources for more information.

A 2DIR spectrum is a two-dimensional frequency correlation plot that probes the vibrational modes of the sample providing information about its structure and dynamics on very fast time scales. 2DIR uses a series of femtosecond laser pulses to interrogate the vibrational modes of the sample. In the case of biological samples these vibrational modes can be connected to detailed structural information that identifies configurational changes or isolates a particular structure or biomolecule of interest. The existence of cross peaks allows the direct observation of interactions between different vibrational modes providing information on the movement of energy through the system or the presence of couplings between vibrational modes. The time resolution of the ultrafast experiments utilized here makes it possible to resolve this structural information on the timescales at which the biomolecular reactions studied here occur making it ideal for the study of the kinetics and dynamics of DNA systems.

In this section we will first detail the origin of the signal measured in 2DIR experiments and the related nonlinear experiments utilized in this work. This will involve a brief discus-

sion of the changes that occur within the system, as a result of the interactions with the laser pulses, that give rise to this signal. Afterwards, the relationship between the peaks in the spectrum and the vibrational modes of the system will be explained. Finally, the relationship between 2DIR and other nonlinear ultrafast measurements will be discussed to understand the different experiments that were conducted and the how information is portrayed in each one.

2.2 Third-Order Nonlinear Polarization and Response Function

To understand the origins of 2DIR we start with simple linear absorption. Semi-classically, absorption can be thought of as a loss of intensity as a result of an electric field emitted by the sample that is out-of-phase with the electric field of the transmitted light. The electric field emitted by the sample is radiated by the macroscopic polarization induced in the sample by the electromagnetic field of the incoming light. The macroscopic polarization **P** is given by the expectation value of the dipole operator $\hat{\mu}$

$$\mathbf{P}(t) = \langle \hat{\mu}\rho(t) \rangle \tag{2.1}$$

where ρ is the density matrix for the system.

For 2DIR we are interested in the third-order polarization induced in the system as a

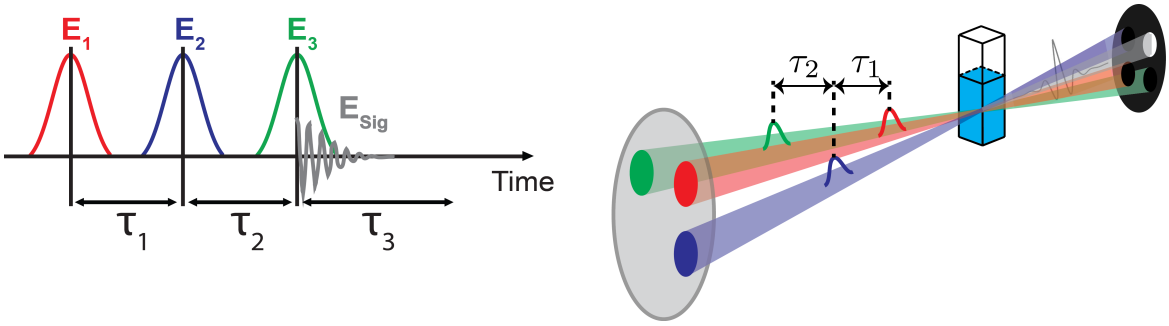

Figure 2.1: Pulse sequence and orientation for a three pulse 2DIR experiment utilizing the Boxcar geometry.

result of interactions with a set of ultrafast infrared pulses. The most basic form of 2DIR spectroscopy involves three laser pulses that are separated by time intervals τ_1 and τ_2. Our focus is on the polarization at some time t during the detection time, τ_3. The pulse sequence is shown in Figure 2.1. Using Equation 2.1 in conjunction with a perturbative expansion of the density matrix that treats the interaction of an electric field with the system as a small time-dependent perturbation to the system Hamiltonian, the equation for the third-order polarization $\mathbf{P}^{(3)}(t)$ is found to be[2,5]

$$\mathbf{P}^{(3)}(t) = \int_0^\infty \int_0^\infty \int_0^\infty \mathbf{R}^{(3)}(\tau_3, \tau_2, \tau_1)\mathbf{E}_3(t-\tau_3)\mathbf{E}_2(t-\tau_3-\tau_2)\times \\ \mathbf{E}_3(t-\tau_3-\tau_2-\tau_1)d\tau_3 d\tau_2 d\tau_1 \quad (2.2)$$

where $\mathbf{E}_\mathrm{N}$ are the incoming electric fields of the three femtosecond IR laser pulses and $\mathbf{R}^{(3)}$ is the third-order response function. In our treatment here we will assume that our femtosecond pulses can be treated as delta functions with impulsive interactions. The third-order response function generates the electric field emitted by the nonlinear polarization and is ultimately what 2DIR aims to measure. The third-order response can be expressed as[2,5]

$$\mathbf{R}^{(3)}(\tau_3, \tau_2, \tau_1) = \left(\frac{-i}{\hbar}\right)^3 \Theta(\tau_3)\Theta(\tau_2)\Theta(\tau_1)\times \\ \langle[[[\hat{\mu}(\tau_3+\tau_2+\tau_1), \hat{\mu}(\tau_2+\tau_1)], \hat{\mu}(\tau_1)], \hat{\mu}(0)]\rho_0\rangle \quad (2.3)$$

where ρ_0 is the equilibrium density matrix and Θ is the Heaviside function, the purpose of which is to enforce the time ordering of the pulses and ensure that the third-order response only occurs after all three electric fields have interacted with the sample.

Now we need a way to visualize how the system evolves as a result of the light-matter interactions that occur during the experiment. This is often done through the use of Feynman diagrams which visually illustrate the light-matter interaction pathways that can occur,

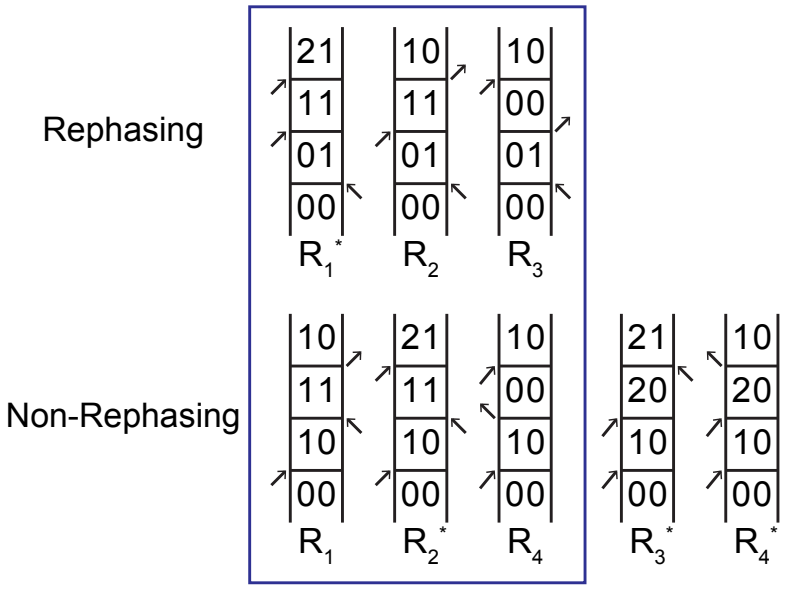

Figure 2.2: The eight Liouville pathways that contribute to the third-order response function. The rephasing and non-rephasing pathways that are experimentally measured are in the blue box while the two quantum coherence non-rephasing pathways are outside the box.

which are known as Liouville pathways. The eight Feynman diagrams that correspond to the possible pathways for the response function given in Equation 2.3 are shown in Figure 2.2. Time runs from the bottom to the top on these diagrams. Horizontal lines indicate a light-matter interaction and the boxes created by the lines show how the density matrix evolves during the time period between interactions. The numbers inside the box are the density matrix elements for that time period. Arrows that point away from the diagram indicate emission while arrows pointing towards the diagram indicate absorption. Of these eight pathways four are unique and the other four are their complex conjugates, denoted by the star in Figure 2.2. The third-order response function can be written as the sum of these pairs

$$\mathbf{R}^{(3)}(\tau_3, \tau_2, \tau_1) = \sum_{i=1}^{4} \mathbf{R}_i - \mathbf{R}_i^* \tag{2.4}$$

demonstrating that it contains all of the information on the evolution of the system through each of these pathways. Under the assumption that our femtosecond pulses can be

treated as delta pulses, the nonlinear polarization is directly proportional to the response function. Presuming that the electric field that is emitted due to the macroscopic polarization is detected in the time domain, the 2DIR spectrum is in practice obtained by taking real part of a two-dimensional Fourier transform of the response function

$$S_{\text{2D}}(\omega_3, \tau_2, \omega_1) = \mathbb{R}\left(\int_0^\infty \int_0^\infty \mathbf{R}(\tau_3, \tau_2, \tau_1) e^{i\omega_1\tau_1} e^{i\omega_3\tau_3} d\tau_1 d\tau_3\right) \tag{2.5}$$

where ω_1 and ω_3 correspond to Fourier transform pairs of τ_1 and τ_3. An important note on notation here is that S_{2D} refers to the real part of the 2D surface while $\tilde{S}_{\text{2D}}$ refers to the complex 2D surface.

2.3 Time Ordering of Interactions and Phase Matching

The eight Louiville pathways can be broken down into two categories, rephasing pathways and non-rephasing pathways demonstrated in Figure 2.2. Pathways are considered to be rephasing when the phase evolutions during τ_1 and τ_3 carry opposite sign whereas for non-rephasing pathways the phase evolutions during τ_1 and τ_3 carry the same sign. These phase evolutions are related to the sign of the wavevector associated with the incoming field. Arrows pointing to the right indicate a positive wavevector while arrows pointing to the left indicate a negative wavevector. Within the non-rephasing group there are two pathways, $\mathbf{R}_3^*$ and $\mathbf{R}_4^*$, that are known as double quantum coherence pathways since they reach a doubly excited state after the first two interactions, and they are the only pathways that do so. The 2DIR spectra collected as described by Equation 2.5 is referred to as the 2DIR correlation spectrum. This is the most useful spectra to acquire since combining the rephasing and non-rephasing spectra produces a spectrum with a purely absorptive line shape. Acquiring this requires collecting both the rephasing and non-rephasing pathways which brings up the concept of phase matching.

Phase matching is a common method for selecting specific pathways based on the

phase and time ordering of the incoming pulses.[5] Each of the incoming electric fields $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ has a corresponding wavevector $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$. Utilizing the boxcar geometry shown in Figure 2.1 the goal is to arrange the pulses in time and space in such a way that the signals from the rephasing and non-rephasing pathways can both be obtained along the same wavevector. This allows both to be collected without needing to realign the spectrometer or use a second detector. Looking at the orientation of the pulses in Figure 2.1, and knowing that $\mathbf{k}_1$ and $\mathbf{k}_3$ must carry opposite phases to obtain the rephasing pathway, we can see that the phase matching condition $-\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$ will generate the rephasing signal along the desired vector shown in Figure 2.1. To collect the non-rephasing signal we could keep the same time ordering and use the phase matching condition $\mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$, however that would result in the signal being generated in a different direction. Instead we alter the time ordering of the pulses and use the phase matching condition $\mathbf{k}_2 - \mathbf{k}_1 + \mathbf{k}_3$ which has both the proper phases for acquiring the non-rephasing signal and emits the signal in the same direction as the rephasing signal. The two quantum coherence pathways have a different phase matching condition and are not acquired during our experiments. The result of this is that the rephasing and non-rephasing spectra can be collected separately by changing the time ordering of the first two pulses. The rephasing and non-rephasing signals can then be added together to obtain the third-order response function.

2.4 Model Six Level System

It is now useful to look at a model system to identify how the different Liouville pathways shown in Figure 2.2 correspond to the experimental spectra. This provides the most tangible way to understand the useful information that is provided by a 2DIR experiment. The model system described here is shown in Figure 2.3c and contains six vibrational energy levels consisting of fundamentals, overtones, and a combination band. The corresponding cartoon 2DIR plot is shown in Figure 2.3b. In the 2DIR spectrum ω_1, presented

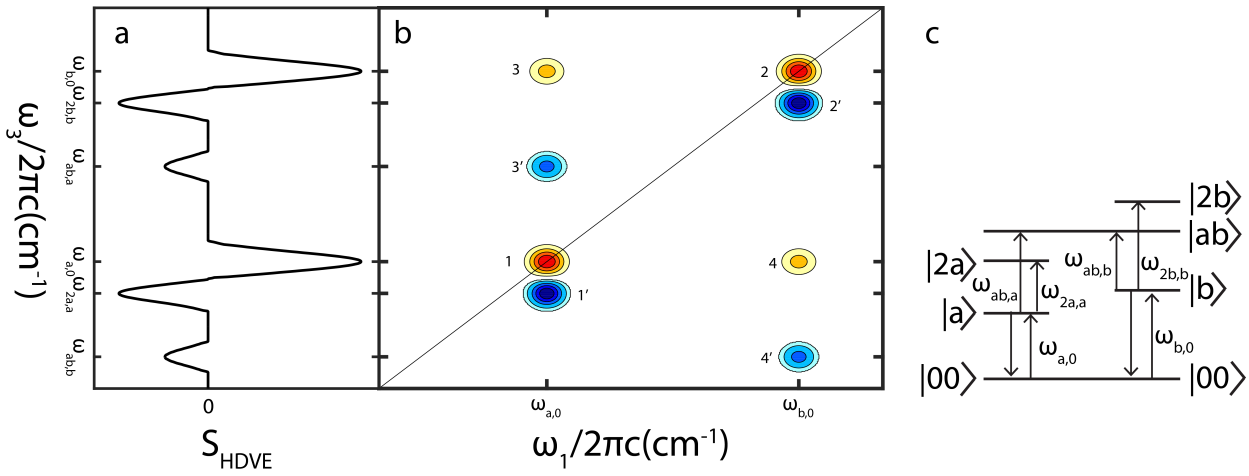

Figure 2.3: Cartoon HDVE (a) and 2DIR spectrum (b) for the model six level system described by the ladder diagram (c).

on the horizontal axis, is the excitation frequency and ω_3, on the vertical axis, is the detection frequency. The ω_1 frequency corresponds to the energy gap of vibrational modes excited by the first electric field. The ω_3 frequency corresponds to the difference in energy between the two states that are occupied after the interaction with the third electric field. Looking at the peaks the first distinction is that there are both positive (red) and negative (blue) peaks. The positive and negative peaks come in pairs as seen in Figure 2.3b, the reason for this will become clear when we assign individual pathways to each peak and see which transitions they contain. The sign corresponds to the transition that is probed by the third pulse. The positive peaks result from two types of pathways known as ground state bleach and stimulated emission. In both cases the states that are occupied after the interaction with the third electric field are the ground state and a singly excited state. The negative peaks result from excited state absorption pathways where the system is in a coherence between a singly excited state and a doubly excited state after the interaction with the third field. The opposite signs are due to the fact that the emitted signal from an excited state absorption pathway has the opposite phase relative to the signal emitted from a ground state bleach or stimulated emission.

Now that we understand the signs of the different peaks we will separate the peaks

according to whether they lie along the diagonal (Peaks 1, 1', 2, and 2') or are off-diagonal peaks (Peaks 3, 3', 4, and 4') which are also known as cross peaks. When examining these peaks in the diagram the axes are labeled according to the frequency of the transition that is probed by the first interaction and the third interaction. In the case of positively signed diagonal peaks the same vibrational mode is probed by the first and third pulses. Looking at the 2DIR and ladder diagram in Figure 2.3 and using peak 1 as an example, both the first and third pulses are probing the 0→a transition, which also explains why the frequency is the same along both axes. Peak 1' is the excited state absorption for this vibrational mode so the first pulse probes the 0→a transition and the third pulse probes the a→2a transition. The reason that the 1' peak is slightly off the diagonal is due to the anharmonicity of the system resulting in the a→2a transition having a slightly lower frequency then the 0→a transition. The difference in frequency between the two peaks provides information about the anharmonicity of the system. The fact that both peak 1 and 1' probe the same vibrational mode with the first pulse demonstrates why the positive and negative peaks exist as a pair.

Cross peaks arise from multiple vibrational modes interacting with the incoming electric fields. These peaks manifest from anharmonic coupling between different vibrational modes, which can arise from a number of interesting physical processes. To give a few examples for the case of DNA, this can occur between modes within a single base such as coupling between the ring mode and carbonyl modes of guanine. They can also occur as a result of coupling between hydrogen bonded bases pairs such as coupling between guanine vibrational modes and cytosine vibrational modes. This latter example is particularly useful as this coupling only occurs when the base pairs are hydrogen bonded meaning this cross peak disappears upon the loss of base pairing. Looking at peaks 3 and 3' as examples, peak 3 arises from the first pulse probing the 0→a transition and the third pulse probing the 0→b transition. This corresponds to the bleaching of the ground state that is shared between the coupled oscillators. For peak 3' the first pulse again probes the 0→a

transition while the third pulse probes the a→ab transition, which is an excitation into a combination band.

2.5 Alternative Nonlinear IR Measurements

There are a few other nonlinear IR measurements that are related to 2DIR, but are all one-dimensional representations of the signal. These are heterodyne dispersed vibrational echo (HDVE), dispersed vibrational echo (DVE), and dispersed pump-probe (DPP). Since these are one-dimensional representations the excitation axis is not resolved making them much faster to acquire since τ_1 is fixed, though ω_1 resolution is of course lost. Starting with the complex 2D surface $\tilde{S}_{\mathsf{2D}}(\omega_3, \omega_1)$ the alternative spectra are given by

$$\begin{aligned} \tilde{S}_{\mathsf{HDVE}}(\omega_3) &= \int_0^\infty \tilde{S}_{\mathsf{2D}}(\omega_3, \omega_1)\, d\omega_1 = \tilde{S}_{\mathsf{2D}}(\omega_3, \tau_1 = 0) \\ S_{\mathsf{DVE}}(\omega_3) &= \left| \int_0^\infty \tilde{S}_{\mathsf{2D}}(\omega_3, \omega_1)\, d\omega_1 \right|^2 = \left| \tilde{S}_{\mathsf{HDVE}}(\omega_3) \right|^2 \\ S_{\mathsf{DPP}}(\omega_3) &= \mathbb{R}\left(\int_0^\infty \tilde{S}_{\mathsf{2D}}(\omega_3, \omega_1)\, d\omega_1 \right) = \mathbb{R}\left(\tilde{S}_{\mathsf{HDVE}}(\omega_3) \right) \end{aligned} \tag{2.6}$$

One useful note is that the DPP is simply the real part of the HDVE. Later on when discussing HDVE this is important to recall as the spectra shown will be the DPP, though we commonly refer to them as HDVE since that is how they are acquired. Another point to note is that the HDVE is a projection of the 2D signal onto the ω_3 axis and collected by fixing τ_1 = 0 fs, by the projection slice theorem. Fixing τ_1 at zero also results in the rephasing and non-rephasing pathways being emitted along the same wavevector[5] allowing both to be acquired simultaneously.

Since the HDVE is the most heavily used in this thesis it is worth revisiting the model six level system shown earlier to see what the corresponding HDVE spectrum would look like. Figure 2.3a shows the HDVE figure determined from the 2D spectrum according to Equation 2.6. In this case we can see many features are still present, despite the

projection. We can still see each of the peaks individually, the only difference being that the positive peaks on the diagonal are combined with the positive cross peaks resulting in only two positive peaks with greater intensity. While in this case there is no significant information lost due to integrating over the ω_1 axis, for spectra of real samples this is usually not the case as there are some peaks that overlap along the ω_3 axis.

One final note on the HDVE, while the details of the experimental methods will be detailed in a later chapter there is one particular aspect that is worth discussing here with regards to the nature of the signal. The HDVE signal is acquired in the frequency domain as the result of the signal being dispersed by a grating onto the detector. The resulting signal measured at each frequency contains only the real part of the HDVE and an additional method is required to obtain the complex HDVE. The method utilized here is Fourier transform spectral interferometry (FTSI). The method has been outlined in detail elsewhere[9,10] and only a brief explanation is provided here. This method utilizes a Kramers-Kronig relation that relates the real and imaginary parts of a complex function which allows the real component to be calculated from the imaginary component and vice versa. Using FTSI the real valued frequency signal obtained by the experiment is inverse Fourier transformed into the time domain resulting in a complex time domain signal that has both positive and negative time components. A Heaviside function is then used to select only the positive time signal which is then Fourier transformed back into the frequency domain resulting in the complex HDVE signal.

2.6 Acknowledgements

I would like to thank Memo Carpenter for his careful reading and thoughtful comments on this chapter.

2.7 References

1. Jonas, D. M. Two-Dimensional Femtosecond Spectroscopy. *Ann. Rev. Phys. Chem.* **2003**, *54*, 425–463.

2. Mukamel, S. *Principles of Nonlinear Optical Spectroscopy*; Oxford University Press: New York, 1995.

3. Cho, M. *Two-Dimensional Optical Spectroscopy*; CRC Press: Boca Raton, 2009.

4. Cho, M. Coherent Two-Dimensional Optical Spectroscopy. *Chem. Rev.* **2008**, *108*, 1331–1418.

5. Hamm, P.; Zanni, M. *Concepts and Methods of 2D Infrared Spectroscopy*; Cambridge University Press: New York, 2011.

6. Khalil, M.; Demirdöven, N.; Tokmakoff, A. Coherent 2D IR Spectroscopy: Molecular Structure and Dynamics in Solution. *J. Phys. Chem. A* **2003**, *107*, 5258–5279.

7. Golonzka, O.; Khalil, M.; Demirdöven, N.; Tokmakoff, A. Coupling and Orientation Between Anharmonic Vibrations Characterized with Two-Dimensional Infrared Vibrational Echo Spectroscopy. *J. Chem. Phys.* **2001**, *115*, 10814–10828.

8. Khalil, M.; Demirdöven, N.; Tokmakoff, A. Obtaining Absorptive Line Shapes in Two-Dimensional Infrared Vibrational Correlation Spectra. *Phys. Rev. Lett.* **2003**, *90*, 047401.

9. Jones, K. C.; Ganim, Z.; Tokmakoff, A. Heterodyne-Detected Dispersed Vibrational Echo Spectroscopy. *J. Phys. Chem. A* **2009**, *113*, 14060–14066.

10. Sanstead, P. J. C. Investigation of DNA Dehybridization through Steady-State and Transient Temperature-Jump Nonlinear Infrared Spectroscopy. Ph.D. thesis, The University of Chicago, 2018.

CHAPTER 3

THEORY AND FORMALISM OF MARKOV PROCESSES AND THE GILLESPIE ALGORITHM

3.1 Introduction

One of the primary goals of this research was to develop a conceptually and computationally accessible model to compliment experimental studies of DNA dynamics and kinetics. It has been well documented that DNA hybridization and dehybridization does not follow a single well defined pathway but rather has a distribution of available pathways through which the process can occur. The process of DNA hybridization and dehybridization is dictated by diffusive motion and random forces that guide the motions and interactions of the individual strands. To model our experimental data we will mathematically represent the association and dissociation of DNA as a stochastic process. While stochastic processes are widely used in mathematics, biology, chemistry, physics, and other fields, some of the underlying mathematics and terminology are likely to be unfamiliar to many. In this chapter the basics necessary to understand the origins of the model are outlined, while the construction of the model itself is detailed in later chapters. We begin with the basics of Markov processes which is followed by an explanation of the algorithm that is utilized to generate stochastic trajectories from a Markov model. The remainder of the chapter is dedicating to outlining additional mathematical methods that will be utilized in the analysis of the model.

A stochastic process is an indexed family of random variables, commonly denoted $X(t)$, where the indexes are a set of times over which the random variable evolves. The set of possible values used to index the random variables is known as the index set and the set of possible values for the random variable is known as the state space. Stochastic processes are used to model numerous systems in a variety of fields that evolve randomly over some period of time. Stochastic processes can have numerous outcomes, due to

their random nature, with individual outcomes known as, among other possible names, a sample function, realization, or trajectory. To put this another way, a trajectory of a random process is generated by allowing the system to iterate through the states in the state space recording each state along the way with an index which is usually the time at which the system is in that state. Since stochastic processes indexed by a set of times are both the most common type of stochastic process, and the type used here, our discussion going forward will consider the index set to be a set of times.

A Markov chain, also referred to as a Markov process, is a particular type of stochastic model that follows the Markov property. The Markov property requires that the next state of a process, and the time at which the system will be in that state, depends solely on the current state and time with no dependence on previous steps in the process. A classic example of a Markov process is a random walk on a 2-D lattice where the system has equal probability of moving to each of the four neighboring states regardless of where the system is. The system can be initially placed at a given position and allowed to move throughout the lattice grid resulting in a number of different random trajectories. Additionally, the next state the trajectory moves to only depends on the state the system currently occupies and is completely independent of how the system arrived at that state.

Markov processes, and stochastic models in general, are often broken down into classifications based on two factors: how they progress forward in time and their state space. The time evolution of a model can occur in either discrete-time or continuous-time. A model can also have either a discrete or continuous state space. It is worth noting that a continuous state space is by definition infinite and uncountable while a discrete state space could be finite or infinite.

The model presented here is based on the states generated by a thermodynamic lattice model, previously developed in our group,[1] resulting in a discrete and countable state space. The model is set up as a continuous-time Markov process because this provides the most natural comparison to the kinetics obtained from the temperature-jump experi-

ment. However, in some instances it will be advantageous to analyze the model within the context of discrete time steps. We will now examine the theory of Markov processes and the rules they follow. Later on the theory behind methods for directly analyzing the Markov model without needing to generate trajectories is presented.

3.2 Theory of Markov Processes

Our goal here is to examine the case of a discrete state Markov process in continuous-time to understand what dictates the evolution of a system following the Markov property. We are interested in a random variable $X(t)$, which at any given time is in one of the states in the state space, and how it evolves as a function of time. We start by defining the transition probability for going from step $n-1$ to step n in a Markov process. The transition probability is given by[2]

$$\Pr\left(x_n, t_n | x_{n-1}, t_{n-1}\right) \tag{3.1}$$

which is the probability of being at state x_n at time t_n given that the system was at x_{n-1} at time t_{n-1} where n and $n-1$ denote steps along the trajectory. A Markov process where the transition probabilities do not evolve as a function of time is referred to as stationary. One can show that this implies the transition probability only depends upon the time difference $t_n - t_{n-1}$. A process respecting these two properties is also known as a time-homogeneous Markov chain. This is true for the model discussed in this work and moving forward we will use the notation $p_{ij}(t)$ for the probability of going from state i to state j in time t. The matrix whose elements are $p_{ij}(t)$ is sometimes known as the transition probability matrix or the transition matrix and will be denoted by $\mathbf{P}$,[2] though it is critical to not confuse this with the transition rate matrix that will be introduced later on. The elements $p_{ij}(t)$ must satisfy two conditions

$$p_{ij}\left(t\right) \geq 0 \tag{3.2}$$

$$\sum_{j} p_{ij}(t) = 1 \tag{3.3}$$

Equation 3.2 requires that all probabilities be non-negative and Equation 3.3 enforces that upon transitioning the model must go to one of the existing states. The term non-negative, and the term non-positive that will be used later, is used intentionally because a value of zero is allowed. Note that the case where $i = j$ must also adhere to these rules and is not necessarily zero.

Our next goal is to introduce the transition rate matrix which is an important component of the method that will be used to generate stochastic trajectories in our kinetic model. Since the transition rate matrix is a critical component of the model it is important to build both a mathematical and intuitive understanding of its elements. To do so we start by introducing the Chapman-Kolmogorov equation which when applied to a Markov process with a discrete state space takes the following form[3]

$$p_{ik}(t+\tau) = \sum_{j} p_{ij}(t)\, p_{jk}(\tau) = \sum_{j} p_{ij}(\tau)\, p_{jk}(t) \tag{3.4}$$

which breaks down a transition from i to k into intermediate transitions ij and jk and replicates the probability of going from i to k by summing over all possible intermediates j. Rewriting this using $\tau = \Delta t$ where $t \gg \Delta t$ gives us

$$p_{ik}(t+\Delta t) = \sum_{j} p_{ij}(t)\, p_{jk}(\Delta t) \tag{3.5}$$

$$p_{ik}(t+\Delta t) = \sum_{j} p_{ij}(\Delta t)\, p_{jk}(t) \tag{3.6}$$

which are the forward and backward master equations respectively[3] and can be thought of as either taking a big first step followed by a tiny second step or a tiny first step followed by a big second step. For the purpose of the work presented here we will only utilize the forward master equation, however for the sake of completeness and the fact that future

work could potentially utilize the backward master equation we will continue to include it in this discussion.

We will now approximate $p_{ij}(\Delta t)$ for small Δt as[2]

$$p_{ij}(\Delta t) = \delta_{ij} + l_{ij}\Delta t + \mathcal{O}\left(\Delta t^2\right) \tag{3.7}$$

where δ_{ij} is the Kronecker delta function, $\mathcal{O}\left(\Delta t^2\right)$ describes the error term due to factors on the order of Δt^2 and smaller, l_{ij} are the entries of the transition rate matrix **L** which has units of per time. For the model presented here the units are s^{-1}. Note that Equation 3.7 results in $p_{ij}(0) = \delta_{ij}$ as expected since the probability of a system that starts in state i being found in state i after a time step of zero is one. The entries in the transition rate matrix describe the rate at which a continuous-time Markov chain moves between states, such that l_{ij} describes the rate at which the process transitions from state i to j. The entries of the transition rate matrix can also be thought of as the time derivative of $p_{ij}(t)$ taking the limit as Δt approaches zero, which we will demonstrate, and discuss further, shortly. The transition rate matrix is commonly also referred to as **C** or **Q**, however to avoid confusion with other variables, we have elected to utilize the formalism of Vanden-Eijnden[4], which refers to these elements as l_{ij}.

To ensure that the approximation given by Equation 3.7 is consistent with the rules for p_{ij} given by Equations 3.2 and 3.3, the following conditions for l_{ij} must be true

$$l_{ii} \leq 0 \tag{3.8}$$

$$l_{ij} \geq 0 \quad \forall i \neq j \tag{3.9}$$

$$\sum_j l_{ij} = 0 \tag{3.10}$$

Equation 3.8 dictates that diagonal elements must be non-positive. Equation 3.9, which utilizes the symbol $\forall$ which means "for all", dictates that all off diagonal elements must be

non-negative. Equation 3.10 dictates that each row of the transition rate matrix must sum to zero. Another consequence of these conditions that is important to note is that

$$l_{ii} = -\sum_{j \neq i} l_{ij} \tag{3.11}$$

These diagonal elements are related to the amount of time needed to exit a state in the model, the calculation of which will be described later on in Section 3.3. As mentioned previously the elements of the transition rate matrix can be thought of as the time derivative of $p_{ij}(t)$ in the limit of Δt approaching zero. To demonstrate this we start by taking Equation 3.5 and subbing in the approximation for $p_{ij}(\Delta t)$ from Equation 3.7 which yields

$$p_{ik}\left(t + \Delta t\right) = \sum_{j} p_{ij}\left(t\right)\left(\delta_{jk} + l_{jk}\Delta t + \mathcal{O}\left(\Delta t^2\right)\right) \tag{3.12}$$

We then subtract off $p_{ik}(t)$ from both sides and divide through by Δt. Looking at the first term in the sum and noting that $\sum_j p_{ij}\left(t\right)\delta_{jk} = p_{ik}\left(t\right)$ results in the expression

$$\frac{p_{ik}\left(t + \Delta t\right) - p_{ik}\left(t\right)}{\Delta t} = \sum_{j} p_{ij}\left(t\right) l_{jk} + \mathcal{O}\left(\Delta t\right) \tag{3.13}$$

Taking the limit as Δt approaches zero gives us the derivative at time t

$$\frac{dp_{ik}}{dt}\left(t\right) = \sum_{j} p_{ij}\left(t\right) l_{jk} \tag{3.14}$$

Now that we have the derivative form of the master equation we can clarify our understanding of the transition rate matrix elements. We want to examine what happens to Equation 3.14 in the limit of t approaching zero. We know from Equation 3.7 that $p_{ij}\left(0\right) = \delta_{ij}$ and

applying this to equation 3.14 we get

$$\lim_{t\to 0} \frac{dp_{ik}}{dt}(t) = \sum_j \delta_{ij} l_{jk} = l_{ik} \tag{3.15}$$

There are two cases to consider here, the first being if the system starts in state i, what is the probability that it is still in state i after time t, which is denoted $p_{ii}(t)$. The second case is if the system starts in state i what is the probability that the system is in state j after time t, denoted $p_{ij}(t)$. In the first case, if $i = k$ in Equation 3.15 we note that the probability of being in state i cannot increase with increasing time, which results in a derivative that must be non-positive at time zero. This means that the derivative in Equation 3.15 must also be non-positive in agreement with Equation 3.8. Alternatively, if $i \neq k$, with increasing time the probability of finding the system in state k can only remain the same or increase with time which results in a non-negative derivative in Equation 3.15 in agreement with Equation 3.9. This demonstrates that an alternative way to think about the elements of the transition rate matrix is as the derivatives of the transition probabilities in the limit of t approaching zero. Finally we can rewrite Equation 3.14 into the matrix form of the forward master equation

$$\frac{d\mathbf{P}}{dt} = \mathbf{PL} \tag{3.16}$$

The same steps can be done for the backward master equation resulting in its matrix form

$$\frac{d\mathbf{P}}{dt} = \mathbf{LP} \tag{3.17}$$

3.3 The Gillespie Algorithm

Once the Markov model has been constructed it is necessary to devise a way to generate the stochastic trajectories. This is achieved through the incorporation of Monte Carlo methods. Like Markov models, Monte Carlo methods are often broken down into two

classes, discrete-time and continuous-time. Considering that our ultimate goal is to generate trajectories for the purpose of analyzing biomolecular kinetics and dynamics derived from experiment, we utilize the Gillespie algorithm, a method for generating trajectories from a continuous-time Markov process that was initially proposed by Daniel Gillespie.[5,6]

The Gillespie algorithm utilizes the transition rate matrix **L** to generate stochastic trajectories. In each step of the process the algorithm determines two factors from the transition rate matrix, the next state in the trajectory and how long the system will spend in the current state before moving to the next state, which we will refer to as the exit time. Generating stochastic trajectories requires that the next state and the exit time are generated randomly from a probability distribution function defined by the model. In this section we will demonstrate the origins of the Gillespie algorithm, where the steps in the algorithm come from, and the method by which the exit time and the next state in the trajectory can be determined from random numbers.

The ultimate goal is to derive a function that describes the probability of moving to a particular state j at some time τ. We will first focus on determining the statistical distribution for the exit time, which is the time at which the trajectory will leave a given state, k, for the next state, which provides the time step for each step in the algorithm. This is referred to as the first passage time in some contexts, however since we utilize that term to also describe the time it takes for the system to first reach the fully formed dimer state from the monomer state, or vice versa, we will continue to refer to the time it takes for the process to first leave any given state as the exit time. To determine the exit time for each step in the algorithm from a random number we need to determine the functional form of the probability distribution of possible exit times. To determine the form of the probability distribution we first calculate the probability that if the system is initially in state k it is still in state k at some time τ later. The probability of being in state k at time $\tau + \Delta\tau$ is equal to the probability of being in state k at time τ, $p_k(\tau)$, times the probability of not moving during the time interval $\Delta\tau$, $p_{kk}(\Delta\tau)$ which we get from Equation 3.7. This would then

give us

$$p_k(\tau + \Delta\tau) = p_k(\tau)(1 + l_{kk}\Delta\tau + \mathcal{O}(\Delta t)) \tag{3.18}$$

which can be rewritten as

$$p_k(\tau + \Delta\tau) = p_k(\tau)\left(1 - \sum_{m \neq k} l_{km}\Delta\tau + \mathcal{O}(\Delta t)\right) \tag{3.19}$$

subtracting off $p_k(\tau)$, dividing through by $\Delta\tau$, and taking the limit as $\Delta\tau$ goes to zero yields

$$\frac{d}{d\tau}p_k(\tau) = -\sum_{m \neq k} l_{km}p_k(\tau) \tag{3.20}$$

Solving the differential equation knowing that we are in state k at time zero then gives us

$$p_k(\tau) = e^{-\sum_{m \neq k} l_{km}\tau} \tag{3.21}$$

which means the probability of leaving state k for the first time in the interval $0 \leq \tau \leq T$, or in other words the probability that the exit time is less than T, is

$$\text{Pr}(\text{exit } k \text{ before time } T) = 1 - e^{-\sum_{m \neq k} l_{km}T} \tag{3.22}$$

Note that this is the cumulative distribution function for an exponential distribution.[2] This tells us that the first exit time is exponentially distributed with the parameter being the sum of all rates out of the occupied step.

Now that we have an expression for the exit time from our initial state k we need to determine what state the system moves to. We take a similar approach by looking to determine the form of the distribution that describes the probability of moving to each possible state. However, there are some changes to the process since the state space is discrete rather than continuous as is the case for time. We again start in state k and want to find the probability of going to a specific state j in the time interval $\tau + \Delta\tau$. This

probability can be expressed as

$$p_{kj}\left(\tau+\Delta\tau\right)=p_{k}\left(\tau\right)l_{kj}\Delta\tau \tag{3.23}$$

Where $p_k(\tau)$ is the probability of being in state k at time τ and $l_{kj}\Delta\tau$ is the probability of going to state j in the time $\Delta\tau$. At this point we need to introduce the concept of a probability density function. The probability density function for random variable X evaluated at a specific value x is given by $f_X(x)$ and has the following three properties[2]

$$f_X(x)\geq 0 \qquad \forall x \tag{3.24}$$

$$\int_{-\infty}^{\infty} f_X(x)dx=1 \tag{3.25}$$

$$\mathsf{Pr}\left(X\leq x\right)=\int_{-\infty}^{x} f_X(y)dy \tag{3.26}$$

Equation 3.24 states that the probability density function is non-negative at each point. Equation 3.25 requires that the total area under the curve must sum to unity. Equation 3.26 dictates the probability of X having a value less than or equal to x, which also explains why the integral in Equation 3.25 must equal one since the value of X must lie in the interval $(-\infty,\infty)$.[2] To find $\mathsf{Pr}\left(a<X\leq b\right)$ one must simply utilize a and b as the bounds of the integral given by Equation 3.26. Intuitively for a very small value of dx one can interpret $f_X(x)dx$ as the probability of x being found in the interval $[x,x+dx]$. This intuitive understanding will be demonstrated through deriving the joint probability density function for our system.

As mentioned previously we desire a function that describes the probability of moving to state j at time τ, which can be found using a probability density function. A probability density function with more than one random variable, in this case there are two, j and τ, is known as a joint probability density function and must follow the same rules. Using similar notation a joint probability density function with two random variables will be denoted as

$f_{X,Y}(x, y)$.

Swapping in our variables j and τ the joint probability density function is defined as the function such that[2]

$$\Pr(a < j \leq b, c < \tau \leq d) = \int_a^b \int_c^d f_{J,T}(j, \tau) d\tau dj \tag{3.27}$$

which means that the integral of the probability density function is the probability that j is in the interval $(a, b]$ and that τ is in the interval $(c, d]$. Equation 3.27 is the most general definition which utilizes integrals, but note that in our case j is discrete so the equation is properly written with a summation over j instead of the integral.

Now we will determine the joint probability distribution for going to state j at time τ. Looking at Equation 3.23 the right hand side of that equation is the probability of going to a specific state j in the time window $\tau + \Delta\tau$. Looking at the left hand of the definition of the joint probability density, Equation 3.27, we see that this is equivalent to Equation 3.23 if the bounds on j are the same, because we are only interested in a single potential final state, and the bounds for time are τ and $\tau + \Delta\tau$. Thus we can plug the right hand side of Equation 3.23 in for the left hand side of Equation 3.27 resulting in

$$p_k(\tau) l_{kj} \Delta\tau = \int_\tau^{\tau+\Delta\tau} f_{J,T}(j, \tau) d\tau \tag{3.28}$$

Utilizing the assumption that $\Delta\tau$ is very small and taking the derivative of both sides yields

$$p_k(\tau) l_{kj} = f_{J,T}(j, \tau) \tag{3.29}$$

which is the joint probability density function for j and τ. This also demonstrates the intuitive explanation of the $f_X(x)dx$ described earlier. We can now substitute in our value

for $p_k(\tau)$ given by Equation 3.21 and multiplying through by $\frac{\sum_{m\neq k} l_{km}}{\sum_{m\neq k} l_{km}}$ yields

$$f_{J,T}(j,\tau) = \frac{l_{kj}}{\sum_{m\neq k} l_{km}} \sum_{m\neq k} l_{km} e^{\sum_{m\neq k} l_{km}} \tag{3.30}$$

where the right side is the probability density function for a continuous exponential distribution representing the exit time and the left side is a discrete probability distribution for the probability of going to a given state j. It can also be shown that

$$\sum_{j\neq k} \int_0^\infty f_{J,T}(j,\tau) d\tau = 1 \tag{3.31}$$

where we sum over all $j \neq k$ since when exiting state k the system is not allowed to reenter state k and integrate over all valid times τ.

Now that the joint probability density function has been obtained we can introduce the Gillespie algorithm[5,6] for simulating trajectories for a stochastic model. How the steps in the algorithm are determined from the joint probability density function will be explained after the steps have been defined. The algorithm is laid out in the following steps:

1. Initialize the model by occupying a single state k at $t = 0$ $(p_n(0) = \delta_{nk})$
2. If this is not the first iteration of the algorithm designate the currently occupied state as k
3. Generate two uniform random numbers in the interval $(0, 1)$, denoted R_1 and R_2
4. Calculate the exit time τ according to

$$\tau = \frac{-\ln R_1}{\sum_{m\neq k} l_{km}} \tag{3.32}$$

5. Set j to be the smallest integer that satisfies

$$\frac{1}{\sum_{m\neq k} l_{km}} \sum_{\substack{m=1,\\ m\neq k}}^{j} l_{km} \geq R_2 \tag{3.33}$$

6. Update the model by occupying state j and increasing the time such that $t = t + \tau$

7. Return to step 2 or terminate the simulation if a predetermined end criteria is met

In the context of the model for DNA association and dissociation presented here the exit criteria for an association trajectory is that the fully formed dimer state has been reached, and in the case of a dissociation trajectory it is that the monomer state has been reached. In steps 4 and 5 we are utilizing what is referred to as the "direct method"[5] for determining the next state in the trajectory and the exit time τ, the basis of which will now be described.

To understand the direct method we start with the joint probability density function and note that for independent variables, which we have here,[3] it can be broken down into two single variable probability density functions.[5] Our goal now is to generate random variables τ and j from their respective probability density functions. We will start with the probability density function for determining the exit time τ given by

$$f_T(\tau) = \sum_{m\neq k} l_{km} e^{-\sum_{m\neq k} l_{km}\tau} \tag{3.34}$$

To understand the method for generating τ we must introduce the cumulative distribution function, denoted $F_T(\tau)$, which is the function of random variable T that when evaluated at τ is the probability that T will have a value less than or equal to τ. The cumulative distribution function gives the area under the PDF from 0 to τ, note that the interval $(-\infty, 0)$

is neglected since we cannot have negative time in this context, such that[2]

$$F_T(\tau) = \int_0^\tau f_T(x)dx \tag{3.35}$$

We now look to draw a random number, R, from the interval $(0, 1)$ and determine the corresponding value of τ from the distribution such that

$$\tau = F_T^{-1}(R) \tag{3.36}$$

We note that the inverse of the cumulative distribution function is known as the quantile function. One can easily calculate the cumulative distribution function

$$F_T(\tau) = \int_0^\tau \sum_{m \neq k} l_{km} e^{-\sum_{m \neq k} l_{km} x} dx = 1 - e^{-\sum_{m \neq k} l_{km} \tau} \tag{3.37}$$

We first note that we can swap out $1 - R$ for R because in the case where R is uniformly distributed on $[0, 1]$ then $1 - R$ must also be uniformly distributed on $[0, 1]$. Making this change in addition to substituting in $F_T(\tau) = R$ and solving for τ gives us the inverse

$$\tau = F_T^{-1}(R) = \frac{-\ln R}{\sum_{m \neq k} l_{km}} \tag{3.38}$$

which is the expression for calculating τ from our probability distribution function based on a random number R utilized in step 4 of the algorithm.

The method for determining the integer j that defines the state that the system moves to follows a similar process with slight tweaks to account for the fact that the probability density function

$$f_J(j) = \frac{l_{kj}}{\sum_{m \neq k} l_{km}} \tag{3.39}$$

is discrete. In the discrete case since $f_J(j)$ is normalized, it is clear that our cumulative

distribution function is simply

$$F_J(j) = \sum_{x=0}^{j} f_J(x) \tag{3.40}$$

where $F_J(j_0)$ is the probability that J will be less than or equal to j_0. Since we are in the discrete case we cannot simply find the inverse of F like we could in the continuous case. Instead the method is to draw a random number R and find the value of j that satisfies[5]

$$F_J(j-1) < R \leq F_J(j) \tag{3.41}$$

Substituting in the cumulative distribution function and utilizing the probability density function results in

$$\sum_{x=0}^{j-1} \frac{l_{kx}}{\sum_{m \neq k} l_{km}} < R \leq \sum_{x=0}^{j} \frac{l_{kx}}{\sum_{m \neq k} l_{km}} \tag{3.42}$$

to solve for j only the right side is needed which is the expression for determining j in step 5 of the algorithm.

3.4 Calculating the Steady State Distribution

The steady state distribution, also known as the stationary distribution, is a vector whose entries, once normalized to sum to one, form a probability distribution that does not evolve with time. As a result this can also be thought of as the equilibrium distribution for a Markov process. The method for determining the steady state distribution for a continuous-time Markov process will now be described as it will be utilized later in the analysis of the model.

We start by noting that an irreducible Markov process has a positive steady state distribution if and only if all of its states are positive recurrent.[7] A Markov process is irreducible if every state is accessible from every other state, including the ability to return to the initial state. A state j is accessible from state i if and only if there exists some integer $n > 0$ such that $p_{ij}^n > 0$. It is positive recurrent if the mean recurrence time, the time

it takes for a process in state i to return to state i, is finite. The Markov model for DNA association and dissociation that is presented in this work can be shown to satisfy both of these conditions so it must have a steady state distribution. Thus it is worth taking the time to understand how to calculate it. To understand how to calculate the steady state distribution we start with the left eigenvector equation for the transition rate matrix

$$\mathbf{XL} = \lambda \mathbf{X} \tag{3.43}$$

where **X** is the left eigenvector, which takes the form of a row vector, **L** is the transition rate matrix and λ is the eigenvalue. The steady state distribution can be written as

$$\lim_{t \to 0} p_{ij}(t) = \pi_j \tag{3.44}$$

where π is a vector containing the steady state solution whose elements π_j are the equilibrium probability for each state, which must be normalized such that they sum to one.

In the case of our model that utilizes the transition rate matrix **L** we are looking for the distribution for which the rates in and out of each state sum to zero. Thus, the steady state distribution is the eigenvector corresponding to an eigenvalue of zero. Note that this is different than looking for the steady state distribution to the transition probability matrix where the solution corresponds to an eigenvalue of one. As a result the steady state distribution satisfies the equation

$$\boldsymbol{\pi}\mathbf{L} = 0 \tag{3.45}$$

which can also be written as the sum

$$\sum_j \pi_j l_{jk} = 0 \tag{3.46}$$

To understand why the steady state distribution corresponds to an eigenvalue of zero we look at the forward master equation in the form given by Equation 3.14 and take the limit

of $p_{ij}(t)$ as t goes to infinity, described by Equation 3.44. This gives us

$$\frac{dp_{ik}}{dt}(t) = \sum_j p_{ij}(t)\, l_{jk} = \sum_j \pi_j l_{jk} = 0 \tag{3.47}$$

which shows that the steady state distribution corresponding to the transition rate matrix is the left eigenvector of the transition rate matrix that corresponds to an eigenvalue of zero. The fact that the time derivative of the transition probabilities is zero also shows that this distribution does not evolve.

3.5 Absorbing Markov Chains

Describing the system as an absorbing Markov chain provides a useful framework for analyzing the model without the need to generate individual trajectories. An absorbing Markov chain is any Markov chain that has one or more absorbing states, which are states where the probability of leaving the state is zero. To formulate the model as an absorbing Markov chain the concept of an embedded Markov chain must first be discussed. The embedded Markov chain is a discrete-time Markov chain formed by converting the transition rate matrix of a continuous-time Markov chain into a transition probability matrix which will be denoted by **S**. The elements of **S**, denoted s_{ij} are calculated according to

$$s_{ij} = \begin{cases} 0 & \forall i = j \\ \frac{l_{ij}}{\sum_{i \neq k} l_{ik}} & \forall i \neq j \end{cases} \tag{3.48}$$

where the elements s_{ij} are the conditional probabilities of transitioning into state j given the system is in state i. As expected, since upon leaving state i the system must move to another state j, the rows of **S** sum to one. What we have done here is take the transition rate matrix and turn it into a form that states the probability of going to each other state. There is one slight difference between an embedded Markov chain and the trans-

ition probability matrix for a discrete-time Markov chain that is worth highlighting. In an embedded Markov chain the diagonal elements are zero forcing each step to move to a different state, whereas the transition probability matrix for a discrete-time Markov chain can have nonzero diagonal elements.

Now that the transition probability matrix for the embedded Markov chain has been constructed we can rearrange it into the canonical form for an absorbing Markov chain where we designate the monomer and fully formed dimer states as absorbing states. A quick note on terminology, it is necessary to differentiate transient states in a Markov chain from transient states in an absorbing Markov chain. In the case of a standard Markov chain, a state is transient if the probability of returning to a state is less than one, whereas for an absorbing Markov chain a transient state is simply any state that is not absorbing. Since the monomer and fully formed dimer states are designated as absorbing states, the diagonal elements for the rows representing the monomer and fully formed dimer states are equal to one and the off-diagonal elements in those rows are set to zero since the system is unable to leave theses states. The monomer and fully formed dimer states are selected as the absorbing states since our primary interest is analyzing statistics of reaching these states from intermediate states. However, in some cases only one absorbing state will be designated. The dissociation reaction can be analyzed by only designating the monomer as an absorbing state while the association reaction can be analyzed by only designating the fully formed dimer state as an absorbing state.

To describe the canonical form for an absorbing Markov chain consider an absorbing Markov chain with r absorbing states and t transient states. To construct the transition probability matrix for an absorbing Markov chain the rows and columns are rearranged such that the first t rows and columns of the matrix are the transient states and the last r rows and columns are the absorbing states. This results in the canonical form of the

transition probability matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{3.49}$$

where **Q** is a t-by-t matrix containing the probability of moving from transient state i to transient state j, **R** is a t-by-r matrix that contains the probability of moving from transient state i to absorbing state j. The elements of **Q** and **R** are drawn from the embedded Markov chain conditional probability matrix **S**. **I** is a r-by-r identity matrix, since the probability of remaining in an absorbing state is one. Finally, the r-by-t zero matrix exists since it is impossible to transition from an absorbing state into a transient state. The entries of this matrix p_{ij} are still the probability of moving from state i to state j as they were previously. We then note that $\mathbf{P}^n$ contains the matrix elements p_{ij}^n which are the probability of being in state j after n discrete steps given that the system started in state i.

We can utilize **Q** to construct the fundamental matrix **N** which will provide useful insights while interpreting the results of the model. The elements n_{ij} are the expected number of times that the process will be in state j given that it starts in state i, prior to reaching an absorbing state. The fundamental matrix is calculated from **Q** as follows

$$\mathbf{N} = \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots = (\mathbf{I} - \mathbf{Q})^{-1} \tag{3.50}$$

Note that **I** here has the same dimensions as **Q** and is not the same identity matrix that shows up in the lower right of the canonical form of the transition matrix **P**. To prove this statement we first point out that for an absorbing Markov chain, the probability that the system will eventually end up in an absorbing state is one, which can be stated as

$$\lim_{n \to \infty} \mathbf{Q}^n = 0 \tag{3.51}$$

The next step is to demonstrate that the matrix $(\mathbf{I} - \mathbf{Q})$ is invertible. First, we note that by the invertible matrix theorem a matrix **A** is invertible if the equation $\mathbf{Ax} = 0$, where **x** is

an arbitrary vector, has only the trivial solution $\mathbf{x} = 0$. This means that it must be shown that $(\mathbf{I} - \mathbf{Q})\,\mathbf{x} = 0$ which can be rearranged into $\mathbf{x} = \mathbf{Qx}$. We then note that multiplying both sides by $\mathbf{Q}$ gives us $\mathbf{Qx} = \mathbf{Q}^2\mathbf{x}$ which means $\mathbf{x} = \mathbf{Q}^2\mathbf{x}$. By induction this proves that $\mathbf{x} = \mathbf{Q}^n\mathbf{x}$ and taking the limit as n goes to infinity, and looking at Equation 3.51, we see that $\mathbf{x} = 0$ proving that $(\mathbf{I} - \mathbf{Q})$ is invertible. Now we will verify that Equation 3.50 is correct. If $\mathbf{N}$ is the expected number of times a system will visit j given that it starts in i and the probability of starting in state i and being in state j after n steps is given by $\mathbf{Q}^n$ it follows that $\mathbf{N} = \sum_n \mathbf{Q}^n = \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots$ which simply leaves demonstrating that this is equal to $(\mathbf{I} - \mathbf{Q})$. We start with the expression

$$(\mathbf{I} - \mathbf{Q})\left(\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots + \mathbf{Q}^n\right) = \mathbf{I} - \mathbf{Q}^{n+1} \tag{3.52}$$

and then multiply both sides by $\mathbf{N}$ using the definition that $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$ on the left hand side

$$\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots + \mathbf{Q}^n = \mathbf{N}\left(\mathbf{I} - \mathbf{Q}^{n+1}\right) \tag{3.53}$$

and then taking the limit as n goes to infinity to demonstrate that the expression in Equation 3.50 is correct.

In addition to directly providing useful insights into how the system evolves according to the Markov model the fundamental matrix is also used to calculate the absorption probabilities for an absorbing Markov chain. The matrix containing the absorption probabilities will be denoted as $\mathbf{B}$ which is a t-by-r matrix with entries b_{ij} that are the probability of being absorbed by state j if starting in state i. Note that the rows of $\mathbf{B}$ must sum to one because, as mentioned previously, the probability that the process will end up in an absorbing state is one as the number of steps approaches infinity.

To determine $\mathbf{B}$ we must consider all pathways between a given transient state i and absorbing state j. The probability of transitioning from a transient state to an absorbing state is given by the elements of $\mathbf{R}$ denoted r_{ij}. In addition to transitioning directly from

transient state i to absorbing state j we must also consider that the system can pass through intermediate transient states k. The probability of transitioning between transient states i and k is given by the elements of **Q** denoted q_{ik}. The probability of going from transient state i to absorbing state j through intermediate transient state k is given by $q_{ik}^n r_{kj}$ where n is the number of steps needed to reach state k from state i. Summing over all possible intermediate states k and number of steps n results in the following expression for **B**

$$b_{ij} = \sum_k \sum_n q_{ik}^n r_{kj} \tag{3.54}$$

where summing over all n accounts for all the different possible number of moves that can be taken to reach state k from i and summing over all k accounts for all possible intermediate transient states between i and j. Since any matrix raised to the zero power is the identity matrix Equation 3.50 shows us that

$$\sum_n q_{ik}^n = n_{ik} \tag{3.55}$$

and thus

$$\mathbf{B} = \sum_k n_{ik} r_{kj} = \mathbf{NR} \tag{3.56}$$

which can then be used to calculate the probability that any state is absorbed by the monomer state versus the dimer state when those two states are the absorbing states in an absorbing Markov chain.

3.6 Transition Path Theory Analysis

We will now introduce transition path theory (TPT) as a method for extracting mechanistic information about DNA association and dissociation from the kinetic model. A major point of emphasis for the analysis utilizing TPT is understanding the mechanism by which the association and dissociation barrier crossings occur. Our kinetic model, like

many others, spends a vast majority of its time in the thermodynamically stable states and barrier crossings are rare events. There is an extensive history of developing methods for observing rare events in simulations including, but not limited to, transition pathway sampling,[8,9] forward flux sampling,[10–13] and umbrella sampling.[14] Our primary goals are identifying the likely transition paths that the system undergoes during barrier crossing events while also identifying dynamical bottlenecks and the identity of configurations in the transition state ensemble. Many other methods for rare event sampling are capable of extracting this information from continuous-time Monte Carlo simulations. TPT was selected because of our interest in extracting this information and understanding the mechanism in the state space of the system, rather than the path space. Additionally, it was selected due to the ability of TPT to extract this information directly from the transition rate matrix of the Markov model, without needing to generate additional trajectories with varying initial conditions.

TPT directly analyzes the transition rate matrix, the matrix **L** introduced in Section 3.2, of a Markov model and provides numerous interesting insights without the need to run stochastic trajectories. TPT is regularly used in the literature to examine transition paths between select states in a Markov state model[4,15–17] and is particularly common with regards to the study of protein folding.[18–20] One particularly useful result it provides is determining and ranking the dominant association and dissociation pathways for a system modeled by a Markov state model to better understand the mechanisms by which these processes occur. It is important to note that in the work presented here we apply this under the assumption that the system is at equilibrium. This is in contrast to the trajectories that are run on a non-equilibrium system since the initial monomer concentration is taken at the initial temperature prior to the introduction of the temperature-jump pulse rather than the temperature at which the system evolves. However, the pathways derived from TPT analysis can still provide key insights into the mechanisms the trajectories follow. The analysis isolates and ranks different mechanistic pathways by comparing the reactive flux

through each pathway. In some cases in the literature the flux is also referred to as the probability flux or the probability current. For the purposes of this discussion we will closely follow the notation and terminology of Metzner, Schütte, and Vanden-Eijnden.[4]

To determine the reactive flux for a pathway we first need to determine the flux for every possible move within the model. In this context the reactive flux between two states is the flux that contributes to the overall pathway of interest. To generalize the equations and reduce confusion we will refer to a general pathway that proceeds from *A* to *B* where *A* and *B* are sets of states. Mathematically all of our equations will be written utilizing set notation for correctness. However, in the case of the work presented here there will only ever be a single state in both *A* and *B* and the discussion with respect to the methods application to the model will be presented as such. In the association case *A* contains the monomer state and *B* contains the fully formed dimer state, with the two flipped for the dissociation case. While the construction of our specific model is discussed in Chapter 7 Figure 7.23 in Appendix 7A shows the states and allowed moves for a simple sequence. This can help to visualize how the individual moves, whose reactive fluxes are being calculated, combine to form the pathways that TPT is analyzing.

The flux going from state i into state j along an overall pathway going from *A* to *B* is given by

$$f_{ij}^{AB} = \begin{cases} \pi_i q_i^- l_{ij} q_j^+ & if\ i \neq j \\ 0 & if\ i = j \end{cases} \tag{3.57}$$

where f_{ij}^{AB} is the flux between i and j that contributes to the overall pathway of interest, π_i is the probability of being in state i, q_i^-, known as the backward committor, is the probability that a process arriving in state i last came from *A* rather than *B*, l_{ij} is the transition rate from i to j, and q_j^+, known as the forward committor, is the probability that the process starting in j will first reach *B* rather than *A*. The calculation of q_i^- and q_j^+ will be described shortly. The probability of occupying a state, π_i is calculated from the steady state distribution for the transition rate matrix, the calculation of which is discussed in Section 3.4.

It is important to note that the steady state distribution is calculated from the transition rate matrix which, as will be discussed in Chapter 7, is built and parameterized to replicate our non-equilibrium temperature-jump experiments. As a result the steady state distribution is not directly comparable to either the population distribution of the system at the initial or final temperature. However, it is used for the analysis to ensure that the condition of detailed balance holds, or that the system being analyzed is at equilibrium even if it is not the physical equilibrium at either the initial or final temperature. For the condition of detailed balance to hold the following expression must be true

$$\pi_i l_{ij} = \pi_j l_{ji} \tag{3.58}$$

In this case it is also true that

$$q_i^+ + q_i^- = 1 \tag{3.59}$$

which states that a process that starts in state i must eventually reach either state *A* or *B*.

To prove Equation 3.59 we must introduce the formal method for solving both the forward and backward committors. While only the forward or backward committor needs to be solved with this method since the other can be solved for using Equation 3.59 the method for calculating both will be provided for completeness. For the purpose of this discussion we will continue with more general language and discuss the calculation of the committors for a system with a set of states denoted as *S* and looking at pathways between two subsets of states within *S* denoted *A* and *B*. We start with the definition of the forward committor which is defined as the values of $q^+ = \left(q_i^+\right)_{i \in S}$ that satisfy[4]

$$\begin{cases} \sum_{k \in S} l_{ik} q_k^+ = 0 & \forall i \in (A \cup B)^c \\ q_i^+ = 0 & \forall i \in A \\ q_i^+ = 1 & \forall i \in B \end{cases} \tag{3.60}$$

where we introduce additional mathematical symbols. The $\cup$ symbol is known as the union symbol and $A \cup B$ refers to elements that belong to either *A* or *B*. The $\in$ symbol means "in" and the superscript c refers to the complement of the set which is all elements not in the set. The line $\forall i \in (A \cup B)^c$ is then read as "for all i in the compliment of *A* and *B*" or "for all i that are not in either *A* or *B*" and the line $\forall i \in A$ is read as "for all i in *A*". To determine q^+ we then solve the following set of linear equations

$$\mathbf{U}q^+ = \mathbf{v} \tag{3.61}$$

where the matrix **U** and vector **v** are given by

$$u_{ij} = l_{ij} \qquad \forall i, j \in (A \cup B)^c \tag{3.62}$$

$$v_i = -\sum_{k \in B} l_{ik} \qquad \forall i \in (A \cup B)^c \tag{3.63}$$

Or in other words if there is a single state *B*, such as the case where *B* is simply the fully formed dimer state, **v** is a vector that contains the negative elements of the column associated with state *B*. Solving the linear system of equations given by Equation 3.61 then provides the values of q^+.

Turning our attention to the backward committor values q^- we start with the definition of the backward committor which is defined as the values of $q^- = \left(q_i^-\right)_{i \in S}$ that satisfy[4]

$$\begin{cases} \sum_{k \in S} \tilde{l}_{ik} q_k^- = 0 & \forall i \in (A \cup B)^c \\ q_i^- = 1 & \forall i \in A \\ q_i^- = 0 & \forall i \in B \end{cases} \tag{3.64}$$

With a steady state distribution denoted as π, calculated as discussed earlier, we define

$$\tilde{l}_{ik} = \frac{\pi_k l_{ki}}{\pi_i} \tag{3.65}$$

If balanced detection does not hold we follow a similar procedure to above by solving the linear set of equations

$$\mathbf{U}q^{-} = \mathbf{v} \tag{3.66}$$

where the matrix **U** and vector **v** are given by

$$u_{ij} = \tilde{l}_{ij} \qquad \forall i,j \in (A \cup B)^c \tag{3.67}$$

$$v_i = -\sum_{k \in A} \tilde{l}_{ik} \qquad \forall i \in (A \cup B)^c \tag{3.68}$$

Or in other words if there is a single state in *A*, such as the case where *A* is the monomer state, **v** is a vector with the negative elements of the column associated with state *A*. As before solving the system of equations given by Equation 3.66 provides the values of q^{-}. In the case where detailed balance holds we first note that by looking at Equations 3.58 and 3.65 we see that $\tilde{l}_{ik} = l_{ik}$ and Equation 3.64 becomes

$$\begin{cases} \sum_{k \in S} l_{ik} q_k^{-} = 0 & \forall i \in (A \cup B)^c \\ q_i^{-} = 1 & \forall i \in A \\ q_i^{-} = 0 & \forall i \in B \end{cases} \tag{3.69}$$

A quick calculation of which demonstrates that $q^{-} = 1 - q^{+}$ demonstrating that the forward and backward committors for a given state will sum to one if detailed balance holds.

Now that we have determined the flux between two states we can calculate the net

flux between i and j that contributes to the *A* to *B* transition according to

$$f_{ij}^{+} = \max\left(f_{ij}^{AB} - f_{ji}^{AB}, 0\right) \tag{3.70}$$

The overall flux for a pathway is then defined as the minimum of the net flux values for each step along the pathway. The step with the minimum net flux is referred to as the bottleneck step. The dominant pathway is defined as the pathway with the largest overall flux, or in other words the pathway whose bottleneck step has the largest net flux among bottleneck steps.

Determining and ranking the pathways requires an algorithm to determine what steps are bottlenecks and rank them according to their net flux values. Additionally, it is likely that each bottleneck step contributes to multiple different pathways, requiring the determination of the most dominant pathway among all pathways that share that bottleneck, all of which have the same overall flux. The first step requires finding the bottleneck with the largest flux among all possible pathways. A bisection algorithm to achieve this was written based on the one proposed by Metzner, Schütte, and Vanden-Eijnden.[4]

After determining the bottleneck for the dominant pathway the other steps in the pathway need to be determined. This must be done considering the fact that there are likely to be multiple pathways that share this bottleneck. This requires a second algorithm which is also based on one proposed by Metzner, Schütte, and Vanden-Eijnden.[4] To determine the most dominant pathway with a bottleneck we look to maximize the net flux for each remaining step in the pathway. To achieve this the pathway is broken into two parts, the initial state to the first state in the bottleneck, and the second state in the bottleneck to the final state. Each of these pathways is treated independently and the first algorithm is utilized to find the bottleneck of each one. Those pathways can then be split up in the same way and the process continues until each step in the pathway has been filled. Utilizing this recursive method ensures that the maximum flux value for each step in the process is

achieved resulting in the dominant pathway containing the given bottleneck. Now that the dominant pathway has been found additional pathways must be isolated. To do this we adapt the algorithm proposed by Noé et al.[20] that is designed to determine the pathways with the largest overall flux in descending order. After finding the first pathway the algorithm subtracts off the pathway flux from the net flux of each step in the pathway setting the net flux for the bottleneck step to zero and reducing the magnitude of the net flux for all other steps. This new net flux matrix is then used to find the next pathway utilizing the same method. Subtracting off the overall pathway flux from each step along the pathway after the pathway is determined ensures that the same pathway is not found again. This does however result in the fact that only one pathway, the most dominant one, is found for each bottleneck and once a step is determined to be a bottleneck step for a pathway it cannot appear in any subsequent pathways.

3.7 Acknowledgements

I would like to thank Greg Kimmel and Paul Sanstead for their careful reading and thoughtful comments on this chapter.

3.8 References

1. Sanstead, P. J.; Tokmakoff, A. A Lattice Model for the Interpretation of Oligonucleotide Hybridization Experiments. *J. Chem. Phys.* **2019**, *150*, 185104.
2. Logan, J. D. *Applied Mathematics*; John Wiley & Sons: Hoboken, 2006.
3. Gardiner, C. W. *Handbook of Stochastic Methods*; Springer: Berlin, 1985.
4. Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7*, 1192–1219.
5. Gillespie, D. T. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comput. Phys.* **1976**, *22*, 403–434.

6. Gillespie, D. T. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **1977**, *81*, 2340–2361.

7. Serfozo, R. *Basics of Applied Stochastic Processes*; Springer Berlin: Heidelberg, 2009.

8. Pratt, L. R. A Statistical Method for Identifying Transition States in High Dimensional Problems. *J. Chem. Phys.* **1986**, *85*, 5045–5048.

9. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

10. Allen, R. J.; Frenkel, D.; ten Wolde, P. R. Simulating Rare Events in Equilibrium or Nonequilibrium Stochastic Systems. *J. Chem. Phys.* **2006**, *124*, 024102.

11. Allen, R. J.; Warren, P. B.; ten Wolde, P. R. Sampling Rare Switching Events in Biochemical Networks. *Phys. Rev. Lett.* **2005**, *94*, 018104.

12. Allen, R. J.; Frenkel, D.; ten Wolde, P. R. Forward Flux Sampling-Type Schemes for Simulating Rare Events: Efficiency Analysis. *J. Chem. Phys.* **2006**, *124*, 194111.

13. Allen, R. J.; Valeriani, C.; ten Wolde, P. R. Forward Flux Sampling for Rare Event Simulations. *J. Phys. Condens. Matter* **2009**, *21*, 463102.

14. Warmflash, A.; Bhimalapuram, P.; Dinner, A. R. Umbrella Sampling for Nonequilibrium Processes. *J. Chem. Phys.* **2007**, *127*, 154112.

15. Weinan, E.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503.

16. Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Illustration of Transition Path Theory on a Collection of Simple Examples. *J. Chem. Phys.* **2006**, *125*, 084110.

17. Weinan, E.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Ann. Rev. Phys. Chem.* **2010**, *61*, 391–420.

18. Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of *ab Initio* Protein Folding for a Millisecond Folder NTL9(1-39). *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.

19. Noé, F.; Fischer, S. Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struc. Biol.* **2008**, *18*, 154–162.

20. Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011–19016.

CHAPTER 4

EXPERIMENTAL METHODS

4.1 Introduction

This chapter discusses the basics of the instrumentation and experiments utilized in the research presented in this thesis. The portions of this chapter dedicated to nonlinear measurements are closely related to the theory and formalism laid out in Chapter 2, particularly with respect to the acquisition of nonlinear signals. The experiments and instrumentation discussed have been described elsewhere in detail [1–3] and full descriptions are not provided here. The main purpose of this chapter is to establish a functional understanding of how the instruments are designed and utilized. The instruments will be briefly discussed while highlighting important aspects of their operation and data collection.

The thermodynamic measurements were conducted utilizing a Fourier transform infrared (FTIR) spectrometer connected to a recirculating chiller for temperature control. The kinetic measurements were conducted utilizing an ultrafast two-dimensional infrared (2DIR) spectrometer, known by us as the boxcar as it is our only system that utilizes the boxcar geometry, which is a component of the overall temperature-jump spectrometer. The other component is the temperature-jump laser that induces the temperature perturbation while the boxcar tracks the changes in the sample as it evolves.

In this chapter we will first discuss the preparation of samples for both the thermodynamic and kinetic experiments and tricks that improve the data collection. This will be followed by a brief discussion of the FTIR temperature ramp setup that was used to obtain the thermodynamic data. Then our focus will shift to the ultrafast setup utilized to obtain the kinetic data. We will discuss the main components of the instrument and the commercial systems used to generate the $\approx$6 μm mid-IR and 2 μm temperature-jump pulse. We will also briefly walk through important components of these instruments' operation. This will include essential aspects of setting up the instrument in addition to an overview

of how the data are collected. Finally we conclude by discussing highlights of the data processing that occurs after data collection.

4.2 Sample Preparation

The first step in sample preparation is to deuterate both the nucleic acid and the buffer components to swap out any labile hydrogens in the sample. This is essential since the H_2O bend vibrational mode, centered at $\approx$1650 cm^{-1}, has an intense absorption[4] in the 1500-1700 cm^{-1} region that contains the DNA vibrational modes of interest and any H_2O present would make it hard to observe the DNA signal. Deuterium oxide has a significantly weaker absorption in this region, due to a redshifted D_2O bend-libration combination band centered at $\approx$1550 cm^{-1},[4] that can be removed during data processing. During the deuteration step the DNA is purified using Amicon Ultra 3 kDa centrifugal filters or Sartorius Setim Biotech Vivaspin 2 2 kDa centrifugal filters, depending on the sample molecular weight, to remove impurities that may remain from the manufacturing process.

A typical sample contains approximately 35 µL of sample containing the DNA at a concentration of 2 mM in a sodium phosphate buffer. Before loading the sample into the sample cell it is useful to briefly spin it in a centrifuge. This forces any particulates that may be present to the bottom of the vial allowing sample to be taken from the top of the vial that is generally free of particulates. While this is not particularly important for FTIR measurements it significantly reduces the risk of scatter washing out nonlinear signal. For both FTIR and nonlinear measurements the sample is then loaded into a home-built sample cell that sandwiches the sample between two calcium fluoride windows, ideal due to their low absorbance across most of the mid-IR range, separated by a Teflon spacer creating a 50 µm path length. When loading the cell care should be taken to avoid the presence of bubbles. In the case of FTIR temperature ramps the presence of bubbles can speed up evaporation which can affect measurements. It is less important for nonlinear

measurements which have a much smaller focus, though it is still preferred to not have to adjust the sample cell to shift the focus away from bubbles. The windows are held together by a brass jacket which is used for its high thermal conductivity which aids in temperature control.

4.3 FTIR Temperature Ramp

FTIR temperature ramp experiments were conducted on a Bruker Tensor 27 FTIR spectrometer hooked up to a recirculating chiller for temperature control. The spectra were obtained utilizing a macro program that automates the process by controlling both the chiller temperature and the FTIR acquisition software. A standard temperature ramp series ramps the chiller temperature from 0-96 °C in 3 °C steps with an FTIR spectrum obtained after each temperature step. This provides a good balance between collecting sufficient data points for analysis while keeping the acquisition time reasonable. It is important to ensure that there is a sufficient waiting time between the chiller reaching each set point and acquiring the FTIR spectrum to ensure that the sample has equilibrated at the new temperature. The necessary time can vary depending on the exact settings used for the particular ramp. As time passes a relatively intense absorption due to the HOD bend vibration grows in at $\approx$1460 cm^{-1} which begins to overwhelm the DNA absorption that can be difficult to remove via processing if it becomes too large. As a result care should be taken to minimize the exposure of the sample to the air to minimize the initial amount of hydrogen present. Additionally, at high temperature evaporation starts to occur which, if significant enough, results in a noticeable absorption decrease in the DNA peaks.

There are two main components of the data processing to prepare spectra for analysis. The first is determining the actual sample temperature at each chiller set point. To determine the correct temperature an additional temperature ramp is done on a sample cell containing water with a thermocouple attached to the window. The automated temper-

ature ramp program is run with the thermocouple temperature logged when each spectrum is taken. While the temperature cannot be determined during the same run as the data since the thermocouple blocks the window, conducting both runs under the same conditions provides a reasonable value of the sample temperature. The other major component is subtracting off the absorption due to D_2O and HOD. This is done by conducting two reference temperature ramps: one with D_2O and the other with a 2% HOD solution. The resulting spectra can then be scaled to the sample spectrum at a point where only HOD and D_2O absorption are present, to account for any differences in absorption due to slight differences in path length, IR intensity, or other slight variations between measurements, allowing the reference spectrum to be subtracted off leaving only the DNA FTIR spectrum at each temperature point.

4.4 Boxcar Spectrometer

4.4.1 Mid-IR Generation

The generation of our ultrafast mid-IR pulses broadly occurs in three steps. The first step is the generation of a 90 fs 795 nm pulse which is done by a titanium sapphire (Ti:Sapph) regenerative amplifier (Libra, Coherent) with a 1 kHz repetition rate. To generate a pulse of sufficient power a 795 nm seed pulse is amplified using chirped pulse amplification.[5] The seed is first stretched out in time by a grating before entering the optical cavity with a Ti:Sapph rod that contains a population inversion generated by 527 nm light from a pump laser. The chirped seed pulse makes multiple round trips through the cavity and is amplified at each step by stimulated emission from the Ti:Sapph rod before the amplified pulse is ejected from the cavity. At this point it is recompressed by a grating compressor resulting in a 795 nm pulse that is approximately 90 fs in duration.

This pulse then undergoes optical parametric amplification (OPA) to generate two frequencies of light that are then used to generate the mid-IR pulse through difference

frequency generation (DFG). The commercial OPA (TOPAS C, Light Conversion) uses a two-step process to generate pulses of light at two center frequencies, referred to as the signal and idler. The signal and idler frequencies sum to the frequency of the 795 nm pulse that enters the OPA and their difference is the frequency of the desired mid-IR. In the first OPA step, known as the pre-amplification stage, the signal is generated at approximately 1.4 μm. In the second stage of the OPA the signal is mixed with remaining 795 nm light from the regenerative amplifier to amplify the signal and generate the idler, which is approximately 1.8 μm. Both the signal and idler can be tuned to generate the exact frequency of mid-IR that is desired. As a reference, a signal and idler of exactly 1.4 and 1.8 μm respectively produces a mid-IR pulse centered at $\approx$1587 cm^{-1}. In practice, for the purpose of studying the dynamics and kinetics of canonical DNA duplexes, our mid-IR pulse is commonly tuned to be centered at $\approx$1630 cm^{-1} with approximately 300 cm^{-1} of bandwidth. After the amplification stage the signal and idler exit the TOPAS and enter the DFG. The DFG is home-built and a full description of it can be found elsewhere.[2] The signal and idler beams are separated to allow their relative timing to be adjusted before they are recombined colinearly in a $AgGaS_2$ crystal with the proper timing and phase matching condition to maximize the generation of the desired mid-IR light. At the conclusion of this process approximately 10 μJ of mid-IR light is generated with average pulse energy fluctuations below 1%, ideal for nonlinear measurements.

4.4.2 Interferometer

Before leaving the DFG the mid-IR is overlapped with a visible HeNe tracer beam that propagates collinearly with the mid-IR throughout the remainder of the instrument. This is for the purposes of alignment since the mid-IR light is not visible to the naked eye. The mid-IR and HeNe are overlapped by reflecting the HeNe off of a germanium plate that transmits the mid-IR. The mid-IR is properly overlapped with the HeNe, to ensure that they propagate collinearly, by sending both beams through an iris into a power meter at

two positions, one close to the DFG and one several meters away. The HeNe is centered on each iris and the power meter is used to maximize mid-IR throughput at both positions. It is essential to use a position both close to, and far away from, the DFG to ensure that the beams continue to propagate collinearly all the way to the sample area.

The boxcar has a home-built interferometer, which is fully described and diagrammed elsewhere.[1,2] The interferometer uses ZnSe beamsplitters to split the incoming mid-IR into five pulses, the three pulses that interact with the sample to generate the signal, the Local Oscillator (LO) pulse, the purpose of which will be described when detection is discussed, and the tracer pulse, which follows the signal pathway for the purposes of alignment and is not used in the experiment. The LO and tracer pulses contain less than 1% of the mid-IR light that enters the interferometer while the other three beams share the remaining light roughly equally. Once separated each beam follows its own path with its own set of optics in the interferometer allowing each one to be independently controlled. All four beam paths have their own retroreflector all of which, except k_3, the third beam in the signal generating pulse set, are mounted on a motorized stage that provides precise control of the time delays between each pulse. After the retroreflector the LO beam passes through another beamsplitter where the reflected light exits the box as the LO and the transmitted light becomes the tracer. The tracer is usually blocked, however when unblocked it becomes the fourth corner of the boxcar geometry. This means that it follows the path of the signal which can be useful for the purposes of aligning the balanced detection optics. Each beam then passes through its own wave plate and polarizer to provide polarization control. In the case of the experiments presented here all four beams have the same polarization. Before exiting the interferometer k_2, the second beam when collecting the rephasing signal and the first beam when collecting the non-rephasing signal, passes through the chopper.

4.4.3 Sample Detection

The four beams, and the tracer if needed, then enter the sample area, which is described in detail elsewhere.[1] The three signal generating pulses enter the sample detection area in the boxcar geometry, as shown in Figure 2.1, with the tracer beam occupying the fourth corner of the box if necessary. The beams are aligned onto a gold parabolic mirror that focuses them into the sample. The LO beam comes into the box separately and is aligned to reflect off of the gold parabolic just outside of the box made by the signal generating beams and is also focused into the sample. A coarse spatial alignment of the beams is done using a set of irises prior to the gold parabolic to properly align them into the boxcar geometry. A more precise alignment, ensuring the beams all focus to the same spot, is conducted by placing a 50 μm pinhole at the sample position and aligning all four beams through the pinhole by maximizing the throughput of each beam on a single channel detector. Once the beams are aligned spatially they need to be overlapped in time as well. Since k_3 is not mounted on a motorized stage all other beams will use it as a reference. Two beams are scanned against each other to find the point in time where constructive interference is maximized, which provides the time at which the beams are overlapped. First k_2 is scanned against k_3 and time zero for k_2 is set to be the time at which they overlap. Then k_2 is scanned against k_1 to set k_1 to the same time zero. Scanning k_1 against k_2 rather than k_3 helps to minimize timing errors in τ_1, the time delay between the first and second pulses as introduced in Chapter 2. Finally the LO can be scanned against k_3 resulting in all four beams sharing the same time zero at the sample.

With the beams properly overlapped in time and space we can move onto detection. The detector is a mercury cadmium telluride (MCT) detector that contains two vertically displaced stripes each with 64 pixels. The signal is detected utilizing a balanced detection scheme which provides a significant increase in the signal-to-noise ratio. Balanced detection works by overlapping the signal and the LO on an anti-reflection (AR) coated beamsplitter resulting in two paths along which the signal and LO both propagate collin-

early. The beamsplitter is AR coated on one side to restrict reflections to only occur at the uncoated face. The LO and signal approach from different directions resulting in the signal reflecting off of the front face of the beamsplitter while the LO reflects off of the back face. Due to the difference in refractive index between air and ZnSe the signal reflection off of the front face has a π phase shift relative to the transmitted LO that it is overlapped with. The LO reflection off of the back face of the uncoated side does not undergo a phase shift and has the same phase as the transmitted signal. This results in the signal collected on each stripe being

$$\begin{aligned} I_1\left(\omega_3.\tau_2,\tau_1,\tau_{\mathsf{LO}}\right) &= \left|\mathbf{E}_{\mathsf{sig}}\left(\omega_3,\tau_2,\tau_1\right)+\mathbf{E}_{\mathsf{LO}}\left(\omega_3,\tau_{\mathsf{LO}}\right)\right|^2 \\ I_2\left(\omega_3.\tau_2,\tau_1,\tau_{\mathsf{LO}}\right) &= \left|\mathbf{E}_{\mathsf{sig}}\left(\omega_3,\tau_2,\tau_1\right)-\mathbf{E}_{\mathsf{LO}}\left(\omega_3,\tau_{\mathsf{LO}}\right)\right|^2 \end{aligned} \tag{4.1}$$

Describing the electric field as a plane wave as $\mathbf{E} = Ae^{i\phi}$ where A is the amplitude and ϕ is the phase. The amplitude and phase of the signal depend on ω_3, τ_2, and τ_1 where as the amplitude and phase of the LO depend on ω_3 and τ_{LO} which will be dropped from here on for simplification. Expanding each term out gives us

$$\begin{aligned} I_1\left(\omega_3.\tau_2,\tau_1,\tau_{\mathsf{LO}}\right) &= A_{\mathsf{sig}}^2+A_{\mathsf{LO}}^2+2A_{\mathsf{sig}}A_{\mathsf{LO}}\cos\left(\phi_{\mathsf{sig}}-\phi_{\mathsf{LO}}\right) \\ I_2\left(\omega_3.\tau_2,\tau_1,\tau_{\mathsf{LO}}\right) &= A_{\mathsf{sig}}^2+A_{\mathsf{LO}}^2-2A_{\mathsf{sig}}A_{\mathsf{LO}}\cos\left(\phi_{\mathsf{sig}}-\phi_{\mathsf{LO}}\right) \end{aligned} \tag{4.2}$$

where the cross term contains the desired phase and amplitude information. These equations designate the detected signal when all three signal generating beams are present, meaning the chopper is in the open position. We will designate this as I_1^{o} and I_2^{o} where the superscript o designates that k_2 passes through the chopper. In the case where the chopper is blocking k_2 the signal detected on each stripe of the MCT array is

$$I_{1/2}^{\mathsf{c}}\left(\omega_3,\tau_{\mathsf{LO}}\right) = \left|\mathbf{E}_{\mathsf{LO}}\left(\omega_3,\tau_{\mathsf{LO}}\right)\right|^2 = A_{\mathsf{LO}}^2 \tag{4.3}$$

since the third order signal cannot be generated along the detected signal path without all three beams interacting with the sample. Here the superscript c designates the chopper is in the closed position and both stripes detect the same signal. Moving forward these will simply be designated as I_1^{c} and I_2^{c} for simplicity. In practice the data are acquired according to the following equation[3,6]

$$\tilde{S}_{\mathsf{exp}}\left(\omega_3, \tau_1\right) = \left[\left(\frac{I_1^{\mathsf{o}} - I_2^{\mathsf{o}}}{I_1^{\mathsf{o}} + I_2^{\mathsf{o}}}\right) - \left(\frac{I_1^{\mathsf{c}} - I_2^{\mathsf{c}}}{I_1^{\mathsf{c}} + I_2^{\mathsf{c}}}\right)\right] \tag{4.4}$$

where the term in the right is essentially zero but helps to remove effects due to scatter and other shot to shot variations. Plugging Equations 4.2 and 4.3 into equation 4.4 results in

$$\tilde{S}_{\mathsf{exp}}\left(\omega_3, \tau_1\right) = \frac{2A_{\mathsf{LO}}A_{\mathsf{sig}}\,\mathsf{cos}(\phi_{\mathsf{sig}} - \phi_{\mathsf{LO}} + \tau_{\mathsf{LO}}\omega_3)}{A_{\mathsf{LO}}^2} \tag{4.5}$$

where the A_{sig}^2 term has been neglected since it is very small relative to the A_{LO}^2 term. Multiplying through by A_{LO}^2, which is also done on the fly by the control software, provides the final signal.

4.4.4 Temperature-Jump Spectrometer

The temperature-jump spectrometer is an extension of the boxcar where a 10 ns 1.98 μm pulse induces a rapid temperature increase in the sample and the resulting structural changes are monitored utilizing the boxcar spectrometer. The temperature-jump laser is a flashlamp pumped Q-switched Nd:YAG (YG981C, Quantel) that generates 1064 nm pulses that are approximately 10 ns in duration at a repetition rate of 20 Hz. These pulses are then frequency doubled by second harmonic generation to 532 nm. This light pumps an optical parametric oscillator (OPO) (Opotek), which, similar to the OPA, generates a signal and idler whose frequencies sum to the frequency of the incoming light, though the method by which this occurs is slightly different. The primary difference between the two being that the OPA first generates a seed which is then amplified in the second stage while

also generating the idler whereas the OPO is self-seeded. The OPO generates a 1.98 μm idler, the signal is discarded, which is used to generate the temperature perturbation. The trigger for the flashlamps and the Q-switch in the temperature-jump laser is generated by a delay generator (DG535, Stanford Research) which itself is triggered by the signal delay generator (SDG) that controls all of the timing electronics for the boxcar syncing the timings for both systems. A schematic of the timing electronics and delays that control both the regenerative amplifier and the temperature-jump laser can be found elsewhere.[1]

The 1.98 μm pulse pumps the overtone of the OD stretch vibrational mode of the D_2O solvent. The resulting vibrational excitation quickly relaxes back to the ground state causing the sample to heat up on the scale of the ≈10 ns temperature-jump pulse width. The overtone of the OD stretch is used since only about 10% of the light is absorbed which results in a more even heating of the sample as the power of the pulse is not significantly reduced as it passes through the sample.

Before discussing spatially overlapping the temperature-jump pulse with the mid-IR pulses at the sample, in addition to setting the timing between the two pulses, it is useful to discuss the profile of the temperature-jump experiment, which is shown in Figure 4.1. Due to the difference in the repetition rates between the two lasers there are 50 mid-IR

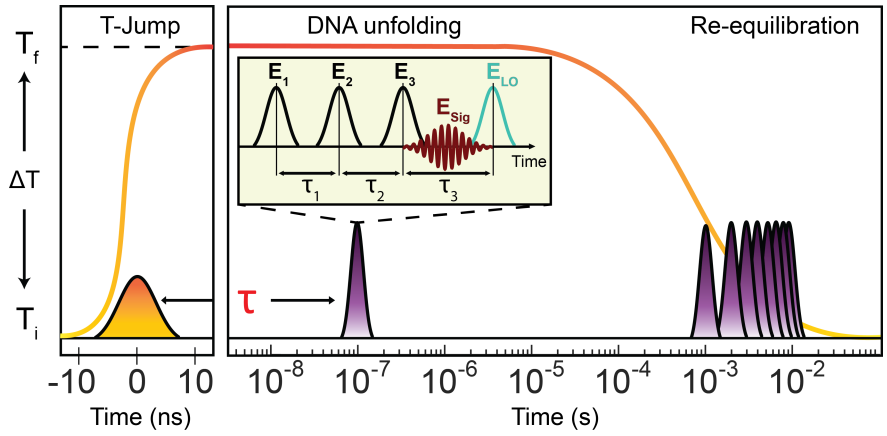

Figure 4.1: Profile of a temperature-jump experiment. The orange pulse represents the temperature-jump pulse while the purple pulses represent the mid-IR pulse sequence, shown in the insert, that tracks the samples response to the temperature perturbation. The temperature-jump time delay τ is adjusted to sample the entire temperature profile.

pulse sequences for each temperature-jump pulse. For simplicity the four mid-IR pulse sequence will be referred to as a single shot, since it originates from a single shot of the Ti:Sapph laser, and the set of 50 mid-IR shots will be referred to as a shotset. Looking at the profile of the experiment in Figure 4.1 we designate time zero to be the time at which the temperature-jump pulse arrives at the sample, which is designated by the orange pulse in the diagram. The mid-IR shots are the purple peaks on the plot with the insert demonstrating that each is split into the four individual pulses. The curve that changes color from orange to red and back to orange demonstrates the temperature profile of the sample. The user controlled time delay between the temperature-jump pulse and the first mid-IR shot is designated as τ. Designating the time delay between the temperature-jump pulse and each mid-IR shot in the shotset as τ_j where j designates the j^{th} shot, $\tau_j = \tau + (j-1)\cdot 1$ ms where j = 1, 2, ..., 50. Since the temperature plateau, during which the sample is evolving at the final temperature, lasts for less than a millisecond before starting to cool back to the initial temperature only the first mid-IR shot for each shotset will fall within the temperature plateau. The 49th and 50th shots occur at a time where the sample has returned to the initial temperature and re-equilibrated. These shots are referred to as the equilibrium shots and are used during processing to determine the change in signal relative to equilibrium at each time point. A series of τ points are collected to sample the temperature profile. For shots j = 2-48 the j^{th} shots are often averaged together across multiple τ points since the change in τ is insignificant in the millisecond time regime in which these shots occur.

After the temperature-jump pulse is routed into the sample area it must be spatially overlapped with the mid-IR pulse at the sample. The temperature-jump pulse is first visually overlapped with the HeNe tracer as a coarse alignment. After the coarse alignment a temperature dependent change in LO transmission of the D_2O bend-libration combination band centered near 1555 cm^{-1} should be observed. Next a fine alignment of the spatial overlap is conducted by adjusting the position of the temperature-jump pulse to maximize

the LO response. The next step is to set the zero point for τ as the point at which the temperature-jump pulse arrives at the sample at same time as the first mid-IR shot in the shotset. This is done by determining the magnitude of the LO response at the top of the plateau and then adjusting the delay that triggers the temperature-jump laser such that the LO response of the first mid-IR shot is half of the maximum magnitude of the LO response. The time at which the LO response is half of the maximum is designated as time zero for τ.

Data acquisition for a temperature-jump experiment generally follows the method utilized by the boxcar as described in Section 4.4.3 with a few distinctions. The first is that τ_1 = 0, meaning that we are acquiring the HDVE rather than a 2D spectrum. This is done to conserve time in cases where the ability to resolve the ω_1 axis is not necessary. Acquiring a full HDVE data set for a single initial temperature for a given sample requires somewhere in the range of four to eight hours. While it is possible to take a full transient 2DIR spectrum it takes much longer. Acquiring the HDVE provides the ability to collect more initial temperatures and τ delays for each sample, or more samples, in the same amount of time.

The second distinction is that chopping every other mid-IR shot during acquisition and processing sequential open and closed shots together according to Equation 4.4 does not work since the neighboring chopped and unchopped shots have different values of τ. To get around this the chopper phase is flipped during data acquisition. The signal is first acquired with the chopper open for the even shots and closed for the odd shots and then the chopper undergoes a π phase shift and the signal is acquired with the chopper closed for the even shots and open for the odd shots. This results in obtaining a signal with both the chopper open and closed for each value of τ which can then be used to obtain the signal as described in Section 4.4.3.

4.4.5 Data Processing

After the data is collected additional processing is done with Matlab scripts. Our focus will be on processing the transient HDVE data as that is what is used in this thesis. The first aspect of data processing is to recover the complex spectral interferogram which is done using the FTSI method that was described in Section 2.5. The next step is to correct any errors that may have occurred in τ_{LO} over the course of the experiment. This is done by comparing the equilibrium shots for each shotset to a pump probe spectrum which is equivalent to the real part of the HDVE spectrum as discussed in Section 2.5. Prior to the acquisition of the temperature-jump data a pump probe spectrum is collected using the chopped k_2 beam and the LO with a delay of 150 fs, the same as the τ_2 delay in the temperature-jump experiment. For each shotset the equilibrium shots are fit to the pump probe spectrum in the frequency domain with a phase correction value as the fit parameter. The best fit is found by a nonlinear least squares fitting algorithm and the resulting phase correction is applied to the entire shot set. To generate the final transient HDVE difference spectrum the equilibrium spectra are subtracted off from the spectra at each time point before dividing through by the maximum value of the equilibrium spectrum. The resulting final transient HDVE spectrum for each time point has a y-axis that is the percent change in signal at time τ relative to the equilibrium signal normalized to the maximum value of the equilibrium spectrum.

4.5 Acknowledgements

I would like to thank Paul Sanstead for his careful reading and thoughtful comments on this chapter.

4.6 References

1. Sanstead, P. J. C. Investigation of DNA Dehybridization through Steady-State and Transient Temperature-Jump Nonlinear Infrared Spectroscopy. Ph.D. thesis, The University of Chicago, 2018.

2. Stevenson, P. Membrane and Membrane Protein Dynamics Studied with Time-Resolved Infrared Spectroscopy. Ph.D. thesis, Massachusetts Institute of Technology, 2017.

3. Jones, K. C. Temperature-Jump 2D IR Spectroscopy to Study Protein Conformational Dynamics. Ph.D. thesis, Massachusetts Institute of Technology, 2012.

4. Max, J.-J.; Chapados, C. Isotope Effects in Liquid Water by Infrared Spectroscopy. III. H_2O and D_2O spectra from 6000 to 0 cm^{-1}. *J. Chem. Phys.* **2009**, *131*, 184505.

5. Strickland, D.; Mourou, G. Compression of Amplified Chirped Optical Pulses. *Opt. Commun.* **1985**, *56*, 219–221.

6. Jones, K. C.; Ganim, Z.; Tokmakoff, A. Heterodyne-Detected Dispersed Vibrational Echo Spectroscopy. *J. Phys. Chem. A* **2009**, *113*, 14060–14066.

CHAPTER 5

ANALYSIS METHODS

5.1 Introduction

This chapter introduces the analysis conducted on the experimental thermodynamic and kinetic data. We will first discuss the methods utilized to extract thermodynamic parameters from the temperature ramp FTIR experiments. These parameters are useful both for understanding the energetics of the association and dissociation processes, but are also a necessary part of the kinetic analysis for the transient temperature-jump experiments. Thermodynamic analysis starts with obtaining the melting curve that tracks the loss of DNA base pairs as a function of temperature. Two different methods of varying complexity for obtaining the melting curve will be discussed providing some flexibility in how it can be obtained. After obtaining the melting curve a seven parameter fit is applied to extract the thermodynamic parameters. Analysis of thermal melting curves can take multiple different forms that depend on how various parameters, such as the melting temperature, are defined since a variety of definitions exist. The resulting expressions for the parameters and the fit itself also differ based on the system being studied, an example being different expressions for self-complimentary and non-self-complimentary DNA duplexes. Here we will present the fitting used in this work to extract the thermodynamic parameters from the melting curve of a duplex made up of self-complimentary monomers that is assumed to dissociate and associate as a two-state process.

This is followed by a discussion of the methods for analyzing the data obtained from the transient temperature-jump experiments. The methods discussed here are widely used both in the literature and in our research group so only a brief discussion is contained here. The first step upon obtaining a completed temperature-jump data set is to confirm the magnitude of the temperature jump for each initial temperature. This is used to determine the final temperature at which the system is evolving. Once the final temperature is known

the kinetic information can be extracted and analyzed. Two different methods for extracting relevant kinetics will be discussed: analyzing the results in the time domain and utilizing an inverse Laplace transform method to translate the data into the rate domain for analysis. Neither method is inherently better than the other, instead they provide different avenues for examining the data and both methods can be useful depending on the specific context. Finally, the methods and mathematics behind the deconvolution of the association and dissociation rates from the overall observed rate, and the assumptions that are made to greatly simplify the process, will be discussed. Looking at not only the association and dissociation rates but also the shape and functional form of the signal obtained both in the time and rate domains is the first step towards understanding the kinetics and dynamics of the samples.

5.2 Thermodynamic Analysis

5.2.1 Obtaining the Melting Curve

Two methods for obtaining the thermodynamic melting curve from an FTIR temperature ramp series, an example of which is shown in Figure 5.1, will be discussed here. While other methods for analyzing the thermodynamics exist, the methods presented here cover a wide range of potential applications of interest. The first method is simpler and analyzes a specific frequency of interest, often the frequency with the maximum absorbance for a given feature. The more complex method considers a range of frequencies whether that includes multiple frequencies within a single peak or a wide range of frequencies spanning multiple features.

The first approach is to plot a signal trace at a specific frequency as a function of temperature. This is analogous to a method commonly used to study DNA thermodynamics in the UV where the melting curve is commonly the absorbance at approximately 260 nm as a function of temperature.[1–3] One key difference relative to UV is that the 260

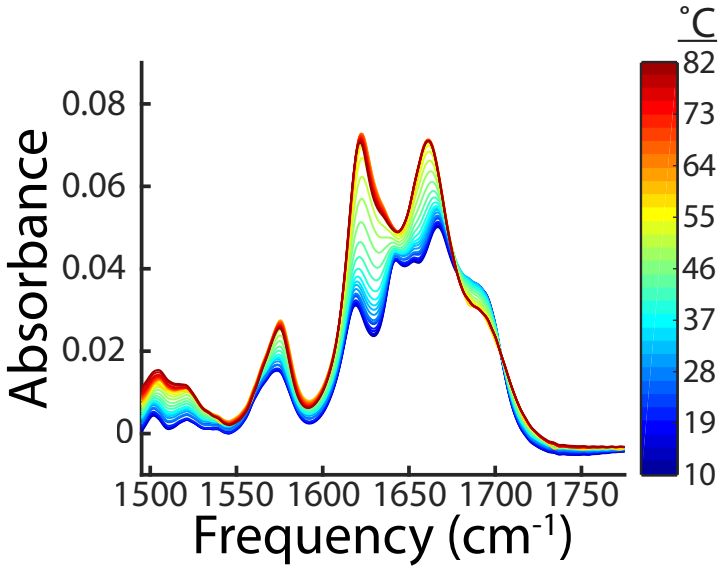

Figure 5.1: FTIR temperature ramp series for the sequence 5'-CATATATATG-3' from 10-82 °C with a spectrum taken every 3 °C.

nm peak contains contributions from all DNA base pairs whereas a single IR frequency in the range examined here will not necessarily contain strong contributions from all four bases and may only have contributions from a single base. This is a potential downside of analyzing a single frequency in the IR since, depending on the sequence composition, it might not accurately represent the overall melting of the duplex. Rather it results in a melting curve that is primarily reporting on either the loss of A:T base pairs or G:C base pairs. However, this can also be useful in some contexts. If the melting curves of A:T base pairs are distinguishable from G:C base pairs this can provide some insight into how the sequence melts. It is worth considering that this can be interpreted as a violation of the two-state approximation since it implies that different base pairs within the same sequence are not melting at the same time.

To help counteract the limitations of using single frequency slices it is advantageous to have a complimentary method that considers a wider frequency range. This could involve considering all frequencies within a single peak or incorporating multiple peaks across a larger frequency range. This can be accomplished through the use of singular value decomposition (SVD) which, while beneficial, does come with increased complexity. A brief

description of SVD specifically focused on its application in this work is provided here to explain the origin of the thermodynamic melting curves, for those interested in a complete description of SVD the minireview by Hendler and Shrager provides a good starting point for further reading.[4] Applying SVD starts by defining a matrix **A** which contains all of the spectra from the temperature ramp such that each row contains the spectra at a given temperature. As a result **A** has w columns, where w is the number of frequencies in the spectra and t rows where t is the number of temperatures spectra were collected at. SVD breaks down the matrix **A** into components according to

$$\mathbf{A}_{wxt} = \mathbf{U}_{wxs}\mathbf{S}_{sxs}\mathbf{V}^{\mathsf{T}}_{sxt} \tag{5.1}$$

where **U** is the set of orthanormal vectors of the column space of **A** which contain the spectral information for each component, **V** is the set of orthanormal vectors of the row space that contain the melting profile for each component and **S** contains the singular values for each component which provide information on the relative significance of each component. Computational languages, such as MATLAB, Python, and R, commonly have built in functions for performing SVD on a matrix. Since we are interested in the melting profile for the system we will focus on the vectors of **V** which contain this information. The first component corresponds to an average spectrum that is roughly static as a function of temperature. The second component contains the dominant spectral changes caused by increasing temperature, which are the result of duplex melting. For a system that is assumed to melt in a two-state fashion the only changes observed should be the dimer to monomer transition with no intermediates present. Thus, in theory, if the system is truly two-state the remaining components would be expected to be essentially noise, though in practice this is not the case. Regardless, within the two-state approximation made here we assume that the normalized second SVD component directly reports on the fraction of intact base pairs relative to the total number of base pairs. This means that the second

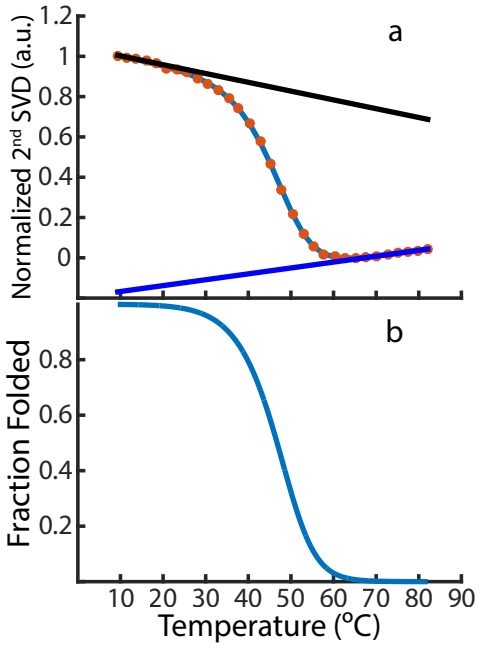

Figure 5.2: (a) Normalized second SVD component (orange dots) with the fit (light blue line) and both upper (black line) and lower (dark blue line) baselines and (b) the resulting fit for 5'-CATATATATATG-3'.

vector in the **V** matrix, after normalization, is taken as the melting curve to be fit. The second SVD component with the fit being applied to it is shown in Figure 5.2a.

5.2.2 Melting Curve Analysis

The ultimate goal of analyzing the melting curves is to obtain the thermodynamic parameters for the system and determine the monomer concentration at each temperature which is required for the analysis of the temperature-jump experiments. This is accomplished by fitting the melting curve assuming that the dissociation can be described as a two-state process during which all strands in the system are either a monomer or in a fully formed duplex where every base pair is intact.[5] All of the sequences studied here are self-complimentary and as such the analysis will be derived for this case. The non-self-complimentary case follows a similar derivation and the relevant equations are contained

in Appendix 5A for reference. Under this assumption the reaction can be written as

$$\mathrm{D} \underset{k_\mathrm{a}}{\overset{k_\mathrm{d}}{\rightleftharpoons}} 2\mathrm{M} \tag{5.2}$$

where D represents the fully formed dimer state and M represents the monomer state. The equilibrium constant for this reaction, which we define with respect to the dissociation, is

$$K = \frac{[\mathrm{M}]^2}{[\mathrm{D}]} = e^{\frac{-\Delta G^0_\mathrm{d}}{RT}} \tag{5.3}$$

where the square brackets indicate concentrations. We now define the total concentration of strands in the system, C_T, as

$$C_\mathrm{T} = 2[\mathrm{D}] + [\mathrm{M}] \tag{5.4}$$

Since each duplex contains two single strands the value of C_T is constant regardless of the ratio of monomers to dimers in the system.

We now define the fraction of intact base pairs for the system under the two-state assumption, where all intact base pairs are contained within a fully formed dimer, as

$$f_\mathrm{D} = \frac{2[\mathrm{D}]}{C_\mathrm{T}} \tag{5.5}$$

since the melting curve reflects the fraction of intact base pairs for the system at a given temperature this is the value that will be fit, which means an expression for f_D based on the thermodynamic parameters of interest is required. The first step is to derive an expression for the equilibrium constant, K, as a function of C_T and f_D. Using Equation 5.5 in combination with Equations 5.4 and 5.3 produces

$$K = \frac{2C_\mathrm{T}\left(1 - f_\mathrm{D}\right)^2}{f_\mathrm{D}} \tag{5.6}$$

This is useful since C_T is a known quantity and the thermodynamic parameters of interest

can all be determined based on their relations to the equilibrium constant. Solving for f_{D}, taking the negative root of the quadratic equation, results in the following expression for f_{D}

$$f_{\mathsf{D}} = \frac{4C_{\mathsf{T}} + K - \sqrt{K^2 + 8KC_{\mathsf{T}}}}{4C_{\mathsf{T}}} \tag{5.7}$$

We now need to determine the thermodynamic quantities of interest from the equilibrium constant. We start by relating the Gibbs free energy to the equilibrium constant

$$\Delta G = \Delta G^0 + RT \ln K \tag{5.8}$$

and noting that ΔG = 0 at equilibrium results in

$$\Delta G^0 = -RT \ln K \tag{5.9}$$

which we can break down into the enthalpy and entropy according to

$$\Delta G^0 = \Delta H^0 + T\Delta S^0 \tag{5.10}$$

At this point we need to formally define the melting temperature, T_{m}, within the context of this work. While many different definitions of the melting temperature exist, it will be considered here to be the temperature at which f_{D} = 0.5. The melting temperature will be utilized as a reference state for determining ΔH^0 and ΔS^0 at any temperature. To do this we invoke the definition of heat capacity

$$C_p = \frac{dH}{dT} = \frac{TdS}{dT} \tag{5.11}$$

and integrate with the bounds T_{m} and T, making the assumption that ΔC_p is constant resulting in

$$\Delta H^0(T) = \Delta H^0\left(T_{\mathsf{m}}\right) + \Delta C_p\left(T - T_{\mathsf{m}}\right) \tag{5.12}$$

$$\Delta S^0(T) = \Delta S^0\left(T_\mathsf{m}\right) + \Delta C_p \ln\left(\frac{T}{T_\mathsf{m}}\right) \tag{5.13}$$

Substituting Equations 5.12 and 5.13 into Equation 5.10 results in the final expression for ΔG^0

$$\Delta G^0(T) = \Delta H^0\left(T_\mathsf{m}\right) + T\Delta S^0\left(T_\mathsf{m}\right) + \Delta C_p\left(T - T_\mathsf{m} - T\ln\left(\frac{T}{T_\mathsf{m}}\right)\right) \tag{5.14}$$

where $\Delta H^0\left(T_\mathsf{m}\right)$ and ΔC_p are fit parameters. At this point all that remains is an expression for $\Delta S^0\left(T_\mathsf{m}\right)$.

Rather than having $\Delta S^0\left(T_\mathsf{m}\right)$ be a fit parameter it can be calculated from the fit parameters $\Delta H^0\left(T_\mathsf{m}\right)$ and T_m. It can be seen in Equation 5.6 that at T_m the equilibrium constant is equal to C_T. Setting the right hand side of Equation 5.9 equal to the right hand side of Equation 5.10 and solving for ΔS^0 at T_m provides the following expression for $\Delta S^0\left(T_\mathsf{m}\right)$ as a function of known quantities and the fit parameters $\Delta H^0\left(T_\mathsf{m}\right)$ and T_m

$$\Delta S^0\left(T_\mathsf{m}\right) = \frac{\Delta H^0\left(T_\mathsf{m}\right) + RT_\mathsf{m}\ln C_\mathsf{T}}{T_\mathsf{m}} \tag{5.15}$$

We can now determine f_D from the fit parameters $\Delta H^0\left(T_\mathsf{m}\right)$, T_m, and ΔC_p according to Equation 5.7 solving for K through Equations 5.14 and 5.9.

In practice the experimental melting curves have slanted baselines for both the upper and lower baselines that must be accounted for, which is done by fitting the second SVD component to the equation

$$V2 = f_\mathsf{D}\left(S_D - S_M\right) + S_M \tag{5.16}$$

where $V2$ is the resulting fit to the melting curve, which is referred to $V2$ since it often fits the vector in the matrix **V** that corresponds to the second SVD component. S_d and S_m are the upper (dimer) baseline and lower (monomer) baseline respectively. Both S_d and

S_{m} are linear and thus simply determined by two parameters, a slope and an intercept. The upper and lower baselines can be seen along with the second SVD component in Figure 5.2a while the resulting values of f_{D} can be seen in Figure 5.2b. This results in four additional fit parameters for a total of seven used to fit the melting curve.

5.3 Kinetic Analysis

5.3.1 Calculating Temperature-Jump Magnitude

The magnitude of the temperature change is determined by the transient response of the D_2O bend-libration combination band in the local oscillator (LO) spectrum. This primarily tracks the change in the solvent transmission as a result of the temperature perturbation and subsequent cooling back to the initial temperature. This transient response also provides the thermal profile of the sample over the course of the experiment. To determine the magnitude of the temperature jump the percent change in LO transmission is compared to a reference of the percent change in transmission in the same peak between FTIR D_2O spectra taken at known temperatures. The first step is to determine the percent change in LO transmission between a time point at the top of the temperature profile and the equilibrium initial temperature. Since the signal obtained from the temperature-jump experiment at each time point has already been referenced to the equilibrium signal, which is taken care of during the data processing as mentioned previously, the signal at a time point at the top of the plateau provides this necessary percent change in transmission. It simply requires a method for determining the magnitude of this signal at each frequency measured, which can be obtained through fitting the solvent response at each frequency to a known function.

The thermal profile of the solvent response is well fit to a stretched exponential which has the form

$$f(\tau) = Ce^{-(\frac{\tau}{t})^{\beta}} \tag{5.17}$$

where C is a scaling factor that accounts for the magnitude of the solvent response, τ is the time point in the experiment, t is the timescale for the temperature relaxation, and β controls the extent to which the function is stretched such that $0 < \beta \leq 1$ with a value of one resulting in a standard exponential function and smaller values increasing the degree of stretching. For the purposes of determining the magnitude of the temperature jump the scaling parameter is the parameter of interest since it reports on the magnitude of the percent change in transmission for the solvent response relative to the equilibrium transmission. An important note on fitting the solvent response is that early time points need to be removed due to the effects of cavitation waves that form as a result of the rapid heating of the sample. These pressure waves also affect the transmission of the sample, which is observed in the LO trace as a function time for a single frequency as a sharp rise that can be observed in the vicinity of 100 ns. To avoid any artefacts in the temperature calculation that could arise from this the fits to the solvent response do not incorporate early time points, a reasonable cut off point is around 200 ns, though this can be adjusted sample to sample as necessary.

The magnitude of the thermal response from the stretched exponential fit can now be compared to a known standard, which is the absorbance of D_2O as a function of temperature obtained from FTIR. The percent change in transmission between two temperatures T_f and T_i is obtained via the equation

$$\Delta\mathrm{trans}\,(\%) = 100\left(\frac{10^{-A(T_\mathrm{f})} - 10^{-A(T_\mathrm{i})}}{10^{-A(T_\mathrm{i})}}\right) \tag{5.18}$$

where A is the IR absorbance obtained from linear FTIR experiments as a function of temperature. The D_2O reference spectrum is taken at one degree temperature steps and calibrated to ensure it serves as an accurate reference. Since the initial temperature for the temperature-jump experiment is known all that remains is to determine the final temperature that results in the percent change in transmittance, as calculated by Equation

5.18, closest to the parameter C determined from the fit to the LO solvent response as determined by Equation 5.17. This is then carried out for every frequency measured in the temperature-jump experiment resulting in a T_f value for each frequency. To minimize the effect of noise on the calculation these values are averaged together to produce the final value of T_f.

5.3.2 Time Domain Analysis

Analyzing the temperature-jump data in the time domain is the first step in the analysis since it simply requires taking frequency slices of the transient HDVE spectrum at frequencies of interest. Figure 5.3a shows an example of a transient HDVE spectrum that shows where frequency slices are taken to generate time traces that show the signal response for the guanine and adenine ring modes. As mentioned when assigning the peaks in the IR DNA spectrum in Chapter 1 these peaks are isolated from the signals generated by other base pairs resulting in the ability to independently track the response, due to the temperature perturbation, of A:T and G:C base pairs. Taking the frequency slice results in the time domain traces seen in Figure 5.3c.

Extracting the relevant kinetic parameters is done through fitting the time traces, which enforces a functional form onto the data. Not only does the fit provide the timescales but the functional form that fits best provides the first piece of mechanistic insight into the system. In this section we will discuss the different functional forms that are used to fit the kinetic traces and simple interpretations of the insights that can be drawn from them.[6] This section is intended to orient the reader to how to generally interpret these results; a more complete analysis of the kinetics, dynamics, and mechanisms for all of the different samples examined will be included in later chapters.

The signal rise in the time trace associated with DNA that dissociates in a two-step all or nothing process should be well fit by a single exponential function whereas more complicated reactions, potentially due to multiple processes occurring simultaneously with

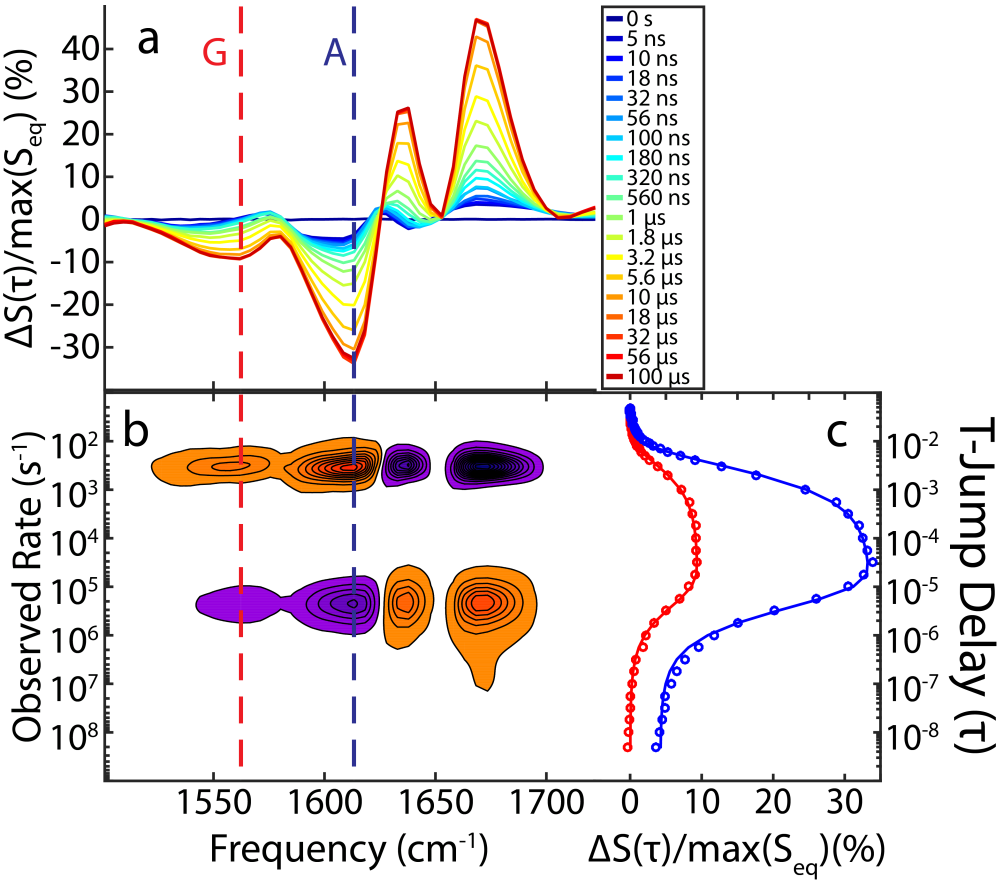

Figure 5.3: Results from the 5'-CATATATG-3' 41-54 °C temperature-jump. The transient HDVE spectra for times up to 100 µs (a) and the corresponding rate map (b). The dashed lines correspond to the frequencies with the maximum signal for the guanine (red) and adenine (blue) ring mode excited state absorptions, which are the time traces plotted in (c).

different timescales or different mechanisms entirely, will deviate from an exponential fit.[6] To fit the trace it is broken down into two sections the initial rise that is caused by the loss of base pairs in the duplex followed by the signal decay that occurs at longer times due to rehybridization as the temperature of the system returns to the initial temperature. The rehybridization portion of the signal trace is well fit to a stretched exponential function since the temperature re-equilbration, which is the dominant factor driving the rehybridization, is well fit to a stretched exponential as mentioned previously. While in theory this could allow the direct observation of the hybridization reaction the analysis of this region is significantly complicated by the fact that the temperature of the sample is evolving during this portion

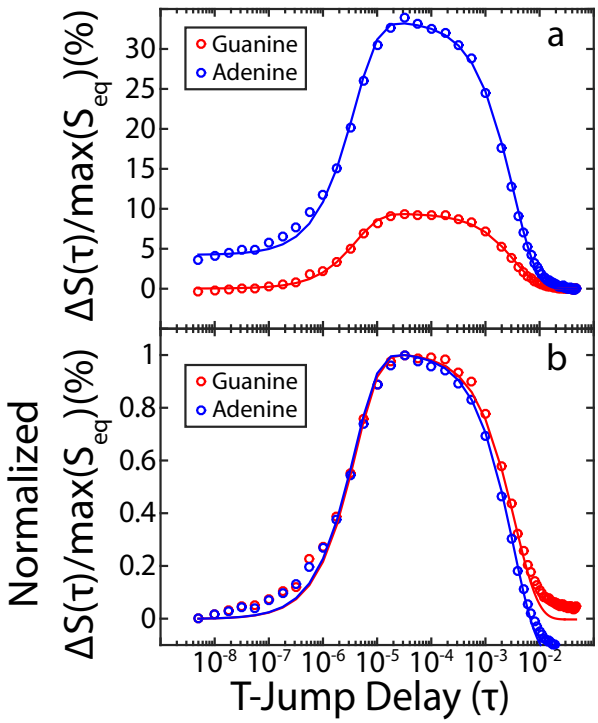

Figure 5.4: (a) Time trace and fit for the adenine and guanine ring mode response to the 41-54 °C temperature-jump for 5'-CATATATG-3'. (b) Normalized time trace and fits such that the first time point lies at zero on the y-axis and the largest signal is equal to one.

of the experiment. Since the hybridization of base pairs is convolved with the temperature relaxation it is extremely difficult to extract reliable information out of this portion of the data.

This leaves the early time portion of the data that observes the dissociation of the duplex in response to the temperature jump. The simplest case observed is that of a standard two-state all or nothing dissociation which is best demonstrated by shorter sequences with G:C end caps, an example of which is shown in both Figure 5.3c and 5.4a. Evidence for the two-state mechanism appears in a few different forms in Figure 5.4. The first is that both the adenine and the guanine response are well fit to a single exponential rise, best seen in the unaltered data in Figure 5.4a indicating that the loss of base pairing is occurring as a two-step process. Additionally the timescales for the rise of the guanine and adenine base pairs are nearly identical, as can be seen in the normalized data in Figure 5.4b, indicating that A:T and G:C base pairs are lost at essentially the same time

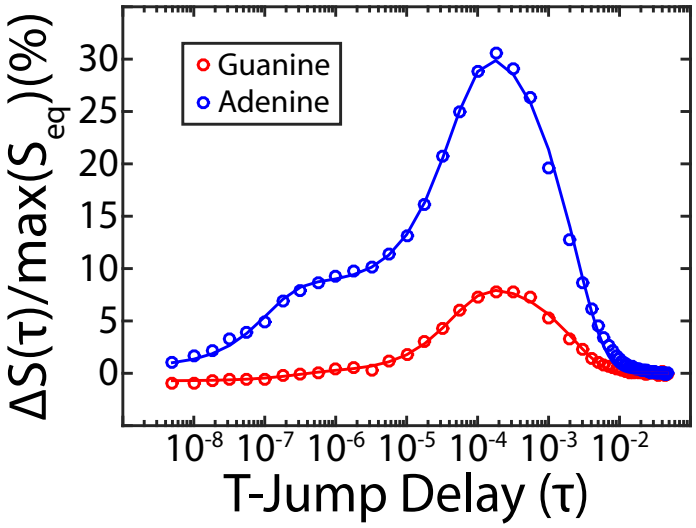

Figure 5.5: (a) Adenine and guanine ring mode time traces for the 5'-ATATGCATAT-3' 46-60.2 °C temperature-jump. The signal rise is fit to a biexponential function for adenine and a single exponential for guanine. Adapted from Ref 7.

which supports the all or nothing dissociation picture.

Another common form utilized to fit the rise of the signal is the use of two exponentials. This has been primarily in sequences with G:C base pairs in the middle flanked by A:T regions.[7] As seen in Figure 5.5 the guanine response follows a standard single exponential rise while the adenine response is best fit by a biexponential function. This functional form can be interpreted as two resolvable processes each occurring in a relatively two-state manner with a single timescale for each process. In the context of the sequence mentioned here this has been interpreted as fast fraying of terminal A:T base pairs, a conclusion strongly supported by the contrast between the biexponential rise in the adenine signal and the single exponential rise of the guanine signal. Additionally the timescale for the guanine response is in reasonable agreement with the second timescale in the adenine response meaning those two processes occur at roughly the same time.[7]

The third common functional form for the rise in signal is a stretched exponential. This can be observed in Figure 5.6 which shows adenine ring mode time domain traces for sequences of two different lengths. In Figure 5.6 the dashed line is the stretched exponential fit to each time trace and the solid line is the exponential fit. The shorter

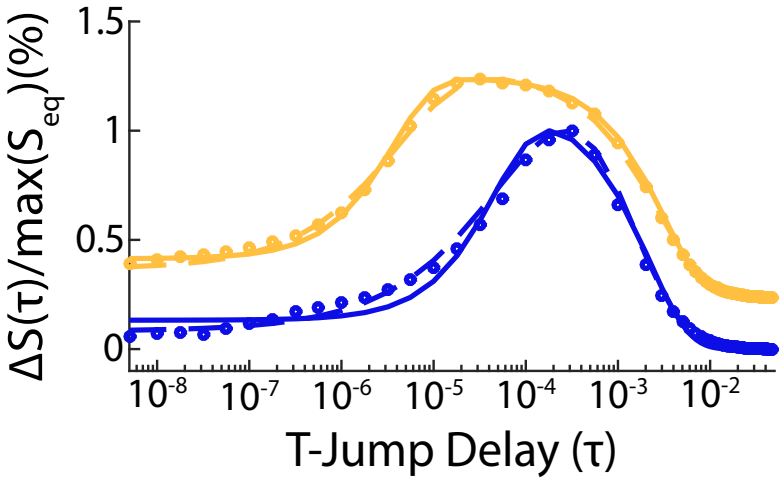

Figure 5.6: Comparison of the adenine ring mode time traces (circles), exponential fits (solid lines), and stretched exponential fits (dashed lines) for temperature-jumps on the sequences 5'-CATATATG-3' (yellow) and 5'-CATATATATATATG-3' (blue) with a final temperature of approximately 53 °C. Both traces are normalized to their maximum value and offset by a value of 0.4 for clarity.

sequence is well fit to the single exponential rise, such that the stretched exponential fit is not a significant deviation from the exponential fit. Whereas for the longer sequence the stretched exponential fit demonstrates a significant improvement relative to the single exponential fit. A stretched exponential is most commonly interpreted to be the result of having a mixture of different processes occurring with different timescales resulting in the signal becoming stretched out in time. There are numerous reasons why processes with different timescales could be occurring simultaneously. One possible example is a system with a unified overall mechanism but a heterogeneous initial population. Another example is a system with a homogeneous initial population but the reaction proceeds via a downhill mechanism through a continuous ensemble of configurations. Both of these examples lead to a broad distribution of timescales and the stretched exponential form as a result.[6]

5.3.3 Rate Domain Analysis

An alternative method to analyzing the kinetic data in the time domain is to transform the data into the rate domain and examine the rate distribution. An example rate map, corresponding to the transient HDVE spectrum shown in Figure 5.3a, is shown in Figure 5.3b. These plots are oriented such that faster rates are at the bottom of the y-axis which results in time progressing from the bottom of the plot to the top. As a result the peaks corresponding to the dissociation as a result of the temperature perturbation occur below the peaks for the association. Figures 5.3b and 5.3c are oriented such that their y-axis are aligned which can help visualize the relationship between viewing the data in the time domain and the rate domain. The main advantage of utilizing this representation is that a functional form does not need to be assumed to extract the kinetic information from the system. It can instead be extracted directly from the rate map. While the identity of the functional form of the time domain plot does provide useful information about the system, having to enforce a functional form can impact the kinetic parameters obtained from the fit. Additionally, that same information is still observed in the rate domain by examining the shape of the peaks in the rate map and in particular the evolution of the peaks along

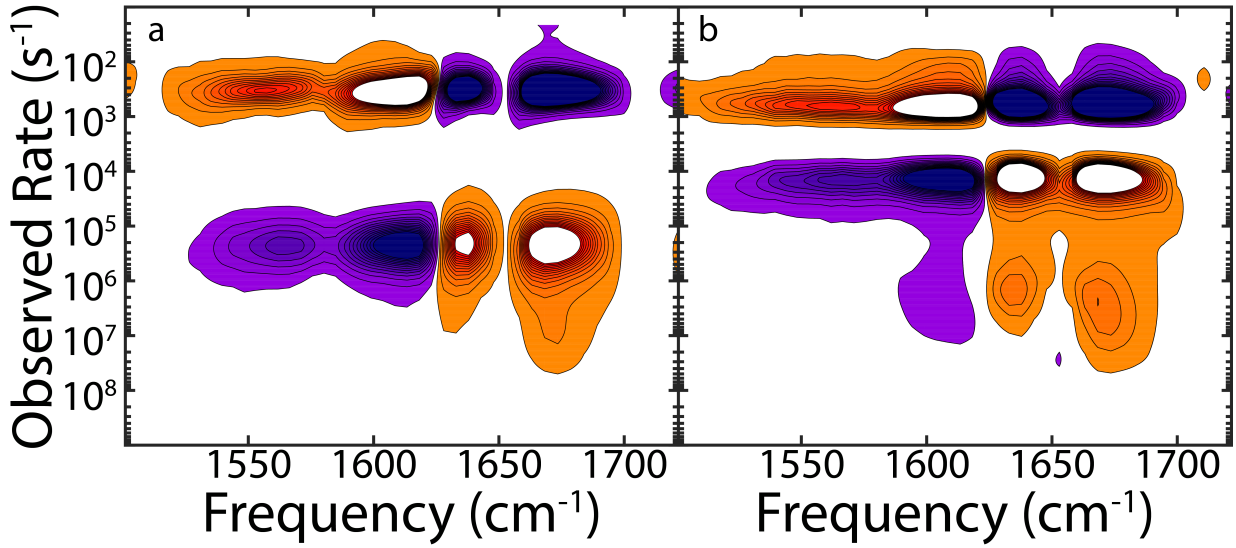

Figure 5.7: Rate maps displayed at 4x magnification for the (a) 5'-CATATATG-3' 41-54 °C and (b) 5'-CATATATATATATG-3' 40-54.8 °C temperature-jumps.

the y-axis, as seen in Figure 5.7. Figure 5.7b contains a tail going out to faster rates for three of the four peaks indicating the presence of faster rates which are not observed in Figure 5.7a. This is analogous to the longer sequence being better fit by the stretched exponential in Figure 5.6. Before discussing the method for extracting the observed rate from the rate domain, that is analogous to the observed time constant extracted from the time domain, the method for transforming the transient HDVE spectrum into the rate map will be briefly discussed.

The transient HDVE spectrum is transformed into the rate domain using a maximum entropy implementation of a numerical inverse Laplace transform. The method has been thoroughly described elsewhere[8,9] and as such the discussion of the method here will focus heavily on its application and implementation.

The goal of the method is to obtain the distribution of rates denoted $g(\lambda)$ that satisfies the equation

$$I(t) = \int_0^\infty g(\lambda) e^{-\lambda t} d\lambda \tag{5.19}$$

which is the definition of a Laplace transform.[10] The Laplace transform is an integral transform that takes a function of a real variable and transforms it into a function of a complex variable. In our case $I(t)$ is obtained from experiment so the corresponding function $g(\lambda)$ needs to be found. This requires a numerical inverse Laplace transform to be carried out on $I(t)$ which contains experimental noise making this process an ill-conditioned problem. To accomplish this we first note that since the data covers many orders of magnitude it makes more sense to work in the log space. We also rewrite the integral for the Laplace transform to be a finite sum resulting in

$$I(t) = \sum_{j=1}^{N} f_j e^{-\lambda_j t} \Delta \left(\log \lambda_j \right) \tag{5.20}$$

where N is the number of data points taken in the experiment. We now turn to the maximum entropy method (MEM) for determining f_j, ultimately by looking to maximize the

parameter Q. Before introducing Q two terms must be introduced, the information entropy and an expression for the normalized mean square error between the model and the data. The information entropy is defined as

$$S = -\sum_{j=1}^{N} f_j \left[\ln\left(\frac{f_j}{F_j}\right) - 1 \right] \tag{5.21}$$

where F_j is called the prior distribution and can be used to incorporate any known information about the rates. In our case we presume no known knowledge about the rate distribution and give each F_j a value of $1e^{-4}$ which is also the starting guess for f_j. The expression for the normalized mean square error between the model and the data is given by

$$\chi^2 = \frac{1}{N} \sum_{j=1}^{N} \frac{\left[I_f\left(t_j\right) - I_e\left(t_j\right)\right]^2}{\sigma_j^2} \tag{5.22}$$

where the subscripts f and e denote the fit value and the experimental value respectively and the term σ_j is the noise variance associated with the jth data point which is supplied by the user. The value of $I_e\left(t_j\right)$ is taken from experiment and the value of $I_f\left(t_j\right)$ is determined for each iteration by Equation 5.20. We can now define Q as

$$Q = S - \eta\chi^2 \tag{5.23}$$

where η is the Lagrange multiplier that is selected to satisfy the constraint that χ^2 = 1. In practice the optimization algorithm that determines the values of f_j will be minimizing the function

$$-Q = \eta\chi^2 - S \tag{5.24}$$

where η is initially set to the mean of σ_j and Q is optimized for this value. The value of η is then increased and the process repeats itself until it terminates when the value of χ^2 is one and returns the value of f_j. An important observation about the MEM in-

verse Laplace transform method is that it must be done individually for each frequency measured in the experiment and that the calculation of the rate distribution for a given frequency is independent from all other frequencies. This makes the code for this method an excellent candidate for parallelization. Utilizing computing nodes with a large number of workers significantly increases the efficiency which makes using such a system highly recommended.

Once the rate maps have been determined using the MEM inverse Laplace transform method the observed rate constant from the process can be determined. The observed rate constant is determined from a weighted average across the main dissociation peak for each feature. This method has two main advantages relative to determining the observed rate constant via fitting to signal traces in the time domain. The first is that there is no assumed functional form of the kinetic response which means that the rate can be determined without enforcing a particular mechanistic description on the system. The second advantage is that it considers the rate across the entire peak rather than just looking at a single frequency. This provides a more consistent observed rate constant because it can account for some degree of experimental noise. It also provides an easy method for estimating the error in the observed rate by considering the amplitude weighted standard deviation in the rate across the peak.

5.3.4 Two-State Kinetics

The observed rate constant for a transient experiment is a convolution of both the forward and backward rates for the system. Deconvolving the observed rate constant into the forward and backward rates for a second order process is a complex problem to solve analytically. However, this problem can be simplified significantly by approximating our experiment as a small amplitude perturbation. This requires making the assumption that the system is at equilibrium and the population changes that occur as a result of the temperature perturbation are relatively small.

The derivation of an expression for the observed rate constant in terms of the association and dissociation rates for the reaction given by Equation 5.2 starts with deriving an expression for the equilibrium concentration of the monomer and dimer prior to the introduction of the temperature perturbation. We start with the time derivatives for the monomer and dimer concentrations[11]

$$\frac{d\,[\mathsf{M}]}{dt} = [\mathsf{D}]\,k_{\mathsf{d}} - [\mathsf{M}]^2\,k_{\mathsf{a}} \tag{5.25}$$

$$\frac{d\,[\mathsf{D}]}{dt} = [\mathsf{M}]^2\,k_{\mathsf{a}} - [\mathsf{D}]\,k_{\mathsf{d}} \tag{5.26}$$

and note that at equilibrium both of them are equal to zero. From this, the definition of the total strand concentration given by Equation 5.4, and the definition of the equilibrium constant $K = \frac{k_{\mathsf{d}}}{k_{\mathsf{a}}}$ we can derive the following expressions for the monomer and dimer concentrations at equilibrium

$$[\mathsf{M}]_{\mathsf{eq}} = \frac{1}{4}\left(-K + \sqrt{K^2 + 8KC_{\mathsf{T}}}\right) \tag{5.27}$$

$$[\mathsf{D}]_{\mathsf{eq}} = -\frac{1}{8}\left(-K + 4C_{\mathsf{T}} + \sqrt{K^2 - 8KC_{\mathsf{T}}}\right) \tag{5.28}$$

We now introduce the small perturbation assumption and define the concentrations for the monomer and dimer after the perturbation, which alters the populations only slightly away from equilibrium, as

$$[\mathsf{M}] = [\mathsf{M}]_{\mathsf{eq}} + [\mathsf{m}] \tag{5.29}$$

$$[\mathsf{D}] = [\mathsf{D}]_{\mathsf{eq}} + [\mathsf{d}] \tag{5.30}$$

where the lower case denotes the small changes in concentration that occur due to the perturbation. We now want to determine $\frac{d[\mathsf{d}]}{dt}$ which can then be used to derive the observed rate constant. The same method can be done using $\frac{d[\mathsf{m}]}{dt}$ and it will produce the same expression for the observed rate constant so it will not be explicitly shown here. We

start by adding together the time derivatives for $[\mathrm{D}]_{\mathrm{eq}}$ and $[\mathrm{d}]$ giving us the expression

$$\frac{d\,[\mathrm{D}]_{\mathrm{eq}}}{dt} + \frac{d[\mathrm{d}]}{dt} = k_{\mathrm{a}}\left([\mathrm{M}]_{\mathrm{eq}} + [\mathrm{m}]\right)^2 - k_{\mathrm{d}}\left([\mathrm{D}]_{\mathrm{eq}} + [\mathrm{d}]\right) \tag{5.31}$$

Incorporating Equation 5.26 yields

$$\frac{d[\mathrm{d}]}{dt} = 2k_{\mathrm{a}}\,[\mathrm{M}]_{\mathrm{eq}}\,[\mathrm{m}] - k_{\mathrm{d}}[\mathrm{d}] + k_{\mathrm{a}}[\mathrm{m}]^2 \tag{5.32}$$

Knowing that C_{T} cannot change as a result of the perturbation means that $[\mathrm{m}] = -2[\mathrm{d}]$ so by substitution we get

$$\frac{d[\mathrm{d}]}{dt} = -4k_{\mathrm{a}}\,[\mathrm{M}]_{\mathrm{eq}}\,[\mathrm{d}] - k_{\mathrm{d}}[\mathrm{d}] + 4k_{\mathrm{a}}[\mathrm{d}]^2 \tag{5.33}$$

At this point we drop the second order term due to our assumption that our perturbation only slightly changes the overall concentration. Solving the differential equation that remains after the second order term is dropped gives us

$$[\mathrm{d}](t) = Ce^{-\left(4k_{\mathrm{a}}[\mathrm{M}]_{\mathrm{eq}} + k_{\mathrm{d}}\right)t} \tag{5.34}$$

where C is a constant. Which gives us the observed rate constant[11]

$$k_{\mathrm{obs}} = 4k_{\mathrm{a}}\,[\mathrm{M}]_{\mathrm{eq}} + k_{\mathrm{d}} \tag{5.35}$$

Using Equations 5.3, 5.4, and 5.5 Equation 5.35 can be rewritten as

$$k_{\mathrm{a}} = k_{\mathrm{obs}}\left(K_{d,f} + 4C_{\mathrm{T}}\left(1 - f_{\mathrm{D,i}}\right)\right)^{-1} \tag{5.36}$$

where $f_{\mathrm{D,i}}$ denotes the fraction of molecules in the duplex state at the initial temperature, and $K_{\mathrm{d,f}}$ is the dissociation equilibrium constant at the final temperature T_{f}.

5.4 References

1. Mergny, J.-L.; Lacroix, L. Analysis of Thermal Melting Curves. *Oligonucleotides* **2003**, *13*, 515–537.

2. Schwinefus, J. J.; Menssen, R. J.; Kohler, J. M.; Schmidt, E. C.; Thomas, A. L. Quantifying the Temperature Dependence of Glycine—Betaine RNA Duplex Destabilization. *Biochemistry* **2013**, *52*, 9339–9346.

3. Schwinefus, J. J.; Baka, N. L.; Modi, K.; Billmeyer, K. N.; Lu, S.; Haase, L. R.; Menssen, R. J. L-Proline and RNA Duplex *m*-Value Temperature Dependence. *J. Phys. Chem. B* **2017**, *121*, 7247–7255.

4. Hendler, R. W.; Shrager, R. I. Deconvolutions Based on Singular Value Decomposition and the Pseudoinverse: A Guide for Beginners. *J. Biochem. Biophys. Methods* **1994**, *28*, 1–33.

5. Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry: Part III: The Behavior of Biological Macromolecules*; W.H. Freeman and Company: San Francisco, 1980.

6. Sabelko, J.; Ervin, J.; Gruebele, M. Observation of Strange Kinetics in Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 6031–6036.

7. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved Through Infrared Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138*, 11792–11801.

8. Sanstead, P. J. C. Investigation of DNA Dehybridization through Steady-State and Transient Temperature-Jump Nonlinear Infrared Spectroscopy. Ph.D. thesis, The University of Chicago, 2018.

9. Kumar, A. T. N.; Zhu, L.; Christian, J. F.; Demidov, A. A.; Champion, P. M. On the Rate Distribution Analysis of Kinetic Data Using the Maximum Entropy Method: Applications to Myoglobin Relaxation on the Nanosecond and Femtosecond Timescales. *J. Phys. Chem. B* **2001**, *105*, 7847–7856.

10. Logan, J. D. *Applied Mathematics*; John Wiley & Sons: Hoboken, 2006.

11. Nölting, B. *Protein Folding Kinetics: Biophysical Methods*; Springer: Berlin, 2006.

Appendix 5A: Equations for the Thermodynamic and Kinetic Analysis of Non-Self-Complimentary Sequences

In this appendix the equations for analyzing non-self-complimentary sequences are provided. Since the derivation for both the thermodynamics and kinetics follow the same general procedure as the self-complimentary case the derivations will not be provided and instead the equations that differ from the self-complimentary analysis are provided. Again following the two-state assumption for the reaction

$$\mathrm{D} \underset{k_{\mathrm{a}}}{\overset{k_{\mathrm{d}}}{\rightleftharpoons}} \mathrm{M}_1 + \mathrm{M}_2 \tag{5.37}$$

where M_1 and M_2 are the two monomers. For the purposes of this analysis we will make the assumption that $[\mathrm{M}_1] = [\mathrm{M}_2]$ such that if the system is fully duplexed there are no remaining unpaired monomers. In this case the equilibrium constant is given by

$$K = \frac{[\mathrm{M}_1]\,[\mathrm{M}_2]}{[\mathrm{D}]} = e^{\frac{-\Delta G^0_{\mathrm{d}}}{RT}} \tag{5.38}$$

In this case C_{T} is defined as as

$$C_{\mathrm{T}} = 2[\mathrm{D}] + [\mathrm{M}_1] + [\mathrm{M}_2] = 2[\mathrm{D}] + 2[\mathrm{M}] \tag{5.39}$$

where $[\mathrm{M}] = [\mathrm{M}_1] = [\mathrm{M}_2]$.

The fraction of intact base pairs for the system under the two-state assumption is the same as the self-complimentary case given in Equation 5.5. The expression for the equilibrium constant as a function of C_{T} and f_{D} in the non-self-complimentary case is

$$K = \frac{C_{\mathrm{T}}\,(1 - f_{\mathrm{D}})^2}{2 f_{\mathrm{D}}} \tag{5.40}$$

From this the expression for f_{D} is determined to be

$$f_{\mathsf{D}} = \frac{C_{\mathsf{T}} + K - \sqrt{K^2 + 2KC_{\mathsf{T}}}}{C_{\mathsf{T}}} \tag{5.41}$$

The expressions for $\Delta H^0(T)$, $\Delta S^0(T)$, and $\Delta G^0(T)$ are unchanged from the self-complimentary case leading to the expression of $\Delta G^0(T)$ given in Equation 5.14.

All that remains is to determine the expression for $\Delta S^0(T_{\mathsf{m}})$. First we must note that based on the definition that $f_{\mathsf{D}} = 0.5$ at T_{m} and Equation 5.40 the equilibrium constant at T_{m} is given by $K(T_{\mathsf{m}}) = \frac{C_{\mathsf{T}}}{4}$. Using this we get

$$\Delta S^0(T_{\mathsf{m}}) = \frac{\Delta H^0(T_{\mathsf{m}}) + RT_{\mathsf{m}} \ln\left(\frac{C_{\mathsf{T}}}{4}\right)}{T_{\mathsf{m}}} \tag{5.42}$$

Following a similar derivation to the self-complimentary case the equation for the observed rate constant for a two-state system assuming a small amplitude perturbation is given by

$$k_{\mathsf{obs}} = k_{\mathsf{d}} + k_{\mathsf{a}}\left([\mathsf{M}_1]_{\mathsf{eq}} + [\mathsf{M}_2]_{\mathsf{eq}}\right) \tag{5.43}$$

CHAPTER 6

LENGTH-DEPENDENT MELTING KINETICS OF SHORT DNA OLIGONUCLEOTIDES USING TEMPERATURE-JUMP IR SPECTROSCOPY

Portions of this chapter have been published and are reprinted with permission from:

Menssen, R. J.; Tokmakoff, A. Length-Dependent Melting Kinetics of Short DNA Oligonucleotides Using Temperature-Jump IR Spectroscopy. *J. Phys. Chem. B* **2019**, *123*, 756-767.

6.1 Abstract

In this work we utilize Fourier transform infrared (FTIR) and temperature-jump (T-jump) IR spectroscopy to investigate the melting thermodynamics and kinetics of a series of five DNA sequences ranging from 6 to 14 base pairs long. IR spectroscopy is well suited for the study of DNA because of its ability to distinguish base specific information and the nanosecond time resolution of the T-jump apparatus can access the relevant range of kinetics. Eyring analysis of a two-state model examines both the activation enthalpy and entropy providing new insight into the energetic driving forces and physical processes behind the association and dissociation while also helping to clarify the commonly observed negative activation energy. Global analysis of the thermodynamic and kinetic data applying a linear dependence of activation barriers on oligo length provides a holistic result by producing reasonable agreement between our data and existing nearest neighbor thermodynamic parameters blending the experimental results with established predictive models. By studying the trends in the thermodynamics and kinetics as a function of length this work demonstrates a direct correlation between the effects additional dinucleotides

have on the kinetics and the nearest neighbor parameters for those dinucleotides. This result further supports the development of a kinetic analog to the thermodynamic nearest neighbor parameters.

6.2 Introduction

One of the main goals of this research is to begin to break down and understand the different variables that impact DNA association and dissociation mechanisms. Numerous variables including, but not limited to, length, temperature, sequence, and salt concentration are known to impact both the kinetics and dynamics of DNA reactions and underlying driving forces.[1–15] While it may seem basic, there is a surprising amount that is not understood about the fundamentals of DNA association and dissociation reactions. Historically some of this has been due to available techniques as only more recently have modern computational techniques developed methods that are able to produce detailed simulations of these reactions. Experimental techniques continue to lag behind the computational methods as simulations have predicted a number of rich and interesting dynamics that experiments have yet to observe.[1,3]

Many experimental studies remain focused on the kinetics of the reactions looking at aspects such as the rates and energetic barriers for the processes.[2,16] While these are interesting and we ourselves also study the kinetics we feel that the real prize is understanding the dynamics of the process. The dynamics focus more on how the reaction proceeds looking at the mechanisms of all processes that occur, not just the overall monomer to dimer reaction. Additional processes such as the diffusion to capture of two monomers, the fluctuations that occur during critical nucleus formation, and fast dynamics during the dissociation such as fraying and bubble formation are all of significant interest to our research group. The use of ultrafast IR spectroscopy is a perfect match for understanding these dynamical questions due to its ability to provide greater structural resolution compared to

other label free techniques.

These topics motivated the study of the length series, the experimental studies that are described in this chapter. The effect of length on the association and dissociation of DNA has not received significant attention from experimental studies using modern techniques. Additionally, with respect to canonical DNA duplex dynamics and kinetics our group has focused more on short DNA oligos that are well described by a two-state mechanism. However, to start to access the rich and complicated dynamics we are interested in, our focus needs to shift to longer sequences where these dynamics are known to occur.[17,18] Examining the length series bridges this gap as, using similar sequence construction, we can examine how the dynamics and kinetics are affected by increasing length. An additional advantage of the length series is, due to the kinetics and dynamics evolving with length, the data set not only motivated the development of the kinetic model but also served as a useful core data set for comparison during development. While there is significant work to be done to fully understand the rich and complex dynamics and kinetics of DNA association and dissociation, focusing on a single variable, length, provided an approachable way to build the necessary tools, both experimental and computational. These tools allow us to start to pull apart and understand the different variables that influence DNA association and dissociation and dive into the fundamental energetic driving forces, mechanism, and dynamics that occur.

6.3 Experimental Methods

6.3.1 Sample Preparation

DNA oligonucleotides with the sequence 5'-C(AT)$_n$G-3' (n = 2-6) and lengths L = 6-14 (i.e. number of base pairs in the single strands) were purchased from Integrated DNA Technologies (IDT) and purified using Amicon Ultra 3 kDa centrifugal filters or Sartorius Setim Biotech Vivaspin 2 2 kDa centrifugal filters depending on the sample molecular

weight. To prepare for IR spectroscopy, DNA samples were then H/D exchanged in D_2O (Cambridge Isotopes) and lyophilized. Samples for both the FTIR and T-jump were measured in a deuterated 50 mM sodium phosphate buffer with an additional 240 mM NaCl and 18 mM $MgCl_2$ and a pH of 7.2. All samples were run at a concentration of 2 mM and a NanoDrop UV/vis spectrometer (Thermo Scientific) was used to ensure sample concentration consistency. Prior to measurement, samples were annealed by heating to 95 °C and gradually cooling to room temperature for 15 minutes. For both the FTIR and T-jump measurements samples were placed between two 1 mm CaF_2 windows with a 50 μm path length formed with a Teflon spacer. The sample was then mounted in a home-built brass sample cell that was temperature controlled by a recirculating chiller.

6.3.2 Temperature Ramp FTIR

For the temperature ramp FTIR measurements the chiller was ramped from 0 to 96 °C in 3 °C steps with a 60 second equilibration time at each point. The sample temperature was calibrated by attaching a thermocouple to the CaF_2 window to determine the temperature at the sample relative to the chiller set point. Spectra were recorded on a Bruker Tensor 27 FTIR spectrometer. The raw FTIR spectra where then processed by subtracting off the D_2O and HOD spectra.

6.3.3 Temperature-Jump Measurements

T-jump kinetic measurements were made with an ultrafast nonlinear IR spectrometer with a center wavelength of 6.2 μm and 1600 cm^{-1} bandwidth synchronized electronically to a nanosecond T-jump laser. The spectrometer and data acquisition methods have been described in detail elsewhere.[19,20] Briefly, the spectrometer collects a heterodyne detected vibrational echo (HDVE) spectrum. The real part of the HDVE spectrum is closely related to a transient absorption spectrum and can be read in the same way. The positive and negative signals are the ground state bleach (GSB) and excited state absorption

(ESA) which originate from 0 to 1 and 1 to 2 vibrational transitions respectively. All transient T-jump spectra we report are differences between the HDVE spectrum measured at a given delay time after a T-jump pulse (τ), and the equilibrium HDVE spectrum acquired at the initial temperature prior to the T-jump (S_0) : $\Delta S(\omega, \tau) = S(\omega, \tau) - S_0(\omega)$.

The T-jump laser was used to jump the sample temperature by approximately 15 °C from an initial equilibrium temperature T_i to a final temperature T_f within ≈10 ns. This transient temperature jump is maintained until the sample thermally re-equilibrates on a time-scale of ≈2 ms. The initial temperatures were selected to sample a minimum of four temperatures across the melting transition of each oligo while ensuring the kinetics fall within the window between 10 ns and 2 ms that the instrument can observe. T_i was maintained by the chiller connected to the brass sample cell. T_f was determined by comparing the transient response of the D_2O solvent as a result of the temperature-jump pulse to changes in intensity observed in equilibrium FTIR measurements of D_2O at different temperatures.

6.4 Results and Discussion

6.4.1 Equilibrium Melting

The self-complementary sequences utilized in this study were selected to ensure they followed simple two-state melting behavior, by choosing relatively short lengths that would be unlikely to form bubbles or hairpins and putting G:C base pairs at each end which limits the likelihood of terminal fraying.[4] The melting profile and underlying thermodynamics of the monomer-dimer transition were determined from temperature-dependent FTIR measurements between 1500 cm^{-1} and 1750 cm^{-1}. An example is shown in Figure 6.1a. The DNA vibrational modes in this frequency range contain contributions from both in-plane ring vibrations, predominately at frequencies below 1650 cm^{-1}, and carbonyl stretches, predominately at frequencies above 1650 cm^{-1}, that are sensitive to DNA hydrogen bond-

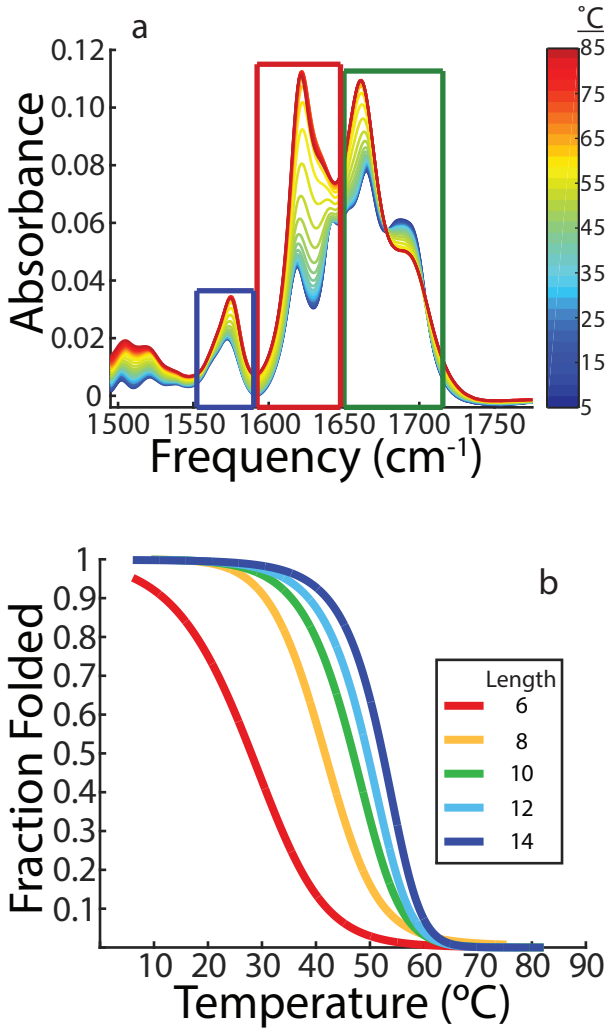

Figure 6.1: (a) FTIR temperature ramp for 5'-CATATATATATATG-3'. The boxes highlight the peaks for the guanine ring mode (blue), adenine ring mode (red), and the overlapping region (green). (b) DNA melting curves obtained from a fit to the second SVD component of the temperature dependent FTIR data.

ing and base stacking interactions.[21,22] Other than the shoulder at 1690 cm^{-1} the features in this frequency range are suppressed by the hybridization of DNA bases resulting in an increase in signal as the DNA double helix melts. We focus on peaks at 1556 cm^{-1} and 1625 cm^{-1} which arise from guanine and adenine ring mode vibrations respectively. These peaks are used to independently resolve the loss of G:C and A:T base pairs. Above 1630 cm^{-1} the spectrum becomes more congested with overlapping contributions for thymine, guanine, and cytosine.

To determine the melting curve from the global changes in the spectrum, singular value decomposition (SVD) analysis was applied to the FTIR temperature series.[23,24] The second SVD component contains the dominant spectral changes caused by increasing temperature, which are the result of duplex melting. The dimer is assumed to melt in a two-state fashion so the only changes observed should be the dimer to monomer transition with no intermediates present. This leads us to assume the normalized second SVD component directly reports on the fraction of intact base pairs relative to the total number of base pairs.

The resulting melting curves are shown in Figure 6.1b. There are two observations to make with the melting curves. The first is that the curves shift to higher temperature as length increases and the amount that the curves shift decreases with increasing length. The second observation is that under our experimental conditions the two shortest sequences do not have a full low temperature baseline. For L = 8 the baseline is not fully established which introduces some error into the fitting algorithm which is partially responsible for the deviation from the nearest neighbor (NN) result that is observed for this sequence. For L = 6 the issue is more pronounced such that the sample never fully duplexes at low temperature as seen in Figure 6.1b. The fact that the duplex state is not fully established results in the lack of a low temperature baseline which causes a more significant deviation from the NN parameters. However, it is worth noting that previous studies[2] have also found discrepancies with the NN parameters at very short lengths due to the fact that the NN parameters were obtained by fitting data to larger duplexes with L > 9 suggesting that this could also be contributing to the discrepancy that we observe.

The equations used to fit the melting curves are derived and discussed in Section 5.2.2. To obtain thermodynamic parameters from melting curves, we make the van't Hoff assumption that ΔH and ΔS are independent of temperature, and fit the temperature-dependent duplex fraction $f_{\mathsf{D}}(T)$, Equation 5.7, using two independent parameters from the model, the dissociation enthalpy $\left(\Delta H^0_{\mathsf{d}}\right)$ and T_{m}, and four additional parameters that

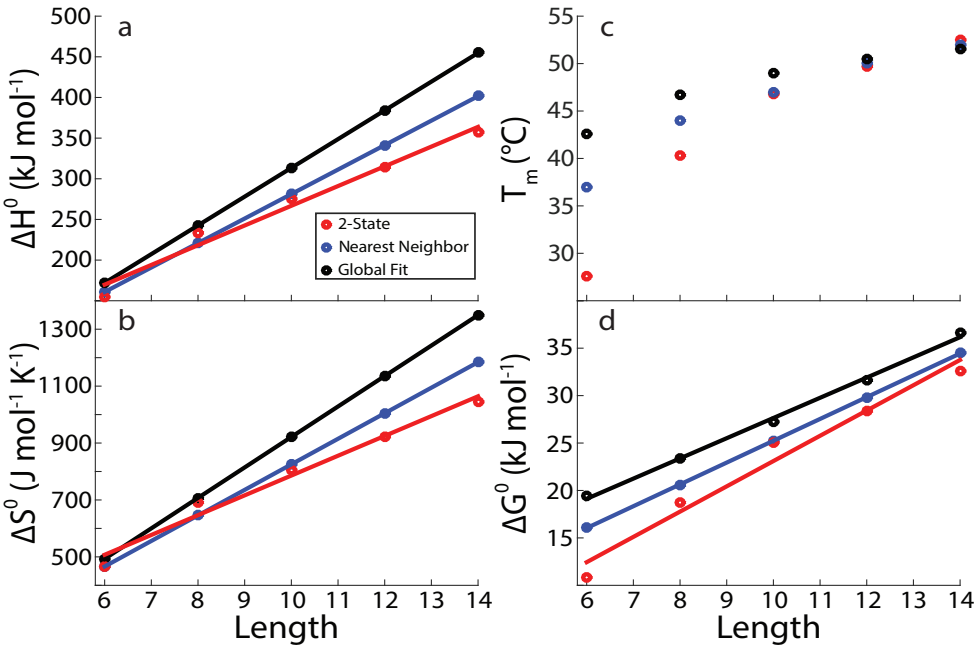

Figure 6.2: Values for the (a) enthalpy, (b) entropy, (c) T_m, and (d) free energy of dissociation at 37 °C for the two-state thermodynamic model (red), the nearest neighbor model (blue), and the global fit (black).

define the baselines for the high and low temperature regimes. ΔS^0_d follows from Equation 5.15, ΔG^0_d from Equation 5.10, and K_d from Equation 5.3. The resulting analysis of the equilibrium melting of the oligomers is summarized in Table 6.1 and plotted in Figure 6.2. All values for the NN parameters were corrected to account for the buffer salt concentrations.[25]

Plots of both the enthalpy and entropy as a function of length, shown in Figure 6.2a and 6.2b respectively, are roughly linear as expected. The resulting slopes for the two-state thermodynamic fit show that the enthalpy and entropy increase by 24 kJ mol^{-1} and 70 J mol^{-1} K^{-1} respectively for each additional base pair. Adding two base pairs, in this case both an AT and a TA dinucleotide, increases the ΔH^0_d by 48 kJ mol^{-1} and ΔS^0_d by 140 J mol^{-1} K^{-1} which is in reasonable agreement with salt-corrected[25] NN predictions of 60.2 kJ mol^{-1} and 179 J mol^{-1} K^{-1} respectively.[26] Finally, for the free energy of dissociation, according to the fit two additional A:T base pairs add 5.3 kJ mol^{-1} to the free energy of the duplex. This is in reasonable agreement with the NN prediction of 6.1 kJ mol^{-1}.[26] It is worth

Table 6.1: Length-dependent thermodynamic parameters for sequences 5'-C(AT)$_n$G-3' where n = 2-6 obtained from two-state analysis of melting curves, nearest neighbor calculations[a], kinetic Eyring analysis, and global fit analysis.

length		6	8	10	12	14
T_m (°C)	global fit	42.6	46.7	49.0	50.5	51.5
	nearest neighbor[a]	37	44	47	50	52
	melting curve	27.6	40.3	46.8	49.7	52.5
ΔG^0_d (kJ mol^{-1})[b]	Eyring analysis	16.8	21.1	24.8	29.7	35.6
	global fit	19.4	23.4	27.2	31.6	36.6
	nearest neighbor[a]	16.1	20.6	25.2	29.8	34.5
	melting curve	10.8	18.7	25.0	28.4	32.6
ΔH^0_d (kJ mol^{-1})	Eyring analysis	163	219	275	345	408
	global fit	172	243	313	384	455
	nearest neighbor[a]	161	221	281	341	402
	melting curve	155	233	275	314	357
ΔS^0_d (J mol^{-1} K^{-1})	Eyring analysis	472	638	808	1017	1201
	global fit	493	707	922	1136	1349
	nearest neighbor[a]	466	646	825	1005	1184
	melting curve	465	691	806	921	1046

[a]Nearest neighbor values are calculated from Ref 26 utilizing salt corrections from Ref 25.
[b]ΔG^0_d is calculated at 37 °C.

noting that all of the two-state results do demonstrate a slight non-linearity, especially at the shortest lengths. This suggests that the assumptions made in the two-state model or the additive nature of the NN model may be breaking down for these short lengths. This is consistent with previous work that showed discrepancies between NN predictions and experimental results at short lengths which suggested it may be due to the fact that the NN parameters were obtained by fitting to longer sequences with $L > 9$.[2]

For the equilibrium melting measurements shown in Figure 6.2 (red), the y-intercept is roughly zero. More precisely, in each case the fit crosses the x-axis at a length cor-

responding to somewhere between -1 and 2 base pairs. Because a change in sign in the thermodynamic parameters demonstrates a change from favorable to unfavorable or vice versa one would expect to cross the x-axis at the length where stable duplexes are no longer able to form, which roughly matches what is observed here. The fact that the y-intercept is not exactly zero demonstrates the fact that there are other factors that contribute to DNA thermodynamics outside of the dinucleotide contributions themselves. An example of this is observed in the NN parameters for initiation and the symmetry penalty.[26] The thermodynamic value at the y-intercept can be thought of as corresponding to the free-energy of a hypothetical duplex with zero bound base pairs but occupying the same molar volume as the fully base paired duplex.[27] Additionally, the fact that the thermodynamic parameters are all roughly zero at a length of zero base pairs helps to reinforce the picture that DNA thermodynamics are linear as a function of length.

6.4.2 Temperature-Jump Melting Kinetics

To study the kinetics, a minimum of four T-jump measurements, each with a jump magnitude of roughly 15 °C, were done on each length with varied T_i that sampled across the melting transition to allow for kinetic analysis. The resulting IR spectra allow the loss of base pairing as a result of the temperature perturbations to be tracked throughout the window of time between approximately 10 ns and 2 ms.

An example series of transient IR spectra following the T-jump for L = 6 is shown in Figure 6.3a. The spectrum shows positive and negative peaks that arise from 0-1 and 1-2 vibrational transitions, but they can be assigned by correspondence to the peaks observed in the FTIR absorption spectrum. Of the four most intense features the two negative peaks at approximately 1560 cm^{-1} and 1610 cm^{-1} correspond to the guanine and adenine ring modes respectively and the two positive peaks correspond to a mixture of guanine, cytosine, and thymine vibrations.

To illustrate the temporal form and length dependence of the kinetics, Figure 6.4

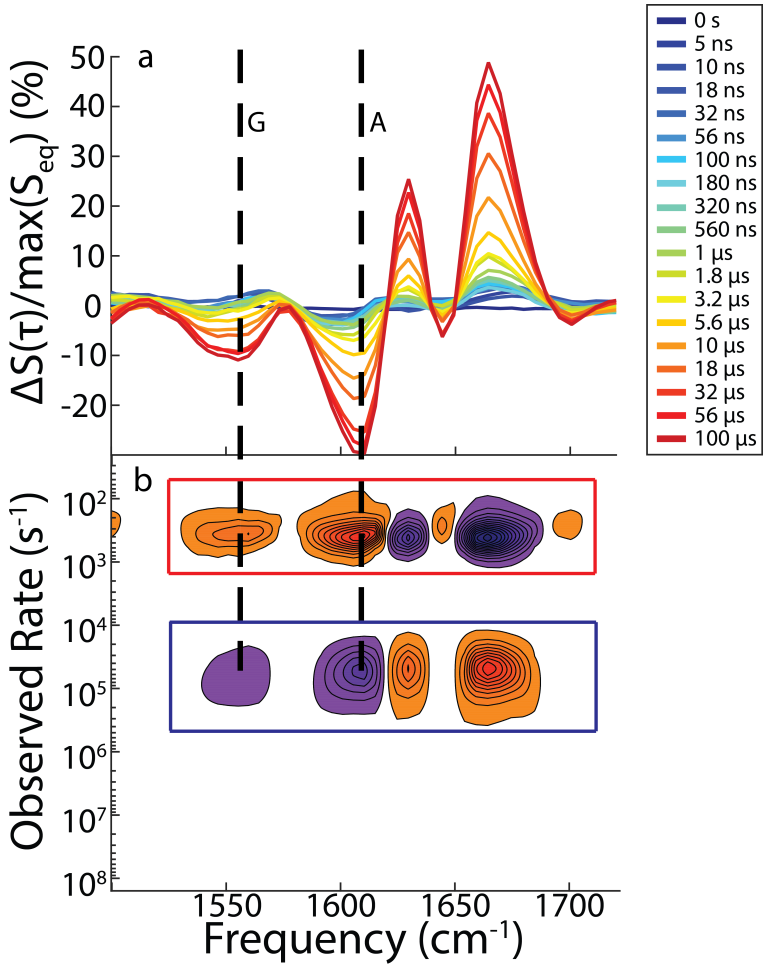

Figure 6.3: 5'-CATATG-3' (a) Transient IR spectrum for the T_i = 25 °C to T_f = 40 °C T-jump for delays between 0 and 0.1 ms, with increasing time delay as colors go from blue to red (re-equilibration not shown). (b) Rate distribution where purple denotes a loss in signal associated with that rate and orange denotes an increase in signal. The dotted black lines are a guide to the eye to show how the two are connected and highlight the guanine and adenine ring modes.

shows the time-dependent changes to the adenine ring mode intensity as a function of temperature-jump delay (τ) and oligo length. Each intensity trace is also superimposed with an exponential and stretched exponential fit. Also shown is the time-dependent temperature of the sample as it thermally re-equilibrates from T_f to T_i through thermal diffusion, which is relatively constant to $\tau \approx$100 µs before relaxing with a ≈2 ms time constant. As a result, we ensure that all data presented here have relaxation time scales $<<$2 ms. Note the kinetics in Figure 6.4 are compared at a fixed temperature of approximately T_f = 53 °C

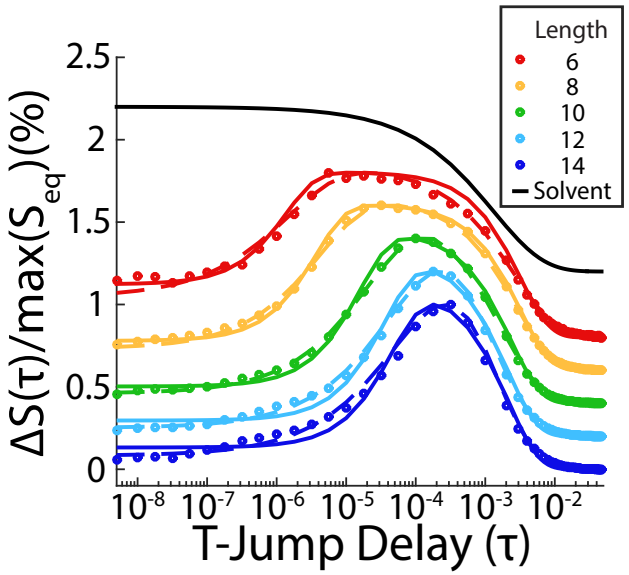

Figure 6.4: Normalized time domain traces for each length of sequence type 5'-C(AT)$_n$G-3' (n = 2-6) at approximately ω = 1610 cm^{-1} and T_f = 53 °C. Each trace is offset by 0.2 in order to facilitate comparison. The time-dependent re-equilibration of the solvent temperature is plotted in black. For each length the raw data (o), an exponential fit (-) and a stretched exponential fit (- -) are plotted.

and a roughly 15 °C jump magnitude, so the region of the melting curve sampled in each temperature jump shown in Figure 6.4 varied with length. The melting curves shown in Figure 6.1b can be used to determine what region of the melting curve was sampled for each of the jumps shown in Figure 6.4.

In the data we observe a growth of the signal between τ = 100 ns and 100 μs that reports on the melting of the duplex, followed by a drop in signal at longer times which reflects the convoluted temperature re-equilibration of the sample and re-hybridization of the oligos. The decrease in the observed melting rate, k_{obs}, as length increases is observed as the rise in signal associated with the dissociating duplexes shifting to longer times. Comparing the exponential and stretched exponential fits to the rise of the signal indicates that the kinetics deviate from exponential and are better represented by stretched exponential relaxation with a stretching exponent that decreases from 0.67 to 0.58 between 5'-C(AT)$_3$G-3' and 5'-C(AT)$_6$G-3'. This increasingly non-exponential behavior likely arises from a distribution of rates resulting from a heterogeneous initial population or the pres-

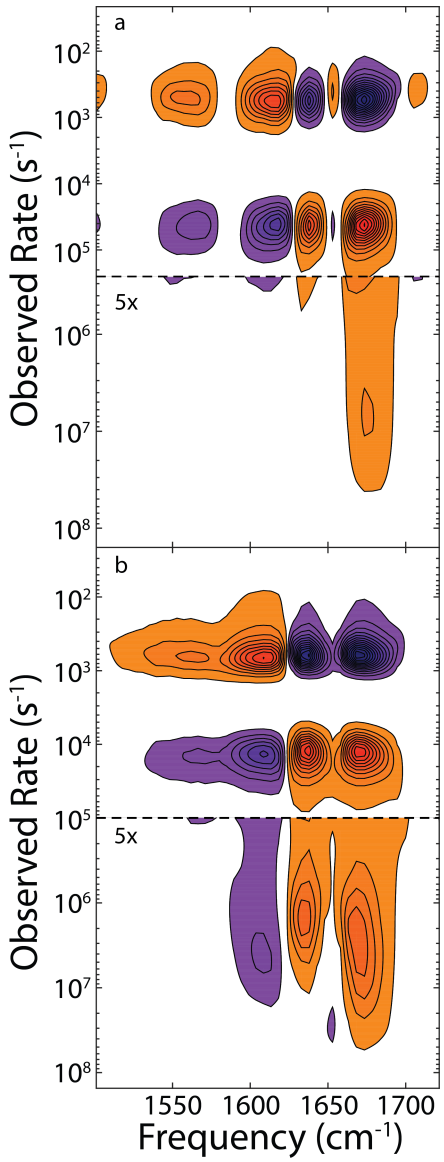

Figure 6.5: Rate maps for (a) 5'-CATATG-3' temperature jump from T_i = 20 °C to T_f = 33 °C and (b) 5'-CATATATATATATG-3' temperature jump from T_i = 40 °C to T_f = 55 °C. The region containing the fast response for both sequences is shown with 5x magnification to highlight the difference between the two sequences.

ence of more complicated dynamics.

Without a clear functional form that can be consistently applied to all of the time traces, we turn to an alternative method to analyze the relaxation kinetics. A maximum entropy implementation of an inverse Laplace transform was used to obtain a relaxation rate distribution map for each of the frequencies of the transient IR spectrum.[23,28] Additional

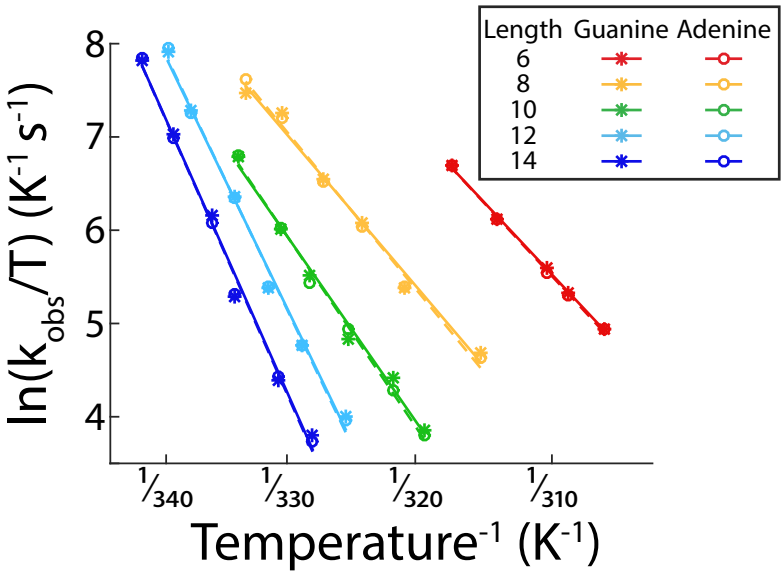

Figure 6.6: Eyring plot of the observed rate constant for the adenine and guanine ring modes for each 5'-C(AT)$_n$G-3' sequence where n = 2-6.

details about the method are provided in Section 5.3.3. Figure 6.3b shows the resulting rate spectrum corresponding to Figure 6.3a, with the fastest rates at the bottom of the rate axis. Peaks in the rate distribution can be separated out into two distinct regions: the dissociation kinetics – highlighted in blue – that correspond to the microsecond kinetics, and the slower re-equilibration regime – highlighted in red. A purple (orange) peak in the rate distribution corresponds to a decrease (increase) in signal, i.e. a decrease (increase) of positive signal or an increase (decrease) of negative signal. In Figure 6.3b, and for other samples, we observe that there is a single common peak in the rate distribution for each IR detection frequency, indicating that all spectral features respond in a correlated manner, as expected for two-state kinetics in which the dissociation of all base pairs is synchronous. As relaxation kinetics become more stretched, the rate distributions broaden – sometimes considerably as shown in Figure 6.5 – but a single well-defined peak for the observed dissociation rate is always apparent.

The observed rate, k_{obs}, for both the adenine ring mode ESA ($\approx$1610 cm^{-1}) and the guanine ring mode ESA ($\approx$1560 cm^{-1}) were determined by the first moment (weighted average) of the dissociation peak in the rate spectrum. The observed rates for the adenine

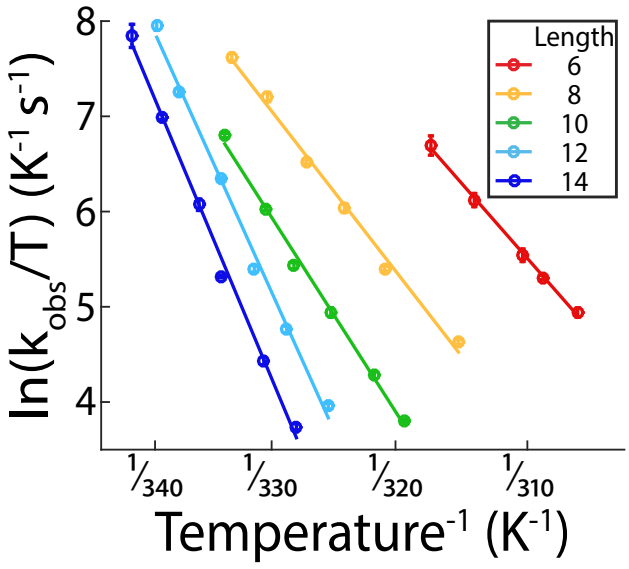

Figure 6.7: Eyring plot of the adenine ring mode observed rate constant for each 5'-C(AT)$_n$G-3' sequence where n = 2-6. Error bars reflect the amplitude weighted standard deviation of the maximum rate for all detected frequencies.

and guanine ring modes are almost identical under all conditions, as seen in Figure 6.6; due to this, our analysis will focus on the adenine ring mode response. The resulting adenine ring mode rate shows a dependence on both temperature and length which is reflected in the Eyring plot, ln $\left(\frac{k_{\text{obs}}}{T_{\text{f}}}\right)$ vs. $\frac{1}{T_{\text{f}}}$, in Figure 6.7. The linear behavior indicates that k_{obs} increases exponentially as temperature increases. With increasing length, k_{obs} decreases and the slope of the line, which reports on the activation enthalpy, increases. The linear behavior in this plot is virtually indistinguishable from an Arrhenius plot, ln k_{obs} vs. $\frac{1}{T_{\text{f}}}$, as shown in Figure 6.8

6.4.3 Two-State Analysis of Kinetics

Next we analyzed the data using the two-state kinetic model that was described in Section 5.3.4. For sequences 5'-C(AT)$_n$G-3' where n = 4-6 both of the thermodynamic parameters needed for kinetic analysis, the duplex fraction at the initial temperature and the dissociation equilibrium constant at the final temperature, are taken from the thermodynamic fits of the melting curves described in Section 5.2.2. For 5'-C(AT)$_n$G-3' where n

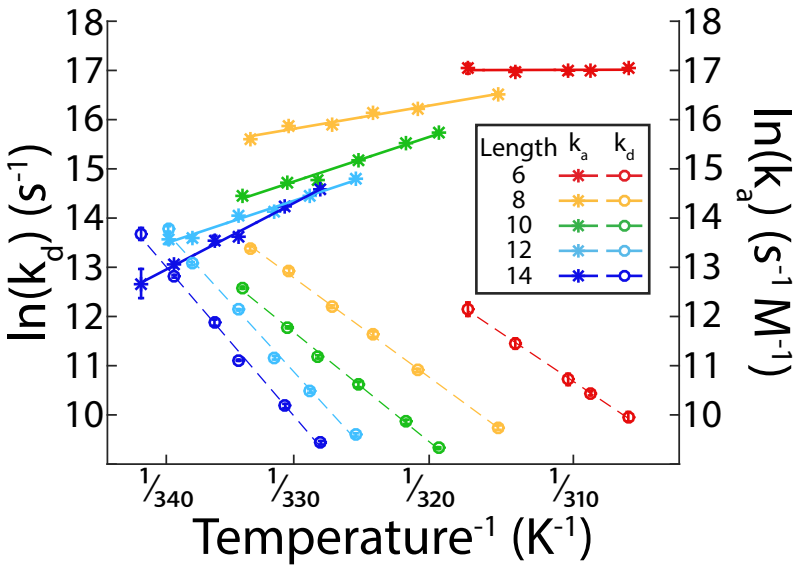

Figure 6.8: Arrhenius plots for the adenine ring mode association and dissociation rates for each 5'-C(AT)$_n$G-3' sequence where n = 2-6.

= 2-3 the values were determined by NN parameters,[26] to avoid possible errors due to the poor low temperature baseline in their melting curves. While the NN parameters may not be perfect in this length regime; between the lack of low temperature baselines in the melting curve and the possibility that the two-state assumption may be breaking down we believe utilizing the well-established NN parameters will provide the most reliable analysis for these sequences. With k_{a} determined from Equation 5.36, we obtain k_{d} from Equation 5.35. The results are shown in Figure 6.9 as an Eyring plot, and are seen to follow linear trends. A comparison between Figure 6.9 and the observed rate constant in Figure 6.7 shows that k_{obs} is heavily dominated by k_{d}.

As our initial attempt to determine the association and dissociation barriers for duplex dissociation, we constructed Arrhenius plots for k_{a} and k_{d} from which the activation energy, E_{A}, and pre-exponential factors, A, for both are determined by a linear fit and are shown in Table 6.2. The Arrhenius plots obtained, shown in Figure 6.8, show the same length and temperature dependent trends as Figure 6.9. Our results show thermally activated kinetics with a positive dissociation barrier and negative association barrier, in reasonable agreement with other published values.[3,14,15] The presence of a negative ac-

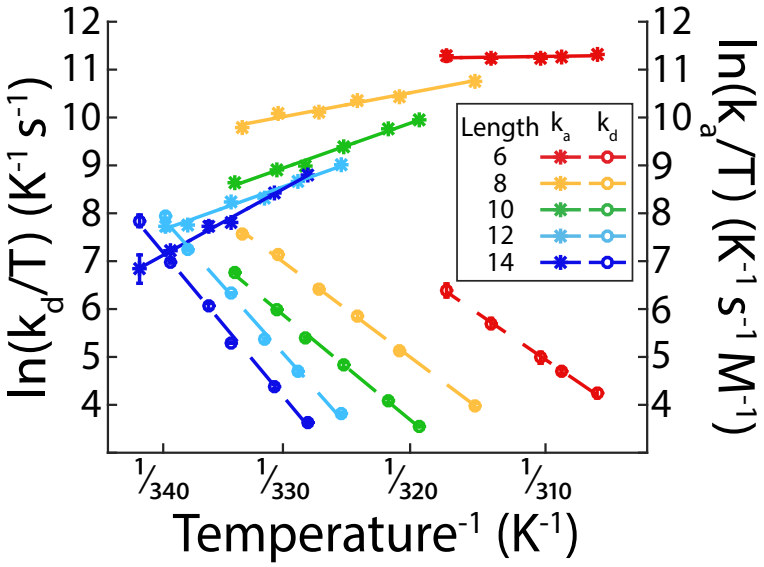

Figure 6.9: Eyring plot for the adenine ring mode association and dissociation rates for each 5'-C(AT)$_n$G-3' sequence where n = 2-6.

tivation barrier indicates that the association process in Equation 5.2 does not represent a fundamental kinetic step.

To avoid the ambiguity over the physical interpretation of the negative activation energy and to better understand the role entropy changes play in the reaction barrier, the data was analyzed using the Eyring equation[29,30]

$$\ln\left(\frac{k}{T}\right) = \frac{-\Delta H^{\ddagger}}{R}\frac{1}{T} + \ln\left(\frac{k_B}{h}\right) + \frac{-\Delta S^{\ddagger}}{R} \tag{6.1}$$

This approach allows us to obtain values for the activation entropy and activation enthalpy that can be directly compared to the thermodynamic results.

Linear fits were used to calculate the enthalpy of activation for the association and dissociation, $\Delta H^{\ddagger}_{a}$ and $\Delta H^{\ddagger}_{d}$, and the entropy of activation for association and dissociation, $\Delta S^{\ddagger}_{a}$ and $\Delta S^{\ddagger}_{d}$. These values are presented in Figure 6.10 and Table 6.3. Our first observation is that the parameters determined by analyzing the melting kinetics through the adenine ring mode and guanine ring mode are identical within our error bars. As expected $\Delta H^{\ddagger}_{a}$ and $\Delta H^{\ddagger}_{d}$ are essentially identical to the Arrhenius activation energies shown in

Table 6.2: Fit parameters for the Arrhenius analysis of the adenine ring mode association and dissociation rates. The activation energies are roughly equivalent to the activation enthalpies from the Eyring analysis given in Table 6.3.

	length	6	8	10	12	14
association	E_A (kJ mol^{-1})	-0.6	-41.3	-80.0	-80.4	-126
	A (M^{-1} s^{-1})	1.9 x 10^{7}	2.1	6.1 x 10^{-7}	3.2 x 10^{-7}	1.9 x 10^{-14}
dissociation	E_A (kJ mol^{-1})	162	177	196	265	282
	A (s^{-1})	9.3 x 10^{31}	4.2 x 10^{33}	1.0 x 10^{36}	4.0 x 10^{46}	9.7 x 10^{48}

Table 6.2 confirming that they contain the same information. The reaction barrier defined by the free energy of activation, which will be denoted $\Delta G^{\ddagger}_{a}$ for the association reaction and $\Delta G^{\ddagger}_{d}$ for the dissociation, are calculated from the enthalpy and entropy of activation and are plotted as a function of temperature in Figure 6.11.

From Figure 6.10, we observe that $\Delta H^{\ddagger}_{a}$, $\Delta H^{\ddagger}_{d}$, $\Delta S^{\ddagger}_{a}$, and $\Delta S^{\ddagger}_{d}$ are all roughly linear with length. For the dissociation they are both positive and increasing with length, whereas for association they are negative and decrease with length. Combining the activation enthalpy and entropy results in a positive free energy barrier to both association and dissociation as shown in Table 6.3 and Figure 6.11. As a result we see that the considerable loss of conformational freedom that results from initiating the duplex formation is partially compensated by an enthalpic benefit from forming favorable contacts in the transition state. The fact that the activation enthalpy and entropy are changing significantly as a function of length shows that the energetics of the transition state have a significant dependence on length. This suggests that changing the length may induce changes in the structure of the transition state, an idea that merits further study in future research. The barriers for duplex dissociation have a more traditional interpretation. $\Delta H^{\ddagger}_{d}$ increases with length as a result of the loss of base pairing and stacking required to reach the transition state,

Table 6.3: Activation free energies, enthalpies, and entropies for the association and dissociation determined from the adenine ring mode Eyring analysis and the global fit.

		length	6	8	10	12	14
dissociation	$\Delta G^{\ddagger}_{\mathrm{d}}$ (kJ mol^{-1})[a]	global fit	49.9	54.2	58.3	63.6	68.0
		Eyring analysis	49.0	54.0	58.1	64.1	66.9
	$\Delta H^{\ddagger}_{\mathrm{d}}$ (kJ mol^{-1})	global fit	146	182	218	255	291
		Eyring analysis	160	175	193	262	279
	$\Delta S^{\ddagger}_{\mathrm{d}}$ (J mol^{-1} K^{-1})	global fit	310	412	515	617	719
		Eyring analysis	358	390	435	638	684
association	$\Delta G^{\ddagger}_{\mathrm{a}}$ (kJ mol^{-1})[a]	global fit	30.5	30.8	31.1	32.0	31.4
		Eyring analysis	32.2	32.9	33.3	34.4	31.3
	$\Delta H^{\ddagger}_{\mathrm{a}}$ (kJ mol^{-1})	global fit	-26.3	-60.7	-95.1	-129	-164
		Eyring analysis	-3.2	-44.0	-82.4	-83.1	-129
	$\Delta S^{\ddagger}_{\mathrm{a}}$ (J mol^{-1} K^{-1})	global fit	-183	-295	-407	-519	-630
		Eyring analysis	-114	-248	-373	-379	-517

[a]Free energy values are calculated at 37 °C

and is partially compensated by the gain in configurational entropy in $\Delta S^{\ddagger}_{\mathrm{d}}$. Since it explicitly accounts for diffusion, Kramers theory is an alternative route to interpreting barrier crossing in solution phase reactions. However, in our experimental analysis the resulting pre-exponential parameters in either theory are equally difficult to interpret microscopically whether they are cast in terms of an attempt frequency and activation entropy or friction coefficient and barrier curvature. Over the temperature range in our experiments, the viscosity of the solvent changes by a small amount ($\approx$1.3) whereas the increase in association and dissociation rates is much higher in most cases. This indicates that diffusion of two strands to encounter is a minor contribution to the overall association barrier, and the primary contribution to the diffusive barrier crossing in Kramers theory is the internal friction experienced by the dynamics of the encounter complex. Indeed, simulations have predicted that the reaction probability for the formation of DNA duplexes is below 1% and

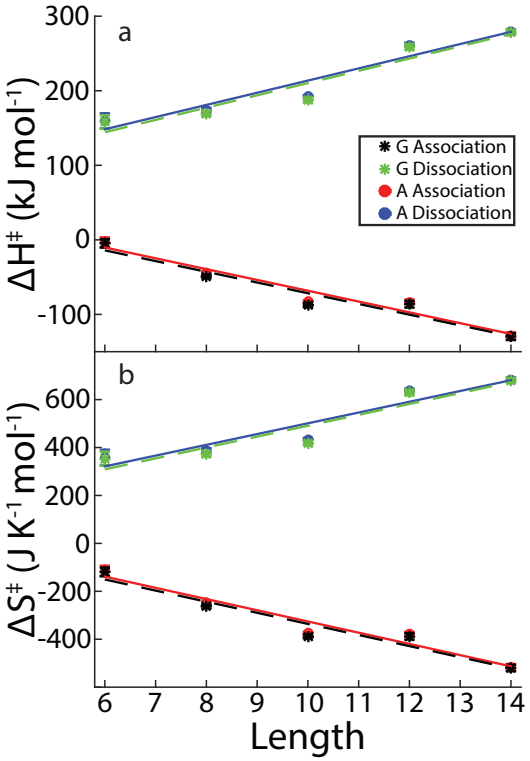

Figure 6.10: Activation enthalpy and entropy of association and dissociation determined from the adenine and guanine ring modes as a function of sequence length.

that the rate limiting step is the contact between the DNA single strands.[1] Similarly, the internal interactions within the dimer state rather than the resistance of the solvent are the dominant contributors to the dissociation barrier.

6.4.4 Global Fit of Thermodynamics and Kinetics

As a further test of the general applicability of the two-state all or nothing model and the linearity of the kinetic parameters we applied a global fit to the thermodynamic and kinetic data under the assumption that the activation enthalpies and entropies were linear in oligo length. Additionally, we wanted to investigate how well the thermodynamics and kinetics could be tied together utilizing an Eyring description of the data and to see if we could describe both the kinetics and the thermodynamics of the system utilizing only a small set of kinetic parameters. The $\Delta H_{\mathrm{a}}^{\ddagger}$, $\Delta H_{\mathrm{d}}^{\ddagger}$, $\Delta S_{\mathrm{a}}^{\ddagger}$ and $\Delta S_{\mathrm{d}}^{\ddagger}$ can be used to describe

the kinetics at any fixed temperature and the thermodynamics can be derived using fundamental relations such as

$$\Delta G_{\mathrm{d}}^{\ddagger} = \Delta H_{\mathrm{d}}^{\ddagger} - T\Delta S_{\mathrm{d}}^{\ddagger} \tag{6.2}$$

and

$$\Delta G_{\mathrm{d}}^{0} = \Delta G_{\mathrm{d}}^{\ddagger} - \Delta G_{\mathrm{a}}^{\ddagger} \tag{6.3}$$

with $\Delta H_{\mathrm{d}}^{0}$ and $\Delta S_{\mathrm{d}}^{0}$ calculated in the same way as $\Delta G_{\mathrm{d}}^{0}$. We then posit that the length dependence of $\Delta H_{\mathrm{a}}^{\ddagger}$, $\Delta H_{\mathrm{d}}^{\ddagger}$, $\Delta S_{\mathrm{a}}^{\ddagger}$ and $\Delta S_{\mathrm{d}}^{\ddagger}$ follow a linear length dependence of the form

$$\Delta H_{d/a}^{\ddagger}(L) = \Delta H_{d/a}^{\ddagger}(0) + L * \delta\Delta H_{d/a}^{\ddagger} \tag{6.4}$$

where the slope $\delta\Delta H^{\ddagger}$ is the change in activation enthalpy for every base added to the sequence, and the intercept $\Delta H^{\ddagger}(0)$ is the activation enthalpy for a hypothetical sequence of length zero. The activation entropies were treated in the same way where $\delta\Delta S^{\ddagger}$ and

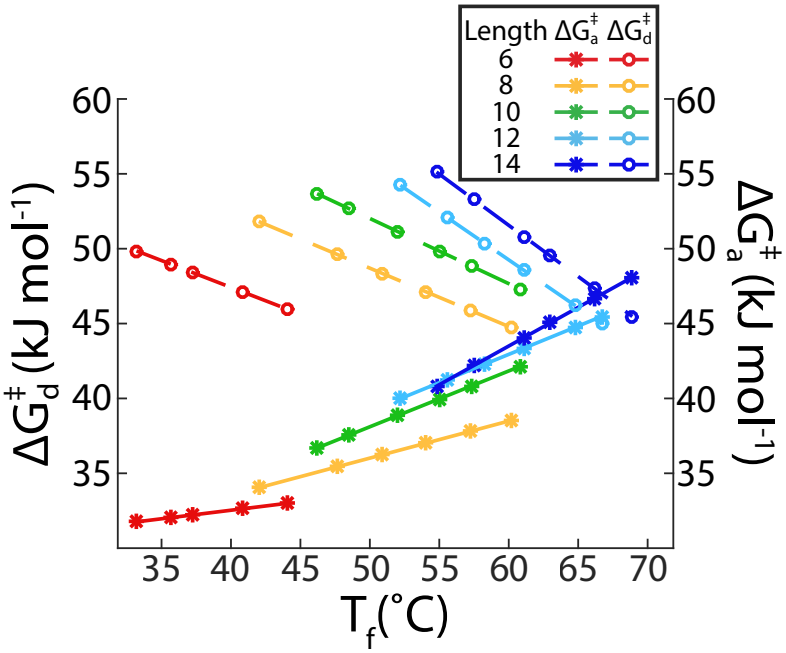

Figure 6.11: Gibbs free energy of activation for association and dissociation determined from the adenine ring mode as a function of T-jump final temperature (T_{f}).

Table 6.4: Global fit parameters compared to the linear fits from the Eyring analysis.

	global fit	Eyring fit
$\delta\Delta H^{\ddagger}_{\mathsf{a}}$ (kJ mol^{-1} bp^{-1})	-17.2	-14.4
$\Delta H^{\ddagger}_{\mathsf{a}}(0)$ (kJ mol^{-1})	76.9	74.4
$\delta\Delta H^{\ddagger}_{\mathsf{d}}$ (kJ mol^{-1} bp^{-1})	18.2	16.4
$\Delta H^{\ddagger}_{\mathsf{d}}(0)$ (kJ mol^{-1})	36.4	48.0
$\delta\Delta S^{\ddagger}_{\mathsf{a}}$ (kJ mol^{-1} K^{-1} bp^{-1})	-0.0559	-0.0466
$\Delta S^{\ddagger}_{\mathsf{a}}(0)$ (kJ mol^{-1} K^{-1})	0.153	0.135
$\delta\Delta S^{\ddagger}_{\mathsf{d}}$ (kJ mol^{-1} K^{-1} bp^{-1})	0.0512	0.0452
$\Delta S^{\ddagger}_{\mathsf{d}}(0)$ (kJ mol^{-1} K^{-1})	0.0022	0.0441

$\Delta S^{\ddagger}(0)$ designate the corresponding slope and intercept respectively for the activation entropies.

The resulting eight parameters, $\delta\Delta H^{\ddagger}_{\mathsf{a}}$, $\delta\Delta H^{\ddagger}_{\mathsf{d}}$,$\delta\Delta S^{\ddagger}_{\mathsf{a}}$,$\delta\Delta S^{\ddagger}_{\mathsf{d}}$, $\Delta H^{\ddagger}_{\mathsf{a}}(0)$, $\Delta H^{\ddagger}_{\mathsf{d}}(0)$, $\Delta S^{\ddagger}_{\mathsf{a}}(0)$, and $\Delta S^{\ddagger}_{\mathsf{d}}(0)$, used in the global fit are given in Table 6.4. These parameters were used to fit the thermodynamic data in the form of the second SVD component from the FTIR temperature ramps for sequences 5'-C(AT)$_n$G-3' where n = 4-6 and the NN derived T_{m} for all lengths. The melting curves were fit by taking the value of $\Delta G^{0}_{\mathsf{d}}$ determined from Equations 6.2 and 6.3 and using that to determine f_{D} using Equations 5.3 and 5.7. The fit to the second SVD component according to Equation 5.16 used the upper and lower baselines determined by the two-state thermodynamic fit described previously. As such the baselines required no additional fit parameters in the global fit. The melting curves for the two shortest sequences were not fit to avoid the limited low temperature baselines skewing the results. The NN T_{m} values were used because, while they were derived from two-state fits to UV melting curves, they are independent of our two-state fit to the melting curve, unlike our T_{m} values, and thus provide an additional data set to fit. To fit the melting

temperature we rearrange Equation 5.15 to obtain

$$T_{\mathsf{m}} = \frac{\Delta H^{\ddagger}_{\mathsf{d}} - \Delta H^{\ddagger}_{\mathsf{a}}}{\left(\Delta S^{\ddagger}_{\mathsf{d}} - \Delta S^{\ddagger}_{\mathsf{a}}\right) - R \ln\left(C_{\mathsf{T}}\right)} \tag{6.5}$$

where $\Delta H^{\ddagger}_{\mathsf{a}}$, $\Delta H^{\ddagger}_{\mathsf{d}}$, $\Delta S^{\ddagger}_{\mathsf{a}}$, and $\Delta S^{\ddagger}_{\mathsf{d}}$ are determined from the relevant versions of Equation 6.4. In addition to fitting the thermodynamics, the observed rate constants from the adenine ring mode for all lengths were also fit. The fit parameters calculated the observed rate constant using the Eyring equation, Equation 6.1, and Equation 5.35 where [M_{eq}] can be determined from the thermodynamic value of f_{D} determined from the global fit parameters as described above in conjunction with Equation 5.4 and the known value of C_{T}. The minimization algorithm independently scaled the residuals for the melting curve fits, observed rates, and T_{m} values to make each residual the same order of magnitude so all three equally contribute to the fit.

Figure 6.12a contains the kinetic results from the global fit and the adenine ring mode observed rate constants from experiment for comparison. Additionally, $\Delta G^{\ddagger}$, $\Delta H^{\ddagger}$, and $\Delta S^{\ddagger}$ from the global fit and the Eyring analysis are compared in Table 6.3. This demonstrates that the global fit is able to reasonably replicate the experimental kinetics for these sequences, in particular sequences with length greater than ten. Figure 6.12b shows that the second SVD components are well fit by the global fit. The better agreement seen in the kinetics relative to the thermodynamics in Figure 6.12 is most likely because the adjustable parameters used are more closely tied to the kinetics then the thermodynamics. The values for the thermodynamic parameters determined by the Eyring analysis, global fit, NN parameters, and two-state thermodynamic fit are all given in Table 6.1 and plotted in Figure 6.2. The enthalpy and entropy appear to be in relatively good agreement for all lengths, but they do appear to deviate more at longer lengths. However, because these two values directly compensate for each other when determining the observed rate and the fraction of intact base pairs, it is more informative to examine the T_{m} and free energy,

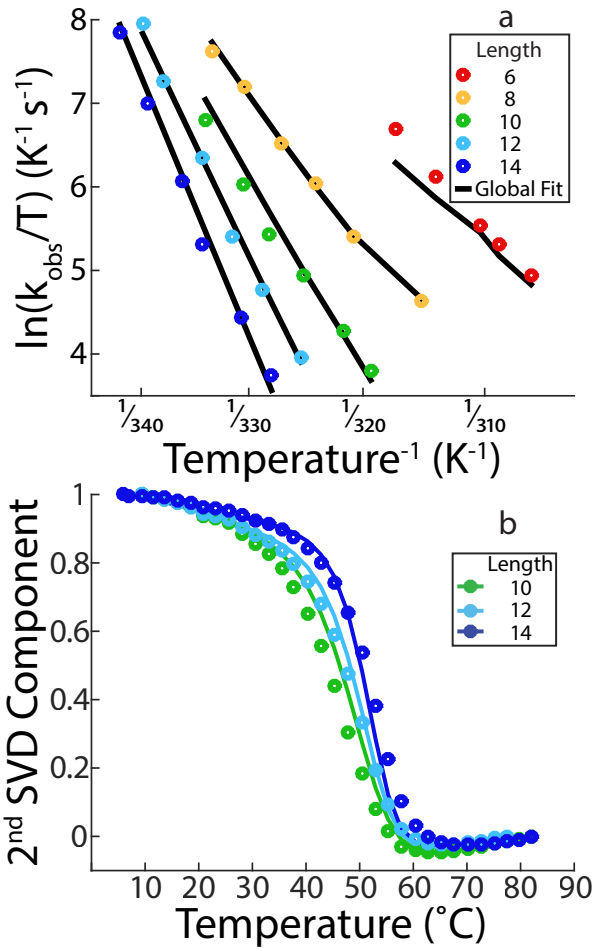

Figure 6.12: (a) Result of the global fit and the adenine ring mode observed rate from the T-jump experiment. (b) The raw second SVD component (o) and the result of the global fit (-) for the three longest sequences.

which are more closely related to the experimental observables. The T_m and free energy display the opposite trend, they are in relatively good agreement at longer lengths, but begin to deviate significantly at lengths shorter than ten base pairs. The worse agreement observed for the shorter sequences could be due to the fact that those second SVD components were not included in the fit. However, because the agreement in the kinetics also appears to be better for the longer sequences it seems more likely that there is a fundamental explanation for the discrepancy observed in the short sequences. The short sequences are reaching the minimum number of bases required to form a stable duplex so it is possible that the assumption that the kinetic parameters are linear as a function

of length may partially break down. This agrees with our earlier observation on the loss of linearity seen in the two-state thermodynamic determined from FTIR in addition to previously mentioned discrepancies between the NN predictions and experimental results.[2] This supports the possibility that the larger discrepancy at lower temperatures is the result of a breakdown in the linearity of the kinetic and thermodynamic parameters. However, overall the global fit is able to reproduce the experimental results with a reasonable accuracy.

As mentioned previously, we use Eyring analysis as the primary interpretive tool out of a desire to reproduce a more complete picture of the energy landscape of DNA hybridization and dehybridization. In previous studies, the negative activation energy for association obtained from the Arrhenius equation is typically cited as reflecting a non-fundamental kinetic step in the form of the pre-equilibrium involved in the formation of the critical nucleus.[3,14] This focus on the enthalpic contribution neglects the significant entropic contribution in the form of the large decrease in activation entropy of association, as expected for assembling a critical nucleus from the free strands. The self-consistency between the kinetic and thermodynamic results illustrates that applying an Eyring analysis to the kinetics of DNA association and dissociation and describing the reaction barrier as an activation free energy provides additional insight into the energetic driving forces of the reaction which produces a robust and physically intuitive description of the process of DNA association and dissociation. Additionally, the fact that the thermodynamic parameters can be determined from the kinetic parameters provides additional validation that the two-state dissociation model is appropriate for sequences within this length regime and for the present sequences.

6.4.5 Linear Scaling of Thermodynamics and Kinetics with Length

Now that the linear scaling of both the thermodynamic and kinetic parameters with length has been established we can dive deeper into the meaning of these trends. First,

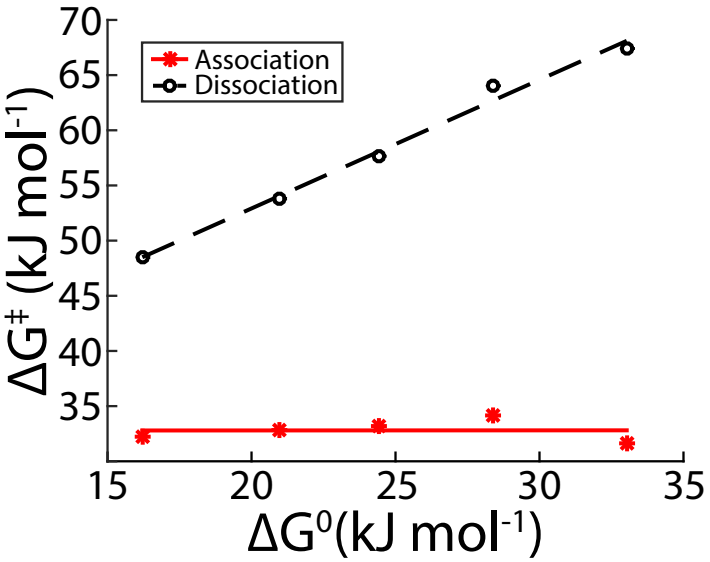

Figure 6.13: The free energy of activation for dissociation and association plotted against the thermodynamic free energy from the FTIR temperature ramp experiments both at 37 °C.

it is worth noting that $\Delta H_a^{\ddagger}$ approaches zero for the shortest sequence, 5'-C(AT)$_2$G-3', of length six which is roughly the lower bound for the formation of stable duplexes. Another interesting note is that although the length-dependent trends observed in Figure 6.10 are roughly linear, we note that the deviations of the linear fit at each length was reproducible across multiple measurements. All results display a discontinuity between 5'-C(AT)$_4$G-3' and 5'-C(AT)$_5$G-3'. While we do not have an explanation for this, it is interesting to note that this length coincides with the 10.5 base pairs per of turn of the B-form DNA double helix.

To further study the relationship between the thermodynamic and kinetic results $\Delta G_a^{\ddagger}$ and $\Delta G_d^{\ddagger}$ were plotted against ΔG^0, shown in Figure 6.13. $\Delta G_d^{\ddagger}$ is strongly correlated to the ΔG^0 (R^2=0.986) in good agreement with existing literature.[16] The change in $\Delta G_d^{\ddagger}$ with respect to ΔG^0 is linear with a slope of about one demonstrating that $\Delta G_d^{\ddagger}$ scales directly with ΔG^0, and further demonstrates the strong ties between the thermodynamics and the kinetics.[16,31]

We will now compare the activation enthalpies in Figure 6.10, the values of which are

shown in Table 6.3, in addition to their linear fits as a function of length, the parameters of which are shown in Table 6.4, with the activation energies from prior studies of small oligo melting utilizing a capacitive discharge temperature-jump apparatus and monitoring changes with UV spectroscopy.[10,14,15] While these studies were conducted with RNA oligos the comparison is still informative none the less. The results from Pörschke et al.[14] and Craig et al.[15], which examined sequences containing only A:U base pairs, are compared graphically with our results in Figure 6.14. They find that the dissociation activation energy is positive and has a significant trend with length. Additionally, comparing their trends in the activation energy of dissociation with respect to length with our trends in the activation enthalpy of dissociation as a function of length demonstrates that the two are in reasonable agreement. These studies also observe that the association activation energy is negative in agreement with our work. While our results appear to have a stronger length dependence there is some ambiguity in the results and how they compare to our work.[14,15] Early studies looking at a variety of sequences including G:C base pairs have found that the dissociation activation energy is also positive and weakly dependent on length.[10] For the association activation energy they found it should be positive and not

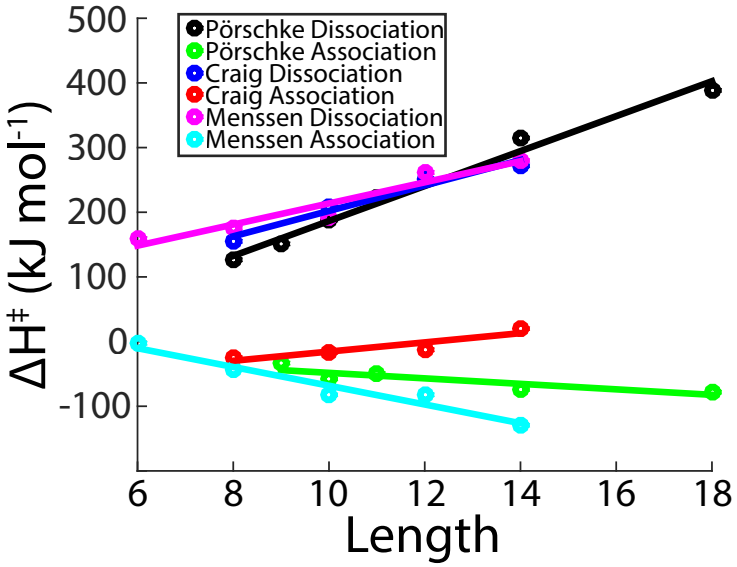

Figure 6.14: Results from studies of RNA sequences in the absence of G:C base pairs by Craig et al.[15] and Pörschke et al.[14] alongside our results.

significantly depend on length, in contrast with our results.[10] More recent results are mixed with previous work from our group finding that G:C containing sequences have negative activation energies[23] while others have found positive activation energies.[16]

When comparing the literature to our work there are two factors that could be causing the discrepancy between Arrhenius and anti-Arrhenius results observed for the association. It has been suggested that the presence of G:C base pairs results in a positive activation energy due to their additional stability relative to A:T base pairs.[10] However, evidence suggests that it is not purely a sequence effect. In this case, similar sequences with identical GC content would have activation energies with the same sign, but comparing our results with results in the literature demonstrates this is not the case.[10,23]

Another explanation is that a temperature effect is responsible for the discrepancy between the Arrhenius and anti-Arrhenius results. It has been demonstrated that the association rate as a function of temperature follows a bell shaped curve with a maximum rate below T_m.[5] This means the temperature at which the rates are determined could cause the discrepancy in the sign of the activation energy. Studies conducted at temperatures below the association rate maximum would be expected to have positive activation energies while studies at temperatures above the maximum would be expected to have negative activation energies with a potential turnover region in between. While this has been experimentally observed,[32] inconsistency in the literature remains as to what the sign of the activation energy should be, even when considering the temperature at which the data was acquired relative to the T_m of the sample. Of the previously mentioned studies that found positive activation energies for sequences containing G:C base pairs one was conducted at temperatures between 3 °C and 45 °C depending on the sequence, but the T_m values for those sequences were also relatively low, between 1 °C and 25 °C, and in all cases the temperatures examined for each sequence were roughly centered around T_m or slightly above meaning they are all above the proposed association rate maximum.[10] The other study with G:C base pairs studied sequences with T_m values between 42.6

°C and 68 °C at temperatures between 6.6 °C and 30.6 °C so the vast majority of their rates were measured below the maximum association rate.[16] Previous work in our group studied sequences with T_m values of 47 °C and 57 °C at temperature ranges of 34-65 °C and 42-70 °C respectively, such that all temperatures were around T_m and above the maximum rate, and found negative activation energies.[23] One study looking at sequences without G:C base pairs that also found predominately negative activation energies looked at sequences with T_m values between 9 °C and 23.5 °C at temperatures between 8.6 °C and 28.6 °C such that in each case the temperatures for that sequence were roughly centered around T_m and were above the association rate maximum.[15] A different study examined sequences of varying length without G:C base pairs at temperatures between 3.4 °C and 32.4 °C and found negative activation energies, however the T_m values at their experimental conditions are not listed making a direct comparison difficult.[14] This demonstrates that while the overall results are inconclusive, they suggest that while temperature likely plays a role it is likely not the only factor responsible for the discrepancy in the sign of the activation energy. Future studies are needed to understand if the sign of the activation energy depends on sequence and temperature and if so determine what that relationship is.

To the extent that $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ are linear functions of length for these short self-complementary oligos, our results indicate that these kinetic parameters are simply additive in the number of A:T base pairs and could be used to predict the kinetics of similar sequences with longer lengths. Such additive relationships underlie the highly successful NN approach to predicting sequence and length dependent thermodynamic parameters. To compare our kinetic parameters with the NN parameters, we note that for both $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ the slopes of the fits to the association and dissociation data in Figure 6.10, are roughly equal in magnitude with opposite signs (See Table 6.4). Adding <u>two</u> A:T base pairs adds both a TA and AT dinucleotide to the sequence resulting in a ±92 J mol^{-1} K^{-1} change to the activation entropy and roughly ±30 kJ mol^{-1} to the activation enthalpy. These values

are similar to the ΔS^0 and ΔH^0 for a single AT dinucleotide from the NN parameters, which are 85 J mol^{-1} K^{-1} and 30 kJ mol^{-1} respectively.[26] As a result adding two dinucleotides to the overall sequence changes $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ for the association and the dissociation reactions by half of what the NN parameters would predict. This suggests that the two results are correlated, but not directly related. Understanding why these values are half of what the NN parameters predict requires additional mechanistic insight into the association and dissociation reactions which could be achieved by examining the results of this work in the context of mechanistic models. We will further examine this relationship later on in this chapter within the context of the nucleation-zipper mechanism.

The consistency between our results and the NN parameters suggests that the NN parameters themselves may be useful for predicting melting and hybridization kinetics for small oligos that follow two-state kinetics.[33] It has been previously noted that the NN parameters can be used to predict dissociation kinetics if the association rate is either known or assumed by calculating the dissociation rate constant from the NN derived equilibrium constant and the assumed association rate.[33] However, it would be preferable to be able to predict the kinetics without making such an assumption, which means an alternative set of parameters, potentially a kinetic analog of the NN parameters based on dinucleotides, is necessary.

It has also been shown that it is possible to develop predictive models making no assumptions about the association rate with sequence and temperature specificity for a variety of sequences of a single length.[34] However, this work was conducted on longer sequences of a single length that were attached to a fluorophore or quenching strands that were as large or larger than the probe and target strands which could impact the observed kinetics. The linear relationships and resulting global fit parameters presented here offer an alternative method for the prediction of both the association and dissociation rate constants, in addition to the thermodynamic parameters, without requiring an assumed value for the association rate constant or the attachment of probes or labels that could affect the

kinetics. This suggests the possibility of predicting kinetics of arbitrary DNA sequences using an equivalent set of NN parameters for kinetics, determined through label free experimental methods, that are able to account for the effect that temperature, sequence, and length all have on the kinetics. Whether or not this is possible will depend on a number of assumptions, including that the melting dynamics of varying oligos follow predictably similar pathways to their transition states and that the simple two-state kinetics of the form investigated here remain valid for different sequences.

6.4.6 Application of Nucleation-Zipper Model

To connect the results presented here to the mechanism of DNA association and dissociation they will be considered further within the context of the nucleation-zipper model. The main focus of this analysis is to investigate the size of the critical nucleus. The critical nucleus is defined as the minimum number of base pairs such that the partially formed duplex is stable and the remaining base pairs rapidly zip up in a sequential and downhill fashion that is orders of magnitude faster than the formation of the critical nucleus.

The expected change in the enthalpy associated with making or breaking base pair dinucleotides is given by the NN parameters.[26,32,33,35,36] In conjunction with the activation enthalpy for association and dissociation determined by the Eyring analysis the NN parameters can be used to determine the number of bases in the critical nucleus for a given sequence length. Based on the definition of the critical nucleus, and the fact that the reaction proceeds downhill from that point, it must lie at or just on the dimer side of the peak of a standard reaction free energy diagram. As a result the size of the critical nucleus can be determined by finding the minimum number of base pairs such that the sum of the enthalpies, given by the NN parameters, is equal to or greater than the activation enthalpy of association.[33] For the dissociation, the number of base pairs that must be broken such that the sum of the NN enthalpies is equal to or greater than the dissociation activation enthalpy can be calculated and the number of intact base pairs at that point can be de-

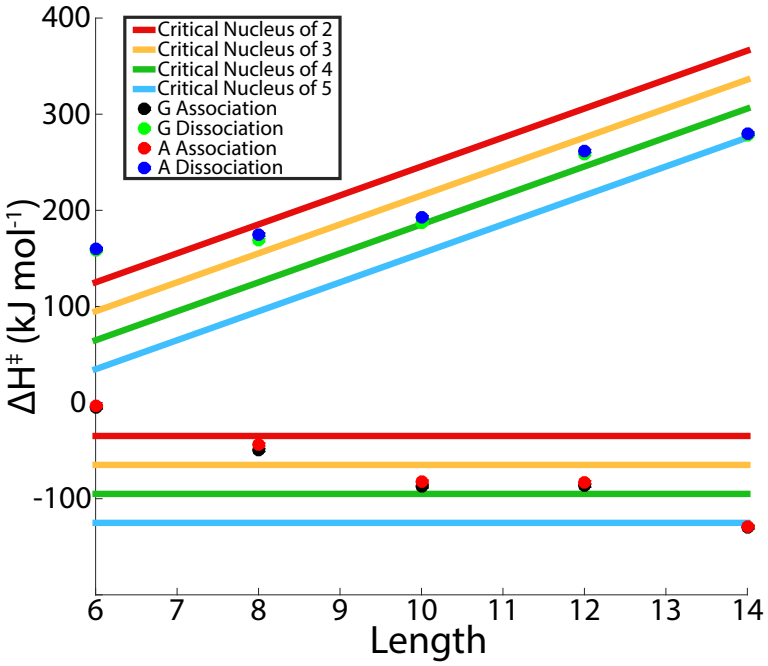

Figure 6.15: Activation enthalpies for association and dissociation from the adenine and guanine ring modes plotted against the predicted activation enthalpy for each sequence length assuming different critical nucleus sizes determined using the nearest neighbor parameters.

termined. The number of intact base pairs at that point will be one less than the number of base pairs in the critical nucleus since it is the largest structure that will dissociate in a downhill fashion back to the monomer state. This means it lies at, or just on the monomer side of, the peak of a standard reaction free energy diagram.

The results of this analysis are shown in Figure 6.15 which contains the activation enthalpies from both the association and dissociation plotted against the predicted activation enthalpy from the NN parameters for a given critical nucleus size at each length. This analysis shows that the number of base pairs in the critical nucleus increases as the overall length of the sequence increases.

It must be acknowledged that this trend cannot be solely attributed to length based on the experimental evidence here since the temperature at which these sequences were studied did increase with length, which can been nicely seen in °C in Figure 6.11 or K^{-1} in Figure 6.7. The temperature difference is more significant at shorter lengths relative

to longer lengths where there is more overlap between the sequences. Coarse-grained molecular dynamics simulations have suggested that the critical nucleus should increase in size with increasing temperature.[3] However, for the purposes of this analysis we will consider the changes observed here to be primarily due to length. The application of the two-state analysis and Eyring analysis presumes an activation enthalpy that is independent of temperature, the validity of which is supported by the linearity in the Eyring plots. As this analysis ties the number of bases in the critical nucleus directly to the activation enthalpy for the sequence we will presume for the purposes of this analysis that the size of the critical nucleus is relatively independent of temperature over the temperature range studied here. This topic will be revisited in Chapter 7 utilizing the kinetic model which is able to independently probe the effect of temperature and sequence length on the critical nucleus.

With that clarified we can return to examining the size of the critical nucleus as a function of sequence length and look to further understand the size increase. Figure 6.15 shows that for every two base pairs added to the overall sequence the critical nucleus increases by a single base pair. This means that the number of base pairs that form during the zippering portion of the reaction must also increase by one. We previously observed that the trends in the entropy and enthalpy of activation as a function of length are correlated to the NN parameters and the value is half of what is predicted by the NN parameters. The additional mechanistic insight gained from the nucleation-zipper model provides an explanation for the factor of two that separates the two values. Adding two base pairs to the sequence results in the addition of two dinucleotides, yet the association and dissociation activation energies only increase by the value of a single dinucleotide. This demonstrates that, for the lengths examined in this study, increasing the overall sequence length by two base pairs increases the size of the critical nucleus and the zippering regime each by a single base pair adding a single dinucleotide. This is in agreement with the results in Figure 6.15. This explains why the activation enthalpies and entropies in our

data increased by the amount predicted by the NN parameters for a single dinucleotide when two base pairs were added to the sequence.

Combining the Eyring analysis with the mechanistic insight from the nucleation-zipper model provides a much clearer picture of the energetic driving forces for DNA association and dissociation. Looking at the increasing critical nucleus size with length suggests that the observed decrease in the activation enthalpy of association is simply due to the exothermic nature of forming a base pair and the fact that longer sequences have more bases in the critical nucleus. The steeper slope for longer lengths in the Eyring plot can be rationalized by the fact that a larger critical nucleus requires more base pairs to be formed and as a result the increased probability of breaking a base pair at higher temperature will have a more significant impact. Additionally, larger critical nuclei will have a larger entropic penalty due to additional bases losing the conformational freedom that they have when unbound. This large negative entropic contribution that increases with length is a significant factor in the association of DNA monomers and offsets the favorable enthalpy of activation resulting in the positive free energy barrier to association. This demonstrates that the barrier to overall DNA association is primarily entropic in nature.

Now that it has been demonstrated that the critical nucleus increases in size as the overall sequence length increases it is worth discussing a possible explanation for this. The stability of the critical nucleus is the result of the favorable enthalpic contribution overcoming the unfavorable entropic component. It is reasonable to assume that after the first G:C base pair all of the remaining A:T base pairs that can make up the critical nucleus will all add similar enthalpic gains that are independent of both position in the sequence and overall length. However, each base pair throughout the sequence is unlikely to have a consistent entropic contribution. A large portion of the entropic loss occurs upon the initial binding event meaning that the initial pairing has the largest entropy penalty. Configurational entropy will also result in longer sequences having a larger entropic penalty upon binding of the first base pair. This means that the initial base pairs have a larger en-

tropy cost at longer lengths than they do at shorter lengths, but the enthalpic benefit is the same for each base pair. This provides an explanation for the increasing size of the critical nucleus. The longer sequences see an increased entropic cost to forming the critical nucleus but the enthalpic gain per base pair does not increase with increasing length. As a result the critical nucleus must increase in size with increasing length so the favorable enthalpic contribution can overcome the entropic cost that increases with increasing length. A similar conclusion in the context of the increasing activation energy of dissociation as a function of increasing sequence length has been noted in the literature.[10]

There are a few aspects of this discussion that are worth highlighting as they will be revisited in the analysis conducted with the kinetic model. The first is in regards to the argument that the critical nucleus size increases with length because additional base pairs are necessary to overcome the larger entropic penalty seen at longer lengths. This is very similar to the argument made in the literature that size increases at higher temperatures because the additional base pairs are needed to stabilize the critical nucleus due to the destabilizing effect of higher temperatures.[3] In reality both of these are likely contributing to our experimental results which we will provide evidence for utilizing the kinetic model. The second aspect is the assumptions that the association always initiates at a G:C base pair when present[10] and that all associations for a given sequence share the same entropic cost. Utilizing the kinetic model we will revisit this to examine the validity of these assumptions while also taking a closer look at the entropic penalty and its dependence on the initiation position. This will demonstrate that it has an even greater impact on the association mechanism than initially realized via the experimental data.

6.4.7 Free Energy Surfaces

With the results of the activation free energies we can also investigate the length-dependence of the free energy landscapes for DNA hybridization. These are shown in Figure 6.16, using the monomer state as the reference state. From the reference state

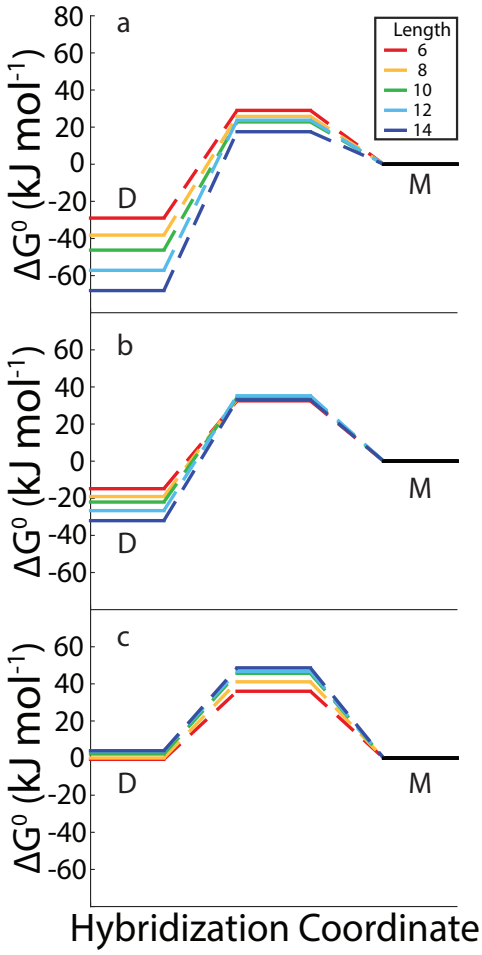

Figure 6.16: Free energy surfaces for each length at 10 °C (a), 40 °C (b), and 70 °C (c) using $\Delta G^{\ddagger}$ values from the Eyring analysis.

the $\Delta G^{\ddagger}_{\mathrm{a}}$ from the Eyring analysis provides the energy difference between the monomer state and the transition state and the $\Delta G^{\ddagger}_{\mathrm{d}}$ from the Eyring analysis provides the energy difference between the transition state and the duplex state. Additional free energy surfaces illustrating the temperature dependence are plotted for sequences 5'-C$(AT)_2$G-3' and 5'-C$(AT)_6$G-3' in Figures 6.17a and 6.17b, respectively. Figure 6.18 complements the free energy diagrams by showing plots of the activation free energy for both the association and dissociation as a function of length at five different temperatures. This alternative representation demonstrates the trends in $\Delta G^{\ddagger}_{\mathrm{a}}$ and $\Delta G^{\ddagger}_{\mathrm{d}}$ more qualitatively and helps to clarify the interpretation of the free energy diagrams.

In both Figure 6.16 and Figure 6.17 the free energy of the transition state reports on

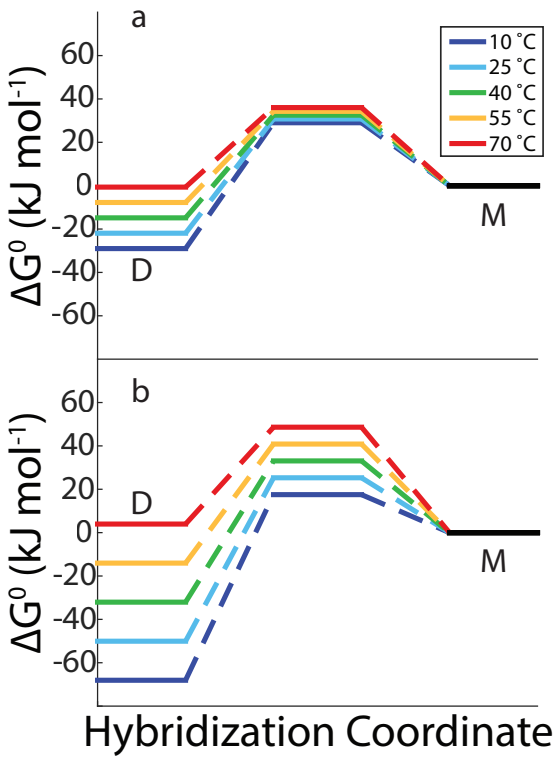

Figure 6.17: Free energy surfaces for 5'-CATATG-3' (a) and 5'-CATATATATATATG-3' (b) at 10 °C, 25 °C, 40 °C, 55 °C, and 70 °C.

$\Delta G^{\ddagger}_{\mathrm{a}}$, because the free energy of the monomer state is referenced at zero. In Figure 6.16a we observe that at low temperature, where association is favored, $\Delta G^{\ddagger}_{\mathrm{a}}$ decreases as length increases. In Figure 6.16c we observe that at high temperature, where dissociation is favored, the trend is flipped. These two trends can also be observed in Figure 6.18. At low temperature the dimer free energy decreases with increasing length while at high temperature it increases with length. However, the trend in the $\Delta G^{\ddagger}_{\mathrm{d}}$ as a function of length does not change and increases with increasing length for all temperatures shown in Figure 6.16. Even though the longest sequence has the lowest energy transition state in Figure 6.16a, it still has the largest $\Delta G^{\ddagger}_{\mathrm{d}}$. Even at temperatures above those shown here the $\Delta G^{\ddagger}_{\mathrm{d}}$ does not definitively flip its trend with length but rather shows no significant change with length as seen in Figure 6.18.

Looking at Figure 6.17 and Figure 6.18 we will now consider the trends in the free energy as a function of temperature for sequences 5'-C$(AT)_2$G-3' and 5'-C$(AT)_6$G-3'. We first

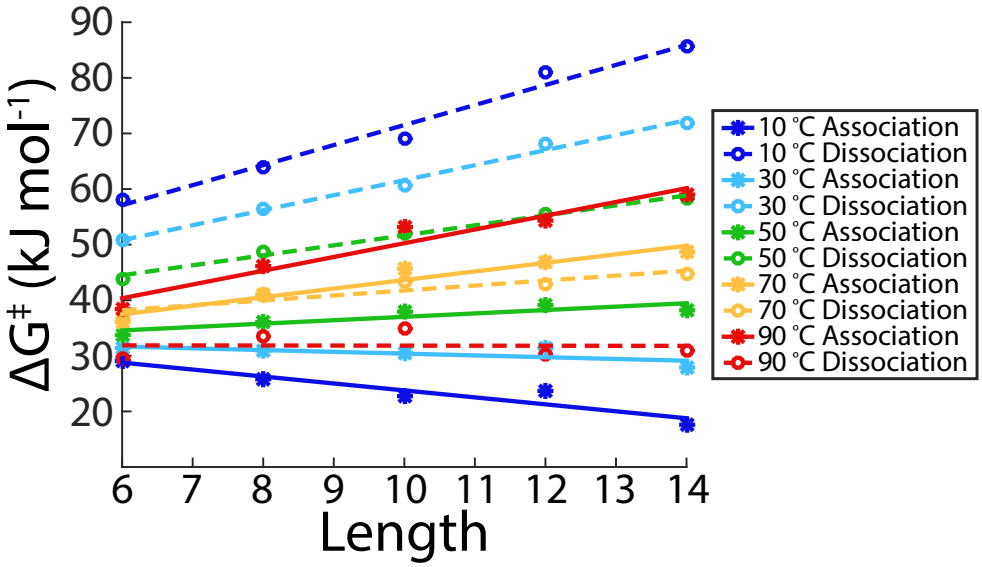

Figure 6.18: Association and dissociation activation free energies determined from the adenine ring mode plotted as a function of length at 10 °C, 30 °C, 50 °C, 70 °C, and 90 °C.

observe that $\Delta G_{\mathrm{a}}^{\ddagger}$ increases roughly linearly with temperature for each length. Additionally, the transition states are more closely spaced for 5'-C(AT)$_2$G-3' relative to 5'-C(AT)$_6$G-3', reflecting that the variation of $\Delta G_{\mathrm{a}}^{\ddagger}$ with temperature for fixed length increases with length, as expected from the linear increase of $\Delta S_{\mathrm{a}}^{\ddagger}$ with temperature. Similarly, $\Delta G_{\mathrm{d}}^{\ddagger}$ decreases with increasing temperature for all length, and we observe a larger temperature dependence at larger length due to the linear increase of $\Delta S_{\mathrm{d}}^{\ddagger}$ with increasing length. This likely reflects the increased conformational freedom that the longer monomers have relative to the shorter monomers. This opposite trend in $\Delta G_{\mathrm{a}}^{\ddagger}$ and $\Delta G_{\mathrm{d}}^{\ddagger}$ results in the larger spacing between the dimer free energies at different temperatures observed for longer lengths.

6.5 Conclusion

In this work we have examined the thermodynamics and kinetics of a series of DNA oligos of lengths ranging from 6 to 14 base pairs. Eyring analysis utilizing a two-state assumption provides additional insight into the energetic driving forces behind the associ-

ation and dissociation of DNA through examining the entropic and enthalpic components to the activation free energy. Eyring analysis demonstrated trends in the activation enthalpy and entropy that are strongly correlated with NN thermodynamic parameters providing a direct link between the thermodynamics and the kinetics. This suggests a kinetic analog of the NN parameters exists that could be used to accurately predict kinetics. However, it is clear that even at these short lengths the reaction may be approaching the limits of the two-state assumption. Further research is needed to study how the kinetics at longer lengths are affected by non-two-state behavior and potentially changing association and dissociation mechanisms. Changes to the dynamics could affect the kinetics in such a way that the predictive power of the NN model, or our global fit parameters, may decrease.

In the work presented here we restricted our analysis to the simplest kinetic model which provides insight into the kinetics of the system and allows comparisons to be made to the literature, where this model is widely prevalent. However, these results open up the opportunity to test them against a wide range of models, such as the kinetic zipper model or Zimm-Bragg model, which we plan in follow up work. Such comparisons will provide further insight and allow additional predictions about the length-dependent kinetics and thermodynamics. Analyzing this data in conjunction with these mechanistic models, with lattice models, or with molecular dynamics simulations should provide additional insight into the dynamics of hybridization and the nature of the transition state in association and dissociation reactions.

6.6 References

1. Hinckley, D. M.; Lequieu, J. P.; de Pablo, J. J. Coarse-Grained Modeling of DNA Oligomer Hybridization: Length, Sequence, and Salt Effects. *J. Chem. Phys.* **2014**, *141*, 035102.
2. Dupuis, N. F.; Holmstrom, E. D.; Nesbitt, D. J. Single-Molecule Kinetics Reveal Cation-Promoted DNA Duplex Formation Through Ordering of Single-Stranded Helices. *Biophys. J.* **2013**, *105*, 756–766.

3. Ouldridge, T. E.; Šulc, P.; Romano, F.; Doye, J. P. K.; Louis, A. A. DNA Hybridization Kinetics: Zippering, Internal Displacement and Sequence Dependence. *Nucleic Acids Res.* **2013**, *41*, 8886–8895.

4. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved Through Infrared Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138*, 11792–11801.

5. Wetmur, J. G.; Davidson, N. Kinetics of Renaturation of DNA. *J. Mol. Biol.* **1968**, *31*, 349–370.

6. Wetmur, J. G. Hybridization and Renaturation Kinetics of Nucleic Acids. *Annu. Rev. Biophys. Bioeng.* **1976**, *5*, 337–361.

7. Pörschke, D. A Direct Measurement of the Unzippering Rate of a Nucleic Acid Double Helix. *Biophys. Chem.* **1974**, *2*, 97–101.

8. Pörschke, D. Cooperative Nonenzymic Base Recognition II. Thermodynamics of the Helix-Coil Transition of Oligoadenylic+Oligouridylic Acids. *Biopolymers* **1971**, *10*, 1989–2013.

9. Pörschke, D. Model Calculations on the Kinetics of Oligonucleotide Double Helix Coil Transitions. Evidence for a Fast Chain Sliding Reaction. *Biophys. Chem.* **1974**, *2*, 83–96.

10. Pörschke, D.; Uhlenbeck, O. C.; Martin, F. H. Thermodynamics and Kinetics of the Helix-Coil Transition of Oligomers Containing GC Base Pairs. *Biopolymers* **1973**, *12*, 1313–1335.

11. Applequist, J.; Damle, V. Theory of the Effects of Concentration and Chain Length on Helix—Coil Equilibria in Two-Stranded Nucleic Acids. *J. Chem. Phys.* **1963**, *39*, 2719–2721.

12. Applequist, J.; Damle, V. Thermodynamics of the Helix-Coil Equilibrium in Oligoadenylic Acid from Hypochromicity Studies. *J. Am. Chem. Soc.* **1965**, *87*, 1450–1458.

13. Williams, A. P.; Longfellow, C. E.; Freier, S. M.; Kierzek, R.; Turner, D. H. Laser Temperature-Jump, Spectroscopic, and Thermodynamic Study of Salt Effects on Duplex Formation by dGCATGC. *Biochemistry* **1989**, *28*, 4283–4291.

14. Pörschke, D.; Eigen, M. Co-operative Non-enzymatic Base Recognition III. Kinetics of the Helix-Coil Transition of the Oligoribouridylic • Oligoriboadenylic Acid System and of Oligoriboadenylic Acid Alone at Acidic pH. *J. Mol. Biol.* **1971**, *62*, 361–381.

15. Craig, M. E.; Crothers, D. M.; Doty, P. Relaxation Kinetics of Dimer Formation by Self Complementary Oligonucleotides. *J. Mol. Biol.* **1971**, *62*, 383–401.

16. Rauzan, B.; McMichael, E.; Cave, R.; Sevcik, L. R.; Ostrosky, K.; Whitman, E.; Stegemann, R. et al. Kinetics and Thermodynamics of DNA, RNA, and Hybrid Duplex Formation. *Biochemistry* **2013**, *52*, 765–772.

17. Altan-Bonnet, G.; Libchaber, A.; Krichevsky, O. Bubble Dynamics in Double-Stranded DNA. *Phys. Rev. Lett.* **2003**, *90*, 138101.

18. Rapti, Z.; Smerzi, A.; Rasmussen, K. Ø.; Bishop, A. R.; Choi, C. H.; Usheva, A. Lengthscales and Cooperativity in DNA Bubble Formation. *Europhys. Lett.* **2006**, *74*, 540–546.

19. Jones, K. C.; Ganim, Z.; Tokmakoff, A. Heterodyne-Detected Dispersed Vibrational Echo Spectroscopy. *J. Phys. Chem. A* **2009**, *113*, 14060–14066.

20. Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A. Transient Two-Dimensional IR Spectrometer for Probing Nanosecond Temperature-Jump Kinetics. *Rev. Sci. Instrum.* **2007**, *78*, 063101.

21. Peng, C. S.; Jones, K. C.; Tokmakoff, A. Anharmonic Vibrational Modes of Nucleic Acid Bases Revealed by 2D IR Spectroscopy. *J. Am. Chem. Soc.* **2011**, *133*, 15650–15660.

22. Banyay, M.; Sarkar, M.; Gräslund, A. A Library of IR Bands of Nucleic Acids in Solution. *Biophys. Chem.* **2003**, *104*, 477–488.

23. Sanstead, P. J.; Tokmakoff, A. Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **2018**, *122*, 3088–3100.

24. Hendler, R. W.; Shrager, R. I. Deconvolutions Based on Singular Value Decomposition and the Pseudoinverse: A Guide for Beginners. *J. Biochem. Biophys. Methods* **1994**, *28*, 1–33.

25. Owczarzy, R.; Moreira, B. G.; You, Y.; Behlke, M. A.; Walder, J. A. Predicting Stability of DNA Duplexes in Solutions Containing Magnesium and Monovalent Cations. *Biochemistry* **2008**, *47*, 5336–5353.

26. SantaLucia Jr, J.; Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 415–440.

27. Manyanga, F.; Horne, M. T.; Brewood, G. P.; Fish, D. J.; Dickman, R.; Benight, A. S. Origins of the “Nucleation” Free Energy in the Hybridization Thermodynamics of Short Duplex DNA. *J. Phys. Chem. B* **2009**, *113*, 2556–2563.

28. Kumar, A. T. N.; Zhu, L.; Christian, J. F.; Demidov, A. A.; Champion, P. M. On the Rate Distribution Analysis of Kinetic Data Using the Maximum Entropy Method: Applications to Myoglobin Relaxation on the Nanosecond and Femtosecond Timescales. *J. Phys. Chem. B* **2001**, *105*, 7847–7856.

29. Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3*, 107–115.

30. Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W.H. Freeman: New York, 1999.

31. Prakash, M. K. Insights on the Role of (Dis)order from Protein–Protein Interaction Linear Free-Energy Relationships. *J. Am. Chem. Soc.* **2011**, *133*, 9976–9979.

32. Chen, C.; Wang, W.; Wang, Z.; Wei, F.; Zhao, X. S. Influence of Secondary Structure on Kinetics and Reaction Mechanism of DNA Hybridization. *Nucleic Acids Res.* **2007**, *35*, 2875–2884.

33. Ohmichi, T.; Nakamuta, H.; Yasuda, K.; Sugimoto, N. Kinetic Property of Bulged Helix Formation: Analysis of Kinetic Behavior Using Nearest-Neighbor Parameters. *J. Am. Chem. Soc.* **2000**, *122*, 11286–11294.

34. Zhang, J. X.; Fang, J. Z.; Duan, W.; Wu, L. R.; Zhang, A. W.; Dalchau, N.; Yordanov, B. et al. Predicting DNA Hybridization Kinetics from Sequence. *Nat. Chem.* **2018**, *10*, 91–98.

35. SantaLucia, J. A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 1460–1465.

36. Sugimoto, N.; Nakano, S.-i.; Yoneyama, M.; Honda, K.-i. Improved Thermodynamic Parameters and Helix Initiation Factor to Predict Stability of DNA Duplexes. *Nucleic Acids Res.* **1996**, *24*, 4501–4505.

CHAPTER 7

THE MECHANISM AND DYNAMICS OF DNA HYBRIDIZATION AND DEHYBRIDIZATION ELUCIDATED BY KINETIC MONTE CARLO SIMULATIONS

7.1 Introduction

The analysis of the equilibrium temperature ramp and transient temperature-jump data provides a great window into the kinetics and underlying energetic driving forces of DNA hybridization and dehybridization. However, the experimental results provide an incomplete picture of the process since deeper mechanistic information is difficult to obtain through experiments alone. Additionally, knowing the rates, or kinetics, provides little insight into the complex mechanistic processes, or dynamics, that occur during DNA association and dissociation. Our goal was to develop a model that can extract dynamic and mechanistic insight from our data in a way that is accessible to experimentally focused researchers both in terms of the model's complexity and computational expense. This model is in no way intended as a replacement for more extensive computational methods such as molecular dynamics (MD) simulations. Rather, by comparing our model to existing models in the literature, we hope to demonstrate that our model is in agreement with respect to the dynamical and mechanistic information that we seek without the additional complexity and computational expense.

A natural starting point for the development of this model is the thermodynamic lattice model that was previously developed by the group.[1,2] The lattice model naturally compliments the experimental results by examining all possible configurations the oligos can adopt and providing insights into the thermodynamics of the system. While the lattice model provides additional insight, there are limits to the conclusions that can be drawn from an equilibrium thermodynamic model. This was the primary motivator for developing

a kinetic model. The thermodynamic model provides the probability of occupying each configuration at equilibrium, but that does not necessarily correlate to the states the reaction is likely to pass through during a hybridization or dehybridization event. The probability of occupying a given state during a single reaction event is not simply determined by the equilibrium free energy for that state, but also depends on the probability of occupying of each of the other states along the possible pathways that lead there. The fact that we want to explore not just the distribution of states, but how the system actually moves through the states leads one directly to the development of a kinetic model.

A kinetic model utilizing Markov state Monte Carlo methods was a natural extension of the thermodynamic lattice model to allow us to further investigate these mechanistic questions. The lattice model provides the kinetic model's state space and the thermodynamic values utilized in the calculation of the reaction rates. Considering the prevalence of the nucleation-zipper mechanism, where base pairs are added sequentially, it made sense to create a model that builds trajectories by stepping through configurations by adding or removing a single base pair. This makes it an optimal system to study using Markov state Monte Carlo methods.

The model, which is introduced and analyzed in this chapter, was applied to a variety of DNA sequences of differing lengths and base pair composition to begin to probe different variables that impact the dynamics and mechanism of DNA hybridization and dehybridization. The main body of experimental results the model is applied to is the length series that was discussed in Chapter 6 and we will continue to refer to these sequences as the length series or CG-ends. This was the most robust data set available and was both the main driving force behind the development of the model in addition to the primary point of comparison for the model's development. Additional data sets are analyzed as well. The sequence 5'-ATATATATAT-3', which we will refer to as AT-all since it is the only sequence studied here that contains only A:T base pairs. This data set will be used sparingly as there are only two temperature points that were collected. Another sequence,

5'-ATATGCATAT-3', will be referred to as GC-core. The purpose of these two sequences was twofold: first, to test how well the model matches experimental results when the number and location of G:C base pairs is altered and second, to better understand how the mechanism of DNA hybridization is affected by changing the sequence composition to a greater degree than was possible without a kinetic model.

Examining this varied set of sequences provides the opportunity to probe numerous different factors that impact the dynamics and mechanism of DNA hybridization and dehybridization. This will also provide insight into the models ability to replicate the different physical processes that occur, particularly the model's ability to replicate experimentally observed non-Arrhenius behavior and early time dynamics. Better understanding the strengths and limitations of the model also provides ideas for future changes to strengthen and improve the model pushing it forward towards new systems of interest beyond these initial investigations.

With regards to probing the dynamics and mechanism of DNA hybridization and dehybridization there is an important limitation that needs to be discussed. The established rules for the model are dictated by the canonical nucleation-zipper model. Since particular mechanistic rules are built into the construction of the model negative results can prove that the mechanism does not represent the physical system but positive results do not definitively prove that the mechanism represents the physical system. This is a result of other mechanistic pathways not being allowed and explicitly tested by the model presented here.

However, more complicated and computationally expensive methods do not have the same restrictions with regards to the possible mechanistic pathways that simulations can follow. Coarse-grained MD simulations do not have any such limitations and have been commonly used to study the dynamics and kinetics of DNA oligos. The OxDNA model and 3SPN.2 model, are two such models that have been utilized to study DNA oligos.[3–9]

Ouldridge et al. used the OxDNA model to examine sequences of 8 and 14 base

pairs. They found that the initial contact between two single strands is stabilized by two to three intact base pairs which is then followed by the remaining base pairs zipping up. They refer to this stabilized configuration as the effective transition state which is enthalpically stabilized by base pairing. At higher temperatures the typical number of base pairs in the transition state increases as more base pairing is required to make duplex formation probable. This increases the activation enthalpy with increasing temperature resulting in non-Arrhenius behavior. They found two reasons for the temperature dependence: at higher temperatures the state with two base pairs itself becomes less stable and new bonds are less likely to form. New bonds are less likely to form because: strands become more unstructured and forming new base pairs generates a smaller free energy gain. Another interesting point is that a free energy diagram built utilizing the OxDNA model has a maximum at a single base pair, in agreement with the lattice model that the kinetic model presented here is built off of, suggesting that the underlying thermodynamics for the two models are in agreement.[4]

The de Pablo group has used their 3SPN.2 model to publish a number of studies of DNA oligos, the findings of three such studies will be highlighted here.[7–9] Utilizing transition path sampling and transition state ensemble analysis they found that DNA re-hybridization is prompted by a distinct nucleation event involving approximately four base pairs.[7,8] The distribution of the transition state ensemble was found to be broader for repetitive sequences than it was for random sequences.[7,8] However, the distributions for a randomized sequence of length 15 and a repetitive sequence of length 14 both had a clear peak in the distribution corresponding to configurations with a size that was about 30% of the overall sequence length.[7,8] Examining sequences with lengths between 10 and 30 base pairs they found that repetitive sequences often observed either sliding mechanisms[7,8] or more complex base pair displacement processes.[9] Homogeneous sequences were also found to commonly follow sliding mechanisms.[9] Random or heterogeneous sequences were most likely to follow the canonical nucleation-zipper mechanism.[7,9] Even

in the case where a mechanism other than the canonical nucleation-zipper mechanism occurs a distinct nucleation event still exists.[7,8] There are two additional findings relevant to the work presented here. For the short oligos studied, ranging from 10 to 30 base pairs, middle to middle nucleation events represented more than 80% of all events.[9] Additionally, as was the case for the OxDNA model,[4] the free energy diagrams generated by the 3SPN.2 model are also very similar to those generated from the lattice model used in the development of our kinetic model, suggesting that the underlying thermodynamics are similar.[8]

In this chapter we will first describe the application of Markov state Monte Carlo methods to DNA association and dissociation reactions. This will be followed by evaluating the ability of the model to replicate the experimental data while also examining trends in the model's parameters to gain some insight into the physical system and the model itself. Finally, we will discuss the insights gained from analyzing the kinetic model with a particular interest in the mechanism by which the sequences associate and dissociate. Comparison with the experimental results provides more detailed mechanistic insight than the experiments can alone since the model distinguishes individual base pairs while the experiments only distinguish G:C and A:T base pairs. The model will also be considered against results in the literature that utilize other methods, particularly MD simulations, to establish the validity of the model and demonstrate how, with regards to the relevant dynamical and mechanistic information, our kinetic model provides many of the same insights but with significantly less complexity. The agreement between our model and the coarse-grained MD simulations discussed previously not only provides significant support for the findings of our model but also mitigates concerns over the interpretation of our findings that result from the mechanistic limitations of our model due to its construction.

7.2 Model Construction

7.2.1 Reaction Scheme

Understanding the construction of the kinetic model starts with understanding the state space and the allowed moves between states. The state space is made up of all possible configurations where all intact base pairs are in-register, meaning they are aligned properly for the formation of the fully formed dimer, and that these intact base pairs form a continuous stretch. Configurations that contain bubbles or out-of-register base pairs are excluded from the model and cannot be occupied. When moving between states only one base pair can form or break during a single step. This means the number of intact base pairs, N_{BP}, must change with every move; moves between configurations with the same N_{BP} are not allowed as that would require both a base pair to form and a base pair to break.

With the exception of the first base pair to form, which can occur anywhere, all subsequent base pair formation must occur adjacent to an already formed base pair. Any move that creates the first base pair between two monomers will be referred to as a nucleation step and any move that creates a base pair next to an already formed base pair will be referred to as a propagation step. Similarly, the only base pairs that can break are the two on the end of the continuous stretch of intact base pairs.

One final aspect about the states that are explored in this kinetic model is that the different possible configurations of the unpaired bases, either frayed ends or monomers, are not explicitly considered. They are however explicitly considered in the lattice model meaning that their energetics are built into the kinetic model. In doing so the model essentially averages over all of the different configurations that the unpaired bases can adopt for a given configuration of intact base pairs. This can be thought of as the model sampling all of the free chain configurations very quickly relative to the making and breaking of base pairs. Now that the state space and rules for moving between states have been outlined

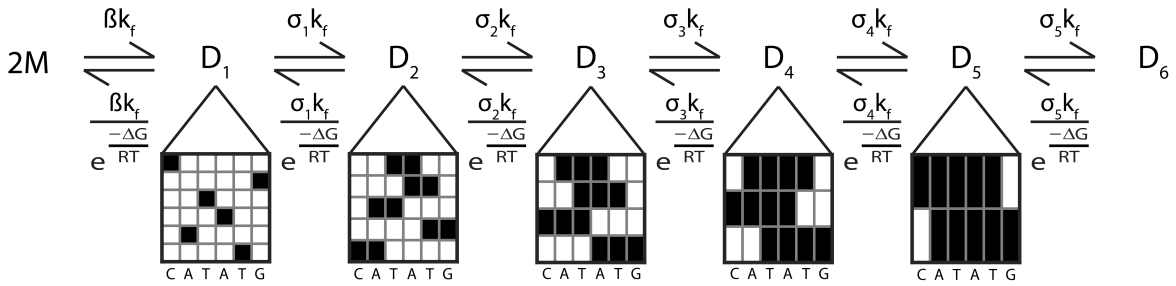

Figure 7.1: Reaction scheme for the kinetic model for the example sequence 5'-CATATG-3'. The boxes below each state show each possible configuration, with each row representing a different possible configuration. A black box represents an intact base pair and a white box represents a broken base pair.

we can look at the full reaction scheme. Figure 7.1 contains the full reaction scheme for the six base pair CG-ends sequence. A short sequence was selected for display to highlight the different allowed configurations for each given N_{BP}. This provides the ability to both highlight what configurations are allowed and to visualize how the model is allowed to move through the system based on the rules provided previously. The general scheme can then be determined simply by extrapolating out for a sequence of any length. While the parameters and rate calculations will be outlined in more detail shortly it is worth highlighting here that the nucleation step is the formation of D_1 and has the rate βk_{f}. The remaining steps are the propagation steps and have a rate of $\sigma_i k_{\mathrm{f}}$. The nucleation step and a series of sequential propagation steps correspond to the two distinct physical processes of nucleation and zippering respectively. The reverse rates are calculated from the ratio of the forward and reverse rates which will be described in the following sections. In the reverse rates R and T, are the ideal gas constant and the temperature at which the system is evolving. The remaining parameters shown in Figure 7.1 are then differentiated by whether they are a thermodynamic parameter or a kinetic parameter and are described in the following sections.

7.2.2 Thermodynamic Parameters

The thermodynamic parameter, ΔG, is calculated from the coarse-grained lattice model, a full description of which has been published elsewhere[1,2] and only a brief overview is included here. At the broadest level the concentration effects associated with the gain in translational entropy that occurs upon the dissociation of the dimer are simulated on a 3D lattice where each site is the size of an individual monomer. Moving a level down, the configurational entropy for each sequence with a minimum of one intact base pair is determined by self-avoiding random walks of beaded polymer chains on a 3D lattice of nucleotide sized sites. At the smallest level the enthalpy of a particular configuration is determined utilizing the NN parameters.[1] The NN parameters used by the lattice model are the "unified" oligonucleotide NN parameters determined by SantaLucia.[10] Finally, there is a single free parameter that is used to account for excess entropy per base pair and its value is selected to ensure that the T_{m} determined by the lattice model matches the value predicted by the NN model.[1] All other parameters are fixed. To determine ΔG for a given move we first must calculate the free energy of a given configuration (δ) the equation for which is

$$G_\delta = -RT \ln\left(p_\delta\right) \tag{7.1}$$

where R is the ideal gas constant and p_δ is the population fraction of configuration δ. For the monomer state p_δ is the fraction of all strands without any intact base pairs and is taken directly from the lattice model. The remaining states are all the configurations with at least one intact base pair, which are referred to as a duplex configurations and are denoted by the subscript D. Note that this language can get confusing since in the case of the two-state assumption a dimer refers specifically to the fully formed duplex since that and the monomer state are the only two states that exist. To alleviate confusion when discussing the model we will refer the state where all base pairs are intact as the fully formed dimer state or the fully formed duplex state. To determine the population fraction of the dimer

states we begin with the partition function

$$q_{\mathrm{D,int}} = \sum_{\delta} W_{\mathrm{D},\delta} e^{\beta E_{\mathrm{D},\delta}} \tag{7.2}$$

where $W_{\mathrm{D},\delta}$ is the degeneracy for a given duplex configuration δ and $E_{\mathrm{D},\delta}$ is the sum of the nearest neighbor dinucleotide enthalpies across the intact dinucleotide subunits of configuration δ. Both of these values are determined by the coarse-grained lattice model. From here the population fraction for each configuration with at least one intact base pair is calculated from

$$p_{\delta} = \Theta_{\mathrm{ext}} \frac{W_{\mathrm{D},\delta} e^{\beta E_{\mathrm{D},\delta}}}{q_{\mathrm{D,int}}} \tag{7.3}$$

where Θ_{ext} is the fraction of all strands with at least one intact base pair. This can then be used in Equation 7.1 to determine the free energy for the configuration.

7.2.3 Kinetic Parameters

In addition to the thermodynamic parameters taken from the coarse-grained lattice model, additional kinetic parameters are utilized in the calculation of the transition rates for the model. These parameters are k_{f}, β, and σ_i. There is also an additional kinetic parameter α that is used in the calculation of σ_i. Two of these parameters, k_{f} and α, are fit parameters, whereas β and σ_i are calculated from these fit parameters and other known quantities.

The first parameter is k_{f}, which has units of s^{-1}, and can be thought of as the "speed limit" for forming a base pair next to an already formed base pair or as the maximum zipping rate.[11–14] We will demonstrate shortly that the rate of formation for a single base pair increases as the number of previously intact base pairs increases until it asymptotically approaches k_{f}. As a result, k_{f} can be thought of as the rate of formation for base pairs at the end of a long sequence of continuous base pairs.

The second parameter, β, is unitless and is used to calculate the rate for nucleation

steps.[11] The β parameter itself is the ratio between the rate for a nucleation step and the maximum zipping rate, k_f. It is important to note that nucleation and zipping are two physically distinct processes. As a result the β parameter does not directly reflect a physical process itself. However, there are a number of physical processes, such as the diffusive motion of the monomers, that impact the value of β. Since bubble states are not allowed, β only factors into the nucleation step and attenuates the rate at which two monomers come together and form the first base pair. To determine the value of β we begin by assuming that the formation of the first base pair can be broken down into two individual steps, the single strands diffusing into the proper orientation and the formation of the base pair once the two monomers are in proximity and properly aligned. With the assumption that these two processes are sequential we can write the timescale for the formation of the first base pair in the sequence as

$$\tau_N = \tau_D + \tau_f \tag{7.4}$$

Where τ_D is the timescale for the two monomers diffusing into proper orientation and τ_f is the timescale for the formation of the first base pair. Due to the assumption that the two base pairs are in proximity to each other and aligned after the first step the formation of the base pair can be presumed to occur at roughly the "speed limit" for base pair formation which is k_f, thus $\tau_f = \frac{1}{k_f}$. It is worth noting here that k_f is treated as a fit parameter when fitting to the experimental data, while the rest of the terms used to calculate β are either known physical quantities or are taken from the thermodynamic lattice model. In the kinetic model the overall rate of formation for the first base pair, which we will denote as k_N, has been defined as βk_f. Utilizing this and Equation 7.4 we can derive the following expression for β

$$\beta = \frac{\tau_f}{\tau_D + \tau_f} \tag{7.5}$$

All that remains is to develop an expression for τ_D.

To approximate τ_D we will utilize the diffusion limited association rate for two identical

spheres with an encounter radius equal to the diameter of the sphere in conjunction with the Einstein relation and Stokes law.[15] The equation for this rate, in units of $\frac{\mathrm{m}^3}{\mathrm{mol}\cdot\mathrm{s}}$ is given by

$$k_{\mathrm{D}} = \frac{8k_{\mathrm{B}}TN_{\mathrm{A}}}{3\eta} \tag{7.6}$$

where k_{B} is the Boltzmann constant, T is temperature in kelvin, N_{A} is Avogadro's number and η is the viscosity of the solution. To then turn this bimolecular association rate into a timescale requires multiplying through by [M], which is the monomer concentration at the initial temperature prior to the arrival of the temperature-jump pulse, and taking the inverse to get

$$\tau_{\mathrm{D}} = \frac{3\eta}{8k_{\mathrm{B}}TN_{\mathrm{A}}[\mathrm{M}]} \tag{7.7}$$

Plugging this expression into Equation 7.5 produces

$$\beta = \frac{8k_{\mathrm{B}}TN_{\mathrm{A}}[\mathrm{M}]}{3\eta k_{\mathrm{f}} + 8k_{\mathrm{B}}TN_{\mathrm{A}}[\mathrm{M}]} \tag{7.8}$$

Calculating β in this way also incorporates the expected concentration dependence for the association of self-complementary DNA single strands.

An important note with respect to β is that in some contexts of the literature it is defined by the σ_i parameters.[16] Or another parameter, often referred to as β apparent, β_{app}, or sometimes still referred to as β, is defined that incorporates a β value as it is defined here in addition to the σ_i parameters to create a single overall attenuation parameter.[11,17,18] In some cases this overall attenuation parameter is also referred to as σ.[19] In other cases instead of differentiating β and σ all of the attenuation parameters are referred to as β_i[12] or σ_i.[13] Thus, it is important to carefully check the definition of the parameters in the literature due to this inconsistency.

The third parameter σ_i, the values of which are contained in the interval $(0, 1]$, attenuates the rate of formation for all base pairs that form next to an existing base pair with

the attenuation decreasing as more base pairs are formed.[11] The subscript i in this case denotes the N_{BP} in the initial state, with respect to the forward, or association, direction, for the move that utilizes that specific σ_i parameter. To put this another way, regardless of if the specific move that is occurring is forming or breaking a base pair, the subscript i denotes the N_{BP} in the state with fewer intact base pairs. When forming a base pair this is the initial state, when dissociating a base pair it is the final state. The reason for this can be visualized by looking at Figure 7.1. The decreasing attenuation with increasing N_{BP} is in line with the conceptual understanding of k_{f} being the rate of formation for a base pair at the end of a long series of intact base pairs. A more complete discussion of the factors that contribute to the value of σ_i as a function of N_{BP} is included later on in this chapter. For the time being we will simply consider it to be primarily due to the additional stability that is associated with the formation of the helical structure that occurs when multiple consecutive intact base pairs exist.[11]

The definition of σ_i requires that the values fall between zero and one and that it starts small, monotonically increases, and asymptotically approaches a value of one. It is worth noting that within the context of this model the value of σ_i is the same for all moves with that i value, regardless of the location of the base pairs within the sequence. To avoid fitting an individual σ_i value for each value of i, which would immediately result in concerns of overfitting, we adopted an alternative description based off a single fit parameter. Based on our definition, and the literature definitions,[11] of σ_i we require that

$$\lim_{i \to \infty} \sigma_i = 1 \tag{7.9}$$

and that σ_i be monotonically increasing with increasing i. The hyperbolic tangent function fits both of these requirements and is a reasonable fit to our intuitive understanding of the

functional form of σ_i resulting in σ_i being defined as

$$\sigma_i = \tanh \frac{\alpha x_i}{1 - x_i} \tag{7.10}$$

where α is a fit parameter in the model that determines how quickly σ_i approaches a value of one and x_i is the normalized value of N_{BP}, $\left(\frac{N_{\mathsf{BP}}}{N}\right)$ where N is the total number of base pairs in the sequence, for the configuration in the move with fewer intact base pairs. The normalized value is used to allow the same function to be used for all sequence lengths.

7.2.4 Calculation of Rate Constants and Assumptions Utilized

This section more clearly defines the origin of the reverse rate and discusses the major assumptions that are made when calculating both the forward and reverse rates for the model. We begin with the assumption that k_{f} is independent of temperature and base pair composition for a given sequence. In other words the value of k_{f} is constant regardless of what temperature the system is evolving at and whether a G:C or A:T base pair is forming. The sequence component of this assumption is commonly made for nucleation-zipper models[11,13,14,16,19–21] since A:T and G:C base pairs are sterically very similar and k_{f} should not significantly depend on stacking interactions.[21]

The assumption that k_{f} is independent of temperature is more contentious in the literature. Models exist that do not include a temperature dependence,[13] while others do by incorporating an activation energy or directly fitting each individual temperature; however, among these models the results are inconclusive. It has been proposed that the activation barrier is small and positive, generally in the range of 1-5 kcal mol^{-1}.[14,16,19,22] This leads to the proposal that the elementary formation of a single base pair adjacent to an intact base pair is diffusion-controlled.[14,16,19] However, caution should be exercised due to studies in the literature examining significantly longer sequences,[19] or fitting as few as two temperatures and acknowledging that under certain experimental conditions

the correct rate as a function of temperature was obtained using an activation energy of zero.[14] Other experimental results, examining sequences with lengths of 8-14 base pairs, demonstrate that k_f varies insignificantly and inconsistently with temperature for a given chain length.[11]

The conceptual understanding of k_f as the ”speed limit” for the formation of a single base pair next to an already formed base pair is consistent with the idea that k_f is the rate for a diffusion-controlled reaction. While a diffusion-controlled reaction would be expected to contain a temperature dependence the resulting barrier is very small, consistent with the 1-5 kcal mol^{-1} previously mentioned.[14,16,19,22] In the work presented here each sequence was studied over a relatively small temperature range with the lowest and highest temperatures being separated by only 10-15 K. Over such a minimal temperature range the change in the solvent viscosity is also relatively minimal. Considering both of these factors within the context of the experimental work that is examined here the impact of the temperature dependence of k_f is considered to be negligible.

Considering the inconsistencies in the literature, and the relative insignificance of a very small activation energy over the temperature range studied here, as a first approximation k_f is assumed to be independent of temperature. This was done in an effort to simplify the model and reduce the number of parameters to alleviate concerns of overfitting, which can easily occur due to the significant number of parameters incorporated into some versions of the nucleation-zipper model.[11,16] Additionally, this allows fitting the model to all temperatures for a given sequence with a single parameter set, rather than fitting a distinct parameter set to each temperature. It is advantageous to do this without needing additional parameters to capture a temperature dependence that is expected to be very small or nonexistent. As a result, fitting two parameters to all temperatures significantly reduces the risk of overfitting.

We will now clarify the calculation of the forward and reverse rates for the model, the rates for forming or breaking a single base pair, before discussing the additional assump-

tions that they require. The only distinguishing factor between different forward rates is the number of previously intact base pairs. As seen in Figure 7.1 the nucleation step proceeds with a rate of βk_{f} and the propagation steps proceed with a rate of $\sigma_i k_{\mathsf{f}}$. The subscript i denotes the N_{BP} prior to the move, such that for the formation of the second base pair the rate is given by $\sigma_1 k_{\mathsf{f}}$. There is no σ_i value associated with the formation of the first base pair so $1 \leq i \leq N-1$ where σ_{N-1} is the σ value associated with the formation of the final base pair.

With the forward rates defined we now look to the calculation of the reverse rates. The forward and reverse rates for moving between any two states, where such a move is allowed, are related by

$$s = \frac{k_i}{k_{-i}} = e^{\frac{-\Delta G}{RT}} \tag{7.11}$$

where s denotes the equilibrium constant following the notation used by many in the literature,[11,19,23] k_i is the forward rate, and k_{-i} is the reverse rate. The value of s is defined as an association equilibrium constant where the forward direction is the formation of a base pair and this convention is used throughout this work. It is important to know that s is unitless. As a result k_{-i} will carry the same units as k_i. Since both β and σ are unitless, k_i and k_{-i} have the same units as k_{f} which are s^{-1}, which is necessary for the transition rate matrix. As a result the reverse rate can be solved by simple rearrangement yielding

$$k_{-i} = \frac{k_i}{s} \tag{7.12}$$

For both the nucleation and propagation type moves, as a result of utilizing the free energy value of each configuration, the reverse rate constants do contain sequence specificity as mentioned previously. This is because the ΔG value depends on the specific configurations of the initial and final states for a given move. Equation 7.12 is then used to calculate the reverse rates seen in the reaction scheme shown in Figure 7.1.

It is worth acknowledging here that the equilibrium constant s only factors into the re-

verse rate meaning this rate carries all of the effect due to the specific identities of the initial and final configurations defined by that particular s. This results in all base pair specificity being carried in the reverse rates, an assumption widely utilized in kinetic Monte Carlo models utilized to study DNA kinetics.[20–22,24–26] An explanation for this is that outside of any diffusion contribution, the formation of a base pair is an elementary reaction. This elementary reaction is predominately driven by steric and structural considerations which would not be expected to carry any significant base pair specificity. With respect to the diffusion contribution, it is incorporated for nucleation steps and its omission in propagation steps was discussed previously with respect to k_{f}.

These rates are inserted into the transition rate matrix **L** as the indices l_{ij}. If l_{ij} is a forward rate it is calculated from the definition based on whether it is a nucleation or a propagation step and then l_{ji} is calculated according to Equation 7.12. There is no correlation between the values of the indices i and j and which state has more intact base pairs such that if $i > j$ the given element l_{ij} could be either a forward or a backward move.

One final assumption that needs to be addressed is that DNA hybridization and dehybridization can be broken down into individual sequential steps making or breaking a single base pair at a time. To consider this we first note that according to the Chapman-Kolmogorov equation, given by Equation 3.4, considering a transition between two configurations with the same N_{BP} as two sequential transitions and summing over all intermediates is the same thing as considering it as one transition over a longer period of time. It is true that making and breaking a base pair could happen simultaneously, particularly the case where a base pair is being simultaneously made or broken on each end of a consecutive stretch of intact base pairs. However, the assumption that association and dissociation can be broken down into rapid individual discrete steps is a hallmark of many models used to study DNA.[11–14,16,20–22,27]

7.2.5 Running Trajectories

The Gillespie algorithm code that generates trajectories is written in C and utilizes inputs both generated by the user and from an accompanying Matlab script. The transition rate matrix, **L**, outlined in the previous section is used by the Gillespie algorithm to generate the trajectories. The steps in the algorithm and the method utilized to select the state that the trajectory moves to and the time step for that move, also known as the exit time, are detailed in Section 3.3. For association (dissociation) trajectories the system starts in the monomer (fully formed dimer) state and runs until reaching the fully formed dimer (monomer) state. In this context the fully formed dimer state refers to the state where every base pair is bound to its native pair. Upon reaching the final state it terminates and the first passage time, time spent in each state, and the states traversed during the barrier crossing are saved. The first passage time in this context is the entire length of the simulation from when the trajectory initiates in the initial state at time zero until it reaches the final state. Logging the states occupied during the barrier crossing event is an important component in the analysis of the model. The barrier crossing for the association (dissociation) is defined as the portion of the trajectory starting with the last time the trajectory was in the monomer (fully formed dimer) state until it reaches the fully formed dimer (monomer) state. If desired the initial and final states can be changed to allow the model to be initiated or terminated at an intermediate state where some, but not all, base pairs are intact.

Once the final parameters were determined for the model a large number of trajectories were run to ensure proper statistics. For the GC-core sequence 5,000 trajectories were run for both the association and dissociation trajectories while for all other sequences 100,000 trajectories were run. These large trajectory sets and the transition rate matrices used to generate them, are the results that are analyzed in this chapter.

7.2.6 Optimization of Parameters to Experiment

To determine the fit parameters α and k_{f}, the model was parameterized against our experimental temperature-jump results for sequences of varying base pair composition and length that have been published previously.[28,29] The parameters for each sequence were fit independently to the observed rate constants from experiment with five or six temperatures included in the fit for each sequence. To compare the simulations to the experimental results a set of association and dissociation trajectories were run for a given set of parameters to determine the mean first passage time for both. The association rate k_{a} was then calculated from

$$k_{\mathsf{a}} = \frac{1}{[\mathsf{M}]\tau_{\mathsf{a}}} \tag{7.13}$$

where [M] is the monomer concentration at the initial temperature prior to the temperature-jump pulse, and τ_{a} is the mean first passage time for association from the model. In the case of the CG-ends sequence the monomer concentration was drawn from the coarse-grained lattice model while for GC-core and AT-all it was drawn from experimental results. However, this distinction is minor as the coarse-grained model has been shown to be in excellent agreement with the experimental results. The dissociation rate, k_{d}, was determined by

$$k_{\mathsf{d}} = \frac{1}{\tau_{\mathsf{d}}} \tag{7.14}$$

where τ_{d} is the mean first passage time for dissociation from the model. The association and dissociation rates were used to calculate the observed rate constant according to a standard two-state kinetic analysis making the assumption that these rates are in response to a weak perturbation, which our temperature jump is assumed to be.[1,28–30] Under this assumption the observed rate constant is given by

$$k_{\mathsf{obs}} = k_{\mathsf{d}} + 4[\mathsf{M}]k_{\mathsf{a}} \tag{7.15}$$

where k_{obs} is the same observed rate constant as the one determined from the experimental data allowing the direct comparison of the two values. The parameters were optimized utilizing a pattern search algorithm that minimized the sum of the squared residuals at each temperature. It is worth noting that these equations are correct for the self-complimentary sequences analyzed here and would need to be altered for the case of non-self-complimentary sequences. The necessary equations for both thermodynamic and kinetic analysis of a non-self-complimentary two-state system are provided in Appendix 5A.

The number of trajectory sets run during each iteration of the fitting algorithm is twice the number of fit parameters, so in the case of fitting k_f and α four trajectory sets must be run each iteration. With thousands of iterations required to optimize the parameters against the experimental results running a large trajectory set each time is not computationally feasible. For this reason the trajectory sets run during the course of the fitting are relatively small, on the order of hundreds of trajectories. Initially a number of optimization routines were run for each sequence with randomized initial parameters until a more concise range in which the parameters were converging was determined. The best parameters from these initial fits were selected and used as the initial parameters for additional optimization routines, using the same method, to determine the final parameters. Once these parameters were determined a full trajectory set was run with these parameters to ensure that the values compared to experiment during the fit were representative of the results of the full trajectory set. This ensured that there was no error due to the small trajectory sets used during the optimization routines.

7.3 Results

7.3.1 Final Parameters and Fit Quality

The parameters returned by the fitting algorithm are given in Table 7.1. The resulting observed rate constants, calculated from the mean first passage time for association and dissociation using the two-state analysis, are compared to those determined by experiment in Figure 7.2 for all sequences except AT-all. AT-all was excluded since only two temperatures are available which makes it a poor metric of fit quality relative to the other sequences. The model is generally in good agreement with the experimental data, particularly at higher temperatures.

Before discussing the fit quality, and the resulting parameters, it is important to discuss the robustness of the fit and the level of confidence in the parameters. There are a number of local minima in the optimization meaning there is not necessarily one clear and unique solution of parameters. The reported values are the best quality fit to the experimental data determined during the optimization. However, there is generally a small set of other parameters that provide fits that could be considered reasonable results since the difference in the results is not necessarily significant. As a result, while the analysis of trends in the fit parameters provides interesting insights, the size of the data set studied,

Table 7.1: Fit parameters returned by the kinetic model for each sequence studied.

sequence	length	k_f (s^{-1})	α
CG-ends	6	5.4344×10^{11}	0.6101
	8	2.0969×10^{11}	1.0790
	10	5.9513×10^{10}	1.5424
	12	4.0526×10^{10}	1.5665
	14	7.4036×10^{9}	2.2705
GC-core	10	3.5993×10^{10}	3.7285
AT-all	10	3.2769×10^{10}	2.8186

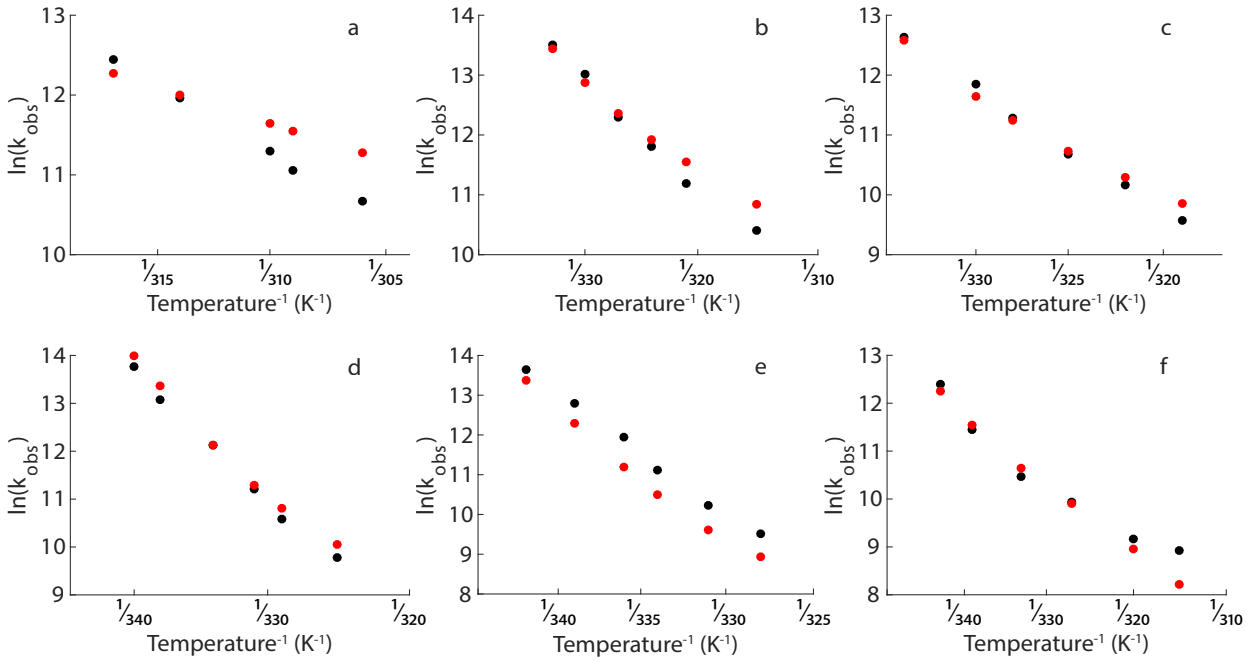

Figure 7.2: Kinetic model observed rate constant (red) compared to the observed rate constant from experiment (black) for (a) 5'-CATATG-3', (b) 5'-CATATATG-3', (c) 5'-CATATATATG-3', (d) 5'-CATATATATATG-3', (e) 5'-CATATATATATATG-3', and (f) 5'-ATATGCATAT-3'

.

both in the number of sequences and the number of temperature points for each sequence, should be expanded in the future to help bolster confidence in the results. Additionally, there are a number of factors and assumptions that go into the parameters utilized in this mathematical model. This means caution should be taken, and the context of the parameters within the model should be considered, while making any connections to real physical processes. It should be noted that the overall mechanistic and pathway information that is provided by the model is not heavily dependent on the exact parameters used as long as they are within range of the reported values. This results in greater confidence with regards to conclusions drawn from the analysis of the overall reaction mechanism relative to the analysis of the individual fit parameters.

One discrepancy between the model and experimental results is the ability of the model to replicate the degree to which the different sequences and lengths demonstrate nonlinear trends in the Arrhenius plots. As demonstrated in Chapter 6 the observed rate

constant is closely related to the dissociation rate constant, where a linear Arrhenius plot is indicative of two-state kinetics dictated by a single temperature independent activation barrier. The model demonstrates a small degree of nonlinearity such that for shorter CG-ends sequences, where the experimental trends are linear, the model does not fully replicate the linear trend resulting in some deviation from the experimental rates at low temperature. The degree of nonlinearity demonstrated by the model appears to be relatively unaffected by sequence length and composition which can be seen by the fact that the model is unable to fully replicate the degree of nonlinearity observed in the GC-core sequence, such that it again deviates from the experimental rates at low temperature. A more thorough discussion of the nonlinear behavior is included in Section 7.4

While discussing the linearity of the rates determined by the model it is worth revisiting the model's construction and the fact that it assumes that DNA association and dissociation can be modeled as sequential steps making and breaking individual base pairs. Initially this construction may seem to be at odds with a system that is known to follow a two-state model, particularly for the shorter CG-ends sequences, which makes this topic worth discussing briefly. In both the case of the model and the experiment the Arrhenius plots for the dissociation rate constant are sightly more linear than the observed rate constant, since some of the nonlinearity in the observed rate is due to the convolution of the association and dissociation rates. However, the difference between the two should be the same for both the model and experimental results. While the model does not generate results that are as linear as experiment for the shortest sequences the curvature of the plots is in reasonable agreement for most of the CG-ends sequences. This demonstrates that even though the trajectories generated by the model are made up a large number of individual steps involving the formation or breaking of a single base pair it is able to reasonably reproduce the two-state behavior of these sequences. However, this should be treated with caution since a two-state assumption is utilized both to calculate the association and dissociation rates from the experimentally determined observed rate

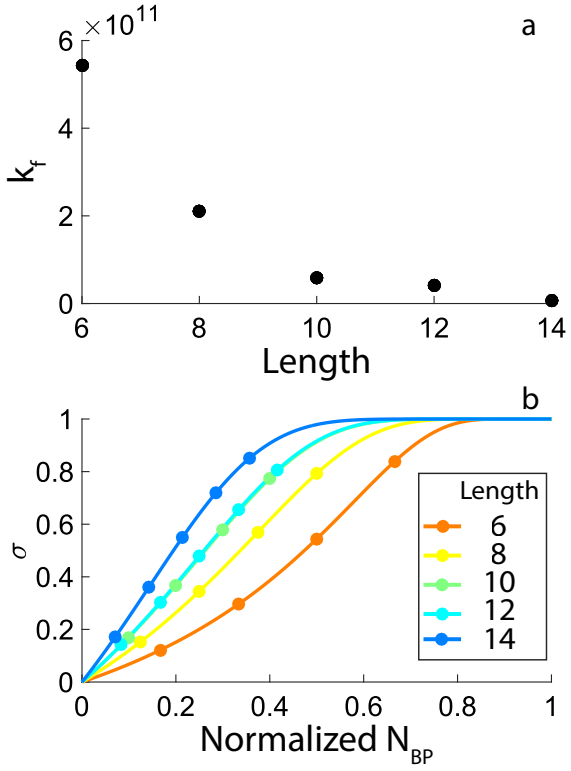

Figure 7.3: The value of k_f as a function of length (a) and the σ values derived from α as a function of the normalized N_{BP} with (•) marking the position of each base pair for a given sequence with an associated σ_i value less than 0.9 (b) for the 5'-C(AT)$_n$G sequences with n = 2-6.

constant and to calculate the observed rate constant from the association and dissociation timescales determined by the model.

Now that overall agreement between the model and experiment has been established it is worth taking a closer look at the resulting fit parameters. Figure 7.3 contains plots of k_f for each length as well as the functional form of σ as a function of N_{BP} normalized by the total number of base pairs for each length. The dots on Figure 7.3b designate the location of σ_i values that are less than 0.9 for each sequence. Base pairs for which the σ_i value is larger are not included for the sake of clarity. In Figure 7.3, k_f clearly decreases as a function of length for the CG-ends sequences and is well fit by a single exponential. However, it is important to note what will be a recurring theme throughout the analysis of the model. It is difficult to disentangle the numerous variables that impact the dynamics

and kinetics of DNA, particularly length, sequence, and temperature. An example of this is that both length and sequence composition affect duplex stability and melting temperature. Sequences with higher melting temperatures must be experimentally studied at higher temperatures making it hard to disentangle the effects of temperature from the effects of either sequence or length. This means that the decreasing trend in k_f as a function of length may also have an underlying temperature component. It is worth pointing out that over the entire temperature range studied across all sequences the viscosity of D_2O changes by approximately a factor of two and as such is not expected to be a significant factor in the observed trend in k_f with sequence length. The potential causes of this trend, both length and temperature, will be further discussed in Section 7.4.

Figure 7.3b shows that the increasing value of α with increasing length seen in Table 7.1 corresponds to the functional form of σ approaching a value of one faster along the normalized x-axis. This results in σ approaching a value of one after approximately 4-5 base pairs regardless of length. It is worth noting that this is approximately half of a full turn of the helix which occurs in 10-11 base pairs. It has been proposed in the literature that only a few intact base pairs are needed to begin to form the double helix structure and obtain the associated stability. As a result σ should approach one in less than a single turn of a helix which is in excellent agreement with our results. An additional note from looking at Table 7.1 is that the values of α for GC-core and AT-all result in σ values that approach a value of one within three and four base pairs respectively. It is interesting that they have the highest α values of all the sequences studied here. While this means that for each N_{BP} value AT-all and GC-core have larger values of σ_i compared to the CG-ends sequence of the same length, they do still approach a value of one in approximately the same number of intact base pairs.

Finally, with respect to β it is worth noting that the values of β calculated utilizing this method are in rough agreement with existing literature values, however those values do cover a large range which, in addition to the wide range of definitions of β that exist in the

literature, makes direct comparisons difficult.[11,12,17,31]

7.3.2 Analyzing Trajectories

Before moving ahead to analyze the trajectories themselves it is worth taking a step back to look at a couple of example trajectories to get a sense for the different aspects that will be shown. Two full dissociation and two full association trajectories are shown in Figure 7.4. Additionally, the final ten nanoseconds of both dissociation trajectories are shown separately. In these figures each dot represents an individual state that the trajectory passes through with the y-axis denoting the number of intact base pairs for the state and the x-axis denoting the time that the trajectory enters that state. Averaging over the

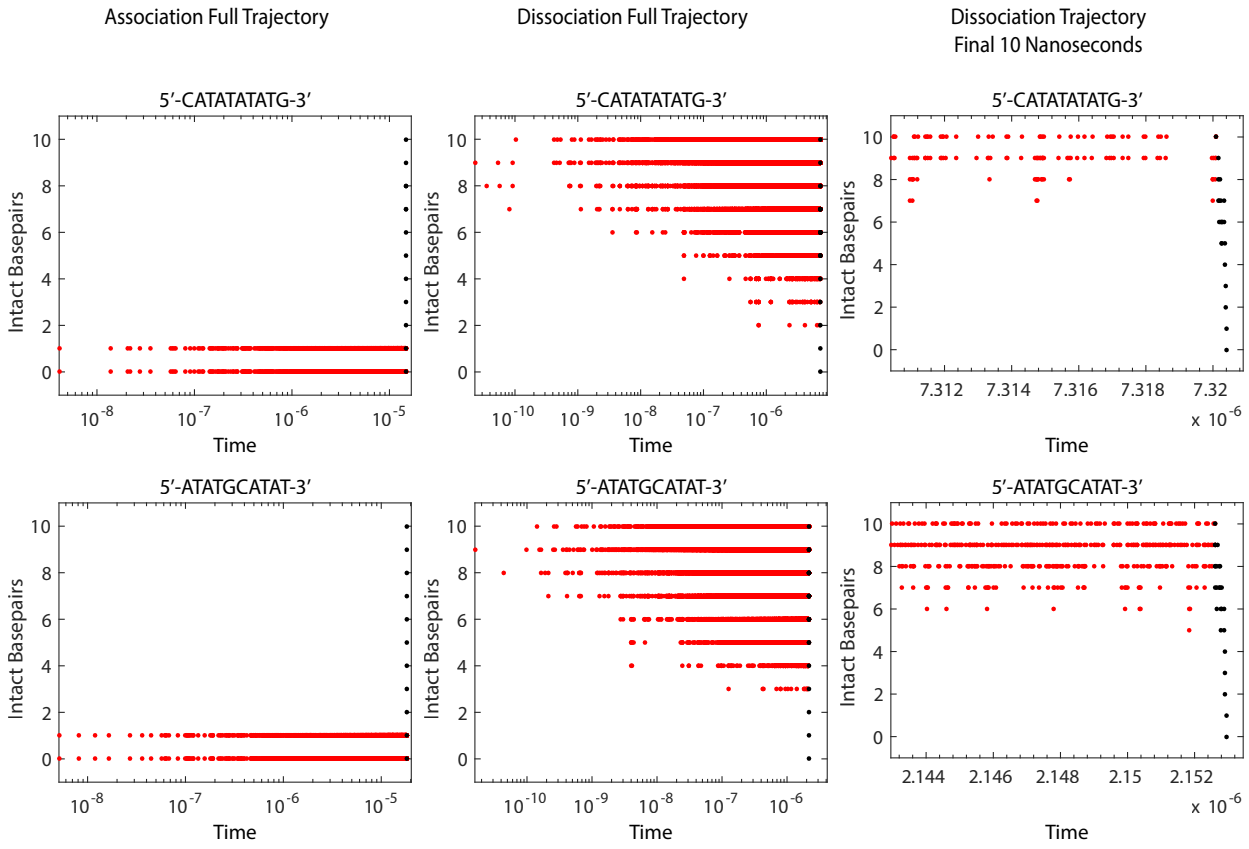

Figure 7.4: Sample association and dissociation trajectories for 5'-ATATGCATAT-3' at 333 K and 5'-CATATATATG-3' at 334 K. The full trajectories are shown in addition to plots highlighting the last 10 nanoseconds of the dissociation trajectories. The black dots in each trajectory represent the time points included in the overall barrier crossing event.

different configurations to just focus on the number of intact base pairs was done in these plots for the sake of clarity. The black dots designate the final barrier crossing event for the trajectory. Here we define the association (dissociation) barrier crossing event as the portion of the trajectory from the last time it is in the monomer (fully formed dimer) state until it reaches the fully formed dimer (monomer) state.

Starting with the dissociation trajectories the first thing to note in Figure 7.4 is the amount of time spent rapidly moving between configurations with different N_{BP} values while regularly returning to the fully formed dimer state. This demonstrates the fraying behavior that is seen to varying degrees across most sequences. The increased fraying behavior for GC-core can be seen by the fact that it both spends more time in states with fewer intact base pairs during the dissociation trajectory relative to the CG-ends sequence and it also reaches states with fewer intact base pairs at an earlier point in time. The increased amount of time GC-core visits states with fewer intact base pairs is particularly clear over the final ten nanoseconds. The fact that in both cases the trajectory regularly returns to the fully formed dimer state throughout the early portion of the trajectory demonstrates the picosecond to nanosecond timescale of the zippering component of the nucleation-zipper model first introduced in Figure 1.1.

The association trajectories in Figure 7.4 help to further highlight the significantly different timescales of the different parts of the nucleation-zipper mechanism seen in in Figure 1.1. The vast majority of the trajectory is spent going between the monomer state and a configuration with a single intact base pair before at the very end forming multiple consecutive base pairs and rapidly zipping up to the fully formed duplex. Due to the vastly different timescales for forming the first base pair, and breaking that base pair, the monomer state is occupied for almost the entirety of the trajectory. This clearly demonstrates the microsecond timescale of the diffusive encounter and pre-equilibrium steps of the reaction which is significantly slower relative to the picosecond to nanosecond timescale of the final zippering which takes up a very small portion of the overall association

trajectories. The time spent in different phases of the association trajectory and how it compares to the physical picture is discussed in more detail later on in conjunction with a closer examination of the rates for forming base pairs, particularly the first base pair, returned by the model.

The analysis here will have two different foci. The first aspect that we will analyze in detail is the barrier crossing event in the trajectories. In Figure 7.4 this portion of the trajectories is highlighted by the black dots. The barrier crossings shown here occur on a timescale of hundreds of picoseconds for the dissociation and a time scale of nanoseconds for the association. These barrier crossing events can involve both forward and backward moves meaning that there is no set number of steps that make up the barrier crossing. However, the minimum number of steps is equal to the number of base pairs in the sequence. For certain sequences it is not entirely uncommon for the barrier crossing event in the trajectory to include two to three times more steps than the minimum amount required.

The second way the trajectories are analyzed is by examining the entire trajectory as a whole. The main focus of analyzing the entire trajectory is examining the fast response observed in experiment for some sequences, particularly GC-core and the longer CG-ends sequences. To do this we will examine aspects of the entire trajectory such as the number of times configurations with each N_{BP} are visited and the average duration of each visit. Doing so will allow us to qualitatively compare the behavior of the dissociation trajectories for each sequence and compare the resulting trends to changes observed in the experimental results to further clarify the dynamics behind the experimental results.

7.3.3 Individual Reaction Pathways

Utilizing TPT analysis, the method for which is described in Section 3.6, individual pathways for barrier crossing events can be isolated and ranked according to the frequency at which they occur. It is important to distinguish what is meant by individual

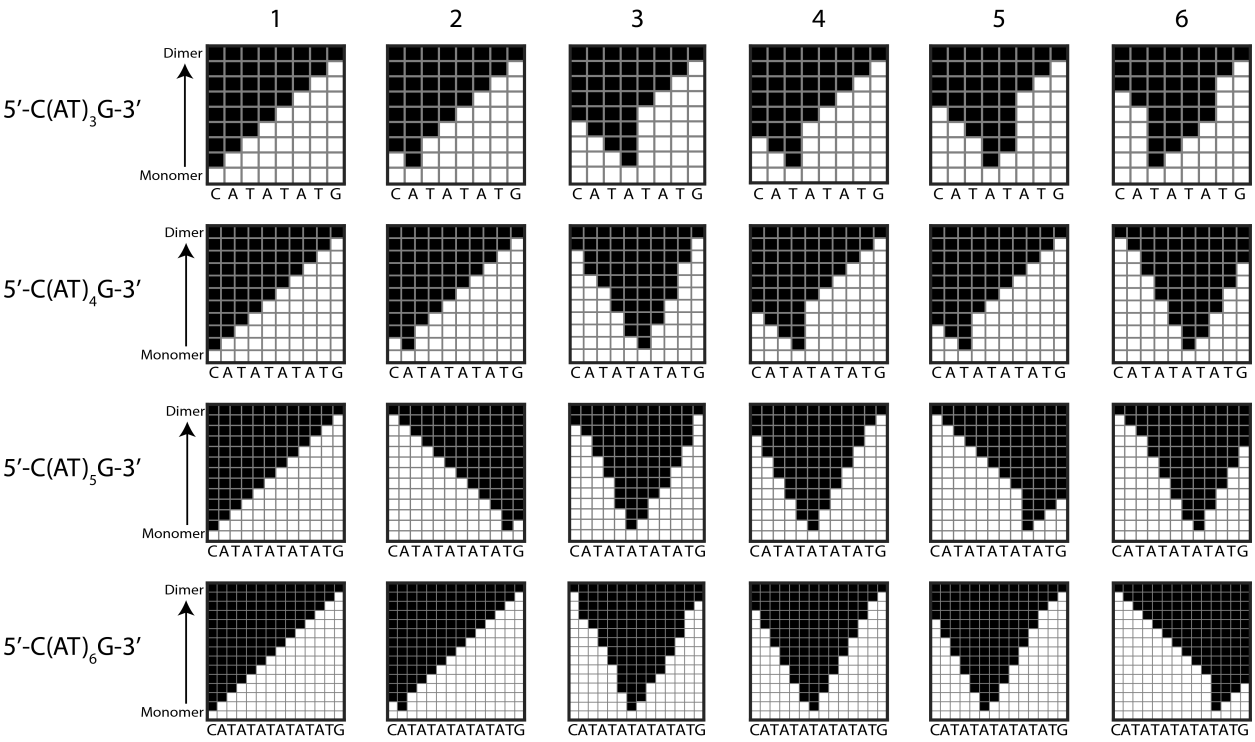

Figure 7.5: Top six pathways for 5'-C(AT)$_n$G-3' sequences with n = 3-6. The pathways are shown at a temperature of 334 K for each sequence except 5'-CATATATG-3' which is 333 K. For each length these pathways are ordered from most probable to least probable from left to right with their ranking denoted by the number above each column. For each length 6-14 these six pathways, and their symmetric partner, make up 87.0%, 69.0%, 57.5%, and 49.5% respectively of the total flux between the monomer state and the fully formed dimer state across all pathways isolated by TPT analysis at the temperatures shown.

pathways from the overall mechanism. Individual pathways are one possible way the system can move through different configurations during either an association or dissociation barrier crossing event. In this context the overall mechanism incorporates the entire distribution of individual pathways and is a more general view of how the system progresses through a barrier crossing event.

Figure 7.5 shows the top six association pathways ordered from left to right according to the probability that an association event will occur along that particular pathway for each CG-ends sequence with lengths of 8-14 base pairs at a temperature of either 333 K or 334 K. The sequences were compared at similar temperatures to isolate mechanistic changes as a function of length. The eight base pair sequence is in the top row of plots

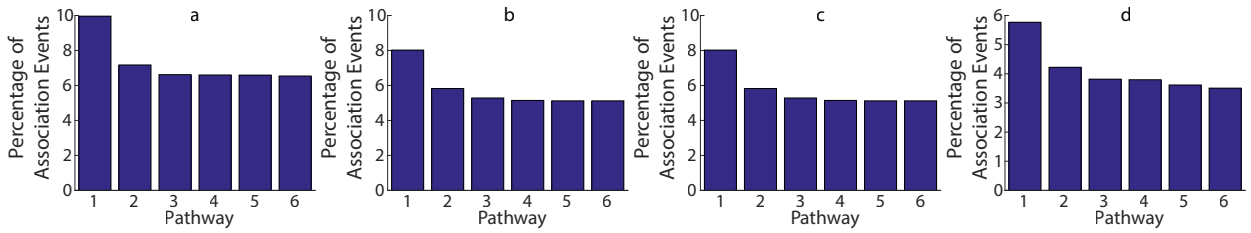

Figure 7.6: Percentage of all association barrier crossing events that occur along each of the top six pathways for (a) 5'-CATATATG-3', (b) 5'-CATATATATG-3', (c) 5'-CATATATATATG-3', and (d) 5'-CATATATATATATG-3' at a sample temperature of (a) 333 K or (b-d) 334 K. Note that each pathway has a symmetric pair that shares the same percentage.

with the sequence length increasing in each subsequent row moving down. Black boxes represent intact base pairs and the plots are read from the bottom up where the first base pair formed is in the second row of the plot and the reaction proceeds up the plot to reach the fully formed dimer state in the top row. It should be noted that this method considers direct pathways and disregards off pathway loops that occur when a trajectory leaves a particular state only to eventually return to that exact state. However, as will be demonstrated later on, this does not significantly impact the ability of the pathways isolated by TPT to represent the pathways of the trajectories generated by the kinetic model. Since these sequences are all self-complimentary each pathway has a symmetric partner that is identical and carries the same probability. For example, the first pathway for each sequence in Figure 7.5 initiates at one end and zips up sequentially across the sequence. The symmetric partner of this pathway is identical except that it starts at the other terminus. For all self-complimentary sequences each pair will be referred to as a unit, for example referring to the two most dominant pathways refers to the two most dominant sets of pathways which is actually four pathways. Figure 7.6 contains the probability that an association event will occur along each particular pathway shown in Figure 7.5 with the pathways numbered according to their ranking, which proceeds from left to right across Figure 7.5. The probability is calculated from the percentage of overall flux between the monomer state and fully formed dimer state that passes through each individual pathway.

The most obvious conclusion to draw from Figures 7.5 and 7.6 is that the two dominant pathways for all lengths initiate at and directly next to the C:G termini respectively. This shows that while there is a distribution of pathways, the simplest pathway for the nucleation zipper picture that proceeds by initiating at one end and sequentially zipping across is the dominant association pathway for this sequence motif. Figure 7.6 shows that there is a significant decrease in probability between the top pathway and the second pathway, followed by a smaller, but still noticeable drop after the second pathway. The difference in the relative probability between the remaining pathways is quite small for all sequences with almost no decrease observed for the shorter sequences. Figure 7.5 also shows that it is advantageous for shorter sequences of this motif to form the termini as quickly as possible. This is best observed in the top four pathways for the shortest sequence all proceeding directly to the closest terminal base pair. For longer sequences proximity to the termini, and forming a terminal base pair early on in the process, becomes less significant. For the two longest sequences, after the two most probable pathways only one of the remaining pathways directly proceeds to a termini. These pathways are the fifth and sixth most probable pathways for the twelve base pair and fourteen base pair sequences respectively. It is also interesting to note that while forming the terminal base pair early on does seem to be favorable, the pathway that forms at the fourth position and proceeds directly to the closest termini is slightly more probable relative to the comparable pathway that initiates at the third position. However, as mentioned previously since these are never one of the top two pathways the difference in probability between them is relatively negligible.

Another interesting observation is that other than forming the first base pair at or next to the termini it is advantageous to initiate near the center of the sequence. In particular, the third most dominant pathway for the three longest sequences all share roughly the same pathway, initiating in the middle and then building out keeping the two frayed ends roughly equal in length until the sequence is fully hybridized. This demonstrates

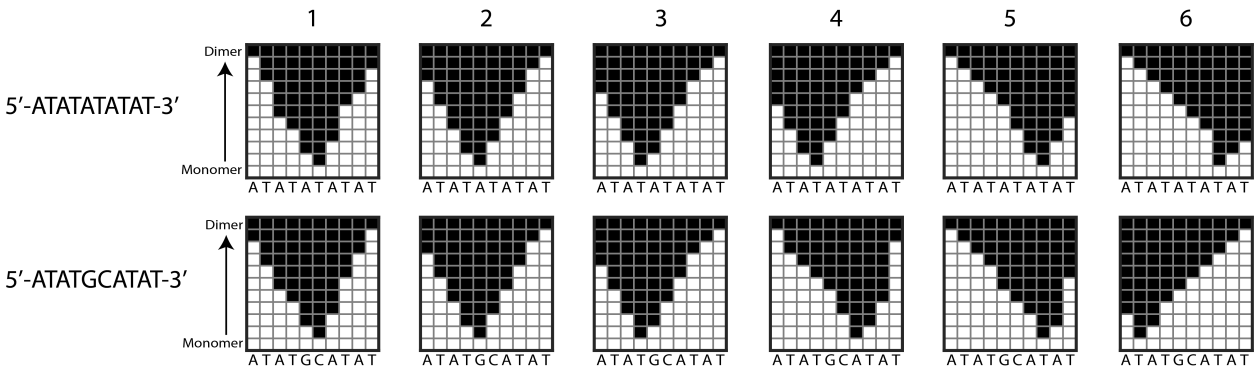

Figure 7.7: Top six pathways for 5’-ATATATATAT-3’ at 308 K (top) and 5’-ATATGCATAT-3’ at 333 K (bottom). Both pathways are ordered from most probable to least probable from left to right with their ranking denoted by the number above each column. At the temperatures shown, these six pathways, and their symmetric partner, make up 64.7% and 74.2% of the total flux between the monomer state and the fully formed dimer state across all pathways isolated by TPT analysis for 5’-ATATATATAT-3’ and 5’-ATATGCATAT-3’ respectively.

that among common pathways there appears to be two main motifs, forming at or next to G:C base pairs, and forming in the center of the sequence and building symmetrically towards the ends. The slight, but consistent, preference for forming at the fourth base pair and proceeding directly to the termini rather than initiating at the third position and doing the same thing also points to a preference for initiating close to the center. We will demonstrate later on in the discussion that pathways that initiate at or near a G:C base pair that forms early on are enthalpically driven, due to the additional stability of G:C base pairs, while pathways that initiate nearer to the center of the sequence are entropically driven. One of the major factors behind the entropic driving force is that the entropy of the configurations that these pathways pass through is more favorable, which has been demonstrated by the thermodynamic lattice model.[2]

The CG-ends pathways can be compared to the AT-all and GC-core pathways which are shown in Figure 7.7 and their corresponding probabilities in Figure 7.8. The two motifs observed in CG-ends are essentially repeated for AT-all and GC-core. Since AT-all does not contain any G:C base pairs only the motif of initiating in the center appears. Figure 7.8 shows for AT-all that there is a relatively small difference between the probability of these

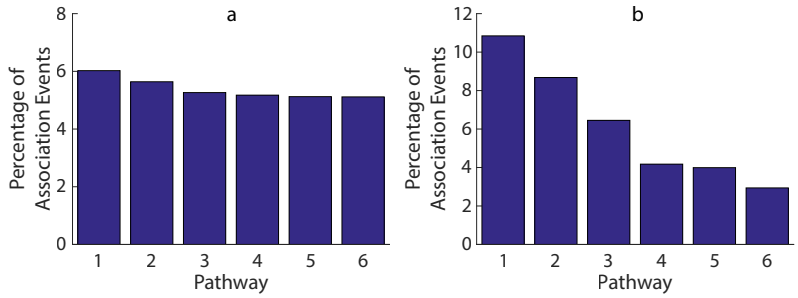

Figure 7.8: Percentage of all association barrier crossing events that occur along each of the top six pathways for (a) 5'-ATATATATAT-3' at 308 K and (b) 5'-ATATGCATAT-3' at 333 K. Note that each pathway has a symmetric pair that shares the same percentage.

pathways and that the top ones do not stand out nearly to the same degree. For GC-core because the G:C base pairs are in the center the two motifs essentially result in the same pathways meaning that these pathways are driven by both enthalpic and entropic driving forces. This results in top pathways that are significantly more dominant than the top pathways for other sequences. The probability of each pathway also drops off more significantly across all six pathways. For the CG-ends sequences there is little difference between the center initiated pathways, AT-all sees a very minor drop across all pathways, and GC-core sequence sees a steep decrease after each of the top three pathways and another noticeable drop after the fifth pathway.

7.3.4 Overall Mechanistic Insights

While the individual pathways are informative on a microscopic scale, it is important to more generally consider the mechanism for monomer-dimer transitions in terms of the overall two-state reaction. However, our focus at this point remains on the barrier crossing event itself. One interesting aspect of the overall mechanism is the probability of initiating a barrier crossing at different positions. Using the pathways isolated by TPT we can determine the percentage of barrier crossings that initiate at each position by summing over all of the pathways, the result of which is shown in Figure 7.9. This can be thought of as the probability that a successful association initiates at a particular position. This

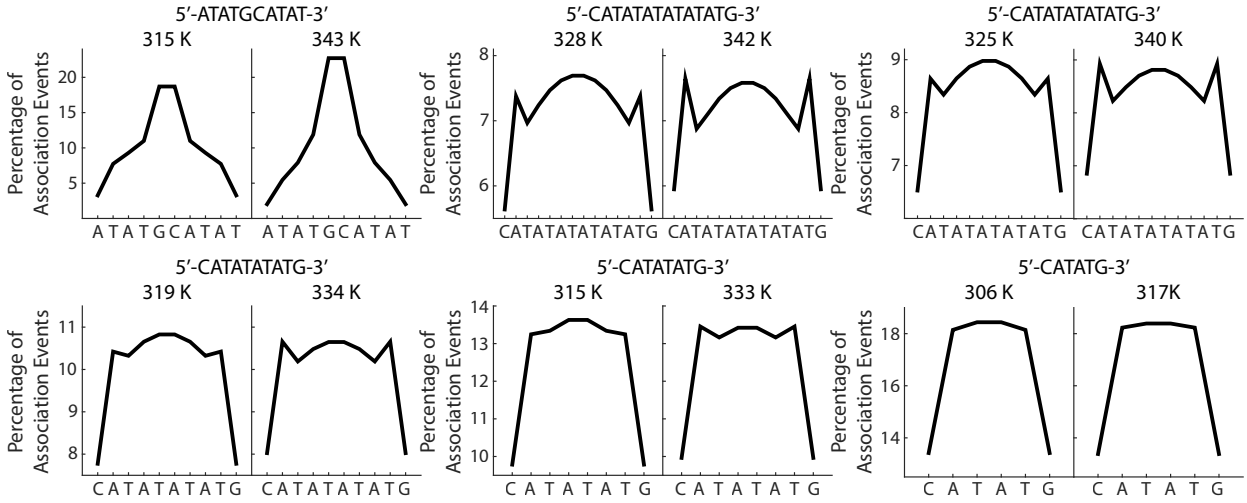

Figure 7.9: Percentage of all association barrier crossing events that initiate at each position for 5'-ATATGCATAT-3' and 5'-C(AT)$_n$G-3', n = 2-6, at the highest and lowest temperatures each sequence was studied at.

can be compared to the trajectories from the kinetic Monte Carlo model by looking at the percentage of all trajectories whose barrier crossing event initiates at each position. The plots generated directly from the trajectories themselves are in excellent agreement with the plots from the TPT analysis as shown in Appendix 7B. This demonstrates that the association pathways isolated utilizing TPT are representative of the association events in the trajectory even though TPT does not isolate every single possible pathway.

It is also interesting to compare these plots to those generated by the last intact base pair for a dissociation barrier crossing event. The last intact base pair is of greater interest than the first base pair to break since the model requires that dissociation initiates at one of the ends, and for self-complimentary sequences the two termini should be essentially identical. The model requires that dissociation initiates at one of the ends because breaking the first base pair in any other position would form a bubble state, which is not allowed by the model. The final intact base pair plots for a dissociation event from both the TPT analysis and the trajectories are in agreement with each other and are also very similar to those for the first intact base pair to form for an association event. This suggests that the association and dissociation barrier crossings can be considered reversible in that they

follow the same mechanistic pathway, just in the opposite direction. While this is not particularly surprising for the TPT analysis since the system is reversible due to utilizing the steady state solution of the transition rate matrix, the trajectories themselves are not run under equilibrium conditions and reversibility is not required. This is an interesting result both because it further supports the idea of microscopic reversibility for the association and dissociation of DNA and also because it provides evidence that enforcing reversibility to simplify the TPT analysis does not impact the methods ability to accurately represent the mechanisms followed by the non-equilibrium trajectories.

One of the more interesting observations gained by examining Figures 7.5 and 7.9 is that while the dominant individual pathway for all CG-ends sequences initiates at a termini those positions are the least likely to initiate a successful association barrier crossing event. The most probable position is actually either next to the G:C termini, the position at which the second most probable individual pathway initiates, or in the center of the sequence depending on temperature. For all sequences except the shortest one, there is also consistently a drop in probability for the third position, relative to the neighboring positions. This is due to the energetic driving forces behind the association reaction. The entropic driving forces preferentially drive barrier crossings that initiate in the center with the benefit decreasing the closer the initiation point gets to the end of the sequence. This is both due to the favorable entropy of the configurations these pathways go through, as mentioned previously, in addition to an additional entropic benefit due to positions in the center having additional pathways available through which association barrier crossings can proceed. These entropic factors explain the dome shape observed in the longer sequences and explains why the fourth position from the end is more probable than the third position. The second position is more probable than the third position because it receives a significant enthalpic benefit from forming next to the G:C base pair at the termini. These energetic driving forces will be discussed in more detail in the discussion section, including an explanation of why the terminal G:C base pairs are such an unlikely initiation position

for CG-ends. The results in Figure 7.9 for CG-ends are in stark contrast to those for the GC-core sequence that demonstrate a very strong preference for initiating in the middle of the sequence, which makes sense when considering the pathways in Figure 7.7 and the overlap between the two main pathway motifs.

The plots in Figure 7.9 show a consistent theme as a function of temperature. For CG-ends sequences a temperature increase increases the probability of initiating a barrier crossing at or next to a G:C base pair with a corresponding drop in the probability of initiating in the center. This would then suggest that the increasing probability of initiating in the center with increasing temperature for G:C core is driven by the location of the G:C base pairs rather than a positional effect.

Another way to investigate the association mechanism is to more directly probe the identity of the transition state or configurations in the transition state ensemble. A common way to analyze Markov state models and isolate structures in the transition state ensemble is to use the committor values, which are calculated for TPT analysis. A common method for determining the transition state ensemble is to select configurations with committor values within some threshold around 0.5, though it should be noted that this analysis is more commonly applied to Markov state models created by binning structures from MD simulations together to create the states.[7,8] This provides more flexibility and control over the states and a more continuous set of states than the model presented here. As such the discussion here will remain broader. Rather than attempting to specifically identify each configuration in the transition state ensemble we will focus on trends in the committor values as a function of length, temperature, and sequence composition to better understand how these variables impact on the makeup of transition state ensemble. We will utilize the forward committor values, which are the probability of going from that particular configuration to the final state, which in the case of association is the fully formed dimer state. The forward committor values are introduced and defined in Section 3.6 Recall that since our TPT analysis presumes reversibility the forward and backward committors sum to one.

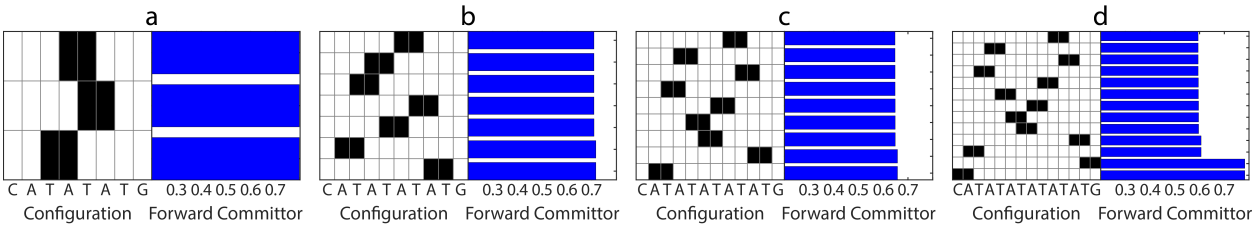

Figure 7.10: All configurations with forward committor values between 0.2 and 0.8 for (a) 5'-CATATATG-3', (b) 5'-CATATATATG-3', (c) 5'-CATATATATATG-3', and (d) 5'-CATATATATATATG-3' at a temperature of (a) 333 K or (b-d) 334 K.

We will start with the CG-ends sequences with 8-14 base pairs at 333 K or 334 K to remove any temperature effects and focus on length. Configurations with a forward committor value between 0.2 and 0.8 are shown in Figure 7.10. In these plots the black and white squares designate intact and broken base pairs respectively. The bar graph to the right of the configurations denotes the forward committor for each of the specific configurations shown. These plots also serve as an example as to why a less rigid description of the transition state ensemble is utilized here. Restricting it to configurations that fall between a narrow window, such as 0.4 to 0.6, would result in at best a limited number of configurations and in some instances there would be none. This is a result of basing the kinetic model on a discrete lattice model whose reaction coordinate is N_{BP}. Even with our expanded range of forward committor values certain pathways will not pass through a configuration with a forward committor value in this range. An easy example of which is the most dominant pathway for sequences of lengths 8, 10, and 12 since the two base pair configuration with an intact terminal base pair does not have a forward committor value in this range.

At these temperatures Figure 7.10 shows that the configurations with forward committor values in this range all have two intact base pairs. The main trend in Figure 7.10 is that, regardless of position, as the length of the overall sequence increases the forward committor values for configurations with two intact base pairs decrease. This is best seen by looking at the configuration with the two central base pairs intact, which is present on all four plots, whose value decreases as the sequences get longer. This is also true for

configurations involving the two base pairs closest to the end that are above 0.8 and off the chart for lengths 8, 10, and 12 but appear for the longest sequence. This shows that as sequences get longer a specific configuration will become less likely to proceed to the fully formed dimer state. This demonstrates that the model predicts that the size of the configurations that make up the transition state ensemble should increase with increasing length.

Our experimental results, discussed in Section 6.4.6 found that the critical nucleus increases in size with increasing length and, while the experiments predict a noticeably larger size increase, it is promising that the model is in agreement with the overall trend. As mentioned previously our experimental results were not able to fully decouple length and temperature. The smaller magnitude of change observed in Figure 7.10, which examines all lengths at roughly the same temperature, provides evidence that the experimental trend does contain contributions from both length and temperature.

The second variable examined with the forward committor values is temperature. Figure 7.11 contains the forward committors for the 14 base pair CG-ends sequence and

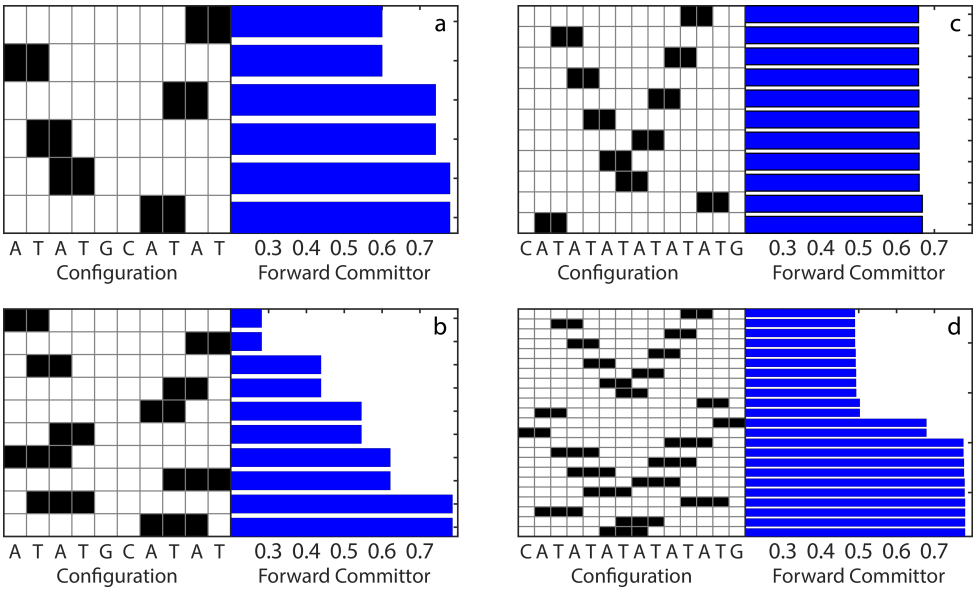

Figure 7.11: All configurations with forward committor values between 0.2 and 0.8 for 5'-ATATGCATAT-3' at 315 K (a) and 343 K (b) and 5'-CATATATATATATG-3' at 328 K (c) and 342 K (d).

GC-core at their highest and lowest temperatures. These sequences were selected because they represent the clearest example of the trends in the configurations with forward committors in the range studied as a function of temperature. Though only these two sequences are shown the trends are representative of all sequences. Like the trend with increasing length, increasing temperature decreases the stability of the configurations resulting in smaller forward committor values. For both sequences in Figure 7.11 when going from low to high temperature the forward committor values for a given configuration decrease. The degree to which is such that configurations with three intact base pairs at high temperature have similar forward committor values to configurations with two intact base pairs at low temperature. Thus, the model clearly demonstrates that the size of the transition state will increase as a function of increasing temperature. This is in good agreement with results from MD simulations in the literature.[4]

This is also in good agreement with our observation above with respect to the experimentally observed increasing critical nucleus size being a function of both length and temperature. Figure 7.11 shows that with an increase as small as 14 K the model predicts the transition state ensemble, as defined here, will include configurations with a single extra base pair. While small, this is still a significantly larger impact compared to the impact of increasing sequence length from 6 to 14. The relatively small impact of both temperat-

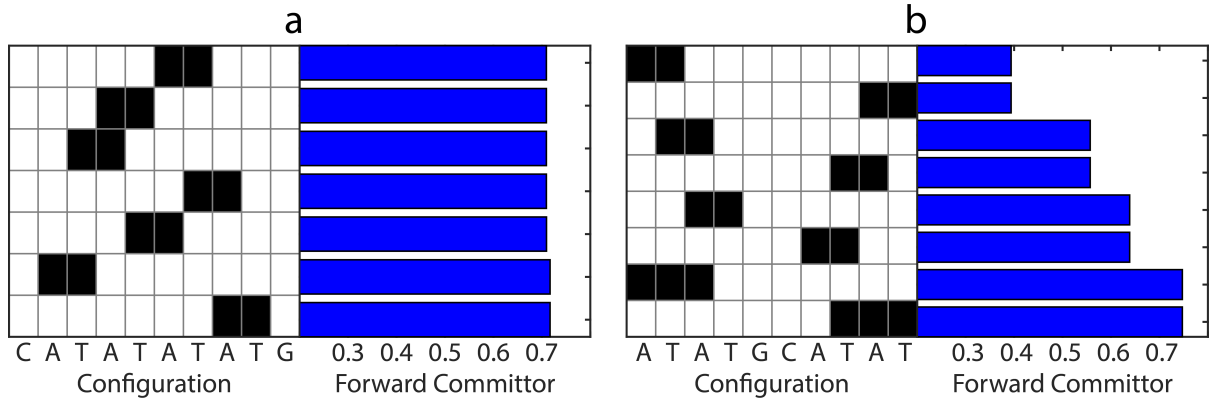

Figure 7.12: All configurations with forward committor values between 0.2 and 0.8 for 5'-CATATATATG-3' at 334 K (a) and 5'-ATATGCATAT-3' at 333 K (b).

ure and length in the model supports the idea that the experimentally observed increase in critical nucleus size with increasing length has both length and temperature contributions.

Finally we will take a look at how the placement of G:C base pairs affects the configurations in the transition state ensemble independent of any temperature or length effects. This can be done by comparing the two sequences shown in Figure 7.12. There are a few differences between the two that show the impact of changing the position of the G:C base pairs. The biggest impact is seen in the GC-core configuration with three intact base pairs that has a forward committor of about 0.7. This shows the effect that shifting the position of the G:C base pairs has on the stability of configurations with only intact A:T base pairs. Even though the %GC is the same for the two sequences there is no configuration for the CG-ends sequence with three intact A:T base pairs. Additionally, in the case of CG-ends the three A:T base pairs could actually be further away from a more stable G:C base pair. Comparing configurations with two intact A:T base pairs at the same position for both sequences we see that in all cases the configurations are more stable for the CG-ends sequence then they are for the GC-core sequence. The sequence effect is significant enough that the CG-ends configuration with two intact A:T base pairs furthest from a stabilizing G:C base pair is more stable than the GC-core configuration with two intact A:T base pairs next to a G:C base pair.

Overall, considering the various sequences, lengths, and temperatures studied here the forward committor values predict that the configurations in the transition state ensemble are made up of approximately 2-3 base pairs. This is in good agreement with the number of base pairs determined by coarse-grained MD simulations for both configurations in the transition state ensemble[7,8] and the critical nucleus.[4] This result is particularly promising considering the relative simplicity of the model presented here.

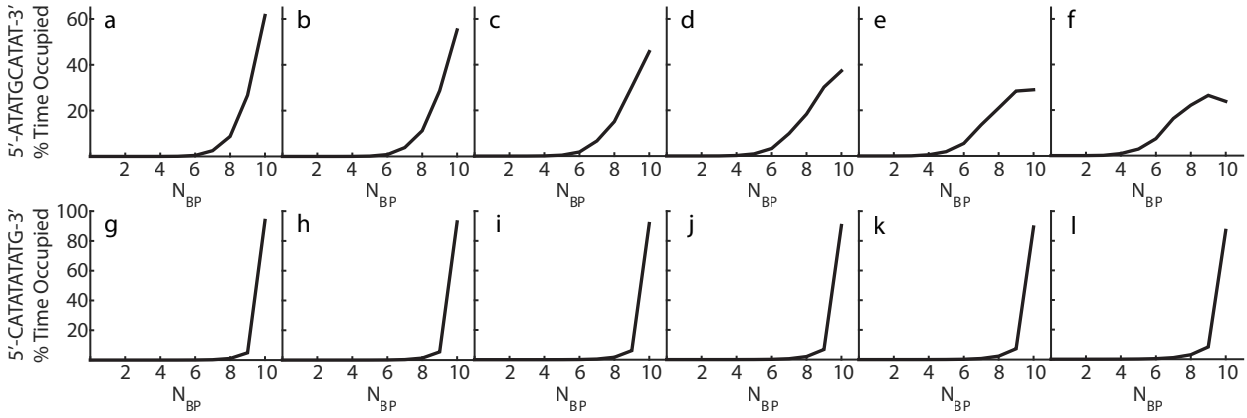

Figure 7.13: Average percentage of time during the simulation that the trajectories spent in states with each N_{BP} for a trajectory starting in the fully formed dimer state for 5'-ATATGCATAT-3' (a-f) and 5'-CATATATATG-3' (g-l). The temperatures for 5'-ATATGCATAT-3' are 315 K (a), 320 K (b), 327 K (c), 333 K (d), 339 K (e), and 343 K (f). The temperatures for 5'-CATATATATG-3' are 319 K (g), 322 K (h), 325 K (i), 328 K (j), 330 K (k), and 334 K (l).

7.3.5 Full Trajectory Analysis

Now we will step back from looking solely at the barrier crossing event and examine the entire trajectory, with a particular eye towards insights the trajectories provide on the experimentally observed fast response. The primary purpose of this is twofold. The first goal is to understand how fraying appears in the model by looking at GC-core, where fraying has been experimentally observed. The second goal is to use this knowledge to gain insight into the fast response that grows in with length in the CG-ends sequences. Analyzing the entire trajectory is difficult since there are thousands of steps in each trajectory as seen in Figure 7.4. However, due to the construction of the model and the fact that all dissociation must initiate at the ends we know that any dissociation must be due to fraying. As a result we do not need to be particularly concerned with the individual configurations and rather simply need to track N_{BP} at each step.

Figure 7.13 contains plots showing the percentage of time spent in states as a function of N_{BP} averaged over all trajectories for GC-core and the ten base pair CG-ends sequence. The CG-ends sequence is used as a point of comparison since it is the same

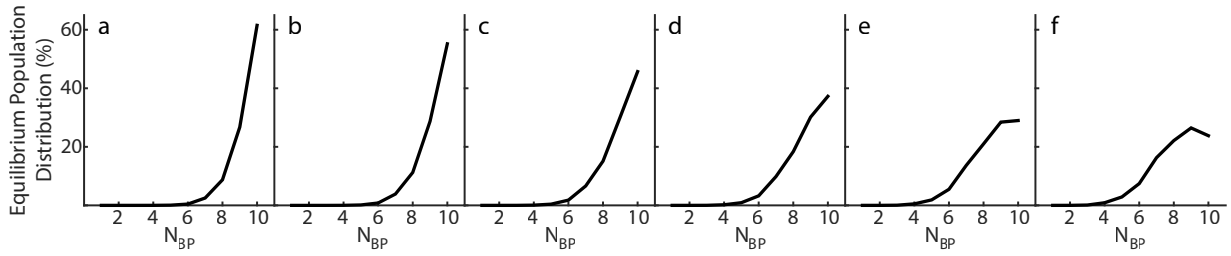

Figure 7.14: Probability of occupying a non-monomer state with a given N_{BP} determined by the thermodynamic lattice model[2] for 5'-ATATGCATAT-3' at 315 K (a), 320 K (b), 327 K (c), 333 K (d), 339 K (e), and 343 K (f).

overall length and relatively little fast response is observed for this sequence. There is a clear distinction between the two sequences. GC-core spends a significantly greater portion of the trajectory in states with multiple broken base pairs, particularly at high temperatures. This demonstrates that the kinetic model replicates the fraying behavior experimentally observed in GC-core while indicating less early time dissociation in the ten base pair CG-ends sequence, also in agreement with experiment.

Not only does the kinetic model agree with the experimental results but also with the thermodynamic lattice model it is an extension of. The plots in Figure 7.13 are an almost exact match to the equilibrium population distribution as a function of N_{BP} from the lattice model, which is shown in Figure 7.14. This demonstrates that over a sufficient number of trajectories the amount of time spent in different states in the kinetic model is primarily dictated by the thermodynamic free energy of the system.

The agreement with experiment is particularly good since the model also demonstrates that for the GC-core sequence primarily A:T base pairs are dissociating at early time. Any configuration with six or more intact base pairs must have both G:C base pairs intact. Among configurations with four or five intact base pairs the vast majority of time is spent in configurations with both G:C base pairs intact. This is demonstrated by looking at the equilibrium probability of occupying each configuration for a given N_{BP} since we have previously established the connection between equilibrium probabilities from the lattice model and the percentage of time spent in each state in the kinetic model. As an example

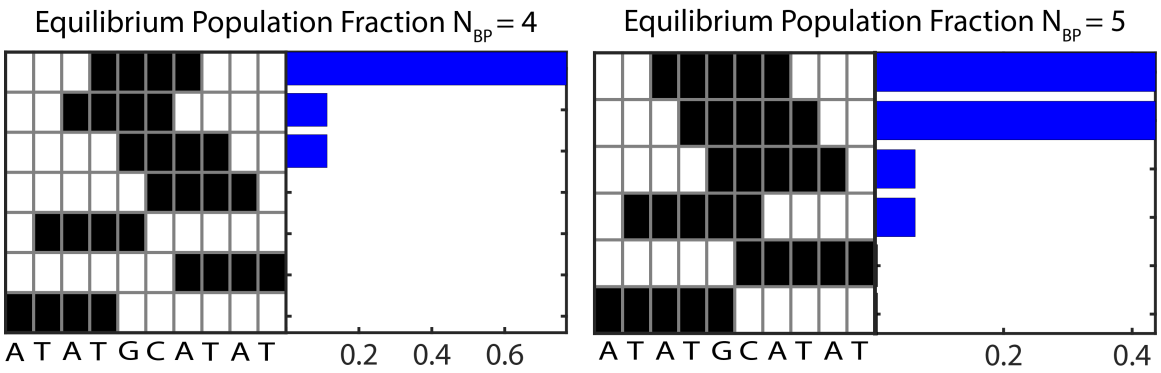

Figure 7.15: Probability of 5'-ATATGCATAT-3' adopting each possible configuration given that the model is in a state with four or five intact base pairs at 343 K. For N_{BP} = 4 and N_{BP} = 5 the probability of occupying a configuration with both G:C base pairs intact is 99.5% and 99.7% respectively.

Figure 7.15 contains the relative probability of occupying each possible configuration with either four or five intact base pairs for GC-core at 343 K. Looking at Figure 7.13, in conjunction with Figure 7.15, shows that very little time is spent in states with fewer than six intact base pairs, and when the model is in those states predominately A:T base pairs are dissociated. This demonstrates that prior to the dissociation barrier crossing almost exclusively A:T base pairs have dissociated, in excellent agreement with experiment.

The GC-core fraying is in stark contrast to the ten base pair CG-ends sequence where the overwhelming majority of time, regardless of temperature, is spent in the fully formed dimer state as seen in Figure 7.13. While the amount of fraying seen for the ten base pair CG-ends sequence is minimal, it is consistent with GC-core in that there is an increase in fraying with temperature. However, the magnitude of this change is significantly smaller.

Now that it has been established that the kinetic model demonstrates the fraying behavior expected for GC-core the same analysis can be applied to the CG-ends sequences to understand what is behind the experimentally observed fast response that grows in with increasing length. Figure 7.16 shows the plots for each length of the CG-ends series at the highest and lowest temperature at which they were experimentally studied. While the trends are small, it is clear that at both temperatures there is a slight increase with length in the time spent in states with broken base pairs. This provides clear evidence that even

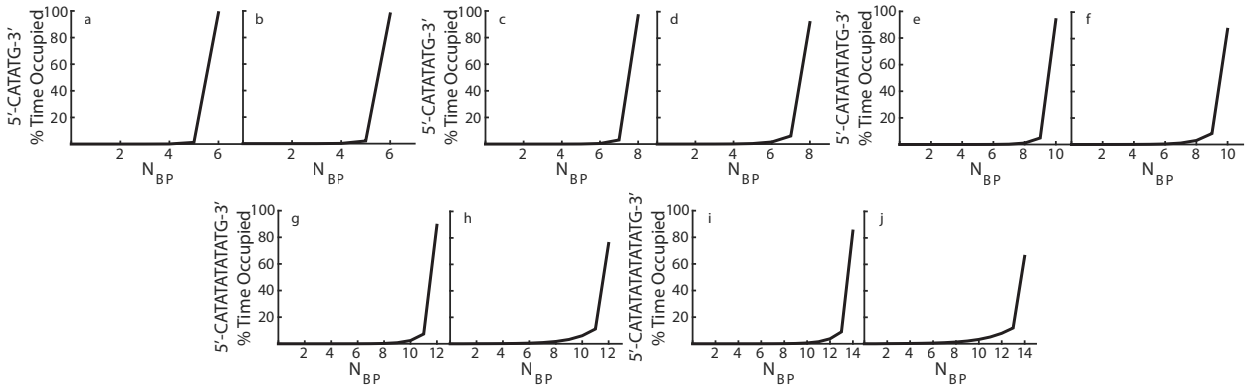

Figure 7.16: Average percentage of time during the simulation that the trajectories spent in states with each N_{BP} at the lowest and highest temperatures studied for each 5'-C(AT)$_n$G-3' sequence as follows: n = 2 306 K (a) and 317 K (b), n = 3 315 K (c) and 333 K (d), n = 4 319 K (e) and 334 K (f), n = 5 325 K (g) and 340 K (h), and n = 6 328 K (i) and 342 K (j).

with the stabilizing G:C base pairs on the termini these sequences become more susceptible to fraying with increasing length. While the trends with length are not nearly of the magnitude observed for GC-core, this is reasonable since the trends observed in the experimental results for the CG-ends sequences are also of smaller magnitude. This suggests that fraying is a likely source of the increasing stretching factor observed in the CG-ends sequences with increasing length.

7.4 Discussion

7.4.1 Kinetic Model Fit and Parameterization

Taking a closer look at the fit quality for the model and trends in the fit parameters provides additional insight into how the model works and its interpretation. The first aspect to consider is the difficulty the model has replicating the changing degree of nonlinearity observed with changing length and sequence. As seen in Figure 7.2 at shorter lengths the experimental data is more linear than the model, while in the case of the GC-core sequence the model is more linear than the experimental data. One potential consideration is that the kinetic model is fit presuming the system is two-state by utilizing the

two-state equation for a small perturbation as described by Equation 7.15, even in cases where the experimental results deviate from linearity. Applying the same assumption for all samples could potentially result in the same degree of nonlinearity in the resulting observed rate constant. Extracting the experimental observed rate constant does not require this assumption. While at first glance this seems to be the obvious explanation, further examination suggests that it is not the only cause.

The distribution of first passage times for the dissociation and association trajectories are both a near perfect exponential distribution, a clear sign of two-state kinetics. However, each individual temperature abiding by a two-state mechanism does not automatically result in a linear Arrhenius plot. A simple system with two states separated by a single barrier will demonstrate non-Arrhenius behavior if that barrier is temperature dependent. This scenario would be consistent with findings in the literature that show that the non-Arrhenius behavior of DNA association is due to a changing activation barrier caused by the size of the critical nucleus increasing with increasing temperature.[4] While this scenario is plausible it does not rule out the possibility that the model returning two-state results for both the association and dissociation for all sequences and temperatures contributes to the difficulty the model has replicating the curvature observed in the GC-core experimental data. It also does not explain why the observed rate constant for the model demonstrates more curvature than the experimental data, most easily observed in the eight base pair CG-ends sequence in Figure 7.2.

The α fit parameter, and corresponding σ_i parameters provide some interesting insight into the interpretation of the nucleation-zipper model. An interesting observation about α is that while it shows a clear trend with length for the CG-ends series, GC-core and AT-all have a significantly larger value than the CG-ends sequence of the same length. As we have demonstrated both GC-core and AT-all prefer to follow very different association pathways, initiating in the center and symmetrically adding base pairs to each side, compared to the CG-ends sequences, where the top pathways initiate at or next to a

terminal G:C base pair. While there is not enough evidence, particularly in the case of AT-all, to make a definitive conclusion this does suggest a link between α and the preferred mechanistic pathways. Additionally, for longer lengths, which also have larger α values, the pathways that initiate at the third or fourth positions and directly form sequential base pairs until a terminal G:C base pair is formed become less favorable relative to pathways initiating in the middle. This trend can clearly be observed in Figure 7.5. While weaker evidence relative to what is observed in GC-core and AT-all this also supports the idea that larger α values may be related to the center initiated pathway motif.

It is also worth considering if GC-core has a larger α value due to the adjacent stronger G:C base pairs, particularly their stronger stacking interactions compared to A:T base pairs.[32,33] It has been previously proposed that stacking is a significant factor in establishing the structure and stability that results in σ approaching one,[11] so it would logically follow that the presence of two G:C base pairs at the most probable initiation sites may result in a larger α. This does not however appear to be the driving factor. The lower α value in the CG-ends sequences could be explained by the fact that the G:C base pairs do not have a neighboring G:C base pair. Additionally, when considering all possible pathways the G:C termini is a relatively unlikely location for initiating a barrier crossing. Furthermore, the idea that the value of α is tied to preferentially forming G:C base pairs early would not provide any obvious explanation for α increasing as a function of length. Also, considering the large α value of AT-all it appears that initiating at the center, regardless of sequence, is linked to a larger α value. If this is indeed true, the fact that GC-core has the largest α value could potentially be due to the preference for center initiated pathways in addition to a small contribution from sequence effects. These two factors may even be linked. It might be as simple as the fact that the location of the G:C base pairs in GC-core drives the most significant preference for initiating at one of the center two positions among sequences studied here, as seen in Figure 7.9. This would mean that the sequence doesn't directly impact the value of α but rather it impacts what pathways are preferred, which in

turn impacts the value of α. While further investigation is required, the results here suggest that mechanism, and potentially sequence to a lesser degree, may have some very interesting effects on α and σ.

Examining the values of k_f for the CG-ends series provides additional insight into the parameters themselves in addition to the model's construction. Two aspects of the parameter will be evaluated. First the values of k_f, and the associated rate of forming the first base pair from the monomer given by βk_f, will be discussed followed by examining trends in k_f as a function of length and temperature. The values of k_f returned by the model presented here range from approximately 5.4×10^{11} to 7.4×10^{9} with the value of k_f decreasing as a function of length. Values of k_f in the literature range cover multiple orders of magnitude ranging from being on the order of 10^6 to 10^9.[11,12,14,19] While these values cover multiple orders of magnitude, as do the values returned by the model presented here, the edges of the two ranges do overlap. Even though the values do show slight agreement with the literature the values are faster than expected, particularly for the shorter lengths. The trends observed with length and temperature, that will be discussed in more detail shortly, suggest that these factors may not be fully accounted for. It is also possible that the magnitude of the rates is a consequence of the construction of the model. Breaking down the association and dissociation into sequential steps of making and breaking individual base pairs, as discussed in Section 7.2.4, is mathematically correct, but may result in a deviation from the physical system. In reality multiple base pairs can diffuse together and form simultaneously, rather than having to do so sequentially. The fact that the model requires the formation of one base pair to occur before the next can start to form could be responsible for the large values of k_f returned by the model. However, further clarity on the role of other potential factors such as length and temperature is necessary before drawing any conclusions with regards to the interpretation of the magnitude of k_f.

While discussing the magnitude of k_f, and its connection to the physical system, it is also important to consider the value of βk_f returned by the model. As mentioned in Section

7.2.3 the value of β incorporates the diffusion limited association rate of two spheres, and since this is orders of magnitude slower than the value of k_{f} this dominates the rate of forming the first base pairs such that it is in close agreement with the diffusion limited association rate of two spheres itself. This demonstrates that the rate of formation for the first base pair has a simple physical interpretation since it is so closely tied to the diffusion limited association rate. The exit time is then determined randomly from an exponential distribution whose parameter is the sum of the rates for all possible moves leaving the monomer state as described in Section 3.3. This results in the exponential distribution having a parameter that is larger than the rate of each individual process for leaving the monomer, by an order of magnitude in the case of a ten base pair sequence. However, this is the stochastically correct method for determining the time at which the trajectory will leave the monomer state and is equivalent to selecting the first reaction that occurs out of all possible reactions each time the trajectory leaves the monomer state.[34] This can help intuitively explain why the exit time for leaving the monomer state is often, but not always, faster than the rate for leaving the monomer state given by βk_{f}. Additionally, the probability of an initial encounter proceeding to the fully formed dimer state has been determined by coarse-grained MD simulations and forward flux sampling to be very small, potentially below 1%.[4,9] This means that a large number of encounters would be expected to occur during the process of two monomers associating to form a fully formed dimer requiring that the formation of the first base pair be multiple orders of magnitude faster than the timescale for the overall association. This however does not fully account for the relatively large rate of formation for the first base pair suggesting that other factors are likely impacting it as well. One such factor that merits further evaluation is the contribution of diffusion to the process. In the model the two monomers must diffuse into proximity each time the first base pair is formed. However, this is not physically realistic. When two monomers are formed due to breaking the only intact base pair they will not necessarily break apart and could reasonably be expected to quickly reform a base pair without a

significant contribution of diffusion to that rate. This would occur significantly faster relative to two monomers that need to diffuse together prior to forming a base pair. Since the model does not distinguish these events, and considers the diffusion contribution to be the same each time, this could potentially factor into the relatively fast rate of forming the first base pair between two monomers returned by the model.

Based on the definition of k_f it would not be intuitively expected to have any significant dependence on length. As such we will now examine this relationship further in an attempt to figure out potential underlying causes of the relationship seen in Figure 7.3a. Base pairs that form with a rate of k_f are at the end of a series of intact base pairs that have adopted the proper double helix configuration and the associated stability. This assumes that σ is purely a function of N_{BP} without considering the number of unpaired bases in the frayed end. If σ approaches a value of one within 4-5 base pairs for all sequence lengths the length of the remaining frayed end must increase with increasing sequence length, the effect of which could be considered.[21] It has been proposed that diffusion plays a significant role in the reaction forming a single base pair and it may even be diffusion-controlled.[14,19] A longer frayed end could slow down diffusion, due to increased drag and a larger mass, and the rate of formation for a single base pair relative to one with the same number of previously intact base pairs but a shorter frayed end. This could potentially explain the decreasing value of k_f with increasing length. If σ_i reaches a value of one too early the model may compensate by reducing k_f, an effect that would increase as the length of the frayed ends increases. This suggests that a better definition of σ_i might consider both N_{BP}, for steric and stability considerations, and the length of the frayed end to account for diffusion. However, over the length of sequences studied here the changes due to slight differences in frayed end length would likely be small and have no significant impact.[21] This makes it unlikely to be the sole cause of the decrease in k_f as a function of length seen here; especially considering that k_f decreases by multiple orders of magnitude.

We must also consider the potential role of temperature since the longer sequences were studied at higher temperature due to their increased thermodynamic stability. The fact that the trend is fit so well to an exponential suggests that the data would be linear on an Arrhenius plot, albeit with a negative activation energy, which is inconsistent with the small positive activation energy expected for a diffusion-controlled process. A negative activation energy for k_f might initially make sense given the well documented negative activation energy for the overall DNA association reaction that is commonly observed, particularly at high temperatures.[11,12,28,29,35] However, both our results and the literature consistently relate the negative activation energy to the early stages of the association mechanism, the formation of the critical nucleus, rather than the elementary rate of formation for a single base pair.[4,11,12,28] Additionally, if we consider a diffusion-controlled reaction to have an activation energy around or below 4-5 kcal/mol[19] we can use that as an estimation of a reasonable magnitude for the activation energy of k_f regardless of sign. In this case an estimation of the activation energy for k_f observed here would be over five times greater than the magnitude of a diffusion-controlled reaction.

One final thought on the values of k_f focuses on the shorter lengths where the most significant decrease in k_f is observed. It is interesting that for these sequences the value of σ approaches one at approximately the fifth base pair for both the six and eight base pair sequences. For longer sequences this remains relatively constant and does not increase further within this length regime. It is interesting to note that as a result of this the two shortest sequences, particularly the shortest one, don't have a significant portion of the reaction that proceeds by zipping at the "speed limit". This does raise some questions over the conceptual definition of k_f as the "speed limit" for base pair formation since the shorter sequences do not undergo rapid zipping at the "speed limit" to nearly the same extent. While no further conclusions can be drawn based on the current information it is interesting to note that the value of k_f appears to level off once it reaches lengths where a number of sequential base pairs are formed at the "speed limit".

The final influence on the k_f value is its use in calculating the β parameter that attenuates the rate of formation for the first base pair. A smaller k_f leads to a larger β so the model may be using k_f to tune β to account for some currently unaccounted for factor. Since β attenuates the rate of formation for the first base, which includes a significant diffusion component, there should be significant temperature and length components, since longer monomers have greater mass. While the calculation of β does carry a temperature and length dependence it may be insufficient and k_f is accounting for this as a result, which could contribute to the observed trend with length.

7.4.2 Energetic Driving Forces Behind DNA Dynamics and Kinetics

Our attention now turns to the driving forces behind the trends observed in the individual pathways and overall barrier crossing mechanisms. Two general motifs were observed, initiating association in the center and initiating at or near a G:C base pair that forms early on. We will now demonstrate that the center initiated motif is entropically driven while the G:C base pair initiated motif is enthalpically driven. These motifs may overlap, resulting in the enthalpic and entropic components driving the same pathways, or they may drive competing pathways.

To demonstrate the entropic nature of the center initiated motif we start with the increased preference for the pathways that follow it, best observed in the top pathway for AT-all in Figure 7.7. The thermodynamic lattice model shows that for configurations with a given N_{BP} the highest entropy state is the one with two frayed ends of equal length, or if an odd number of broken base pairs exists one frayed end is a single base pair longer than the other. The next highest entropy states are those that have two frayed ends but with unequal lengths and the entropy decreases as the difference between the two grows. Finally, the lowest entropy configuration is the one with only a single frayed end. Since the entropy of the system is reduced each time a base pair is formed it is preferred to go to the configuration with the highest possible entropy. This explains the ranking of center

initiated pathways, best seen in AT-all in Figure 7.7.

Since the pathways that initiate at or next to the terminal G:C base pairs in the CG-ends sequences are expected to have a higher entropic cost, the relative preference for these pathways must be enthalpically driven. This is not particularly surprising as G:C base pairs are known to be more stable and it has been previously proposed that they play a role in the early stages of the association process for this reason.[27]

The influence of these enthalpic and entropic driving forces is also seen in Figure 7.9. For GC-core both contribute to the strong dominance of initiating in the center. For the CG-ends sequences the entropic benefit contributes to the dome shape in the center and the enthalpic benefit contributes to the preference for the position next to the G:C termini. However, these factors alone cannot account for how unlikely it is to initiate at the termini which means there must be an additional factor at play with respect to the probability of initiating at each position.

This factor is an additional entropic benefit to initiating near the center of the sequence that does not appear until the entire distribution of pathways is considered. There are more pathways that can initiate in the center compared to positions nearer to the ends. A clear example of this is that only one pathway initiates at each terminus, the most dominant CG-ends pathway that zips straight across, whereas there are numerous pathways initiating in the middle generated by changing the order in which bases are added to both sides. Even though these pathways become increasingly unlikely, when combined together the contribution becomes significant. This provides an explanation for why in Figure 7.9 the dome shape in the center becomes more prominent with increasing length. As sequence length increases combinatorics dictates that the number of pathways available to positions in the center will increase at a faster rate relative to positions closer to the end and there can only ever be one pathway initiating at the termini. The number of available pathways for each position along a sequence follows the binomial distribution and can be found by looking at the row of Pascal's triangle that contains the number of entries equal to the

sequence length. The preference for initiating successful association events in the center is thus the result of both the increased number of pathways and the entropic benefit to each individual pathway. This also explains why the CG-ends termini are surprisingly improbable when considering the full distribution of pathways; while it is the most probable pathway, it is also the only pathway.

Evidence for the energetic driving forces is also observed in how temperature impacts the preference for different initiation positions. These effects can be seen in Figure 7.9. With increasing temperature each CG-ends sequence shows a decrease in the preference for initiating in the center and a corresponding increase for initiating at or next to a G:C base pair. With increasing temperature GC-core's preference for initiating at the G:C base pairs in the center increases, even though both the entropic and enthalpic driving forces drive this preference. Increasing temperature magnifies the contribution, to the association free energy barrier, of the unfavorable entropy due to forming base pairs. This means that both increasing the enthalpic gain, by prioritizing G:C base pairs, and minimizing the entropic penalty, by initiating near the center, would help to minimize the increase in the association free energy barrier with increasing temperature. The fact that for all sequences the probability of initiating at G:C base pairs increases suggests that the additional enthalpic benefit gained from forming G:C base pairs early in the process provides a more significant benefit, with regards to mitigating the effect of increasing temperature, relative to minimizing the entropic penalty by initiating in the center.

Additional evidence for the greater significance of the enthalpic driving force is seen in the ranking of CG-ends pathways in Figure 7.5 where for each length at least the top two pathways are enthalpically driven. Furthermore, these pathways are expected to be very entropically unfavorable but, as a result of the favorable enthalpy, are significantly more probable than the most favorable entropically driven pathway for each length.

7.4.3 Literature Comparison

Before ending the discussion of individual pathways and the overall mechanistic picture it is important to make comparisons to the literature. Coarse-grained MD simulations have found that contacts in the center of the sequence are critical for hybridization, particularly in the case of more randomized sequences where internal rearrangement is not possible.[7,8] For both randomized and repetitive sequences nucleation is biased towards the center[7] and one study found that middle to middle nucleation events represent more than 80% of all those possible for all oligos examined.[9] All of which is in great agreement with our findings.

Considering G:C base pairs, it has been proposed that sequences that contain them are expected to initiate at their position.[27,35] While we do find a preference for forming at or near G:C base pairs, it is still very location dependent and not overwhelming. While the findings for GC-core do show that a large number of initiations will occur at the G:C base pairs it is still less than 50% of all initiations for all temperatures in the range studied here. For CG-ends this number is even lower with initiation at the terminal G:C base pairs making up less than 28% of all initiations for the shortest sequence and less than 12% for the longest sequence. This further demonstrates that, for CG-ends, while initiating at a G:C base pair does appear to result in a dominant individual pathway, when considering the mechanism as a whole the relative significance of that pathway diminishes, particularly for longer lengths.

7.4.4 Identity of Critical Nucleus and Transition State

While the identity of the transition state, and the related and often discussed critical nucleus, has received significant attention in the literature it has remained elusive and difficult to definitively observe, particularly through experimental methods. In this section we will start by further clarifying the relationship between the transition state and the critical nucleus. We will then dive into different angles of analysis that directly probe either the

critical nucleus or the transition state ensemble. In both cases our primary focus is on their size and location while also identifying trends as a function of sequence length and temperature.

Before jumping in it is worth ensuring that the terminology used and the connection between the terms is clear. The critical nucleus, shown in Figure 1.1 is defined as the minimum number of base pairs such that the partially formed duplex is stable and the remaining base pairs rapidly zip up in a sequential and downhill fashion orders of magnitude faster than the formation of the critical nucleus. The transition state is defined as the configuration at the peak of the reaction free energy diagram such that the probabilities of going to the monomer and fully formed dimer states are roughly equal. It is important to recall that there is an ensemble of configurations that fits this definition due to the dynamical nature of the reaction and the multitude of available pathways. Comparing the definitions makes the relationship between them clear. Considering a two-state reaction diagram for a particular pathway the transition state is at the highest free energy point and the critical nucleus is just off the peak on the side of the dimer. Throughout this section the two terms are both used since different analysis methods are focused on one or the other. However, by keeping their relationship in mind, any insights into one can be applied to the other.

While it may seem overly complicated to utilize both of these related, but not identical, reaction intermediates it will hopefully become clear why both are useful components of the analysis. A concrete example of why it is useful in the context of this model to utilize both the critical nucleus and the transition state comes from the committor values used to determine transition state configurations. As mentioned previously, Figure 7.10 demonstrates how there are a number of pathways that do not have a transition state according to the definition used here. However, each pathway must have a configuration that fits the definition of the critical nucleus. While the critical nucleus may not be the most obvious or intuitive point of emphasis within the association process, each association event

must contain a critical nucleus, making it a useful configuration to highlight, particularly for pathways where no transition state exists.

Looking at the individual pathways provides the first insight into the size of the critical nucleus. For both GC-core and CG-ends, pathways that form a G:C base pair in the first or second step are more probable than those that form a G:C base pair later, with the difference being particularly striking when forming the G:C base pair first as can be seen in Figures 7.6 and 7.8. This suggests that the additional enthalpic benefit from forming a G:C base pair is significantly less advantageous after the first two base pairs have been formed. This is also supported by the plots in Figure 7.9 where a decrease in probability is seen between initiating at the second and third base pairs from the end. Changes in Figure 7.9 with temperature provide support as well. At higher temperature as the enthalpic benefit from the G:C base pair becomes more important. For both GC-core and CG-ends, the positions at or next to a G:C base pair increase in probability while all other positions decrease, further demonstrating that forming a G:C base pair in the third step provides relatively less benefit. This suggests that the partially formed duplex is stable prior to the third base pair forming implying a critical nucleus of two base pairs, in good agreement with the literature.[4,11,12] It is also in reasonable agreement with our experimental results discussed in Chapter 6. While there are no clear trends in critical nucleus size with either length or temperature, the forward committor values, which already demonstrated that small trends exist, are better suited for analyzing them and we will do so later on.

Two additional interesting observations can be made based on a critical nucleus size of two. The first demonstrates an interesting connection between the enthalpic benefit from forming G:C base pairs and the critical nucleus. The enthalpic benefit is significantly greater if the critical nucleus is not yet formed. The second observation provides some insight into the relative dominance of the top two CG-ends pathways. These two dominant pathways are the only ones that contain a critical nucleus with a G:C base pair, presuming a critical nucleus size of two. Looking at the additional stability of G:C base

pairs we note that while G:C base pairs do have some additional stability, relative to A:T base pairs, from the extra hydrogen bond, the larger component of the additional stability comes from increased stacking interactions.[32,33] If one assumes that the full benefit from stacking requires the neighboring base pair be intact it would be expected that the two pathways would have roughly the same probability. Since this is not the case, and the difference in probability between the two is quite large, this might suggest that the G:C base pair is gaining additional stability due to stacking interactions with its unpaired neighbor. This would suggest that the frayed end is adopting a relatively structured conformation that allows for some stacking interactions. While this observation is interesting it is worth noting that this model does not resolve the conformation of the frayed end and further investigation is necessary utilizing methods, such as coarse-grained MD, that are better suited for directly probing frayed end conformations.

The analysis of the critical nucleus also provides insight into the relative likelihood of it forming at different positions along the sequence. Presuming that the critical nucleus contains two intact base pairs, looking at the probability of initiating at different positions shown in Figure 7.9 provides some insight into the location of the critical nucleus. This suggests that for the GC-core sequence the critical nucleus has a very high probability of forming near the center while in the case of CG-ends it is likely to be found either near to the center or contain a terminal G:C base pair, with the balance between the two having a temperature dependence. This rationale behind this conclusion follows the same reasoning as the relative probability observing initiations and various locations discussed earlier.

We now shift to analyzing the transition state ensemble by examining the committor values that are shown in Figures 7.10, 7.11, and 7.12. Across all of the different temperatures and sequences the configurations with committor values in the range shown are made up of two or three base pairs which is consistent with the size of the critical nucleus of two base pairs. This is also in good agreement with the size of the most com-

mon configurations in transition state ensembles for similar sequences determined utilizing coarse-grained MD simulations.[7,8]

Now we will use the transition state analysis utilizing the committors to explore how the configurations in the transition state evolve as a function of different variables including temperature, sequence, and length. It is helpful to discuss these trends in conjunction with reexamining the experimental critical nucleus analysis presented in Chapter 6 considering the new insight and context provided by the model. This is done because the analysis of the committors provides the best ability to independently analyze each variable while controlling for the others. The experimental results found that the critical nucleus requires additional intact base pairs at longer lengths because the additional enthalpic gain is needed to overcome the increased entropic penalty from binding longer monomers.

One drawback of the experimental data is that even with the ability of IR spectroscopy to independently resolve A:T and G:C base pairs the kinetic analysis is limited to a relatively broad approach that assumes a two-state mechanism. This means the experimental data is unable to distinguish where along the sequence the critical nucleus forms but, as we have demonstrated, the model can. The experimental analysis presumed that the first dinucleotide to form was the terminal CA dinucleotide, the most likely initiation position according to our thermodynamic model and the literature.[27,35] Since our work suggests that the critical nucleus should be made up of two base pairs or more it is not significant which of the two base pairs in the CA dinucleotide formed first. So the analysis used to determine the size of the critical nucleus from experiment assumes the association mechanism follows one of the top two pathways shown in Figure 7.5 for all lengths. While they are the most dominant their combined probability is still only 20-40% depending on sequence length. However, utilizing the same analysis assuming that the critical nucleus forms in the center and contains no G:C base pairs does not significantly change the predicted size of the critical nucleus. This is because changing a single G:C base pair to an A:T base pair is not a huge effect, though the effect would become more significant in sequences

with more G:C base pairs.

However, since this analysis is conducted with the activation enthalpy determined from a two-state analysis of the data, it assumes the entropic penalty is the same regardless of where the critical nucleus forms. Since the analysis determines the number of dinucleotides necessary to equal or surpass the experimentally determined activation enthalpy, changing the initiation point changes what dinucleotide units are included but not the target enthalpy. The model has demonstrated that the entropic penalty depends on initiation position, which is not considered in the analysis of the experimental data. So while changing the initiation position did not change the size of the critical nucleus, it would be interesting to see if that is still true if the position dependence of the entropic penalty were fully considered. More generally, the ability to determine the activation enthalpy and entropy for individual mechanistic pathways would provide the ability to explore these questions with significantly more detail and provide valuable information on what drives, and distinguishes, the different available mechanistic pathways.

We continue our analysis by examining trends as a function of temperature and how they influence our understanding of the kinetics. Figure 7.11 shows that transition state configuration size increases with increasing temperature. It has also been established that the critical nucleus size is correlated to the activation enthalpy. Increasing the number of base pairs in the critical nucleus increases the activation enthalpy as well. Since activation enthalpy is very closely related to activation energy determined by Arrhenius analysis this demonstrates a direct connection between the commonly discussed non-Arrhenius behavior of DNA association kinetics with the transition state and critical nucleus.[4,35] This also helps explain why, as mentioned in the literature, the pre-equilibrium step is responsible for the negative activation energy of association.[4,11,12,28] The critical nucleus increases in size with increasing temperature, and the individual configurations that are formed along the path to the critical nucleus are less stable and more prone to dissociating back into the monomers. This makes the critical nucleus harder to form at higher temperatures.

More time is spent fluctuating between configurations prior to forming the critical nucleus resulting in a slower rate. Since the zippering portion of the reaction is many orders of magnitude faster regardless of temperature the overall association rate decreases with increasing temperature resulting in the negative activation energy of association.

It is interesting to note that the argument for why the critical nucleus increases in size with increasing temperature is almost identical to the argument for why it should increase with increasing sequence length. In both cases extra base pairs in the critical nucleus are necessary to offset additional instability, whether it is due to higher temperature, or the larger entropic penalty that comes with increased sequence length. The decrease in the forward committor values for configurations with a given number of intact base pairs that occurs as a result of increasing length, independent of temperature changes, is shown in Figure 7.10. The fact that both have similar physical explanations, in addition to the kinetic model suggesting that both trends exist, further supports the conclusion that both independent trends in the model are smaller in magnitude then the trend observed in experiment because the trend in experiment carries contributions from both effects.

The final aspect, as shown in Figure 7.12, is the effect of moving the G:C base pairs around in the sequence. For the GC-core sequence configurations of three base pairs that include a terminal base pair have a relatively high probability of returning to the monomer state. If a configuration with three base pairs includes even a single G:C base pair the probability of going to the fully formed dimer state is very high. This suggests that even within a given sequence, the critical nucleus size may differ depending on where along the sequence it forms and what its base pair composition is.

7.4.5 Fast Dynamics and Fraying

As we have previously established the amount of time GC-core sequences spend, during the trajectories, as a partially formed duplex is representative of fraying. At longer lengths the CG-ends sequences are also observed to spend more time as partially formed

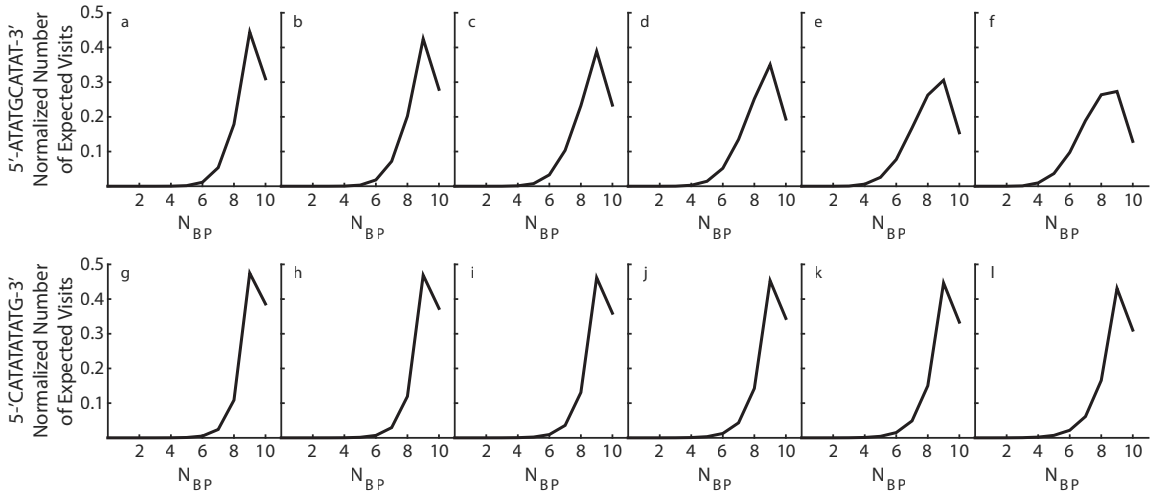

Figure 7.17: Expected number of visits to configurations with each N_{BP}, normalized to the total number of expected visits to all configurations during the trajectory, for a trajectory starting in the fully formed dimer state for 5'-ATATGCATAT-3' (a-f) and 5'-CATATATATG-3' (g-l). 5'-ATATGCATAT-3' was calculated at 315 K (a), 320 K (b), 327 K (c), 333 K (d), 339 K (e), and 343 K (f). 5'-CATATATATG-3' was calculated at 319 K (g), 322 K (h), 325 K (i), 328 K (j), 330 K (k), and 334 K (l).

duplexes, though the distribution of time as a function of N_{BP} has a different form as seen in Figure 7.16. The amount of time spent in a state over the course of a trajectory depends both on the number of times that state is accessed and the duration of each visit. To start to understand the interplay between these factors we can look at the fundamental matrix from the absorbing Markov chain analysis introduced in Section 3.5. Briefly, the fundamental matrix **N** contains the elements n_{ij} which are the expected number of times a trajectory will visit transient state j given that it started in transient state i. To determine the number of visits to each possible configuration for a dissociation trajectory the monomer is set to be the only absorbing state and i is the dimer state. Figure 7.17 shows the values obtained from the fundamental matrix for the GC-core sequence and the ten base pair CG-ends sequence at each temperature. Combining the expected number of visits with the average amount of time spent in states with each N_{BP} value during the course of a trajectory allows the calculation of the average duration of each visit which is plotted in Figure 7.18 for the same sequences.

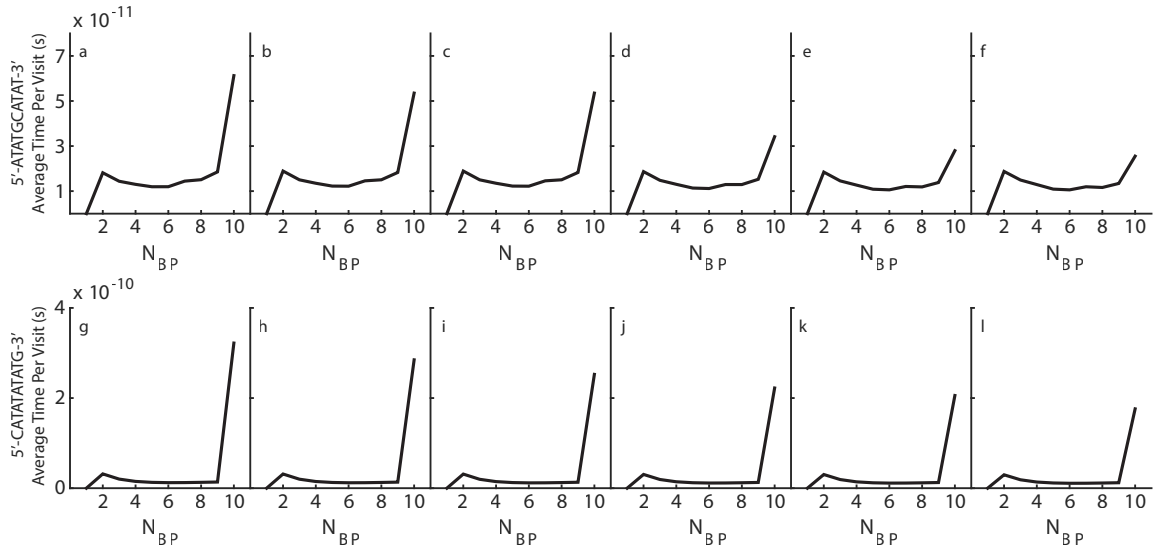

Figure 7.18: Average time per visit, in seconds, to configurations with each N_{BP} for 5'-ATATGCATAT-3' (a-f) and 5'-CATATATATG-3' (g-l). 5'-ATATGCATAT-3' was calculated at 315 K (a), 320 K (b), 327 K (c), 333 K (d), 339 K (e), and 343 K (f). 5'-CATATATATG-3' was calculated at 319 K (g), 322 K (h), 325 K (i), 328 K (j), 330 K (k), and 334 K (l).

Figures 7.17 and 7.18 combine to show that the stark differences seen in Figure 7.13, both as a function of sequence and temperature, are primarily the result of changes in the number of visits rather than their duration. The normalized number of visits for GC-core seen in Figure 7.17 very closely tracks the changes in Figure 7.13. A significant drop in visits is observed for states with nine or ten intact base pairs and a corresponding increase is observed for states with fewer intact base pairs. The same is true for the CG-ends sequence of the same length, while the trends in both Figure 7.13 and Figure 7.17 are much smaller for CG-ends; they still match up quite well.

Looking at Figure 7.18 the average time per visit does not appear to be significantly affected by sequence, at least to the same degree as the number of visits in Figure 7.17. The only real significant difference between GC-core and the ten base pair CG-ends sequence is the average time per visit to the fully intact dimer state and the state with a single intact base pair. In the case of these two states the values between the two sequences differ by an order of magnitude with GC-core being an order of magnitude smaller in both cases. For configurations with a value of N_{BP} ranging from two to nine the average

amount of time per visit to these states is very similar, easily within an order of magnitude. Furthermore, both observe the same relative trend of shorter visits at higher temperature with the fully intact dimer state seeing a substantially larger effect than other configurations. This suggests that the fraying observed in the GC-core sequence is predominately due to an increase in the relative accessibility of the frayed states rather than any significant change in the timescales for entering and leaving them. This in turn further supports the idea that the frayed response is due more to the thermodynamics and probability of occupying a given state throughout the course of a trajectory rather than the kinetics of moving between states. The thermodynamics can be interpreted as changes to the reaction free energy surface, particularly the states in the dimer well, which is in agreement with the free energy surfaces calculated by the thermodynamic lattice model, as previously demonstrated. While this agreement may not seem surprising, since the kinetic model util-

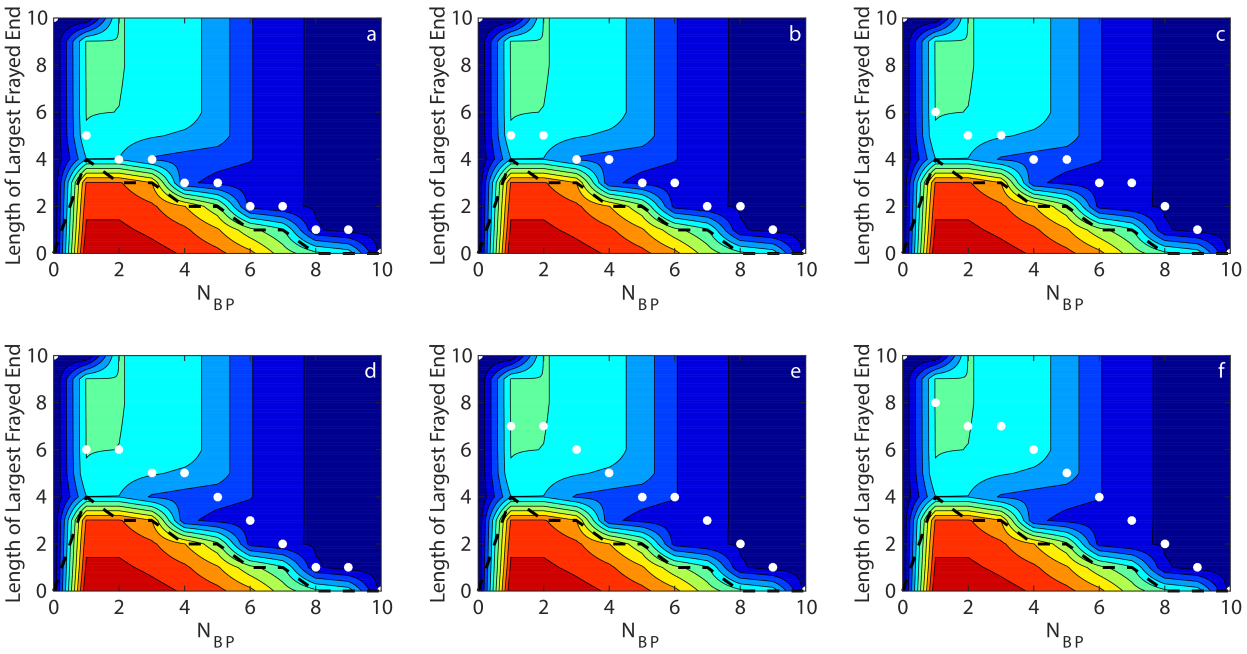

Figure 7.19: Lattice model free energy surfaces at 333 K for 5'-ATATGCATAT-3' where blue denotes the most favorable free energy and red the least. Configurations on or below the black dashed line must include at least one bubble and are therefore not allowed. The white dots represent the top six pathways predicted by TPT in descending order (a-f). The probability of a successful association event occurring along each pathway according to TPT is: 10.84% (a), 8.68% (b), 6.45% (c), 4.18% (d), 3.99% (e), and 2.94% (f).

izes the free energies from the lattice model, the lattice model does not contain any kinetic considerations. Through incorporating the kinetics of the system into the analysis of the fraying dynamics we now have clearer evidence that the thermodynamics are the driving force behind the fast fraying dynamics and role of the kinetics is relatively minimal.

To further examine the role of the thermodynamics as the driving force for the dissociation we can directly examine the free energy surface from the lattice model[2] along with the dominant pathways determined by the TPT analysis which are shown in Figure 7.19. The free energy surface runs from red (largest free energy) to blue (lowest free energy). The white dots represent the pathway determined by the TPT analysis. Any configuration that lies below the dashed black line must include at least one bubble and is thus not included in the kinetic model. The top three pathways, particularly the top two, follow a clear trend where they keep the frayed ends on each side short which also tracks the pathway with the lowest free energy. In particular when the top pathways have one or two intact base pairs they enter a clear valley in the free energy surface where the length of the longest frayed end is 4-6 base pairs and the remaining pathways all have longer frayed ends at this point. This also provides another method for visualizing the dominant unfold-

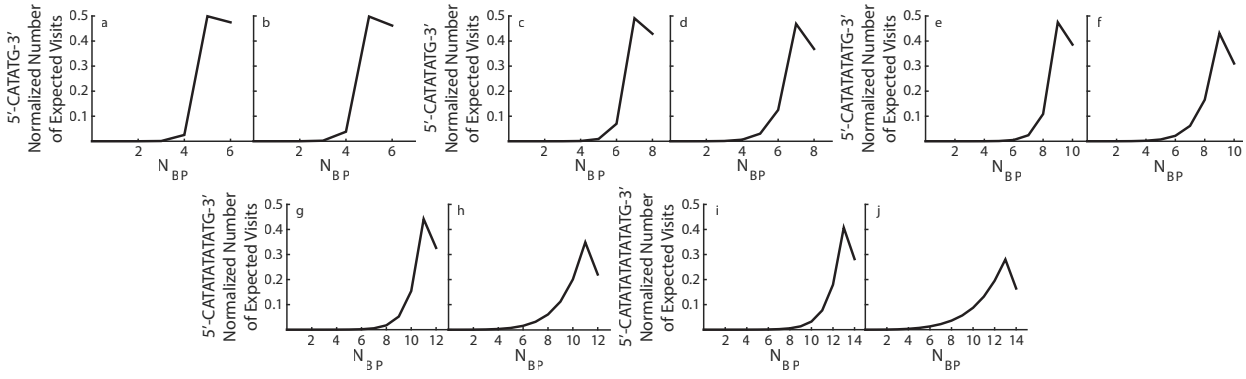

Figure 7.20: Expected number of visits to configurations with each N_{BP}, normalized to the total number of expected visits to all configurations during the trajectory, for a trajectory starting in the fully formed dimer state at the lowest and highest temperatures studied for each 5'-C(AT)$_n$G-3' sequence as follows: n = 2 at 306 K (a) and 317 K (b), n = 3 at 315 K (c) and 333 K (d), n = 4 at 319 K (e) and 334 K (f), n = 5 at 325 K (g) and 340 K (h), and n = 6 at 328 K (i) and 342 K (j).

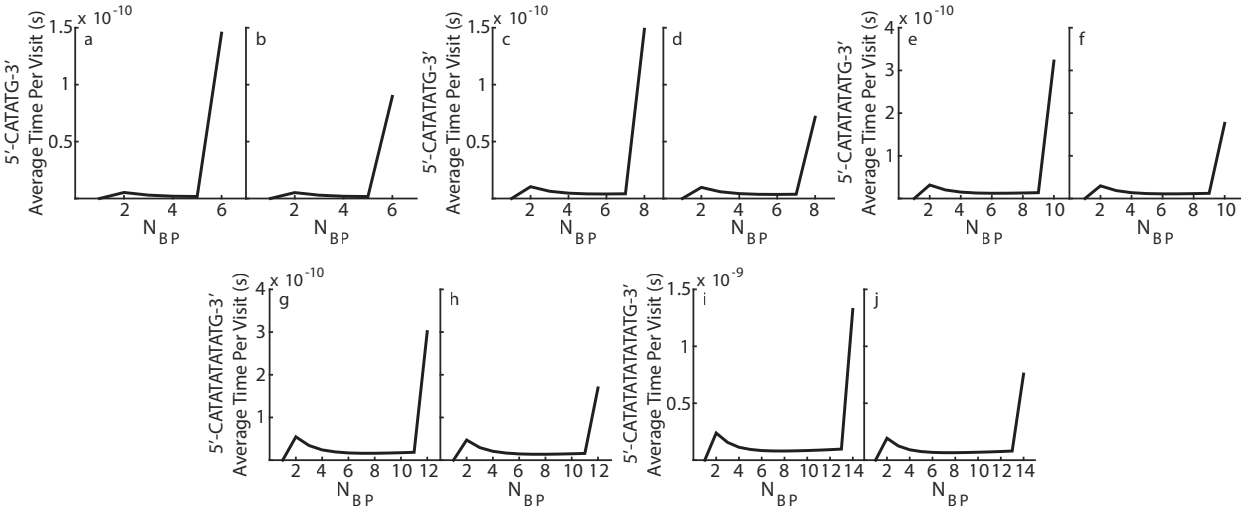

Figure 7.21: Average time per visit, in seconds, to configurations with each N_{BP} at the lowest and highest temperatures studied for each 5'-C(AT)$_n$G-3' sequence as follows: n = 2 at 306 K (a) and 317 K (b), n = 3 at 315 K (c) and 333 K (d), n = 4 at 319 K (e) and 334 K (f), n = 5 at 325 K (g) and 340 K (h), and n = 6 at 328 K (i) and 342 K (j).

ing motif of preferentially maintaining two frayed ends that are of roughly equal length. It can be clearly seen in Figure 7.19 that pathways four, five, and six have longer frayed ends along most, if not all, of the N_{BP} values for the dissociation and also make their final barrier crossing with a longer frayed end, higher up on the y-axis, that clearly has higher free energy values.

The same analysis used for GC-core can now be applied to CG-ends to obtain more insight into their fast response. Figure 7.20 shows the normalized number of visits for the highest and lowest temperatures for each length of CG-ends sequence while Figure 7.21 shows the average duration of each visit both as a function N_{BP}. The changes observed in Figures 7.20 and 7.21 as a function of length can be compared to those in Figures 7.17 and 7.18 as a function of sequence, while also comparing changes as a function of temperature to gain insight into the fast dynamics. The trends observed in Figure 7.21 as a function of length and temperature very closely match those seen in Figure 7.18 for sequence and temperature. In Figure 7.21 the magnitudes vary between plots but they all have the same general shape. The average visit length also slightly decreases with

temperature for all states except the fully intact duplex state which sees a sharp decrease. The trends in Figure 7.20 are a close match to the trends in the average time per visit to states with each N_{BP} in Figure 7.16, which is again comparable to what was observed for GC-core. As length increases there is a relative decrease in the number of visits to the fully intact dimer state and the state with a single broken base pair with a relative increase observed for states with more broken base pairs. Additionally, temperature has a significantly larger effect on longer sequences relative to shorter sequences which is comparable to the comparison between GC-core and the ten base pair CG-ends sequence. The strong correlation between the effect of increasing length and the effect of moving the G:C base pairs to the center suggests that the source of the increasing fast response observed with length is also fraying.

One interesting note between GC-core and the longest CG-ends sequence is that the experimental fast response for GC-core is of the form of a biexponential while CG-ends

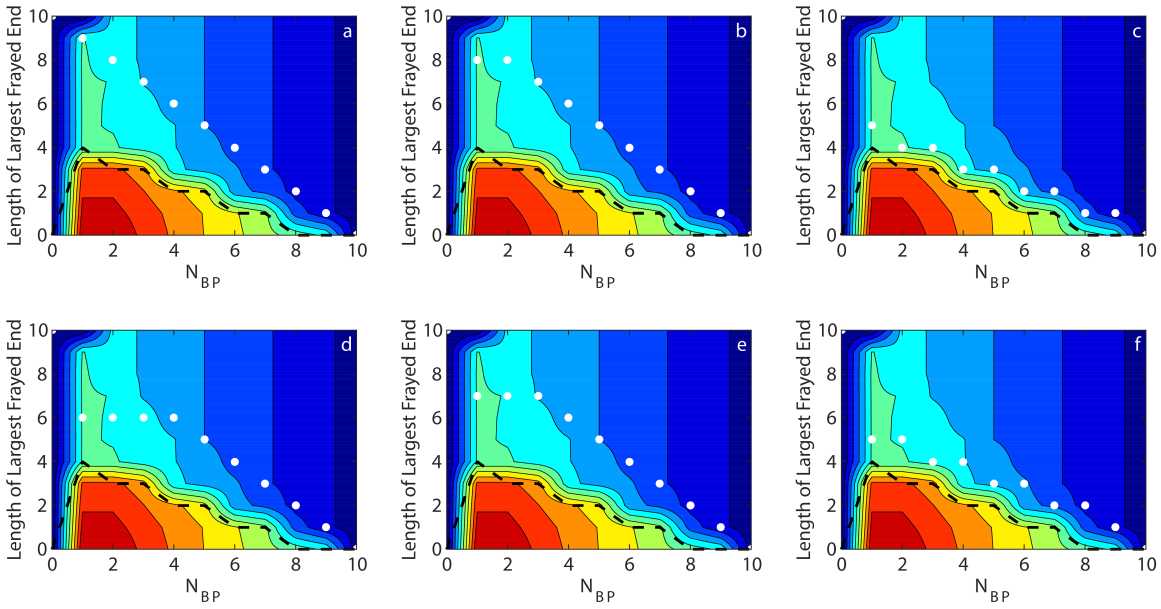

Figure 7.22: Lattice model free energy surface at 334 K for 5'-CATATATATG-3' where blue denotes the most favorable free energy and red the least. Configurations on or below the black dashed line must include at least one bubble and are therefore not allowed. The white dots represent the top six pathways predicted by TPT in descending order (a-f). The probability of a successful association event occurring along each pathway according to TPT is: 8.02% (a), 5.82% (b), 5.28% (c), 5.14% (d), 5.12% (e), and 5.12% (f).

is a stretched exponential. While no conclusive evidence is provided here it is interesting to note that at higher temperatures the shape of the distribution for GC-core in Figure 7.13 significantly changes while for the longest CG-ends sequence it stays relatively exponential in Figure 7.16. Figures 7.17 and 7.20 also mirror this observation. In Figures 7.13 and 7.17 we see that GC-core does stay relatively exponential for the first three or four temperatures before the distribution starts to change shape. It would be interesting to see if pushing to higher temperatures resulted in this change occurring for CG-ends or if the form of the distribution remains consistent. This possible connection between the functional form of the experimental kinetics and the percentage of time spent in states as a function of N_{BP} in the model is interesting. It would imply a connection to the thermodynamic free energy surface and result in new avenues for understanding the fast dynamics.

The lattice model free energy surfaces with the top six kinetic pathways overlaid for the ten base pair CG-ends sequence are shown in Figure 7.22. Just like Figure 7.19 for GC-core these show that the most dominant pathways are again closely related to the thermodynamics. However, in the case of CG-ends the preference for the pathways after the first one may not seem as inherently obvious since the free energy surface is so clearly aligned with the top pathway. In particular the third pathway isolated by the kinetic model seems somewhat counter-intuitive when considering the thermodynamic model. This helps to illustrate the point that while the thermodynamic model can provide insights into what the dominant association pathway may be, gaining additional insights requires a kinetic model.

While the consistencies between the effect of increasing length and changing the sequence suggest that fraying is a significant contributor to the observed fast response in CG-ends there are other factors that should be considered. Four other factors will be considered here as potential sources of the fast response either instead of, or in addition to, fraying. The first alternative factor is bubble formation which is a partial dissociation including only internal base pairs. While the final version of the model does not incorpor-

ate bubble states there are two pieces of evidence from models that we have in addition to experimental evidence that suggest fraying does not significantly contribute to the fast response. The first comes from the thermodynamic lattice model that shows that for sequences within this length regime bubble states have very high free energies, making them unlikely to be accessed. This is observed in both Figures 7.19 and 7.22 by looking at the area outlined by the black dashed line that represents configurations that include bubbles. This area of the free energy surface has a significantly larger free energy than any other area on the surface. Knowing the previously established connection between the thermodynamic free energy and the dominant pathways these plots help to show that it is highly unlikely for dissociation events to occur that involve passing through these high energy bubble states. Early iterations of the kinetic model, which also utilized the Gillespie algorithm but were parameterized differently, included bubble states for short sequences of six and eight base pairs. In these early trials the bubble states were not significantly accessed and there was no observable difference between the results of the kinetic model with and without bubble states. In the CG-ends experimental results the fast response is observed in both the A:T and G:C response. Bubble formation must keep G:C base pairs intact, so even if it occurs it cannot be the only fast dynamics occurring since these dynamics involve dissociating G:C base pairs. Thus, while bubble formation cannot entirely be ruled out it does not appear likely and also cannot be the sole cause of the fast response.

The second possible contributor is an increasingly heterogeneous initial dimer population. Increasing heterogeneity can lead to an increasingly stretched rate distribution since configurations with fewer intact base pairs could dissociate faster. This would increase with length since longer lengths have more stable intermediates causing a more diverse equilibrium population distribution. However, the model suggests this would not generate a stretched exponential rate distribution. Running the model starting in the fully formed dimer state and a large variety of intermediates results in indistinguishable mean first passage times. It is worth pointing out that the model itself struggles to capture the

effect of non-two-state kinetics, so it may be that the model is unable to capture the different timescales. This would likely be tied to the construction of the model as a Markov state Monte Carlo method. According to the forward committor values, configurations with more than three intact base pairs are more likely to go to the dimer state than the monomer state and with more intact base pairs the probability quickly becomes well over 90%. This means that trajectories that start in an intermediate configuration will likely go to the fully formed dimer. At that point, since the model is memoryless, this is the same as starting a trajectory in the fully formed dimer state, but at a time later than time zero. Thus, while the model predicts that a heterogeneous dimer configuration is not responsible for the fast response, it could be due to a limitation of the method used rather than a fully accurate representation of the physical system.

The third possible contributor is similar in that it affects the distribution of rates itself, but rather than being due to a distribution of initial states it is instead the result of a distribution in mechanistic pathways that share common initial and final states. This has been proposed as an explanation for stretched exponential kinetics observed in proteins.[36] The model does cast some doubt on this because the distribution of first passage times is an almost perfect exponential distribution. Since the model does take into account a number of different pathways across the many trajectories this would suggest that an increased distribution of pathways is not responsible for the fast response. However, this is another case where the inability to capture the effect of non-two-state kinetics means that this cannot fully rule out a distribution of pathways as the cause of the observed fast dynamics.

The final possible explanation discussed here is mismatched initial states. In some ways this is a more extreme case of the heterogeneous dimer distribution in that partially formed intermediates in the initial population distribution cause the fast response. These configurations are inherently less stable than the fully formed duplex and when samples are properly annealed the number of mismatched sequences should be very low. However, at longer lengths the repeating AT dinucleotide section in the center does allow for

numerous consecutive mismatched bases to form making them more stable than they are for shorter lengths. Additionally, it would not be unreasonable that over the course of a temperature-jump experiment repeatedly breaking apart and reforming the duplexes for multiple hours might increase the presence of mismatches. Mismatches are not allowed by the model so only the experimental results can provide insight into the potential role of mismatches. While we already discussed why starting from a configuration that is not a fully formed duplex is an unlikely explanation for the increasingly stretched kinetics a mismatched sequence would likely result in a more significant deviation in the kinetics. There are two potential reasons for this, the first is that mismatches by definition cannot have any intact G:C base pairs for the CG-ends sequences since the complimentary base pairs must shift out-of-register. This results in a less stable configuration than a partially formed in-register configuration that can contain a G:C base pair. Additionally, an in-register partially formed configuration can, and for some intermediates likely will, reform the fully formed duplex state. While a partially formed out-of-register structure could add additional base pairs in some situations, it cannot reach the fully formed duplex without either fully dissociating first, or undergoing a more complex mechanism. Thus it has a lower ceiling for stability and could be expected to dissociate faster. However, the fact that there is a fast response for the G:C and A:T base pairs means that out-of-register mismatches cannot be solely responsible for the fast response. This is because a shifted registry state, the most stable possible mismatch for these sequences, can not include intact G:C base pairs.

Ultimately, this results in a position where, even with the model, we are still unable to definitively provide a clear explanation for the cause of the fast response seen in CG-ends that grows in with length. However, we have gained additional insight beyond the experimental results. The kinetic model provides strong evidence that fraying dynamics are occurring and contribute to the CG-ends fast response while also suggesting that it may be the most significant contributor. While the other alternatives cannot be definit-

ively ruled out there is evidence that they are less significant or unlikely to contribute at all. This is especially true for the cases of a heterogeneous dimer population and a distribution of pathways where the model provides some evidence, though inconclusive, that these factors do not contribute to the fast response. Additionally, mismatches, heterogeneous initial populations, and bubble formation would not result in a fast G:C base pair response. Furthermore, in the given length regime bubble states are not particularly accessible as suggested by both the thermodynamic lattice model and early iterations of the kinetic model. This strongly suggests that the driving factor behind the increasing fast dynamics is fraying with a possibility that mismatched sequences might provide an additional small contribution.

7.5 Future Directions

The relative simplicity of the model makes it easily accessible to future researchers while also providing significant versatility for incorporating further improvements. This can come in the form of both the incorporation new experimental data in addition to further improvements to the construction of the model. One aspect that will be critical for future development is expanding the library of experimental data for comparison and fitting. Any work to incorporate additional parameterization will benefit greatly from additional data to avoid concerns of overfitting while also improving the models ability to explore and decouple parameters such as length, sequence, and temperature.

There are a few particularly clear avenues for potential alterations to the construction of the model. The first is a direct consideration of temperature with respect to k_{f}. While preliminary explorations have suggested that the temperature dependence is small and trends so far have been inconsistent a more explicit consideration of the relationship may provide new insights. This would be particularly interesting with regards to the CG-ends sequence where a trend in k_{f} emerged and its dependence on length and temperature

cannot be fully decoupled. Acquiring a larger range of temperatures in the experimental data and altering the way in which k_f is defined in the model could shed new light on this. Another interesting avenue for exploration would be decoupling the β and k_f parameters. This might not only clarify the interpretation of both parameters but also provide more insights into the diffusion to capture component of the association mechanism that is known to be a highly influential and complex portion of the mechanism.

A significant improvement in both our understanding of the model and its parameters could lead to deciphering trends in the parameters as a function of basic variables such as length, sequence, temperature, and additional effects such as salt concentrations. If these trends can be isolated the model could potentially shift from fitting experimental data to become a predictive model. This would have numerous significant benefits beyond the ability to provide insights into sequences without experimental data. Currently, a significant component of the computational expense lies in the fitting and parameterization of the model. Eliminating the need for this would allow new avenues to become accessible. This would open up the possibility of adding in additional allowed configurations such as bubble states or even out-of-register base pairs. This would provide the ability to extend the model out to longer sequences that contain richer and more complex dynamics where these alternative configurations become more relevant. Additionally, more complex mechanisms such as pseudoknot association, which has been observed by coarse-grained MD simulations, would become allowed by the model. Not only would this provide the ability to expand the models out to longer lengths but also could shed some light on the dynamics responsible for the fast response observed for the CG-ends sequences. The ability of the model to replicate such rich dynamics would nicely compliment ongoing experimental technique development into new ways to probe the kinetics and dynamics of DNA which together can continue to provide novel insights into the complex questions surrounding the association and dissociation of DNA duplexes.

7.6 Conclusions

The two parameter Markov state Monte Carlo model presented here is able to reasonably reproduce the experimental results while also producing findings that are in agreement with existing coarse-grained MD simulations that are significantly more complex and computationally expensive. The model shows that the initiation position for a successful association barrier crossing, which corresponds to the location of the critical nucleus and transition state, is driven by two factors. An entropic contribution that preferentially drives initiating at the center of the sequence and an enthalpic contribution that preferentially drives initiating near G:C base pairs if present in the sequence. The effects of these energetic forces, particularly the enthalpic benefit, become far less significant after the formation of the first few base pairs which is in agreement with the canonical nucleation-zipper mechanism and the corresponding critical nucleus. Based on insights gained by looking at the dominant association mechanisms and the relative stability of intermediate configurations the critical nucleus is predicted to be on the order of two to three base pairs, in excellent agreement with predictions from the literature.[4,7,8] Additionally, the model provides evidence that the critical nucleus is expected to increase in size with increasing temperature and sequence length. This is consistent with the insights into the energetic driving forces gained from the model, our own experimental findings on length, and results in the literature on the effects of temperature.[4] With regards to fast dynamics prior to the full dissociation of the duplex, the model recreates the fast fraying dynamics experimentally observed for GC-core and provides further evidence that these dynamics are primarily driven by thermodynamic factors and the reshaping of the free energy surface rather than kinetic factors. Additionally, the model suggests that the origins of the increasing fast response observed with increasing length in the CG-ends series is similar to those observed in the GC-core sequence suggesting that fraying plays a significant role in these dynamics.

7.7 Acknowledgements

I would like to thank Greg Kimmel, Paul Sanstead, and Brennan Ashwood for their careful reading and thoughtful comments on this chapter.

7.8 References

1. Sanstead, P. J. C. Investigation of DNA Dehybridization through Steady-State and Transient Temperature-Jump Nonlinear Infrared Spectroscopy. Ph.D. thesis, The University of Chicago, 2018.

2. Sanstead, P. J.; Tokmakoff, A. A Lattice Model for the Interpretation of Oligonucleotide Hybridization Experiments. *J. Chem. Phys.* **2019**, *150*, 185104.

3. Šulc, P.; Romano, F.; Ouldridge, T. E.; Rovigatti, L.; Doye, J. P. K.; Louis, A. A. Sequence-Dependent Thermodynamics of a Coarse-Grained DNA Model. *J. Chem. Phys.* **2012**, *137*, 135101.

4. Ouldridge, T. E.; Šulc, P.; Romano, F.; Doye, J. P. K.; Louis, A. A. DNA Hybridization Kinetics: Zippering, Internal Displacement and Sequence Dependence. *Nucleic Acids Res.* **2013**, *41*, 8886–8895.

5. Sambriski, E. J.; Schwartz, D. C.; de Pablo, J. J. A Mesoscale Model of DNA and Its Renaturation. *Biophys. J.* **2009**, *96*, 1675–1690.

6. Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; de Pablo, J. J. An Experimentally-Informed Coarse-Grained 3-Site-Per-Nucleotide Model of DNA: Structure, Thermodynamics, and Dynamics of Hybridization. *J. Chem. Phys.* **2013**, *139*, 144903.

7. Sambriski, E. J.; Schwartz, D. C.; de Pablo, J. J. Uncovering Pathways in DNA Oligonucleotide Hybridization via Transition State Analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 18125–18130.

8. Hoefert, M. J.; Sambriski, E. J.; de Pablo, J. J. Molecular Pathways in DNA-DNA Hybridization of Surface-Bound Oligonucleotides. *Soft Matter* **2011**, *7*, 560–566.

9. Hinckley, D. M.; Lequieu, J. P.; de Pablo, J. J. Coarse-Grained Modeling of DNA Oligomer Hybridization: Length, Sequence, and Salt Effects. *J. Chem. Phys.* **2014**, *141*, 035102.

10. SantaLucia, J. A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 1460–1465.

11. Craig, M. E.; Crothers, D. M.; Doty, P. Relaxation Kinetics of Dimer Formation by Self Complementary Oligonucleotides. *J. Mol. Biol.* **1971**, *62*, 383–401.

12. Pörschke, D.; Eigen, M. Co-operative Non-enzymatic Base Recognition III. Kinetics of the Helix-Coil Transition of the Oligoribouridylic • Oligoriboadenylic Acid System and of Oligoriboadenylic Acid Alone at Acidic pH. *J. Mol. Biol.* **1971**, *62*, 361–381.

13. Marimuthu, K.; Chakrabarti, R. Sequence-Dependent Theory of Oligonucleotide Hybridization Kinetics. *J. Chem. Phys.* **2014**, *140*, 175104.

14. Pörschke, D. A Direct Measurement of the Unzippering Rate of a Nucleic Acid Double Helix. *Biophys. Chem.* **1974**, *2*, 97–101.

15. Calef, D. F.; Deutch, J. Diffusion-Controlled Reactions. *Ann. Rev. Phys. Chem.* **1983**, *34*, 493–524.

16. Pörschke, D. Model Calculations on the Kinetics of Oligonucleotide Double Helix Coil Transitions. Evidence for a Fast Chain Sliding Reaction. *Biophys. Chem.* **1974**, *2*, 83–96.

17. Applequist, J.; Damle, V. Thermodynamics of the Helix-Coil Equilibrium in Oligoadenylic Acid from Hypochromicity Studies. *J. Am. Chem. Soc.* **1965**, *87*, 1450–1458.

18. Eigen, M.; Pörschke, D. Co-operative Non-enzymic Base Recognition: I. Thermodynamics of the Helix-Coil Transition of Oligoriboadenylic Acids at Acidic pH. *J. Mol. Biol.* **1970**, *53*, 123–141.

19. Wetmur, J. G.; Davidson, N. Kinetics of Renaturation of DNA. *J. Mol. Biol.* **1968**, *31*, 349–370.

20. Banik, S. K.; Ambjörnsson, T.; Metzler, R. Stochastic Approach to DNA Breathing Dynamics. *Europhys. Lett.* **2005**, *71*, 852–858.

21. Ambjörnsson, T.; Banik, S. K.; Krichevsky, O.; Metzler, R. Breathing Dynamics in Heteropolymer DNA. *Biophys. J.* **2007**, *92*, 2674–2684.

22. Chen, X.; Zhou, Y.; Qu, P.; Zhao, X. S. Base-by-Base Dynamics in DNA Hybridization Probed by Fluorescence Correlation Spectroscopy. *J. Am. Chem. Soc.* **2008**, *130*, 16947–16952.

23. Zimm, B. H. Theory of “Melting” of the Helical Form in Double Chains of the DNA Type. *J. Chem. Phys.* **1960**, *33*, 1349–1356.

24. Schaeffer, J. M.; Thachuk, C.; Winfree, E. Stochastic Simulation of the Kinetics of Multiple Interacting Nucleic Acid Strands. In *DNA Computing and Molecular Programming*, Boston and Cambridge, MA, August 17-21, 2015; Phillips, A., Yin, P., Eds.; Springer: Cham, 2015; pp 194–211.

25. Dunn, K. E.; Dannenberg, F.; Ouldridge, T. E.; Kwiatkowska, M.; Turberfield, A. J.; Bath, J. Guiding the Folding Pathway of DNA Origami. *Nature* **2015**, *525*, 82–86.

26. Dannenberg, F.; Dunn, K. E.; Bath, J.; Kwiatkowska, M.; Turberfield, A. J.; Ouldridge, T. E. Modelling DNA Origami Self-Assembly at the Domain Level. *J. Chem. Phys.* **2015**, *143*, 165102.

27. Pörschke, D.; Uhlenbeck, O. C.; Martin, F. H. Thermodynamics and Kinetics of the Helix-Coil Transition of Oligomers Containing GC Base Pairs. *Biopolymers* **1973**, *12*, 1313–1335.

28. Menssen, R. J.; Tokmakoff, A. Length-Dependent Melting Kinetics of Short DNA Oligonucleotides Using Temperature-Jump IR Spectroscopy. *J. Phys. Chem. B* **2019**, *123*, 756–767.

29. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved Through Infrared Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138*, 11792–11801.

30. Sanstead, P. J.; Tokmakoff, A. Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **2018**, *122*, 3088–3100.

31. Pörschke, D. Cooperative Nonenzymic Base Recognition II. Thermodynamics of the Helix-Coil Transition of Oligoadenylic+Oligouridylic Acids. *Biopolymers* **1971**, *10*, 1989–2013.

32. Protozanova, E.; Yakovchuk, P.; Frank-Kamenetskii, M. D. Stacked–Unstacked Equilibrium at the Nick Site of DNA. *J. Mol. Biol.* **2004**, *342*, 775–785.

33. Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-Stacking and Base-Pairing Contributions into Thermal Stability of the DNA Double Helix. *Nucleic Acids Res.* **2006**, *34*, 564–574.

34. Gillespie, D. T. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comput. Phys.* **1976**, *22*, 403–434.

35. Chen, C.; Wang, W.; Wang, Z.; Wei, F.; Zhao, X. S. Influence of Secondary Structure on Kinetics and Reaction Mechanism of DNA Hybridization. *Nucleic Acids Res.* **2007**, *35*, 2875–2884.

36. Sabelko, J.; Ervin, J.; Gruebele, M. Observation of Strange Kinetics in Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 6031–6036.

Appendix 7A: Kinetic Model Example for a Three Base Pair Sequence

To better visualize the construction of the kinetic model it is useful to demonstrate the possible states, and allowed moves between them, utilizing a basic example. Considering a simple model sequence of three base pairs there are initially eight possible configurations the system can adopt. Utilizing a binary representation for the configurations where a one represents a site with an intact base pair and a zero represents a site with a broken base pair the possible configurations are given in Table 7.2. With the assumption that bubble states, such as the state denoted (1,0,1), are not sufficiently populated to have a significant impact on the simulations we are left with seven remaining configurations that are indexed. From these seven configurations, utilizing the allowed moves described in section 7.2.1, there are 18 allowed moves between configurations for this three base pair sequence. The resulting reaction scheme is shown in Figure 7.23 and shows the 18 allowed moves between the configurations. Now the rates for each allowed moves must be calculated. Calculating the rates as described in section 7.2.4 for each allowed move results in the transition rate matrix, **L** for moving from state i (rows) to state j (columns) where zeros denote a move that is not allowed by the reaction scheme shown in Figure 7.23. For the sake of making the table easier to read the diagonal elements of the matrix, which are equal to the negative of the sum of the off diagonal elements for that row,

Table 7.2: All possible configurations for a sequence with three base pairs.

index	configuration
1	(0,0,0)
2	(1,0,0)
3	(0,1,0)
4	(0,0,1)
5	(1,1,0)
–	(1,0,1)
6	(0,1,1)
7	(1,1,1)

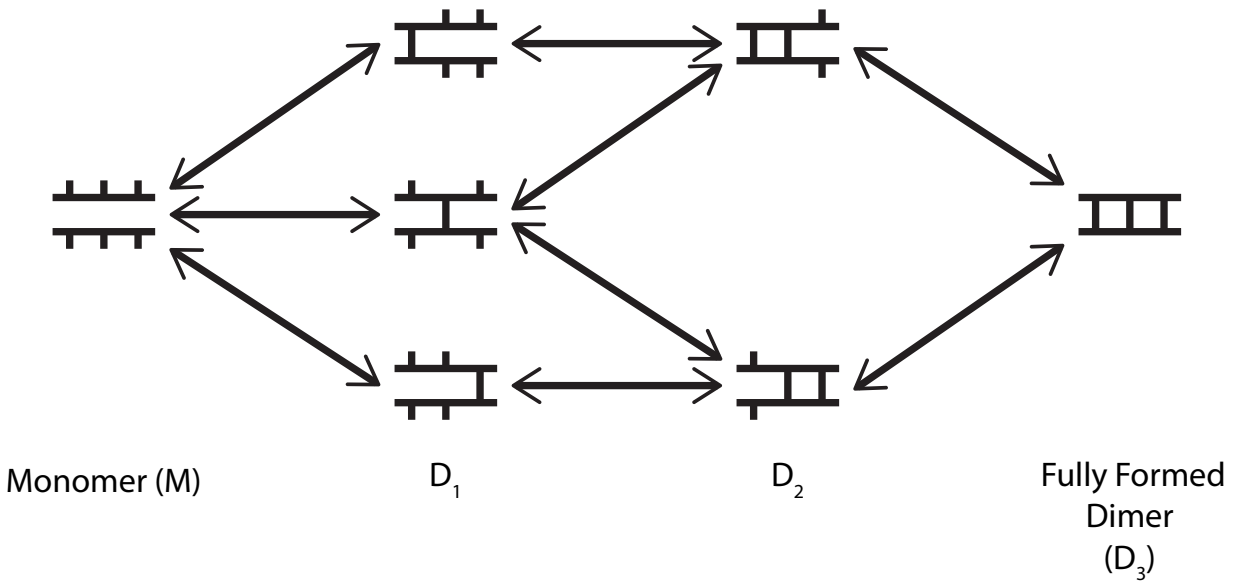

Figure 7.23: Diagram of moves allowed by the kinetic model for a three base pair DNA sequence.

have been replaced with a dashed line. Here G_i refers to the free energy of the configuration which is calculated from the lattice model, R is the ideal gas constant and T is the temperature at which the system is evolving.

Table 7.3: Transition rate matrix for a three base pair sequence.

i \ j	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(0,1,1)	(1,1,1)
(0,0,0)	–	βk_1	βk_1	βk_1	0	0	0
(1,0,0)	$\frac{\beta k_1}{e^{-\frac{G_i-G_j}{RT}}}$	–	0	0	$\sigma_1 k_1$	0	0
(0,1,0)	$\frac{\beta k_1}{e^{-\frac{G_i-G_j}{RT}}}$	0	–	0	$\sigma_1 k_1$	$\sigma_1 k_1$	0
(0,0,1)	$\frac{\beta k_1}{e^{-\frac{G_i-G_j}{RT}}}$	0	0	–	0	$\sigma_1 k_1$	0
(1,1,0)	0	$\frac{\sigma_1 k_1}{e^{-\frac{G_i-G_j}{RT}}}$	$\frac{\sigma_1 k_1}{e^{-\frac{G_i-G_j}{RT}}}$	0	–	0	$\sigma_2 k_1$
(0,1,1)	0	0	$\frac{\sigma_1 k_1}{e^{-\frac{G_i-G_j}{RT}}}$	$\frac{\sigma_1 k_1}{e^{-\frac{G_i-G_j}{RT}}}$	0	–	$\sigma_2 k_1$
(1,1,1)	0	0	0	0	$\frac{\sigma_2 k_1}{e^{-\frac{G_i-G_j}{RT}}}$	$\frac{\sigma_2 k_1}{e^{-\frac{G_i-G_j}{RT}}}$	–

Appendix 7B: Comparing the Percentage of All Association Barrier Crossing Events that Initiate at Each Position Determined by Transition Pathway Theory and the Stochastic Trajectories

This appendix contains plots showing the percentage of all association barrier crossing events that initiate at each position for the CG-ends and GC-core sequences. Each figure contains the values determined by the TPT analysis, also shown in Figure 7.9, and directly from the trajectories. These plots show that the results from the two methods are in good agreement with one another.

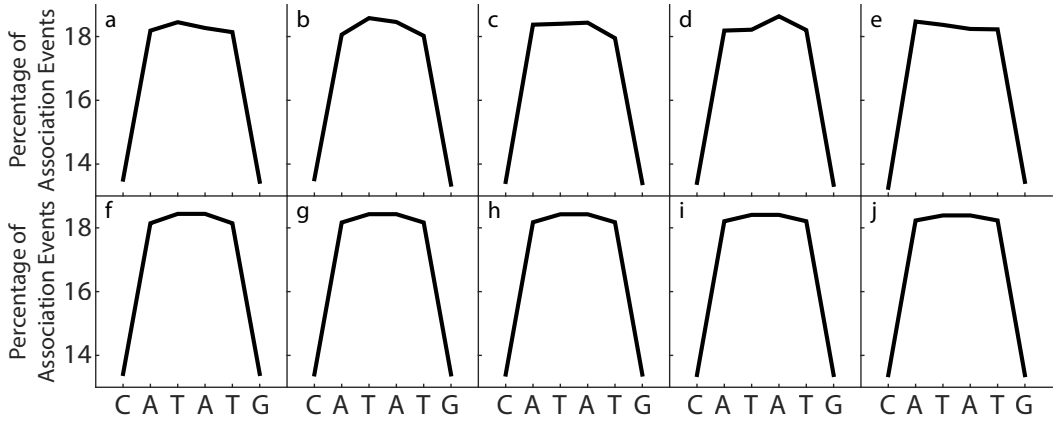

Figure 7.24: Percentage of all association barrier crossing events that initiate at each position for 5'-CATATG-3' from the stochastic trajectories (a-e) and the transition path theory analysis (f-j). The temperatures are: 306 K (a) and (f), 309 K (b) and (g), 310 K (c) and (h), 314 K (d) and (i), and 317 K (e) and (j).

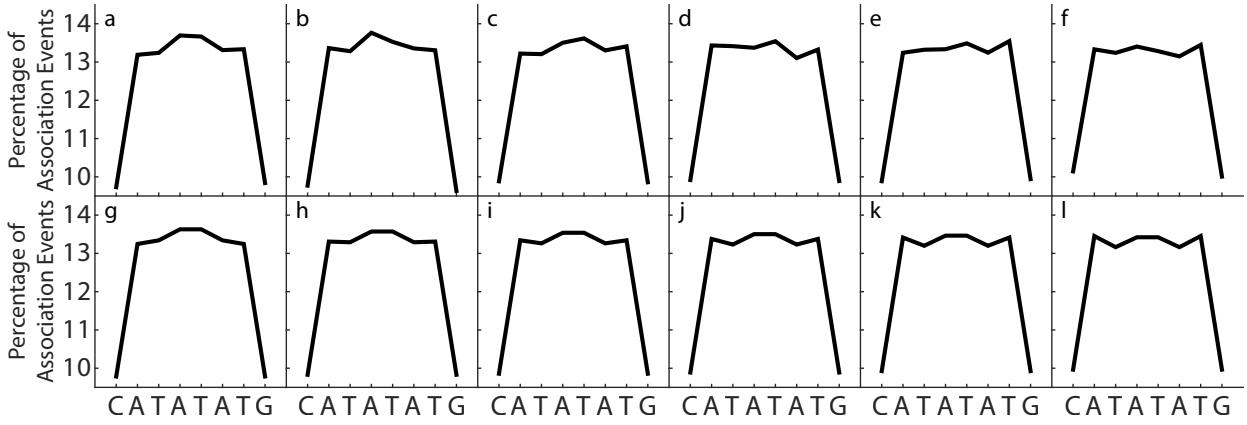

Figure 7.25: Percentage of all association barrier crossing events that initiate at each position for 5'-CATATATG-3' from the stochastic trajectories (a-f) and the transition path theory analysis (g-l). The temperatures are: 315 K (a) and (g), 321 K (b) and (h), 324 K (c) and (i), 327 K (d) and (j), 330 K (e) and (k), and 333 K (f) and (l).

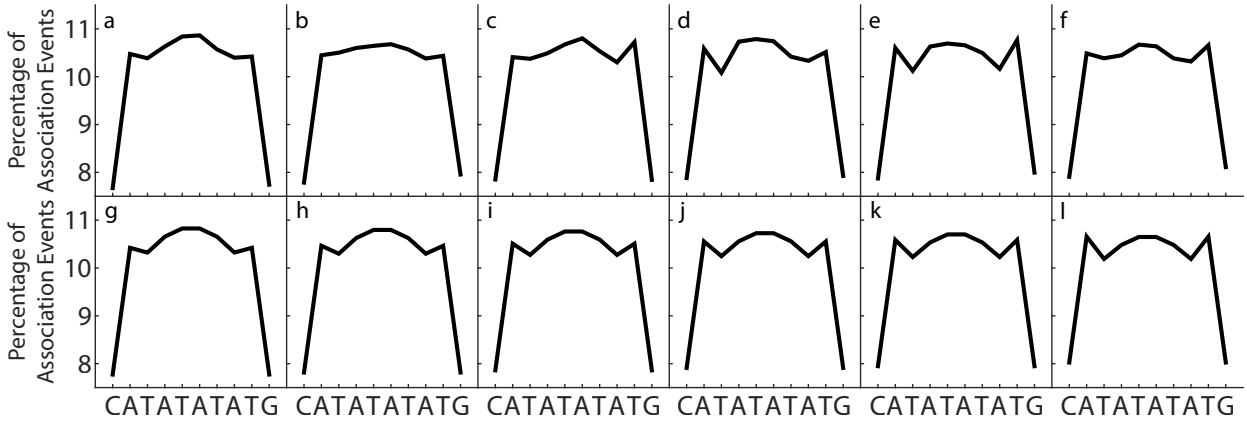

Figure 7.26: Percentage of all association barrier crossing events that initiate at each position for 5'-CATATATATG-3' from the stochastic trajectories (a-f) and the transition path theory analysis (g-l). The temperatures are: 319 K (a) and (g), 322 K (b) and (h), 325 K (c) and (i), 328 K (d) and (j), 330 K (e) and (k), and 334 K (f) and (l).

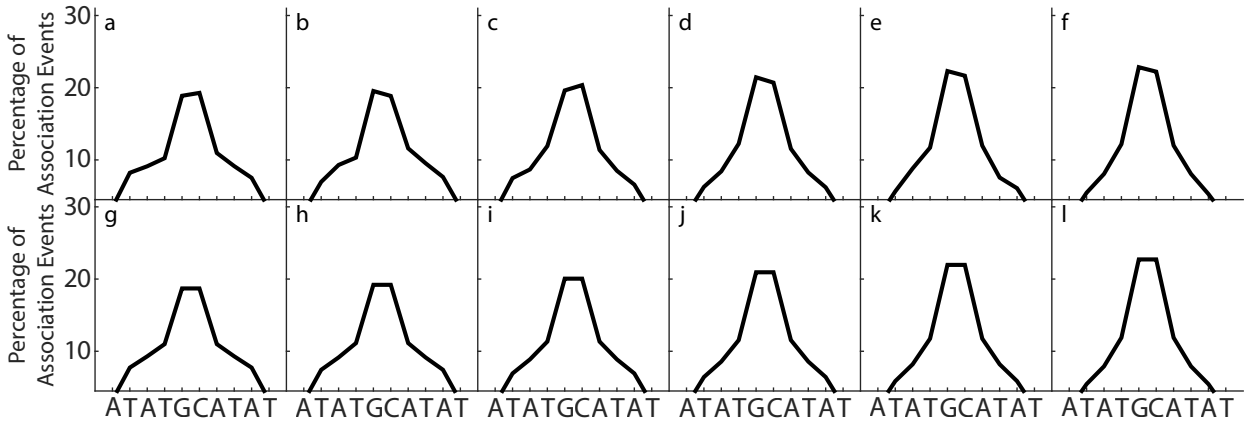

Figure 7.27: Percentage of all association barrier crossing events that initiate at each position for 5'-ATATGCATAT-3' from the stochastic trajectories (a-f) and the transition path theory analysis (g-l). The temperatures are: 315 K (a) and (g), 320 K (b) and (h), 327 K (c) and (i), 333 K (d) and (j), 339 K (e) and (k), and 343 K (f) and (l).

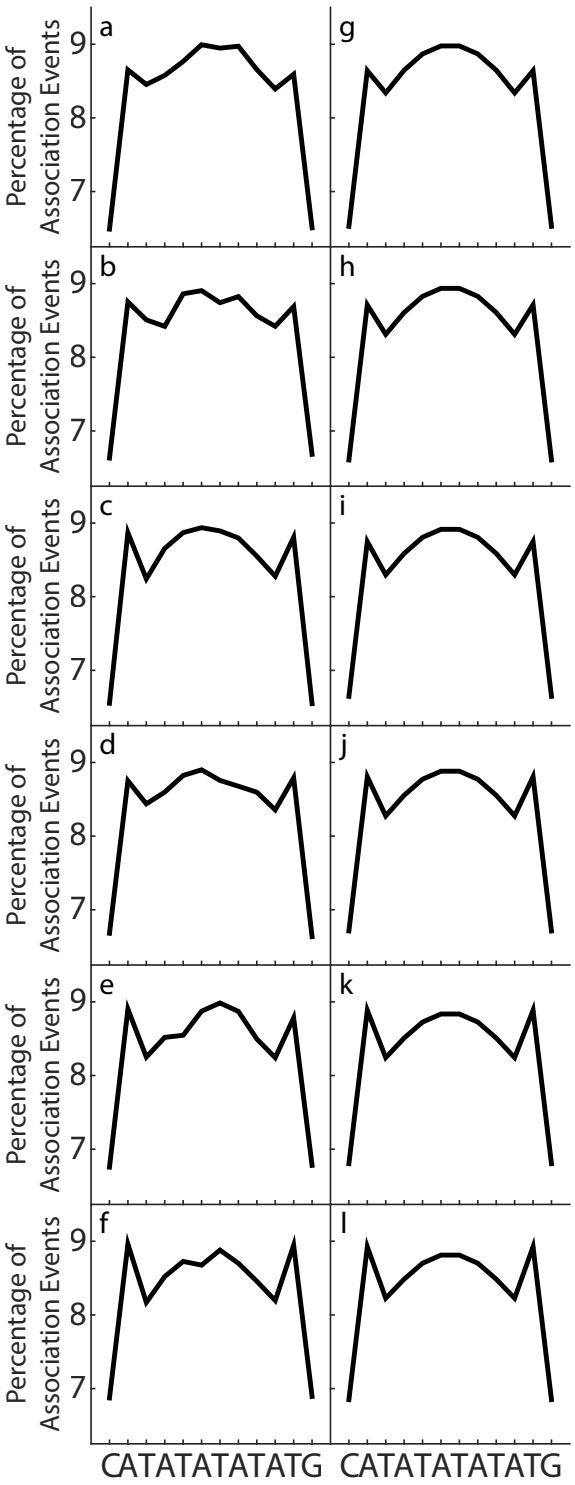

Figure 7.28: Percentage of all association barrier crossing events that initiate at each position for 5'-CATATATATATG-3' from the stochastic trajectories (a-f) and the transition path theory analysis (g-l). The temperatures are: 325 K (a) and (g), 329 K (b) and (h), 331 K (c) and (i), 334 K (d) and (j), 338 K (e) and (k), and 340 K (f) and (l).

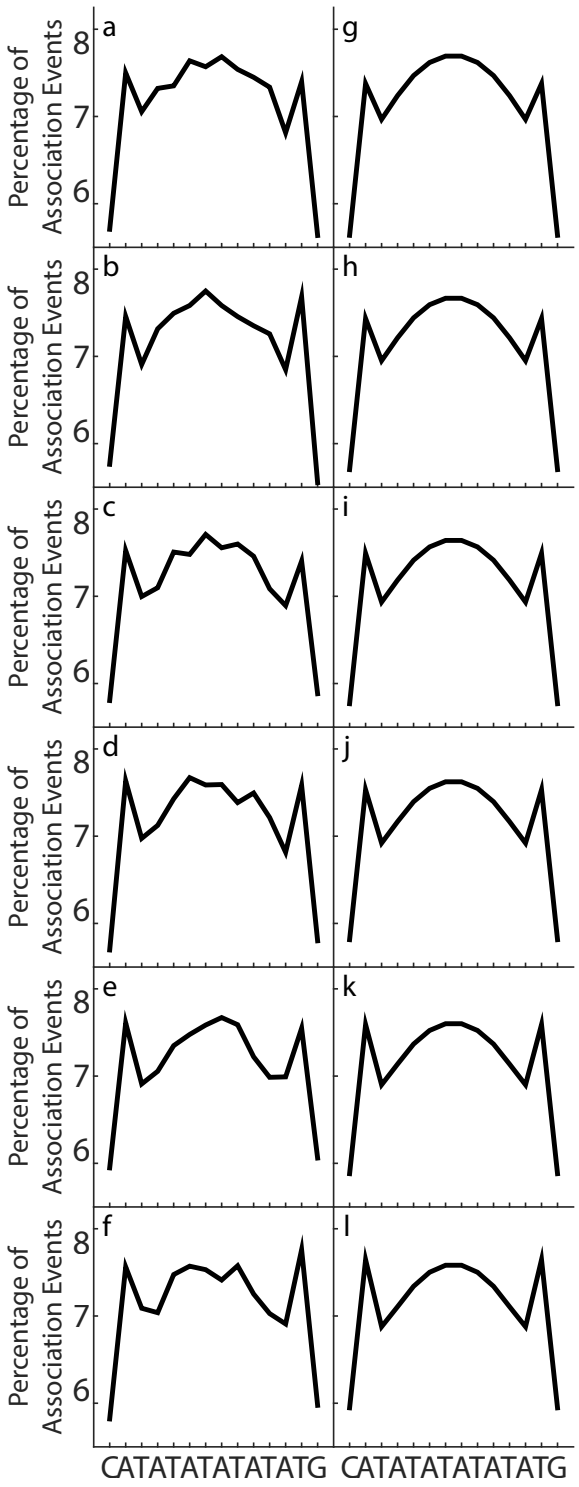

Figure 7.29: Percentage of all association barrier crossing events that initiate at each position for 5'-CATATATATATATG-3' from the stochastic trajectories (a-f) and the transition path theory analysis (g-l). The temperatures are: 328 K (a) and (g), 331 K (b) and (h), 334 K (c) and (i), 336 K (d) and (j), 339 K (e) and (k), and 342 K (f) and (l).