THE UNIVERSITY OF CHICAGO

INVESTIGATION OF DNA DEHYBRIDIZATION THROUGH STEADY-STATE AND TRANSIENT TEMPERATURE-JUMP NONLINEAR INFRARED SPECTROSCOPY

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

PAUL JONATHAN CREGAN SANSTEAD

CHICAGO, ILLINOIS

DECEMBER 2018

Contents

List of Figures

List of Tables

Acknowledgements

The work presented in this thesis would not have been possible without the assistance and support of many people. My advisor, Andrei Tokmakoff, was a calm and consistent guide throughout. He welcomed me into his lab at a time when the group was still in the process of moving to Chicago and the physical lab spaces were not yet finished. I am thankful to have had the opportunity to be a part of the building process and to participate firsthand in the re-establishment of the group in its new home. My view of research has been shaped by Andrei's curiosity driven approach to science as well as his commitment to high standards. I appreciate the freedom he allows as an advisor for students to pursue their scientific interests while at the same time remaining approachable and ever ready to provide guidance.

I am also indebted to the students and postdocs of the Tokmakoff group, both past and present. When I first joined, Ann Fitzpatrick was the only group member, besides Andrei, who was in Chicago. She introduced me to life in a laser lab and was a patient and enthusiastic mentor. Sam Peng founded the DNA project in the Tokmakoff group and I learned an incredible amount from him in the short time that we overlapped about the IR spectroscopy of nucleic acids as well as how to design and approach a research project. Luigi De Marco and Paul Stevenson were excellent colleagues and friends. I have learned much from them both and will always value our time spent together, both inside and outside of the GCIS basement. More recently, I have had the good fortune to work with Brennan Ashwood. His enthusiasm for research and work ethic are an inspiration. I look forward to the bright future of the DNA project and the "Big Lab" in his capable hands. Thanks is also due to the rest of the excellent graduate students and postdocs I have had the

privilege of working with over the past several years, who I can happily say are too numerous to name here. My graduate experience owes much to you all.

I would like to thank my undergraduate advisors, Sunghee Lee and Robert E. Novak, for sparking my interest in science and for providing me with my first opportunities to become involved in research. Their guidance was instrumental in my decision to pursue research in physical chemistry and their continued advocacy has meant much to me.

This thesis would not have been possible without the support of family. My mom, Jeanne Cregan Sanstead, has always been the strongest advocate of my education. The decision to pursue a doctorate degree was in no small part aided by her encouragement and loving support. My siblings, Caitlin, Christopher, and Erinn, have always been there, whether to help forget a rough day or just for a good laugh. My Uncle Paul and his family as well as my grandmother, Mary Cregan, have gone above and beyond over the years and I would not be where I am today without them. Jean Smidt has been an important person in my life since childhood and remains a close family friend. I thank her for her continued support. Finally, I would like to thank Audrey Gallagher for her unwavering encouragement and for sharing so many good times with me over the past five years.

Abstract

Many of the most important functions performed by nucleic acids are highly dynamic, whether in natural biological roles or in the field of DNA based nanotechnology. Despite a secure understanding of the thermodynamics of hybridization, the kinetics and particularly the dynamics remain less well understood. The fundamental structural transition that underlies much of nucleic acid folding is the formation of base pairs mediated by hydrogen bonding between complementary nucleobases and by stacking interactions with neighboring bases along the strand. To advance our understanding, an experimental approach that possesses both high time resolution and structural sensitivity towards these fundamental interactions is required. The work in this thesis develops a strategy for addressing DNA structural dynamics and hybridization kinetics through steady-state and transient temperature jump (T-jump) nonlinear infrared (IR) spectroscopy since the molecular vibrations probed are sensitive to the hydrogen bonding and base stacking interactions that mediate nucleic acid folding. In particular, two-dimensional infrared (2D IR) spectroscopy offers sub-picosecond time resolution and enhanced structural sensitivity through cross-peak information that reveals the coupling between nucleobase vibrations.

By studying a model set of DNA oligonucleotides in which the placement of guanine-cytosine (GC) base pairs is varied in an otherwise adenine-thymine (AT) sequence, an assembly of IR experimental and analysis methods reveals sequence-dependent variation in the ensemble of hybridized duplex structures. A simple statistical lattice model is developed that provides an intuitive interpretation of the experimental results. Transient T-jump experiments that track the dehybridization of the DNA double helix in real-time between nanoseconds to milliseconds resolve essentially barrierless unzipping dynamics as the terminal base pairs fray as well as activated barrier crossing between the duplex and single strand states.

Once validated on studies of model canonical oligonucleotides, the approach developed in the first half of the thesis is applied to investigate naturally occurring non-canonical nucleobases implicated in epigenetic regulation of the mammalian genome. Specifically, modified deoxycytidines that result from methylation of the 5-position of cytosine (mC) followed by successive oxidation of the methyl group to 5-hydroxymethyl- (hmC), 5-formyl- (fC), and 5-carboxyl- (caC) cytosine are involved in the active DNA demethylation cycle, which is central to gene regulation and cellular development. The influence of each of these modified nucleobases on the fundamental biophysical properties of DNA as well as the potential biological implications of such effects remains a topic of ongoing debate. The latter half of the thesis seeks to address some of these unresolved questions. 2D IR measurements reveal that the canonical keto-amino tautomeric form predominates for fC and caC under physiological conditions, ruling out the possibility that the formyl and carboxyl groups shift the tautomeric equilibrium and thereby disrupt base pairing. Proposed weakened base pairing in oligonucleotides containing fC and caC is supported by observations of increased acidity at the cytosine N3 hydrogen bond acceptor site as well as altered stability in fC and caC containing duplexes. Finally, the impact of each of the cytosine derivatives on the kinetic barrier to opening modified base pair domains is characterized by T-jump measurements, revealing a significant reduction in the dissociation barrier for base pairs involving fC and hmC while both mC and caC show a minor reduction in barrier height relative to canonical C. Possible biological implications of these trends are discussed.

Chapter 1

Introduction

1.1 Nucleic Acid Dynamics

Deoxyribonucleic acids (DNA) and ribonucleic acids (RNA) constitute a class of biomolecules, the nucleic acids, which underpin the fundamental molecular foundation of all known living things. Most famously, DNA stores the blueprint that directs both how a living organism maintains its special status far from equilibrium with its surroundings as well as how new life is propagated and assembled from a host of other chemical building blocks. Few would argue that isolated proteins are themselves alive, but add nucleic acids to the mix and a debate emerges as to whether or not a virus, consisting of little more than some genetic material inside a protective coat, qualifies as a living thing. RNA in particular participates in a wide diversity of biological roles, ranging from gene regulation and protein expression to catalyzing biochemical reactions. Because of its unique ability to perform so many of the diverse molecular tasks necessary for life, it has been hypothesized that RNA-based organisms arose before the evolution of proteins and DNA and were likely the prototype of all life on Earth.[1,2] Clearly to understand nucleic acids on a fundamental level is an opportunity to access central insight about life itself. Moreover, the impact of a detailed understanding of nucleic acids extends beyond the realm of biology. Due to the high fidelity and predictability of Watson-Crick base pairing, DNA nanotechnology has developed considerably over the last three decades, including the ability to construct a diversity of

complex structures on the nanoscale and to design dynamic molecular machines such as motors, walkers, and DNA based computers.[3-7]

In light of such broad functional versatility the molecular subcomponents that make up nucleic acids are deceptively simple, especially when compared to the proteins, lipids, and carbohydrates that constitute the other major classes of biological macromolecules. A sugar-phosphate backbone provides a scaffold for a linear sequence of nucleobases drawn from only four options. In the case of DNA these are adenine (A), thymine (T), guanine (G), and cytosine (C). Through base stacking interactions and hydrogen bonding between complementary bases, A to T and G to C, two DNA strands can hybridize, forming a characteristic double helical structure first reported in 1953.[8] This beautiful but static structure is only the beginning of the story. Many of the most essential functions involving DNA are highly dynamic, requiring shifts in the local solvation environment, the opening or closing of base pair domains, or larger scale rearrangements such as bending or supercoiling.[9-13]

The two decades following the initial discovery of the structure and information storage capacity of DNA witnessed a flood of interest in the kinetics and dynamics of nucleic acid hybridization. These pioneering researchers used temperature jump (T-jump) ultraviolet (UV) absorbance spectroscopy,[14-17] rapid mixing,[14,18,19] and imino-proton exchange nuclear magnetic resonance (NMR)[20,21] among other biophysical methods to establish the foundation for much of the present understanding of the mechanism of DNA base pairing and hybridization. In summation of early key findings, the association process was discovered to be a concentration dependent second order process that exhibits non-Arrhenius behavior while the dissociation process is well described by Arrhenius kinetics.

A reaction mechanism consistent with these experimental observations was proposed in which two strands must first diffuse into proximity in a second order process and establish a metastable encounter complex characterized by a few initial base pair contacts. These initial contacts are fleeting and stochastic as the two strands make and break base pairs while diffusing in and out of the proper orientation for local hybridization. A critical nucleus of three to four consecutive base pairs must form before full hybridization can proceed and the formation of such a critical nucleus is thus proposed as the rate limiting step. Following nucleation the remaining base pairs rapidly "zipper" into place, resulting in a fully hybridized double helix. Fig. 1.1 shows a schematic of the proposed nucleation-zipper mechanism of DNA hybridization.

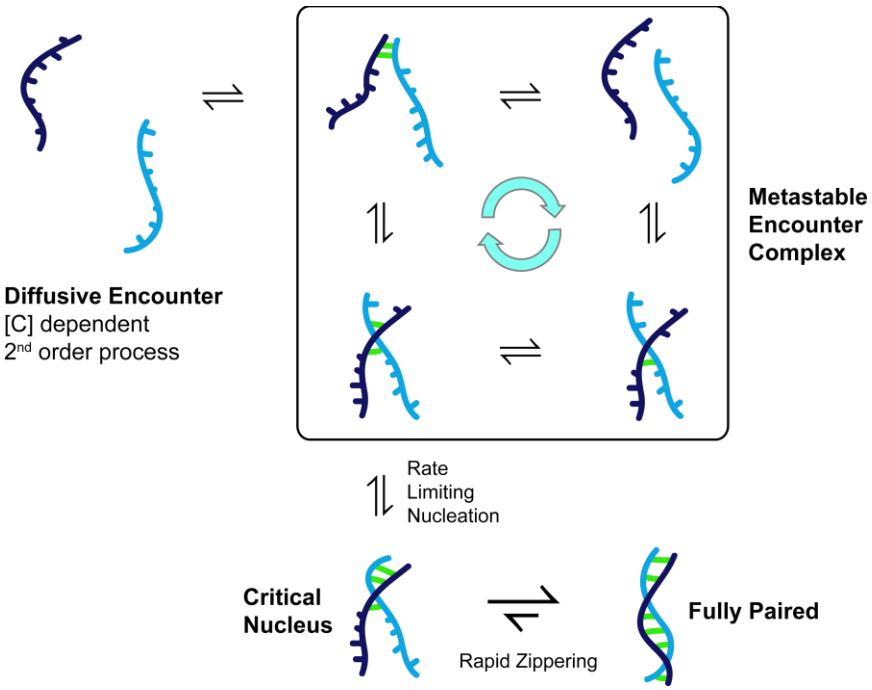

Figure 1.1: Schematic of the canonical nucleation-zipper mechanism of DNA hybridization first proposed to account for early measurements of DNA association and dissociation kinetics.

Subsequent studies have filled in many of the finer details of this hybridization mechanism. Measurements relying on DNA functionalized with fluorescent tags have led the way in characterizing localized base pairing dynamics.[22-24] Single-molecule force pulling experiments have mapped the free energy surfaces and transition states with respect to a mechanically induced dehybridization coordinate.[25,26] In addition to DNA, there has been much interest in the folding of RNA hairpins, which serve as the most basic RNA secondary structural motif. Evidence suggests complex dynamics on a rough energy landscape even for hairpin sequences as short as eight base pairs.[27-29]

Despite significant experimental advances, many aspects of the kinetics and dynamics of nucleic acid hybridization remain poorly understood. For instance, the direct details of the rapid zippering phase of the reaction are largely uncharacterized. As are the mechanistic details of the metastable encounter complex and the identity of the hybridization transition states. Nucleic acid hybridization is sensitive to base sequence, strand length, and cation concentration, but the precise origin of many of these effects have yet to be fully resolved. Within the past decade considerable advances in coarse grained models of DNA have allowed the simulation of the encounter, initial recognition, and zippering dynamics spanning across several decades in time. These simulations predict rich hybridization dynamics even for short oligonucleotides, including various out of register nucleation and internal rearrangement schemes, one-dimensional search of a strand along its complement, and in-register nucleation followed by rapid zippering.[30-32]

Going forward a new approach to studying the kinetics and dynamics of DNA hybridization with the ability to span the requisite picosecond to millisecond timescales without significantly perturbing the structure, thermodynamics, or native base pairing mechanism is required. The work presented in this thesis seeks to develop a steady-state and transient infrared

(IR) spectroscopy approach that satisfies these criteria. Two broad goals of this work can be articulated. First, a toolbox of experimental methods and analysis techniques is established using model DNA oligonucleotides. Applying this approach, previously unexplored aspects of the hybridization mechanism, mainly rapid zippering, are characterized in detail. Second, once this groundwork is established, these tools are applied to study naturally occurring modified nucleobases and oligonucleotides implicated in epigenetic regulation of the mammalian genome. Although much work remains to fully understand DNA hybridization, we hope the strategy presented here will prove generally applicable to the study of a wide range of dynamic questions in nucleic acid folding.

1.2 Infrared Spectroscopy of DNA

1.2.1 Linear and Two-Dimensional IR of the In-Plane Nucleobase Vibrations

Infrared spectroscopy offers several key advantages with regard to the study of the dynamics and kinetics of biomolecules. Most notably, the greatest advantage of IR is its intrinsic high time resolution. In particular, ultrafast nonlinear techniques, such as two-dimensional infrared (2D IR) spectroscopy, can directly monitor dynamics on a sub-picosecond timescale. Multidimensional spectroscopies also provide enhanced information by spreading the spectrum across two frequency axes and directly measuring couplings between resonances. Since infrared light is resonant with vibrational transitions, IR spectroscopy probes the intrinsic vibrations of molecular bonds without the need for any external probe or label. Vibrational transitions are sensitive to the identity of the atoms involved, the nature of the bonds between them, and the collective character of a particular mode of vibration. As a result, IR spectroscopy offers a degree of structural resolution that, when coupled with high time resolution, allows the evolution of some

of the fastest events in biology to be tracked in real time. Further still, since biomolecules such as nucleic acids occupy their electronic and vibrational ground states in many contexts of interest it is advantageous to employ a technique that can probe these states without inducing higher energy transitions.

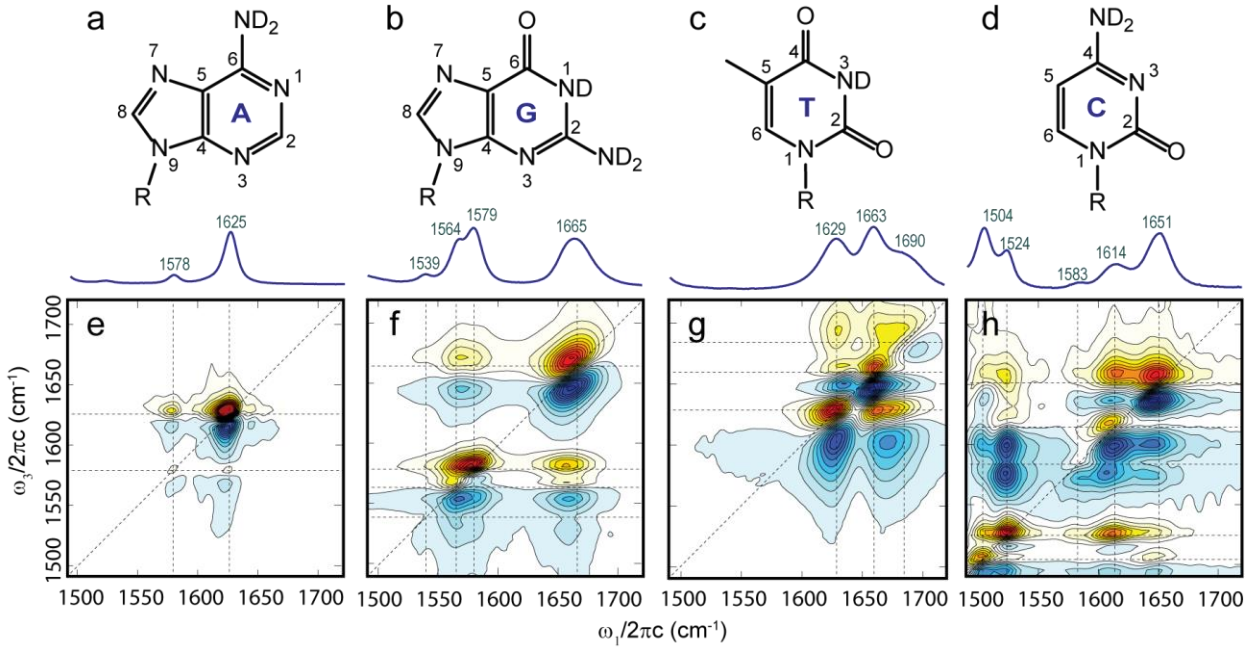

Figure 1.2: The structure of each of the four canonical nucleobases and their associated linear spectra in the 1500-1800 cm^{-1} frequency range for (a) adenine, (b) guanine, (c) thymine, and (d) cytosine. (e-h) The corresponding 2D IR spectrum for each base is plotted directly below the linear spectrum for reference. Adapted from Ref. 36.

Both linear and nonlinear infrared techniques will be employed throughout this thesis to study the structure and dynamics of DNA. The IR spectra of nucleic acids as well as the general mode assignments have been well characterized.[33,34] Across the mid-IR range from 800-1800 cm^{-1} the spectrum can be divided into four frequency regions that are each sensitive to different aspects of nucleic acid structure. Vibrations located in the 1500-1800 cm^{-1} range are characterized by in-plane ring vibrations originating from C=O stretches, C=N stretches, and ND_2 bending. When

studying this frequency range samples are commonly prepared in D_2O rather than H_2O to remove interference from the H_2O bend absorption at 1650 cm^{-1}. Base-ribose vibrations that are sensitive to the glycosidic bond rotation absorb between 1250-1500 cm^{-1}. The 1000-1250 cm^{-1} frequency range contains phosphate-ribose vibrations that report on the conformation of the DNA backbone. Lastly and lowest in frequency, ribose modes in the 800-1000 cm^{-1} range are sensitive to the geometry of the sugar. As a result of this wide range of structural sensitivity, IR spectroscopy is an ideal candidate for studying the conformational dynamics of nucleic acids.

In this thesis, the focus is on the 1500-1800 cm^{-1} frequency range since the in-plane base vibrations are sensitive to the hydrogen bonding and base stacking interactions that mediate Watson-Crick base pairing and DNA hybridization. Furthermore, each of the nucleobases demonstrate a unique vibrational fingerprint in this frequency range. Fig. 1.2 shows the structure of the four canonical nucleobases as well as the associated linear and 2D IR spectrum between 1500-1800 cm^{-1}. The 2D IR spectra of the in-plane nucleobase vibrations have previously been used to characterize the transition energies, vibrational anharmonicities, couplings, and transition dipole strengths as well as relative dipole orientations.[35,36]

1.2.2 Mode Assignments for the In-Plane Nucleobase Vibrations

Critically and in contrast to the historical local-mode assignments of the in-plane nucleobase vibrations of DNA, 2D IR spectroscopy has revealed that there are no simple or intuitive structural correlations to assign the spectral features with respect to the localized vibrations of the constituent molecular bonds. Density functional theory (DFT) calculations of explicitly solvated nucleobases have been used to assign the nature of each of the vibrational modes. Table 1.1 contains a summary of these mode assignments from Ref. 36. Throughout this

thesis, these vibrational mode assignments are used as a guide for interpreting the infrared spectroscopy of the in-plane nucleobase vibrations of DNA.

Table 1.1: Vibrational mode assignments used throughout this thesis for the in-plane nucleobase vibrations of DNA based on the 2D IR and DFT study from Ref. 36.

Nucleotide	Peak	ω_i (cm^{-1})	Vibrations
AMP	**A1**	1625	ν (C^4=C^5, C^5-C^6 out-of-phase), δ (C^2-H), δ (N^6D_2), *Py*
	A2	1578	ν (C^4=C^5, C^5-C^6 in-phase), ν (N^1-C^6), ν (N^3-C^4), ν (N^7=C^8), δ (C^8-H), *Py* + *Im*
GMP	**G1**	1665	ν (C^6=O), δ (N^1-D), *Py*
	G2	1579	ν (C^2=N^3), ν (C^6=O), δ (N^1-D), δ (N^2D_2), *Py*
	G3	1565	ν (C^2=N^3-C^4=C^5), δ (C^8-H), *Py* + *Im*
	G4	1539	ν (C^4=C^5), ν (N^7=C^8), δ (C^8-H), *Py* + *Im*
TMP	**T1**	1690	ν (C^2=O), ν (C^4=O), δ (N^3-D)
	T2	1663	ν (C^4=O), ν (C^5=C^6), δ (N^3-D), δ (C^5H_3), δ (C^6-H)
	T3	1629	ν (C^5=C^6), ν (C^4=O), δ (C^5H_3) , δ (C^6-H)
CMP	**C1**	1651	ν (C^2=O), ν (N^1=C^6), δ (C^6-H)
	C2	1614	ν (N^3=C^4-C^5=C^6), ν (C^2=O), δ (C^5-H), δ (C^6-H)
	C3	1583	
	C4	1524	ν (C^4-C^5), ν (N^1-C^6), δ (C^5-H), δ (C^6-H), δ (N^4D_2)
	C5	1504	ν (N^3=C^4-N^4), ν (C^5=C^6), δ (C^5-H), δ (C^6-H), δ (N^4D_2)

ν : stretching; δ : bending; *Py*: pyrimidine ring vibration; *Im*: imidazole ring vibration

The presence of cross-peaks between every diagonal peak in the 2D IR spectra in Fig. 1.2e-h is reflective of the high degree to which the in-plane nucleobase modes are intrinsically coupled to one another. The formation of a Watson-Crick base pair results in additional cross-peaks between complementary nucleobases mediated through the hydrogen bonding and base stacking interactions of hybridized DNA. Compared to the individual nucleobases, the 2D IR spectra of oligonucleotide and duplex DNA are not as well understood. Initial efforts focused on the carbonyl stretches of GC sequences and found these vibrational modes to be strongly coupled through and delocalized across the hydrogen bonds and base stacks of duplexed DNA.[37]

Complementary theoretical work obtained reasonable agreement with the experimental spectra using coupling constants and basis modes obtained from DFT calculations, but as of yet there is no general theory for mapping a DNA structure to its 2D IR spectrum.[38,39] Studies of hydrated DNA thin films of AT sequences have revealed decelerated water dynamics in the hydration shell and have determined that the mechanism of vibrational relaxation of duplex DNA proceeds through the phosphate backbone coupling to the solvent.[40,41] Beyond isolated DNA, 2D IR spectroscopy has also been used to study ligand binding to nucleic acids, revealing mechanistic details of the association of a fluorescent dye within the minor groove.[42]

Despite experimental progress, future work is still needed to establish a firm theoretical framework for the 2D IR spectroscopy of duplexed DNA. Spectroscopic assignments in this thesis are therefore based on the nucleobases as well as the long-established assignments for the linear infrared spectra of both single stranded DNA and various folded structures.[33,34] For ease of discussion, the vibrations in the 1500-1800 cm^{-1} frequency range can be roughly divided into two categories. From an inspection of Table 1.1, one can see that the vibrational modes above ~1650 cm^{-1} contain strong C=O character. The nucleobase vibrations below ~1650 cm^{-1} have less pronounced C=O character and are instead better characterized as delocalized vibrations involving the ring atoms. The vibrations in the former group can thus be categorized as carbonyl modes while those in the latter can be referred to collectively as ring modes.

1.2.3 Tracking the Dehybridization of DNA across Many Decades in Time

The dynamics that can be studied using equilibrium 2D IR spectroscopy are fundamentally limited by the vibrational lifetime of the system since this will dictate the timescale over which the signal decays. A characteristic lifetime for the in-plane nucleobase vibrations of DNA is ~1 ps.

However, many dynamic events of interest, including the formation/loss of hydrogen bonds and stacking interactions as well as the hybridization of complementary strands occur on a ns-ms timescale. Furthermore, rare stochastic events are difficult to resolve in steady-state ensemble measurements. To address these limitations, an external perturbation can be used to rapidly establish a non-equilibrium state and the relaxation of the population toward a new equilibrium can be monitored to observe the response of the system.[43]

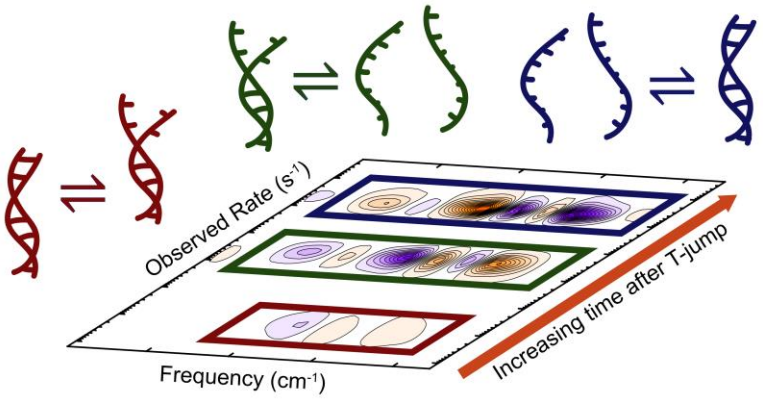

Figure 1.3: Schematic illustration of the resolution of DNA dehybridization dynamics across many decades in time following an optical temperature jump. The rate spectrum spans six orders of magnitude, from nanoseconds to milliseconds. Figure is reused with permission from Ref. 44.

In this thesis transient temperature jump (T-jump) nonlinear IR spectroscopy is employed to study DNA kinetics and dynamics occurring beyond the vibrational lifetime of the nucleobases. A second laser that delivers a ns T-jump up to ~20 °C is synchronized to a 2D IR spectrometer that is employed in a "snap-shot" mode to probe the system at an electronically controlled delay some time after the T-jump event. This approach has been used previously by the Tokmakoff group to study protein folding as well as tautomerization in nucleobases.[45-47]

1.3 Extension to Non-Canonical DNA

Many unresolved questions surrounding the dynamics and kinetics of nucleic acids involve modifications that lie outside of the four canonical nucleobases. Such changes can result from DNA damage and the formation of lesions[48] or, as is often the case for RNA, can represent post transcriptional modifications that are essential for proper function.[49] Not all relevant non-canonical nucleobases are naturally occurring. Synthetic nucleobase analogs are a promising class of pharmaceuticals, particularly for addressing cancer[50] and as antivirals.[47,51]

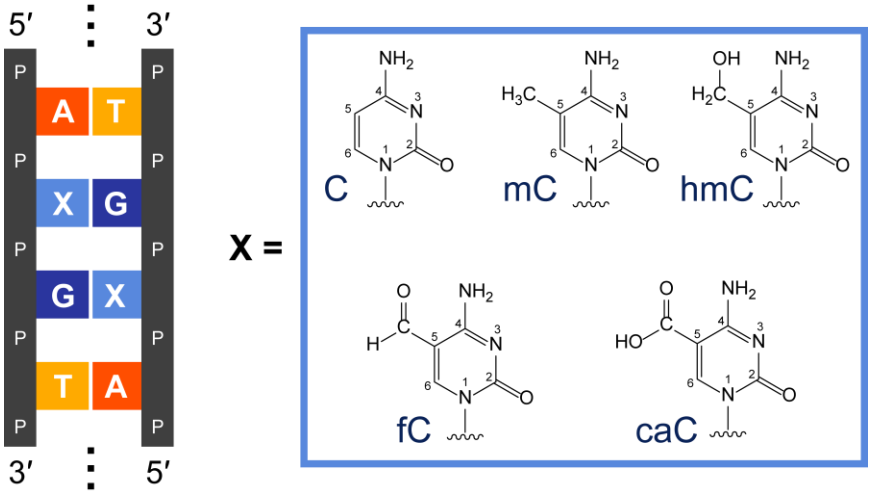

Figure 1.4: Diagram of an XpG domain found in promoter regions of the genome where X is either cytosine (C), 5-methylcytosine (mC), 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), or 5-carboxylcytosine (caC).

The IR spectroscopy-based techniques and analysis methods assembled to study the mechanism of DNA dehybridization in this thesis are also applied to study a class of epigenetically modified cytosines that are involved in the regulation of gene expression. These modifications are found naturally in GC rich tracts of the genome that occur in promoter regions called CpG domains. Here "p" denotes that the bases are linked by the phosphate backbone of a shared DNA

strand. Enzymatically methylating the 5-position, resulting in 5-methylcytosine (mC), has the effect of silencing genes. Until recently, the details of how CpG domains are demethylated, and therefore genes reactivated, remained a mystery. In 2011 an active cytosine demethylation pathway was discovered in which the methyl group of mC is successively oxidized to 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC).[52,53] The fC and caC bases in this oxidation chain can be selectively excised by enzymes and the abasic site is subsequently repaired with canonical C. Fig. 1.4 shows a representation of a XpG domain where X is either C, mC, hmC, fC, or caC. The structure of each of the modified cytosine derivatives is also indicated.

The particular influence of each of the modified cytosine nucleobases on the fundamental biophysical properties of DNA as well as the potential implication of such influences with regard to the active demethylation mechanism remains an ongoing debate.[54,55] Of particular interest in this thesis is the question of how each modification impacts the thermodynamic stability of a CpG domain as well as the kinetic barrier to opening the domain. The final two chapters explore these topics and discuss the possible biological consequences of certain modifications reducing base pairing stability and lowering the barrier to dehybridizing CpG domains.

1.4 Thermodynamics of the DNA Duplex to Single Strand Transition

1.4.1 Generalized Dimer to Monomer Reaction

This section presents the standard thermodynamic picture commonly applied to describe DNA duplex melting. This material serves as an important starting point for the work in this thesis and is included here as a convenient reference. The discussion will focus on self-complementary

DNA sequences since this is the relevant case for the oligonucleotides studied here. To begin, consider the generic dimer to monomer reaction

$$[D] \underset{k_a}{\overset{k_d}{\rightleftharpoons}} 2[M] \tag{1.1}$$

Where [D] is the dimer concentration, [M] is the monomer concentration, and k_d and k_a correspond to the dissociation and association rate constants, respectively. A dimer in this case is defined as any pair of monomer strands sharing at least one intact base pair. In practice the quantity that is most easily set in the laboratory is the total concentration of DNA strands, $[C_{Tot}] = 2[D] + [M]$. At equilibrium, the dimer to monomer reaction is described by the dissociation constant eq 1.2, where the concentrations are assumed throughout to be referenced to the standard state concentration of 1 mol/L.

$$K_d = \frac{[M]^2}{[D]} \tag{1.2}$$

The chemical potential, μ_i per mole of component i, n_i is related to the Gibbs free energy G through eq 1.3.

$$\mu_i = \left(\frac{\partial G}{\partial n_i} \right)_{p,T,\{n_j, j \neq i\}} \tag{1.3}$$

Furthermore, the partial molar free energy in terms of the concentration of component i, c_i is given by

$$\mu_i = \mu_i^\circ + RT \ln c_i \tag{1.4}$$

where μ_i^o is the reference chemical potential for component i at standard state conditions. For the dissociation reaction, the molar reaction free energy can be expressed as

$$\Delta G = \sum_i \nu_i \mu_i = 2\mu_M - \mu_D \tag{1.5}$$

where v_i is the stoichiometric coefficient for component *i*. The molar reaction free energy is related to the dimer and monomer concentrations through

$$\Delta G = \Delta G^{\circ} + RT \ln \frac{[M]^2}{[D]} \tag{1.6}$$

At equilibrium, ΔG = 0 and therefore

$$\Delta G^{\circ} = -RT \ln \frac{[M]^2}{[D]} \tag{1.7}$$

1.4.2 The All-or-None Assumption and van 't Hoff Analysis

With these basic thermodynamic definitions in hand, we will next establish a direct connection to an experimental observable through the thermal melting curve. From the definition of the total concentration $[C_{Tot}] = 2[D] + [M]$ we can define the total fraction of all DNA that exists as monomers $\theta_M = [M]/C_{Tot}$ and the total fraction of dimers $\theta_D = 2[D]/C_{Tot}$. Starting with these definitions and using the expression for the equilibrium constant eq 1.2, one can derive an expression for the dimer fraction in terms of K_d and C_{Tot}.

$$\theta_D = \frac{4C_{Tot} + K_d - \sqrt{K_d^2 + 8K_d C_{Tot}}}{4C_{Tot}} \tag{1.8}$$

It is commonly assumed that the thermal melting curve reports directly on the dimer fraction. In order for this assumption to be true, one must also assume that base pairing follows an all-or-none picture in which all possible base pairs in a given strand are either fully intact or fully broken. In other words, all melting base pairs contributing to the signal can be attributed to the dimer transitioning to monomer. Once a melting curve has been measured, the simplest thermodynamic treatment is a van 't Hoff analysis.[56] This model assumes a temperature independent enthalpy and entropy change. The melting temperature T_m is commonly defined as the temperature at which

θ_D = ½. In the van 't Hoff picture, the free energy of dissociation can be defined in terms of the slope of the melting curve at the melting temperature.

$$\Delta G^\circ(T) = T\ln\left(C_{Tot}\right) + 6RT_m^2\left(1-\frac{T}{T_m}\right)\left(\frac{\partial\theta_D}{\partial T}\right)_{T=T_m} \tag{1.9}$$

1.4.3 Modeling Temperature Dependent Changes in Enthalpy and Entropy

Although the van 't Hoff picture is often adequate to describe the melting of small oligonucleotides, it is expected that more complicated conformational changes will be accompanied by a temperature dependent enthalpy and entropy change. Using the definition of the heat capacity $C_p = (\partial H/\partial T)_{N,P} = (T\partial S/\partial T)_{N,P}$ and assuming a linear temperature dependence,

$$\Delta H^\circ(T) = \Delta H^\circ(T_m) + \Delta C_p\left[T - T_m\right] \tag{1.10}$$

$$\Delta S^\circ(T) = \Delta S^\circ(T_m) + \Delta C_p \ln\left(\frac{T}{T_m}\right) \tag{1.11}$$

Note that ΔC_p is considered to be independent of temperature in this description. The melting temperature, T_m, is taken to be the reference temperature. Given that $\Delta G°(T) = \Delta H°(T) - T\Delta S°(T)$, eq 1.10 and 1.11 can be substituted into eq 1.6 above to obtain an expression for the standard Gibbs free energy.

$$\Delta G^\circ(T) = \Delta H^\circ(T_m) - T\Delta S^\circ(T_m) + \Delta C_p\left[T - T_m - T\ln\left(\frac{T}{T_m}\right)\right] \tag{1.12}$$

At T_m the standard entropy can be expressed as

$$\Delta S^\circ(T_m) = \frac{\Delta H^\circ(T_m) + RT_m\ln\left(C_{Tot}\right)}{T_m} \tag{1.13}$$

In practice optical melting curves often exhibit sloping baselines that obscure a direct fit to the dimer fraction, eq 1.8. Instead, the melting curve $S_{melt}(T)$ is fit to

$$S_{melt}(T) = \theta_D(T)\left[S_{up}(T) - S_{low}(T)\right] + S_{low}(T) \tag{1.14}$$

Where the $S_{up}(T)$ and $S_{low}(T)$ are simultaneous linear fits to the upper and lower sloping baselines, respectively.

1.5 Thesis Outline

The remainder of this thesis is organized as follows. Chapters 2 and 3 contain the necessary theoretical and experimental background. Chapter 4 describes additional analysis methods that lie outside of the standard approach commonly used in steady-state and transient infrared spectroscopy. Chapter 5 presents the details of a statistical lattice model developed to describe the thermodynamics of DNA dehybridization and to characterize the ensemble of hybridized dimer states. Chapter 6 introduces a set of model DNA oligonucleotides that have a uniform AT and GC content but differ in the placement of their GC pairs. Linear and nonlinear IR spectroscopy are used to characterize the sequence-dependent dehybridization thermodynamics and the lattice model is used to interpret the experimental results in terms of the accumulation of frayed states in the melting dimer ensemble. In Chapter 7, a kinetic analysis of two of the model sequences from the previous chapter using transient T-jump nonlinear IR spectroscopy reveals both an activated kinetic process associated with dimer dissociation and downhill dynamics of nanosecond unzipping of the helical termini as a function of nucleobase sequence. After establishing the experimental approach and analysis methods of Chapters 6 and 7 using model oligonucleotides consisting of canonical nucleobases, Chapter 8 turns to nucleobases and oligonucleotides containing modifications relevant to epigenetic regulation and the active cytosine demethylation pathway. The 2D IR spectra of caC and fC confirm that these nucleotides exist in the canonical keto-amino form under physiological conditions rather than rare tautomeric forms that could

disrupt base pairing. The pK_as of the modified nucleobases' titratable groups are assigned using linear IR spectroscopy and DFT calculations. Weakened N3 base pairing as measured by a reduction in pK_a and oligonucleotide duplex stability is hypothesized to contribute to the selective excision of caC and fC over C, mC, and hmC. Chapter 9 builds on the results of the previous chapter and addresses anomalous behavior observed for some of the modified sequences' melting curves. The barrier to opening XpG domains as a function of X = C, mC, hmC, fC, or caC is determined. A significant reduction in the dissociation barrier is observed for the X = hmC and fC sequences while mC and caC exhibit a smaller reduction relative to the canonical sequence. The chapter ends with a discussion of the potential biological implications of these findings.

1.6 References

1. Leslie E, O., Prebiotic chemistry and the origin of the RNA world. *Critical reviews in biochemistry and molecular biology* **2004,** *39* (2), 99-123.

2. Benner, S. A.; Ellington, A. D.; Tauer, A., Modern metabolism as a palimpsest of the RNA world. *Proceedings of the National Academy of Sciences* **1989,** *86* (18), 7054-7058.

3. Seeman, N. C., DNA nanotechnology: novel DNA constructions. *Annual review of biophysics and biomolecular structure* **1998,** *27* (1), 225-248.

4. Pinheiro, A. V.; Han, D.; Shih, W. M.; Yan, H., Challenges and opportunities for structural DNA nanotechnology. *Nature nanotechnology* **2011,** *6* (12), 763.

5. Shin, J.-S.; Pierce, N. A., A synthetic DNA walker for molecular transport. *Journal of the American Chemical Society* **2004,** *126* (35), 10834-10835.

6. Doering, C.; Ermentrout, B.; Oster, G., Rotary DNA motors. *Biophysical Journal* **1995,** *69* (6), 2256-2267.

7. Braich, R. S.; Chelyapov, N.; Johnson, C.; Rothemund, P. W.; Adleman, L., Solution of a 20-variable 3-SAT problem on a DNA computer. *Science* **2002,** *296* (5567), 499-502.

8. Watson, J. D.; Crick, F. H., Genetical implications of the structure of deoxyribonucleic acid. *Nature* **1953,** *171* (4361), 964-967.

9. Brauns, E. B.; Madaras, M. L.; Coleman, R. S.; Murphy, C. J.; Berg, M. A., Measurement of local DNA reorganization on the picosecond and nanosecond time scales. *Journal of the American Chemical Society* **1999,** *121* (50), 11644-11649.

10. Krueger, A.; Protozanova, E.; Frank-Kamenetskii, M. D., Sequence-dependent basepair opening in DNA double helix. *Biophysical journal* **2006,** *90* (9), 3091-3099.

11. Cao, C.; Jiang, Y. L.; Stivers, J. T.; Song, F., Dynamic opening of DNA during the enzymatic search for a damaged base. *Nature Structural and Molecular Biology* **2004,** *11* (12), 1230-1236.

12. Ramstein, J.; Lavery, R., Energetic coupling between DNA bending and base pair opening. *Proceedings of the National Academy of Sciences* **1988,** *85* (19), 7231-7235.

13. Horowitz, D. S.; Wang, J. C., Torsional rigidity of DNA and length dependence of the free energy of DNA supercoiling. *Journal of molecular biology* **1984,** *173* (1), 75-91.

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. *Journal of molecular biology* **1971,** *62* (2), 361-381.

15. Pörschke, D.; Uhlenbeck, O.; Martin, F., Thermodynamics and kinetics of the helix-coil transition of oligomers containing GC base pairs. *Biopolymers: Original Research on Biomolecules* **1973,** *12* (6), 1313-1335.

16. Wetmur, J. G.; Davidson, N., Kinetics of renaturation of DNA. *Journal of molecular biology* **1968,** *31* (3), 349-370.

17. Craig, M. E.; Crothers, D. M.; Doty, P., Relaxation kinetics of dimer formation by self complementary oligonucleotides. *Journal of molecular biology* **1971,** *62* (2), 383-401.

18. Crothers, D. M., The kinetics of DNA denaturation. *Journal of molecular biology* **1964,** *9* (3), 712-733.

19. Ageno, M.; Dore, E.; Frontali, C., The alkaline denaturation of DNA. *Biophysical journal* **1969,** *9* (11), 1281.

20. Patel, D. J.; Canuel, L., Nuclear magnetic resonance studies of the helix-coil transition of poly (dA-dT) in aqueous solution. *Proceedings of the National Academy of Sciences* **1976,** *73* (3), 674-678.

21. McConnell, B.; von Hippel, P. H., Hydrogen exchange as a probe of the dynamic structure of DNA: I. General acid-base catalysis. *Journal of molecular biology* **1970,** *50* (2), 297-316.

22. Chen, X.; Zhou, Y.; Qu, P.; Zhao, X. S., Base-by-base dynamics in DNA hybridization probed by fluorescence correlation spectroscopy. *Journal of the American Chemical Society* **2008,** *130* (50), 16947-16952.

23. Rachofsky, E. L.; Osman, R.; Ross, J. A., Probing structure and dynamics of DNA with 2-aminopurine: effects of local environment on fluorescence. *Biochemistry* **2001,** *40* (4), 946-956.

24. Weiss, S., Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Structural and Molecular Biology* **2000,** *7* (9), 724-729.

25. Woodside, M. T.; Anthony, P. C.; Behnke-Parks, W. M.; Larizadeh, K.; Herschlag, D.; Block, S. M., Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid. *Science* **2006,** *314* (5801), 1001-1004.

26. Strunz, T.; Oroszlan, K.; Schäfer, R.; Güntherodt, H.J., Dynamic force spectroscopy of single DNA molecules. *Proceedings of the National Academy of Sciences* **1999,** *96* (20), 11277-11282.

27. Ma, H.; Proctor, D. J.; Kierzek, E.; Kierzek, R.; Bevilacqua, P. C.; Gruebele, M., Exploring the energy landscape of a small RNA hairpin. *Journal of the American Chemical Society* **2006,** *128* (5), 1523-1530.

28. Ansari, A.; Kuznetsov, S. V.; Shen, Y., Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proceedings of the National Academy of Sciences* **2001,** *98* (14), 7771-7776.

29. Brion, P.; Westhof, E., Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure* **1997,** *26* (1), 113-137.

30. Hinckley, D. M.; Lequieu, J. P.; De Pablo, J. J., Coarse-grained modeling of DNA oligomer hybridization: length, sequence, and salt effects. *The Journal of chemical physics* **2014,** *141* (3), 035102.

31. 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. *The Journal of chemical physics* **2013,** *139* (14), 144903.

32. Šulc, P.; Romano, F.; Ouldridge, T. E.; Rovigatti, L.; Doye, J. P.; Louis, A. A., Sequence-dependent thermodynamics of a coarse-grained DNA model. *The Journal of chemical physics* **2012,** *137* (13), 135101.

33. Banyay, M.; Sarkar, M.; Gräslund, A., A library of IR bands of nucleic acids in solution. *Biophysical chemistry* **2003,** *104* (2), 477-488.

34. Taillandier, E.; Liquier, J.; Ghomi, M., Conformational transitions of nucleic acids studied by IR and Raman spectroscopies. *Journal of Molecular Structure* **1989,** *214*, 185-211.

35. Peng, C. S.; Jones, K. C.; Tokmakoff, A., Anharmonic vibrational modes of nucleic acid bases revealed by 2D IR spectroscopy. *Journal of the American Chemical Society* **2011,** *133* (39), 15650-15660.

36. Peng, C. S. Two-dimensional infrared spectroscopy of nucleic acids: application to tautomerism and DNA aptamer unfolding dynamics. Doctoral Dissertation, Massachusetts Institute of Technology, 2014.

37. Krummel, A. T.; Zanni, M. T., DNA vibrational coupling revealed with two-dimensional infrared spectroscopy: insight into why vibrational spectroscopy is sensitive to DNA structure. *The Journal of Physical Chemistry B* **2006,** *110* (28), 13991-14000.

38. Lee, C.; Park, K.-H.; Cho, M., Vibrational dynamics of DNA. I. Vibrational basis modes and couplings. *The Journal of chemical physics* **2006,** *125* (11), 114508.

39. Lee, C.; Cho, M., Vibrational dynamics of DNA: IV. Vibrational spectroscopic characteristics of A-, B-, and Z-form DNA's. *The Journal of chemical physics* **2007,** *126* (14), 145102.

40. Szyc, Ł.; Yang, M.; Elsaesser, T., Ultrafast Energy Exchange via Water− Phosphate Interactions in Hydrated DNA. *The Journal of Physical Chemistry B* **2010,** *114* (23), 7951-7957.

41. Yang, M.; Szyc, Ł.; Elsaesser, T., Decelerated water dynamics and vibrational couplings of hydrated DNA mapped by two-dimensional infrared spectroscopy. *The Journal of Physical Chemistry B* **2011,** *115* (44), 13093-13100.

42. Ramakers, L. A.; Hithell, G.; May, J. J.; Greetham, G. M.; Donaldson, P. M.; Towrie, M.; Parker, A. W.; Burley, G. A.; Hunt, N. T., 2D-IR Spectroscopy Shows that Optimized DNA Minor Groove Binding of Hoechst33258 Follows an Induced Fit Model. *The Journal of Physical Chemistry B* **2017,** *121* (6), 1295-1303.

43. Nölting, B., *Protein folding kinetics: biophysical methods*. Springer: Berlin, 2006.

44. Sanstead, P. J.; Tokmakoff, A., Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *The Journal of Physical Chemistry B* **2018,** *122* (12), 3088-3100.

45. Chung, H. S.; Khalil, M.; Smith, A. W.; Ganim, Z.; Tokmakoff, A., Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin. *Proceedings of the National Academy of Sciences* **2005,** *102* (3), 612-617.

46. Jones, K. C.; Peng, C. S.; Tokmakoff, A., Folding of a heterogeneous β-hairpin peptide from temperature-jump 2D IR spectroscopy. *Proceedings of the National Academy of Sciences* **2013,** *110* (8), 2828-2833.

47. Peng, C. S.; Fedeles, B. I.; Singh, V.; Li, D.; Amariuta, T.; Essigmann, J. M.; Tokmakoff, A., Two-dimensional IR spectroscopy of the anti-HIV agent KP1212 reveals protonated and neutral tautomers that influence pH-dependent mutagenicity. *Proceedings of the National Academy of Sciences* **2015**, 3229-3234.

48. Pfeifer, G. P., Formation and processing of UV photoproducts: effects of DNA sequence and chromatin environment. *Photochemistry and photobiology* **1997,** *65* (2), 270-283.

49. Helm, M., Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic acids research* **2006,** *34* (2), 721-733.

50. Galmarini, C. M.; Mackey, J. R.; Dumontet, C., Nucleoside analogues and nucleobases in cancer treatment. *The lancet oncology* **2002,** *3* (7), 415-424.

51. Lee, Y.-S.; Kim, B. H., Heterocyclic nucleoside analogues: design and synthesis of antiviral, modified nucleosides containing isoxazole heterocycles. *Bioorganic & medicinal chemistry letters* **2002,** *12* (10), 1395-1397.

52. Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y., Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **2011,** *333* (6047), 1300-1303.

53. He, Y.-F.; Li, B.-Z.; Li, Z.; Liu, P.; Wang, Y.; Tang, Q.; Ding, J.; Jia, Y.; Chen, Z.; Li, L., Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **2011,** *333* (6047), 1303-1307.

54. Hardwick, J. S.; Lane, A. N.; Brown, T., Epigenetic modifications of cytosine: biophysical properties, regulation, and function in mammalian DNA. *BioEssays* **2018,** *40* (3), 1700199.

55. Drohat, A. C.; Coey, C. T., Role of base excision "Repair" enzymes in erasing epigenetic marks from DNA. *Chemical reviews* **2016,** *116* (20), 12711-12729.

56. Marky, L. A.; Breslauer, K. J., Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers: Original Research on Biomolecules* **1987,** *26* (9), 1601-1620.

Chapter 2

Two-Dimensional Infrared Spectroscopy

2.1 Introduction

Two-dimensional infrared spectroscopy (2D IR) is an ultrafast nonlinear spectroscopic technique that interrogates the vibrational potential of a system using a series of femtosecond infrared pulses. When applied to study biophysical systems such as nucleic acids and proteins, the vibrational spectrum is often related to the molecular structure, which can in turn be monitored with intrinsic time resolution of hundreds of femtoseconds. 2D IR thus offers structural sensitivity encoded by the infrared spectrum of the biomolecule as well as high time-resolution that provides a window into many of the fastest events in biology. As a multidimensional spectroscopic technique, 2D IR reports directly on the connections between the vibrational modes of the system and can characterize vibrational couplings, energy transfer, and the dynamics of chemical exchange.[1,2] Furthermore homogeneous and inhomogeneous contributions to the line shape can be resolved and monitored in time to measure the loss of frequency correlation in the system.[3-5]

The theoretical framework that describes 2D IR is well established.[6-8] The intention of this chapter is to provide a brief overview of this framework that will be sufficient to describe what gives rise to the features in a 2D IR spectrum and to discuss the information content of the spectrum as it pertains to molecular structure. The first section presents an outline of the third-order response function formalism that is commonly used to describe 2D IR. Simple model spectra are used to

illustrate how different pathways up and down the vibrational energy ladder contribute to the observed signal as well as the factors that give rise to time dependent changes to line shapes and intensities. A method for reconstructing the dispersed pump probe spectrum from the heterodyne-detected dispersed vibrational echo (HDVE) collected in the boxcar geometry is described. This method is employed in the transient temperature jump experiments detailed in the next chapter.

2.2 The Information Content of a 2D IR Spectrum

2.2.1 Third-Order Response Function Formalism

The standard theoretical description of 2D IR spectroscopy employs a semi-classical description in which the light is treated classically and the vibrational energy levels of the system are quantized. In the most common experimental realization of 2D IR spectroscopy, three short pulses of light separated by variable time delays are used to interrogate the vibrational states of the system and to map out couplings between them. The signal that is ultimately measured is the electric field radiated by the third-order macroscopic polarization $\mathbf{P}^{(3)}$ induced in the system ensemble after a series of interactions with the incident light fields. This polarization will depend on the respective wavevectors $\mathbf{k}_n$ of the incident fields, the time delays between pulses τ_n, and the intrinsic response of the system, $\mathbf{R}^{(3)}$.[6-8]

$$\boldsymbol{P}^{(3)}(\boldsymbol{k}_{sig},t,\tau_2,\tau_1)=\int_0^\infty\int_0^\infty\int_0^\infty \boldsymbol{R}(\tau_3,\tau_2,\tau_1)\boldsymbol{E}_3\left(\boldsymbol{k}_3,\omega_3,t-\tau_3\right)\times \boldsymbol{E}_2\left(\boldsymbol{k}_2,\omega_2,t-\tau_3-\tau_2\right)\boldsymbol{E}_1\left(\boldsymbol{k}_1,\omega_1,t-\tau_3-\tau_2-\tau_1\right)d\tau_1 d\tau_2 d\tau_3 \tag{2.1}$$

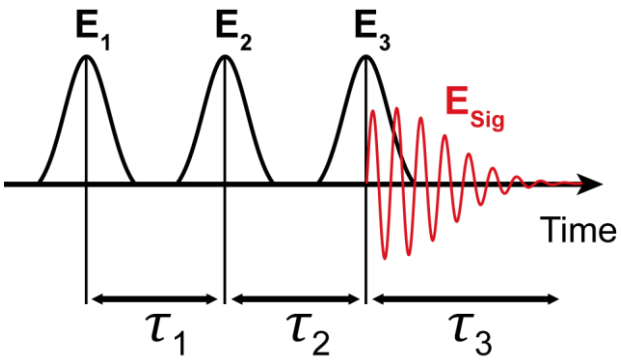

Figure 2.1: Series of three femtosecond infrared pulses separated by variable time delays τ_n used to generate the macroscopic third-order polarization in the sample that radiates the 2D IR signal of interest, $\mathbf{E}_{sig}$. The three time intervals are commonly referred to as the coherence (τ_1), waiting (τ_2), and detection (τ_3) times.

Fig. 2.1 shows a schematic of three pulses delayed relative to each other by the variable time delays defined in the figure caption. The third-order nonlinear response function $\mathbf{R}^{(3)}$ contains the system information of interest. To arrive at an expression for $\mathbf{R}^{(3)}$, a perturbative expansion of the system density matrix in which the light-matter interaction is treated as a small time-dependent perturbation to the system Hamiltonian results in a series of nested commutators that represent a four-point dipole correlation function evaluated at each light-matter interaction.[1,6]

$$\boldsymbol{R}^{(3)}(\tau_3,\tau_2,\tau_1)=\left(\frac{i}{\hbar}\right)^3\left\langle[[[\boldsymbol{M}(\tau_3+\tau_2+\tau_1),\boldsymbol{M}(\tau_2+\tau_1)],\boldsymbol{M}(\tau_1)],\boldsymbol{M}(0)]\rho_0\right\rangle \tag{2.2}$$

Where **M** is the dipole operator and ρ_0 is the equilibrium density matrix for the system eigenstates. The commutator in eq 2.2 can be expanded into eight terms that represent the set of possible light-matter interaction pathways, called Liouville pathways, that describe the evolution of the system density matrix.

$$\boldsymbol{R}^{(3)}(\tau_3,\tau_2,\tau_1)=\sum_{n=1}^{4}\boldsymbol{R}_n(\tau_3,\tau_2,\tau_1)-\boldsymbol{R}_n^{*}(\tau_3,\tau_2,\tau_1) \tag{2.3}$$

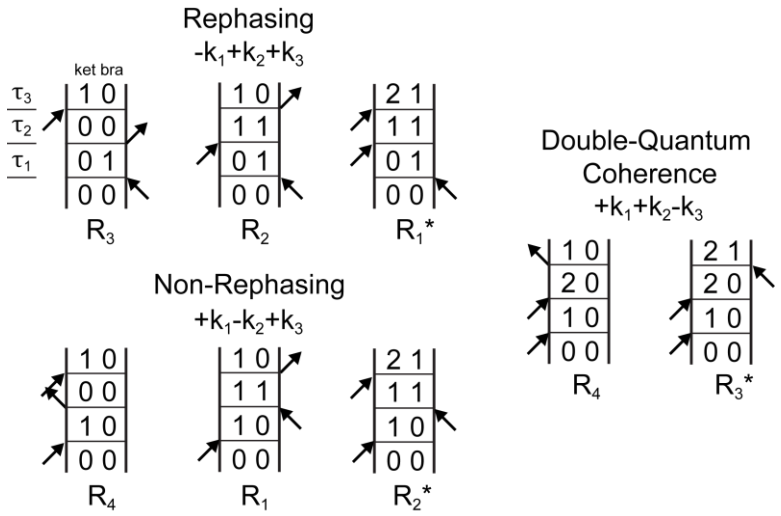

Figure 2.2: Set of double-sided Feynman diagrams that depict the Liouville pathways for a three level system. Time runs from bottom to top. Horizontal bars denote light-matter interactions while the spaces between depict how the system evolves during the intervening time periods. Arrows pointing towards the diagram indicate absorption while arrows pointing away indicate emission.

The response function contains all of the information pertaining to the vibrational and orientational dynamics of the system. Often these degrees of freedom are assumed to be separable, which allows each term in the sum of eq 2.3 to be expressed as a product of a vibrational and orientational term.[7] We will not go into further detail here, but the vibrational term encodes the energies of the system eigenstates, the coherences between states, and the coupling between the system and the surroundings reflected in the dynamics of the 2D line shapes and intensities. The orientational term encodes the dependence of the signal on the polarization of the excitation fields due to the relative orientation between the transition dipoles of the system as well as reorientation of the molecules during the time interval between interactions.

Double-sided Feynman diagrams provide a convenient way to visualize the Liouville pathways that contribute to the third-order signal. Time in the diagrams runs from bottom to top. Horizontal lines represent light-matter interactions while the spaces between lines indicate how

the density matrix evolves during the intervening period between interactions. Arrows pointing towards the diagram indicate absorption while arrows pointing away indicate emission. As an example, the Feynman diagrams for a three-level system are given in Fig. 2.2. Often pathways are grouped according to their phase-matching condition specified by the wavevectors $\mathbf{k}_n$, associated with the $\mathbf{E}_n$ electric fields where n = 1, 2, or 3 and indicates the order of interaction with the system. The diagrams with $-\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$ are referred to as rephasing (R) pathways because, as can be seen in Fig. 2.2, the phase of the coherence during τ_1 is conjugate to that acquired during τ_3 resulting in a photon echo. Pathways with $+\mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$ are referred to as nonrephasing (NR) pathways and the phase is identical during both τ_1 and τ_3 time periods. Finally the phase-matching condition $+\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3$ results in double-quantum coherence pathways, characterized by a coherence between the ground state and second excited state during the waiting period, τ_2.

In the next chapter, two 2D IR spectrometer designs that are employed throughout this thesis are described. Although similar in many respects, one of their major differences lies in their respective beam geometries and it is worth making a note of this here as it relates to phase-matching. One spectrometer employs a pump-probe geometry in which the first and second pulses are collinear and therefore $\mathbf{k}_1 = \mathbf{k}_2$. As a result, the rephasing and nonrephasing pathways are both emitted in the same direction. In contrast, the other spectrometer employs a boxcar geometry in which each of the three signal generating pulses are arranged at the corners of a square. In this case the rephasing and nonrephasing signals are emitted in different phase-matched directions since $\mathbf{k}_1$ and $\mathbf{k}_2$ are unique wavevectors. In practice and for ease of detection, reversing the time ordering of the first two pulses allows for the two types of signal pathways to be emitted in the same direction at the fourth corner of the square, however this still requires that each surface is collected independently. Further details of 2D IR in practice are the topic of Chapter 3.

2.2.2 The 2D IR Spectrum of a Model Two-Level System

The following sections use simple model systems and simulated 2D IR spectra to illustrate how the information content of the third-order response function discussed in the previous section manifests in a 2D IR spectrum. Complexity is added gradually and the resulting influence on the spectrum is discussed. The simplest possible system has only two vibrational energy levels: a ground state and a singly excited state. To model the response function, we adopt the simplifying assumptions of Ref. 5 and also assume a two-dimensional Gaussian line shape. For the purposes of the simple model, the dynamics of the line shape are described by a single normalized frequency correlation function, $C(\tau)$.

$$C(\tau) = \langle \delta\omega(\tau)\delta\omega(0)\rangle \big/ \langle \delta\omega^2 \rangle \tag{2.4}$$

$$\delta\omega(\tau) = \omega(\tau) - \langle \omega \rangle \tag{2.5}$$

The third-order response function $\mathbf{R}^{(3)}$ can be expressed as the sum of a rephasing (−) and nonrephasing (+) response.

$$\boldsymbol{R}^{(3)}(\tau_1,\tau_2,\tau_3) = \boldsymbol{R}_{-}^{(3)}(\tau_1,\tau_2,\tau_3) + \boldsymbol{R}_{+}^{(3)}(\tau_1,\tau_2,\tau_3) \tag{2.6}$$

$$\begin{aligned} \boldsymbol{R}_{\pm}^{(3)} = \exp\left[-\frac{1}{2}\langle \delta\omega^2 \rangle \left(\tau_1^2 + \tau_3^2 \pm \tau_1\tau_2 C(\tau_2) \right) \right] \\ \times \exp\left[-\tau_2 / T_1 \right] \exp\left[i\langle \omega \rangle \left(\pm\tau_1 + \tau_3 \right) \right] \Theta(\tau_1)\Theta(\tau_3) \end{aligned} \tag{2.7}$$

The dephasing dynamics modeled by $C(\tau)$ are assumed to only impact the response during the waiting time period τ_2. In the model, the correlation function takes the form of a single exponential decay with correlation time constant τ_c. A phenomenological population relaxation term characterized by T_1 is added to represent vibrational relaxation. The $\Theta(\tau_n)$ in eq 2.7 are unit step functions that impose causality on the modeled response by setting the signal before light-matter interaction to zero.

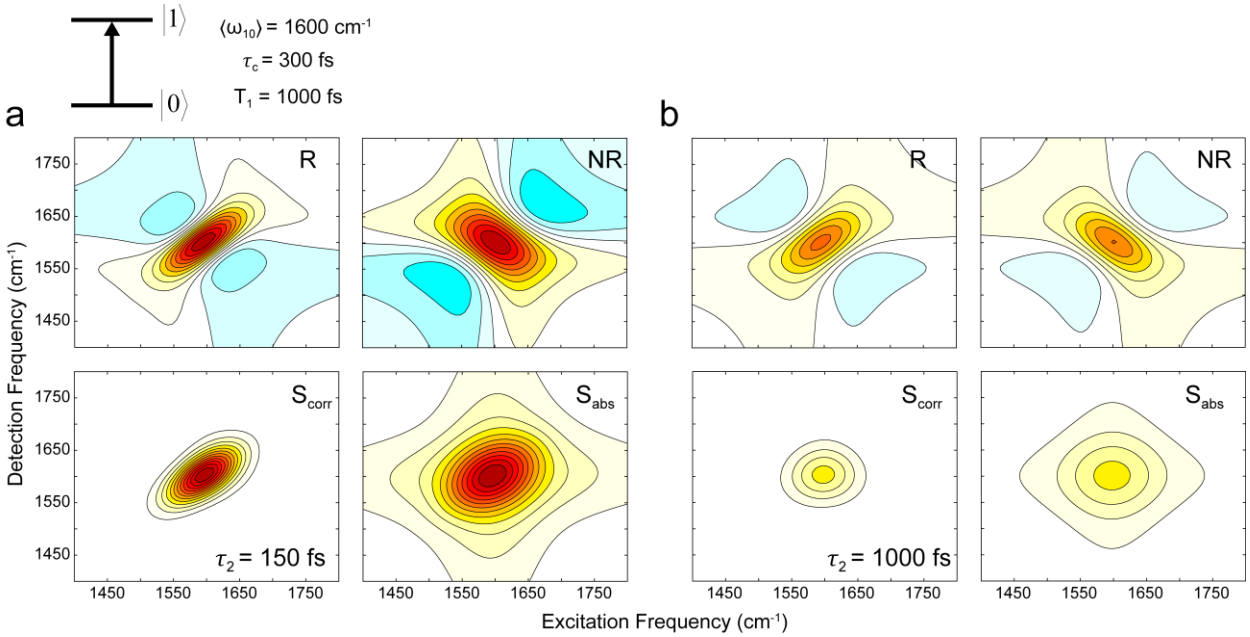

Figure 2.3: The rephasing (R), nonrephasing (NR), correlation (S_{corr}), and absolute value (S_{abs}) surfaces for a model two-level system. The mean frequency is set to 1600 cm^{-1}, the correlation time to 300 fs, and the vibrational lifetime to 1000 fs. (a) Surfaces at τ_2 = 150 fs illustrate elongation along the diagonal indicative of inhomogeneous broadening. (b) By τ_2 = 1000 fs $C(\tau_2)$ has decayed away, as evidenced by the symmetrizing of the line shape. The overall intensity has also decayed due to vibrational relaxation modeled by T_1.

The sign change in the oscillatory term in eq 2.7 illustrates the difference between the rephasing and nonrephasing processes. The opposite sign of τ_1 and τ_3 in the rephasing response gives rise to an echo assuming that the waiting time τ_2, is short enough that $C(\tau_2)$ has not decayed away. In contrast the nonrephasing response has no such conjugate oscillation during these two delay periods and instead the signal decays monotonically. A two-dimensional Fourier transform along both the τ_1 and τ_3 axes results in a more intuitive frequency-frequency representation of the 2D IR spectrum. Fig. 2.3 shows the rephasing and nonrephasing surfaces for a model two-level system with the central frequency set to 1600 cm^{-1}, the frequency correlation time τ_c = 300 fs, and the vibrational lifetime T_1 = 1000 fs. Both the R and NR surfaces clearly contain dispersive wings. However, adding the two surfaces together cancels the dispersive contributions and results in a

purely absorptive 2D IR correlation surface.[9] In practice this approach relies on proper weighting of the two surfaces and the correlation surface is sensitive to relative phase offsets that can mix the real and imaginary components. Practical considerations regarding the collection of the R and NR signals and mitigating experimental phase error are discussed in Chapter 3. Plotting the absolute value power spectrum is a simpler alternative, but at the cost of losing the phase sensitive information. The real absorptive correlation surface is most often reported since this representation of the third-order signal conveys the greatest possible information content.

Fig. 2.3a illustrates the case in which the waiting time is $\tau_2 = 150$ fs such that $\tau_2 < \tau_c$ while Fig. 2.3b shows the spectrum at $\tau_2 = 1000$ fs such that $\tau_2 > \tau_c$. These two scenarios illustrate the ability of 2D IR to separate different line broadening mechanisms. Although not modeled explicitly here, often these contributions to the line shape are discussed as two limiting cases.[10] In the homogenous limit the frequency fluctuations of the system are fast relative to the measurement period. As a result, correlation between the excitation and detection frequencies is quickly lost and the resulting line shape is symmetrically broad. In other words, an oscillator excited at a given frequency subsequently experiences rapid changes in its local environment that cause a shift in the frequency at which it is detected. Conversely, in the inhomogeneous limit frequency fluctuations are static on the timescale of the measurement. An oscillator excited at a given frequency will still be detected at that frequency sometime later. However, different oscillators in the ensemble can still exist in a distribution of distinct local environments that influences their vibrational frequencies. The resulting line shape is elongated along the diagonal where $\omega_1 = \omega_3$. In practice real systems exist between these limiting cases and there are both homogeneous and inhomogeneous contributions to the line shape.

For the case in which $\tau_2 < \tau_c$, the peak in the 2D IR correlation surface shows elongation along the diagonal, characteristic of inhomogeneous broadening. However, by the time $\tau_2 = 1000$ fs the line shape has symmetrized, indicating a loss of frequency correlation in the system. It is clear from Fig. 2.3 that the R surface largely dictates the diagonal linewidth while the NR surface largely dictates the anti-diagonal linewidth. In addition to the dephasing dynamics reflected in the evolution of the line shape as a function of waiting time, the overall intensity of the signal decays as τ_2 increases. This effect is the result of vibrational relaxation, which has been incorporated into the model through an exponential decay with characteristic lifetime T_1. In reality vibrational relaxation in the condensed phase is a complex process involving a multitude of states and relaxation pathways. As a result, it is included here phenomenologically.

2.2.3 Ground State Bleach and Excited State Absorption

All of the systems studied in this thesis overwhelmingly populate the vibrational ground state at equilibrium over the temperature range considered. Therefore the initial state of the system before interaction with a laser pulse is assumed to reflect the vibrational ground state. However, because the series of three light-matter interactions can sequentially excite vibrational energy levels, we must consider transitions between the ground state and first excited state ($0 \rightarrow 1$) as well as transitions between the first and second excited states ($1 \rightarrow 2$), assuming that the vibrational potential is weakly anharmonic and that the harmonic oscillator selection rules apply. A three-level system is the simplest example that allows for all of these possibilities. As depicted in the Feynman diagrams in Fig. 2.2, the first interaction establishes a coherence between the ground state and the first excited state that oscillates at the transition frequency ω_{10} during τ_1.

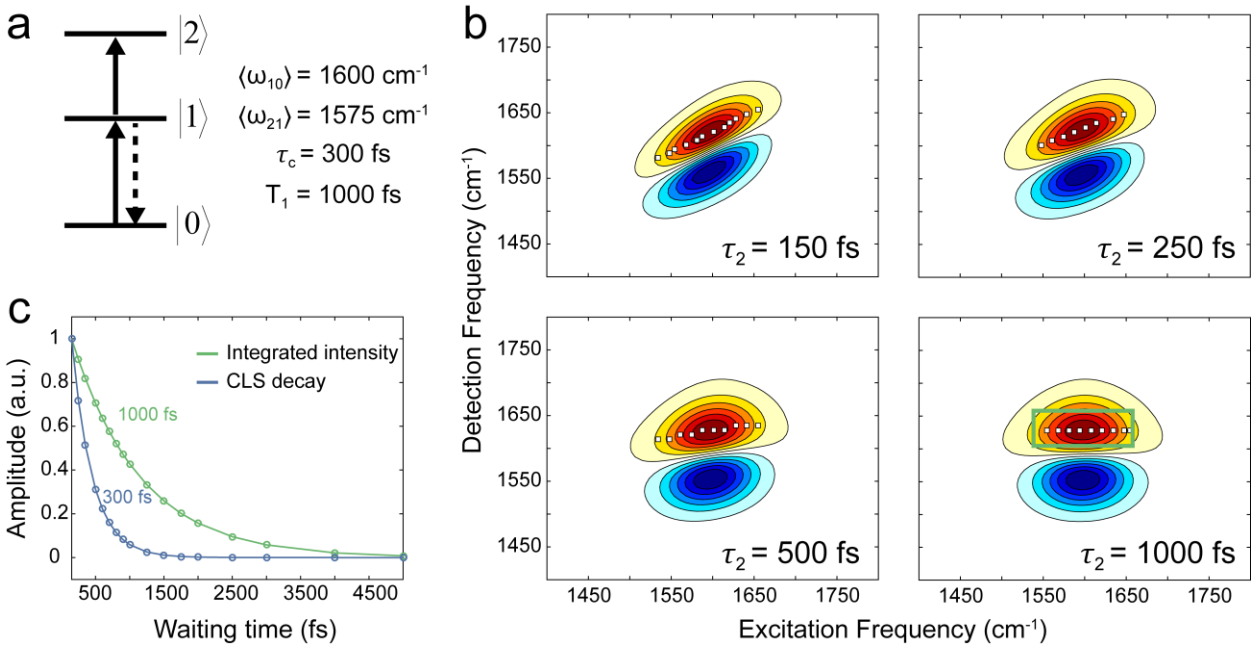

Figure 2.4: (a) Energy level diagram for a model three-level system. The 0→1 and 1→2 transitions are assumed to be perfectly correlated for simplicity. (b) A series of modeled 2D IR surfaces for the three-level system in panel a. The evolution of the line shape with increasing τ_2 reflects the loss of frequency correlation between excitation and detection. (c) Tracking the integrated intensity within the green box in panel b across τ_2 reports on the vibrational relaxation of the system (T_1) while monitoring the decay of the center-line slope reports on the decay of frequency correlation in the system (τ_c).

Following the second interaction there are three possible outcomes. A subset of oscillators populates the first excited state, another subset returns to the ground state population, and a third exists in a coherence between the ground state and second excited state. The third interaction establishes a final coherence that radiates signal oscillating at either ω_{10} or at ω_{21}. Signal pathways that oscillate at ω_{10} during τ_3 can result from the third interaction exciting a ground state population, referred to as a ground state bleach (GSB), or from stimulating emission (SE) from an excited state. Both of these types of pathways will be measured as less absorption (or more transmission) directly on the diagonal, as seen for the model three-level system in Fig. 2.4b. In the

color map used throughout this thesis, GSB/SE features are red and are considered positive in sign by convention.

Signal that oscillates at ω_{21} during τ_3 results from sequential excitation of oscillators that are already in a population in the first excited state or in a double-quantum coherence between the ground and second excited state. These signals therefore represent excited state absorption (ESA) and will be measured as less light at a detection frequency shifted off of the diagonal by the anharmonicity of the potential. Assuming positive anharmonicity, the spacing between levels will decrease as the vibrational energy levels increase, resulting in a shift to lower frequency of ω_{21} relative to ω_{10}. The peak shifted below the diagonal in Fig. 2.4b thus corresponds to the $1\rightarrow2$ transition frequency of the model three-level system. We consider the sign of an ESA peak to be negative by convention and plot these features in blue. As a result of these two types of signal pathways, peaks in a 2D IR spectrum appear as oppositely signed doublets with a GSB/SE feature on the diagonal and an ESA feature shifted along the detection axis by the anharmonicity.

The three-level system in Fig. 2.4 serves as an opportunity to discuss how frequency fluctuations and vibrational lifetime information contained within the third-order response manifest in this slightly more realistic 2D IR spectrum as well as how this information can be extracted in practice. The parameters of the model are identical to the two-level system simulated in the previous section, except that a $1\rightarrow2$ transition has been added and is shifted from ω_{10} by an anharmonicity of $\Delta = 25\ \text{cm}^{-1}$. For simplicity the ESA and GSB features are assumed to be perfectly correlated and the frequency correlation function $C(\tau_2)$ is modeled as a single exponential decay with a time constant $\tau_c = 300$ fs. A single exponential vibrational relaxation process is also added as before with a characteristic relaxation time $T_1 = 1000$ fs.

As discussed in the previous section, elongation along the diagonal indicates inhomogeneous broadening. Frequency memory is lost with increasing waiting time and the line shapes are completely symmetrized when the system is probed sufficiently beyond τ_c. In experimental 2D IR spectra, these line shape dynamics are a reflection of the transition energy fluctuations and spectral diffusion as dictated by changes in configuration or interaction with the surrounding environment. Measuring the decay of the center-line slope (CLS), defined by determining the maximum in the peak along ω_3 as a function of ω_1, is a common method to experimentally access these dynamics.[3,4] The white points in Fig. 2.4b show the center-line at a series of several waiting times. As τ_2 increases the slope of the CLS is observed to decay to zero. Under certain assumptions, the CLS can be related directly to a frequency correlation function that characterizes the amplitude and timescale of frequency fluctuations, which in turn can reveal information about the molecular dynamics of the system.[3,5] Tracking the decay of the CLS for the model three-level system results in the single-exponential decay with a time constant of τ_c = 300 fs plotted in blue in Fig. 2.4c. Unsurprisingly, the supplied frequency correlation function is perfectly recovered by monitoring the CLS decay for this simple model case.

Extracting the detailed microscopic origin of vibrational relaxation from the 2D IR spectrum is more difficult, but in practice phenomenological lifetimes are often measured by tracking the decay of an integrated peak volume as a function of waiting time. The green box in Fig. 2.4b defines the frequency range over which this integration is performed for the model data set. The decay trace plotted in green in Fig. 2.4c shows the loss of amplitude with increasing τ_2. Fitting this time trace to a single-exponential decay results in a time constant of T_1 = 1000 fs, consistent with the vibrational relaxation term that is included in the third-order response function above. In real systems the amplitude decay is often better described by a sum of discrete

exponentials and can demonstrate a more complicated frequency dependence than the uniform single-exponential relaxation modeled here.

2.2.4 Cross-Peaks in 2D IR Spectroscopy

Likely the most defining and recognizable feature of multidimensional spectroscopies is the ability to measure cross-peaks. The model systems discussed so far have only considered a single mean transition frequency. Since these systems have only a single vibrational mode there is no possibility of this mode coupling to other vibrations. The simplest system in which cross-peaks are possible must have at least two vibrational modes that are linked in some way. A six-level system in which there are two states, *a* and *b*, with mean 0→1 transition frequencies $\langle\omega_{a0}\rangle$ and $\langle\omega_{b0}\rangle$ can satisfy these requirements. Fig. 2.5a shows the energy level diagram for such a model system. The transition frequencies for the 1→2 transitions are given by $\langle\omega_{a0}\rangle - \Delta_a$ and $\langle\omega_{b0}\rangle - \Delta_b$ where Δ_n is the anharmonicity associated with state *n*. In addition to the ground, first, and second excited states for *a* and *b*, there is also a combination band with energy relative to the ground state $\langle\omega_{a0}\rangle + \langle\omega_{b0}\rangle - \Delta_{ab}$.

The purpose of this section is to build an intuition for the origin of cross-peaks that will serve as a useful foundation for discussing the remainder of the work in this thesis rather than to elaborate on the origin of every detail in the spectrum in Fig. 2.5b. In reality, the information content of a 2D IR spectrum even for a model six-level system is enormously rich. An excellent real-world example of this complexity is the 2D IR spectrum of dicarbonylacetylacetonato rhodium (I) (RDC) which exhibits a single pair of coupled symmetric and asymmetric carbonyl vibrations that are well described as a six-level system with an energy diagram resembling Fig. 2.5a. Despite this relatively simple energy picture, there are 66 Feynman diagrams that

contribute to the experimentally observed spectrum![2] Clearly the complexity and number of available interaction pathways will quickly balloon for systems with even more coupled states, as is the case for the nucleobases and oligonucleotides studied in subsequent chapters.

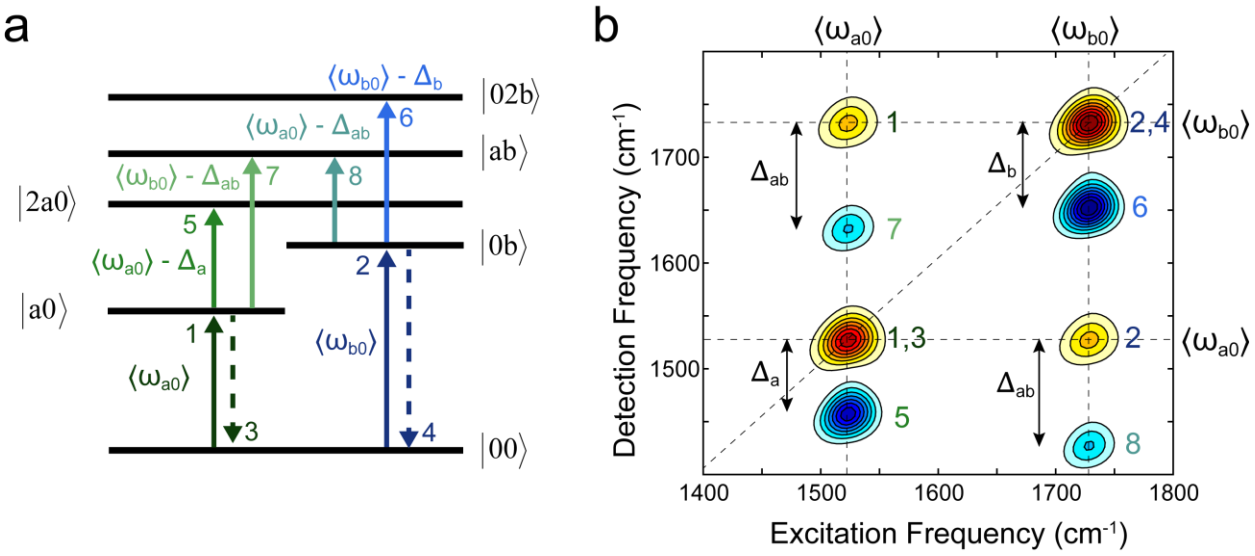

Figure 2.5: (a) The energy level diagram for a model six-level system with the transition frequencies indicated. (b) The simulated 2D IR spectrum corresponding to the six-level system in panel a. The peak numbering refers to the detection frequency at which each feature is observed after excitation at either ω_{a0} or ω_{b0}.

Inspection of the 2D IR spectrum in Fig. 2.5b reveals peaks along the diagonal that are similar in appearance to the oppositely signed doublet discussed for the three-level system above. There are two GSB/SE features centered at $\omega_1 = \omega_3 = \langle\omega_{a0}\rangle$ and at $\omega_1 = \omega_3 = \langle\omega_{b0}\rangle$. Red-shifted along the detection axis relative to each GSB/SE feature is an oppositely signed peak corresponding to the associated ESA for each mode. The magnitude of the frequency shift between the GSB/SE and the ESA is dictated by the anharmonicities Δ_a and Δ_b. The numbering of the peaks in Fig. 2.5b corresponds to the transition frequency at which each feature is detected following initial excitation at either ω_{a0} or ω_{b0}. The appearance of peaks along the diagonal of the 2D IR

spectrum are thus a reflection of probing the system at the 0→1 and 1→2 transitions for each of the vibrational modes in the system.

In addition to the doublets along the diagonal, there are off-diagonal cross-peaks. The appearance of cross-peaks in this case indicates that states *a* and *b* are coupled. In the broadest sense, what it means to say that two vibrational modes are coupled is that the presence of quanta of energy in one mode influences the energy of the other mode. The physical origin of coupling between vibrations can take different forms, including electrostatic and mechanical coupling, but in any case results in doublet features located off of the diagonal. As an example, the GSB of the cross-peak located above the diagonal in Fig. 2.5b is located at $\omega_1 = \langle\omega_{a0}\rangle$ and $\omega_3 = \langle\omega_{b0}\rangle$ and originates from pathways in which a coherence oscillates at the fundamental frequency of *a* during τ_1 and the fundamental frequency of *b* during τ_3. The presence of this cross-peak is a reflection of the depleted ground state population (or bleach) in state *a* caused by the initial excitation at ω_{a0} but observed by probing the system at the fundamental frequency of state *b*.

The associated cross-peak ESA is red-shifted by Δ_{ab} and originates from pathways that involve the combination band in which there are quanta of excitation in both states. In this example, $\Delta_{ab} > \Delta_b > \Delta_a$ and the cross-peak ESA is shifted lower in frequency than the diagonal ESA. A non-zero off-diagonal anharmonicity is required in order to observe a cross-peak between two modes, otherwise the ESA and GSB will overlay and cancel. In other words, a combination band of two modes must be shifted in energy relative to the sum of the fundamentals in order to observe a cross-peak, which amounts to a restatement of the definition of coupling stated above. More generally, some anharmonicity in the vibrational potential is necessary in order to observe a 2D IR signal, since for a perfectly harmonic system all of the GSB/SE and ESA features will perfectly cancel.

2.3 Retrieving Amplitude and Phase Information from Heterodyne-Detected Dispersed Vibrational Echo Spectroscopy

2.3.1 One-Dimensional Representations of the Third-Order Signal

The fullest characterization of the third-order vibrational response of a system can be achieved by measuring the full 2D IR spectrum. In practice, this requires both the excitation and detection axes to be fully resolved. In the most common experimental approach, a mixed time-frequency method of data collection is employed in which τ_1 is scanned in time and a Fourier transform of τ_3 is accomplished by dispersion off of a grating. Further details of data collection and processing are discussed in the next chapter. Scanning the τ_1 delay in order to resolve the excitation axis can be a time consuming process. In many instances, such as when the sample is unstable or when many different experimental conditions must be measured, a more rapid data acquisition is desirable. One-dimensional representations of the third-order signal, such as the dispersed vibrational echo (DVE), heterodyne-detected DVE (HDVE), and dispersed pump-probe (DPP) can be collected more rapidly than the full 2D IR spectrum because the excitation axis is not resolved. The tradeoff, of course, is that information previously spread across two frequency axes is now projected onto a single congested frequency axis. The relationship between the complex 2D IR signal and these 1D signals is given by:

$$\tilde{S}_{HDVE}(\omega_3) = \int_{-\infty}^{\infty} \tilde{S}_{2D}(\omega_1, \omega_3) d\omega_1 = \tilde{S}_{2D}(\tau_1 = 0, \omega_3) \tag{2.8}$$

$$S_{DVE}(\omega_3) = \left| \int_{-\infty}^{\infty} \tilde{S}_{2D}(\omega_1, \omega_3) d\omega_1 \right|^2 = \left| \tilde{S}_{HDVE}(\omega_3) \right|^2 \tag{2.9}$$

$$S_{DPP}(\omega_3) = \mathrm{Re}\left[\int_{-\infty}^{\infty} \tilde{S}_{2D}(\omega_1, \omega_3) d\omega_1 \right] = \mathrm{Re}\left[\tilde{S}_{HDVE}(\omega_3) \right] \tag{2.10}$$

Where a tilde indicates a complex valued quantity and τ_2 is assumed to be set at some fixed delay.

The complex HDVE signal is related to the projection of the complex 2D IR signal onto the detection axis by the projection-slice theorem.[11] The DVE spectrum reports only on the amplitude of the third-order signal and is thus related to the complex magnitude of the HDVE signal. Both HDVE and DPP signals are heterodyned measurements, meaning that the signal is interfered against a local oscillator (LO) field of known phase in order to extract signal phase information. Such an approach is necessary because so-called "square-law" detectors only measure the squared intensity of the signal field. The DPP signal is intrinsically heterodyned and therefore the delay between the LO and the signal is fixed at $\tau_{LO} = 0$. In contrast, an external LO with an adjustable τ_{LO} delay is used to measure the HDVE signal. The DPP spectrum is equivalent to the real part of the HDVE spectrum when $\tau_{LO} = 0$.

2.3.2 Fourier Transform Spectral Interferometry Method of Reconstructing the Complex Spectral Interferogram

Notice that given the complex HDVE spectrum, both the DVE and DPP spectrum can be reconstructed. It is therefore desirable to collect the HDVE spectrum in circumstances where a 1D representation of the third-order signal is required. However, signals measured in the laboratory are real-valued and the challenge of fully characterizing the complex third-order signal requires the clear resolution of both signal amplitude and phase. Fourier transform spectral interferometry (FTSI) is one approach for reconstructing the complex HDVE spectrum from the measured HDVE signal.[12] To demonstrate the FTSI method, Fig. 2.6a shows the real-valued HDVE spectrum for the model three-level system introduced in Fig. 2.4 with $\tau_2 = 150$ fs. Stepping the τ_{LO} delay between -10 and 10 fs in 5 fs steps demonstrates the phase contrast between the LO and signal field. The FTSI method relies on a Kramers-Kronig relation to relate the real and imaginary components of the spectral interferogram.

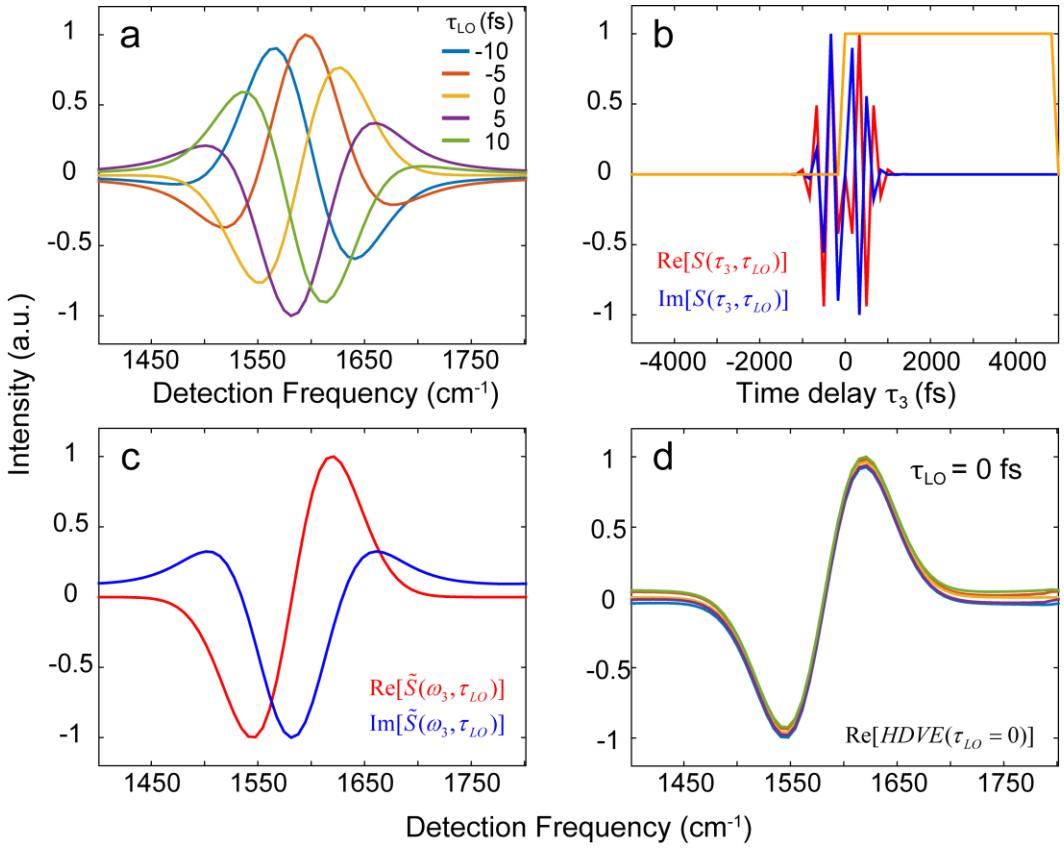

Figure 2.6: (a) The real-valued HDVE signal for the same model three-level system described in Fig. 2.4 for several τ_{LO} delays. The waiting time is fixed at 150 fs. (b) The real and imaginary components after inverse-Fourier transforming the $\tau_{LO} = 0$ spectrum in panel a. The unit step filter plotted in orange is applied to set the negative-time signal to zero. (c) Fourier transformation back to the frequency domain recovers both the real and imaginary components of the complex spectral interferogram. (d) The full set of recovered DPP spectra after applying the FTSI method to the modeled signals in panel a.

The measured real signal, which depends on ω_3 and τ_{LO}, is inverse-Fourier transformed to the time domain, resulting in a complex signal with the property $\tilde{S}(t) = \tilde{S}(-t)^*$. Both positive and negative time delay components are contained within $\tilde{S}(t)$ and both are shifted symmetrically by τ_{LO}. Fig. 2.6b shows the real and imaginary components of the time domain signal for the $\tau_{LO} = 0$ spectrum after inverse-Fourier transformation of the signal. Invoking a Kramers-Kronig assumption, windowing off the negative time components with a unit step function breaks the

symmetry of the time domain signal and Fourier transforming back to the frequency domain recovers the full complex spectrum. Fig. 2.6c shows the real and imaginary components after Fourier transforming the windowed signal in Fig. 2.6b. Applying the same approach for each of the τ_{LO} delays shown in Fig. 2.6a and shifting the supplied phase offset back to $\tau_{LO} = 0$ fs results in the set of real HDVE spectra plotted in Fig. 2.6d, which according to eq 2.10 are equivalent to the DPP spectrum. At this point the set of reconstructed spectra collected across the set of τ_{LO} delays are equivalent and can be averaged together.

2.4 References

1. Hamm, P.; Zanni, M., *Concepts and methods of 2D infrared spectroscopy*. Cambridge University Press: 2011.

2. Khalil, M.; Demirdöven, N.; Tokmakoff, A., Coherent 2D IR spectroscopy: molecular structure and dynamics in solution. *The Journal of Physical Chemistry A* **2003,** *107* (27), 5258-5279.

3. Kwak, K.; Park, S.; Finkelstein, I. J.; Fayer, M., Frequency-frequency correlation functions and apodization in two-dimensional infrared vibrational echo spectroscopy: A new approach. *The Journal of chemical physics* **2007,** *127* (12), 124503.

4. Fenn, E. E.; Fayer, M., Extracting 2D IR frequency-frequency correlation functions from two component systems. *The Journal of chemical physics* **2011,** *135* (7), 074502.

5. Roberts, S. T.; Loparo, J. J.; Tokmakoff, A., Characterization of spectral diffusion from two-dimensional line shapes. *The Journal of chemical physics* **2006,** *125* (8), 084502.

6. Mukamel, S., *Principles of nonlinear optical spectroscopy*. Oxford University Press on Demand: 1999.

7. Sung, J.; Silbey, R. J., Four wave mixing spectroscopy for a multilevel system. *The Journal of Chemical Physics* **2001,** *115* (20), 9266-9287.

8. Jonas, D. M., Two-dimensional femtosecond spectroscopy. *Annual review of physical chemistry* **2003,** *54* (1), 425-463.

9. Khalil, M.; Demirdöven, N.; Tokmakoff, A., Obtaining absorptive line shapes in two-dimensional infrared vibrational correlation spectra. *Physical review letters* **2003,** *90* (4), 047401.

10. Schmidt, J.; Sundlass, N.; Skinner, J., Line shapes and photon echoes within a generalized Kubo model. *Chemical physics letters* **2003,** *378* (5-6), 559-566.

11. Gallagher Faeder, S. M.; Jonas, D. M., Two-dimensional electronic correlation and relaxation spectra: Theory and model calculations. *The Journal of Physical Chemistry A* **1999,** *103* (49), 10489-10505.

12. Jones, K. C.; Ganim, Z.; Tokmakoff, A., Heterodyne-detected dispersed vibrational echo spectroscopy. *The Journal of Physical Chemistry A* **2009,** *113* (51), 14060-14066.

Chapter 3

Experimental Methods

3.1 Introduction

This chapter details the instrumentation and experimental methods employed throughout this thesis. Our focus will be on the homebuilt multidimensional infrared spectrometers used to measure the third order nonlinear signals discussed in the previous chapter. All of the nonlinear spectra reported here were measured on one of two spectrometers. Standard equilibrium measurements were acquired on the compact two dimensional infrared (2D IR) system. This instrument was designed with the purpose of being a robust and user-friendly spectrometer that allows for rapid, straightforward alignment and data acquisition. As a result, the design employs a simplified pump-probe beam geometry and records an interferogram between the pump pulses used to automatically correct for timing errors in the experimental delay between pulses. The first half of this chapter will present the compact 2D IR spectrometer in detail, beginning with the process of generating mid-IR light and the process of overlapping IR light with a visible tracer beam. The interferometer design and operation are discussed as well as practical aspects of the alignment. Finally the collection and processing of 2D IR spectra is described.

Despite the rich information content of equilibrium 2D IR measurements, this technique is limited in several critical ways when applied to problems in biomolecular association and folding. First, the upper bound on the time resolution of the measurement is limited by the vibrational lifetime of the molecules being probed, which for the carbonyl stretches and nucleobase ring

modes discussed here is around a picosecond. Second, the small transition dipoles of vibrational transitions as well as the poor sensitivity of mid-IR detection technology equates to limited sensitivity. As a result, 2D IR measurements are restricted to large ensembles of molecules. Therefore rare stochastic events at equilibrium, such as the spontaneous opening of base pairs in a DNA duplex or the encounter and hybridization of monomer strands to form a double helix, are washed out by the average behavior of the large ensemble and cannot be individually resolved. Introducing an external perturbation that induces a large-scale shift in population is one approach to overcome these limitations.

To study the dehybridization of DNA oligonucleotides, we use a transient temperature jump (T-jump) technique in which a second laser is electronically synchronized with a 2D IR spectrometer which is used to generate a T-jump so that the response of the ensemble following a nanosecond heating event can be tracked by a nonlinear infrared probe. This instrument differs from the compact 2D IR system in that the interferometer adopts a boxcar geometry and a heterodyned balanced detection scheme is employed. The laser that delivers the T-jump pulse operates at a repetition rate fifty times slower than the 2D IR spectrometer, and it is this reduced repetition rate plus the rate of thermal relaxation in the sample that define the upper limit on the time resolution of the transient experiment. In many respects, such as the generation of mid-IR light and the basic approach to acquiring and processing data, the boxcar and compact 2D IR spectrometers are similar, although the boxcar geometry, balanced detection, and the ability to collect transient data adds considerable complexity to the experimental design. The second half of this chapter describes the details of the T-jump spectrometer. The discussion focuses on those aspects in which the boxcar geometry and transient data collection differ from steady-state 2D IR in the pump-probe geometry, both in data collection and processing.

3.2 Compact 2D IR Spectrometer for Measurements at Equilibrium

3.2.1 Mid-IR Generation

Many of the molecular vibrations relevant to nucleic acid folding lie in the 1500-1800 cm^{-1} frequency range, which contains the in-plane ring vibrations and carbonyl stretches of DNA nucleobases that are sensitive to base stacking and hydrogen bonding. The measurement of a 2D IR spectrum requires short pulses of light that are resonant with the molecular transitions of interest. For the frequency range above, a well suited center wavelength is ~6.2 μm. In practice, mid-IR light is generated starting with the 795 nm output of a titanium sapphire (Ti:Sapph) regenerative amplifier (Solstice, Spectra Physics) since Ti:Sapph is an ideal gain medium for the generation of high intensity light pulses only tens of femtoseconds in duration. This near-IR light is then down converted through a series of nonlinear optical processes into the mid-IR frequency range of interest.

To achieve the high peak powers required for generation of sufficient mid-IR light for 2D IR measurements, intense 795 nm pulses are produced through a process of chirped pulse amplification in which the frequencies of a low power femtosecond seed pulse are stretched out in time, or "chirped", so as not to damage the Ti:Sapph gain medium within the optical cavity of the amplifier. The seed laser (Mai Tai, Spectra Physics) operates at an 80 MHz repetition rate with an average power of 1.0 W. An electronically controlled Pockels cell selects a single 795 nm seed pulse by rotating the polarization such that the pulse can enter the regenerative amplifier, which consists of a Ti:Sapph rod situated between two end mirrors that define an optical cavity. The Ti:Sapph gain medium is pumped by 18 W of 527 nm light generated by the second harmonic of an Nd:YLF laser operating at a 1 kHz repetition rate (Empower, Spectra Physics). The pump laser produces a population inversion in the Ti:Sapph crystal which amplifies the seed pulse by

stimulated emission as it passes through the gain medium. After a series of passes through the crystal, a second Pockels cell inside of the cavity switches the amplified pulse out of the amplifier. The pulse is then recompressed using a grating compressor. After stretching, amplification, and compression, the pulses are output by the regenerative amplifier at a 1 kHz repetition rate, have 3.5 mJ of energy, and are 90 fs in duration.

Although these 795 nm pulses are high in energy, short in time, and have ~160 cm^{-1} bandwidth, they are not resonant with the DNA vibrational transitions of interest. Performing an optical parametric amplification (OPA) step followed by difference frequency generation (DFG) produces the desired mid-IR frequencies. A diagram of the experimental layout of the compact 2D IR spectrometer is given in Fig. 3.1, illustrating the sequence of processes by which mid-IR pulses are generated as well as the interferometer, sample, and detection layout, which are discussed in detail in Section 3.2.3. Both the OPA (TOPAS Prime, Light Conversion) and DFG (NDFG, Light Conversion) are commercial instruments and only a general overview of their operation is covered here.

The TOPAS design consists of a pre-amplification stage that produces a small amount of light at a desired signal frequency which is then subsequently amplified in a second nonlinear crystal. The OPA process generates light at two frequencies, called the signal and idler, which sum to the frequency of light used to originally pump the process. For our purposes, the objective is to generate a signal and idler whose difference frequency will result in mid-IR light resonant with DNA nucleobase vibrations. Following regenerative amplification, 60% of the 795 nm light is routed into the OPA where the first optic encountered reflects ~90% of the light to be used in the later power amplification stage. The ~10% transmitted beam is further divided by a second beam splitter. The small fraction which is transmitted is focused into a sapphire plate, generating a white

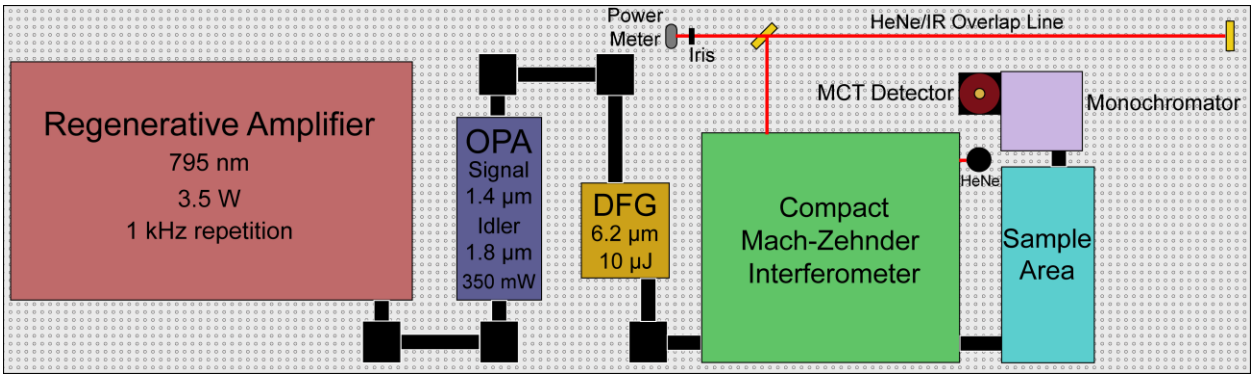

Figure 3.1: Diagram of the compact 2D IR setup. A regenerative amplifier outputs 795 nm pulses which are sent into an OPA to generate a signal and idler wavelength. To obtain pulses in the mid-IR, a DFG process generates light at the difference frequency of the signal and idler. A Mach-Zehnder interferometer produces the sequence of pulses and delays needed to perform a 2D IR measurement. The pulse sequence is focused into the sample in the sample area and the emitted 2D IR signal is dispersed in a monochromator and collected on an MCT pixel array.

light continuum. This white light is focused into a β-barium borate (BBO) crystal along with the remainder of the ~10% 795 nm light transmitted through the first beam splitter and reflected off of the second. Some portion of the white light continuum (1100-1600 nm) is selected and amplified in the BBO crystal by overlapping the desired portion of the continuum and the 795 nm pump spatially and temporally in the crystal. The pre-amplification stage adopts a non-collinear geometry so that everything besides the desired signal wavelength can be easily filtered out. The outcome of the pre-amplification stage is 3-5 μJ of signal, which is then overlapped collinearly in a second BBO crystal with the ~90% 795 nm light that was reflected at the first beam splitter. The power of the signal and idler produced in the amplification stage sums to ~350 mW.

To generate pulses centered around 6.2 μm, the difference frequency between a 1.4 μm signal and 1.8 μm idler is generated in a DFG process. The NDFG splits the collinear signal and idler generated by the OPA and recombines them non-collinearly in a $AgGaS_2$ (AGS) crystal. The non-collinear geometry facilitates the separation of the generated difference frequency from

residual signal and idler. Following this process, ~10 μJ of mid-IR light results. Sub 1% stability (2σ/mean) fluctuations in the average pulse energy are desirable and result in high quality 2D IR surfaces. At each stage of the OPA and DFG process, software controlled delays and crystal mounts can be adjusted to optimize temporal overlap and phase matching angle to maximize the output power and stability.

3.2.2 HeNe IR Overlap

Once mid-IR light of sufficient power and stability has been generated, it is possible to perform a 2D IR measurement. To accomplish this, a sequence of pulses separated by controllable time intervals must be created. The strategy employed for creating 2D IR pulse sequences in both spectrometers discussed here relies on an interferometer with programmable delay stages. However, one major challenge when working with mid-IR light is that the wavelengths of interest lie well outside the range of human vision. This limitation makes it difficult to achieve a quality interferometer alignment. To circumvent this problem, the IR light is overlapped with a visible HeNe tracer beam (633 nm) such that both propagate collinearly. The visible red light can then be used for alignment purposes. Before overlapping the tracer with the IR, the HeNe beam diameter is expanded using a refracting telescope to match the size of the IR beam. Overlap is achieved on a germanium window that transmits 6.2 μm, filters out residual signal and idler, and reflects 633 nm. Two mirrors before the Ge window are used to iteratively overlap the transmitted IR with the reflected tracer on an iris placed near the window and a second iris placed several meters away. A power meter is used to maximize the IR throughput at each point. The HeNe/IR overlap line is several meters (3.5 m for the compact, 9.5 m for the boxcar setup) longer than the total distance of the beam path through the spectrometer to ensure that the two beams remain collinear throughout

the experiment. Unfortunately materials well-suited for the mid-IR are often not the optimal materials for visible light and as a result additional tracer reflections that do not track the IR or a small degree of walk-off between the HeNe and IR over the beam path of the instrument are a reality. It is therefore important to carefully characterize which HeNe reflection traces the IR and to compensate for any accumulated mismatch between the tracer and IR.

3.2.3 2D IR in the Pump-Probe Geometry

The compact 2D IR spectrometer is designed to be a simple and robust instrument. With this purpose in mind, a Mach-Zehnder interferometer is used to split and delay mid-IR pulses and a simplified pump-probe beam geometry is employed. This design has several advantages in that the footprint is compact, the spectrometer is straightforward to align, and there are only two independent beams. Furthermore the signal is intrinsically heterodyned by the k_3 probe pulse, obviating the need for an external local oscillator to recover signal phase information.[1] As introduced in Chapter 2, we will refer to the individual signal generating fields by their wavevector, $\mathbf{k}_n$ where n specifies the pulse ordering. The rephasing (R) and nonrephasing (NR) signal pathways are collected simultaneously in the pump-probe geometry since the collinear pump pulses are indistinguishable, thus eliminating the need to independently collect, process, and sum R and NR surfaces to obtain the absorptive 2D IR spectrum.

Another advantage of the Mach-Zehnder design is that the second output from the interferometer can be recorded on a separate channel to obtain an interferogram between the collinear $\mathbf{k}_1$ and $\mathbf{k}_2$ pump pulses. This interference can be used to automatically correct phase ambiguity introduced by errors in the pulse timings due to imperfections in the delay stage accuracy. The details of this correction will be discussed in Section 3.2.6. There are several

disadvantages to the design of the compact 2D IR spectrometer as well, but in the interest of optimism these will be framed as advantages of the boxcar geometry in Section 3.3. Fig. 3.2 shows a schematic of the interferometer and sample area of the instrument.

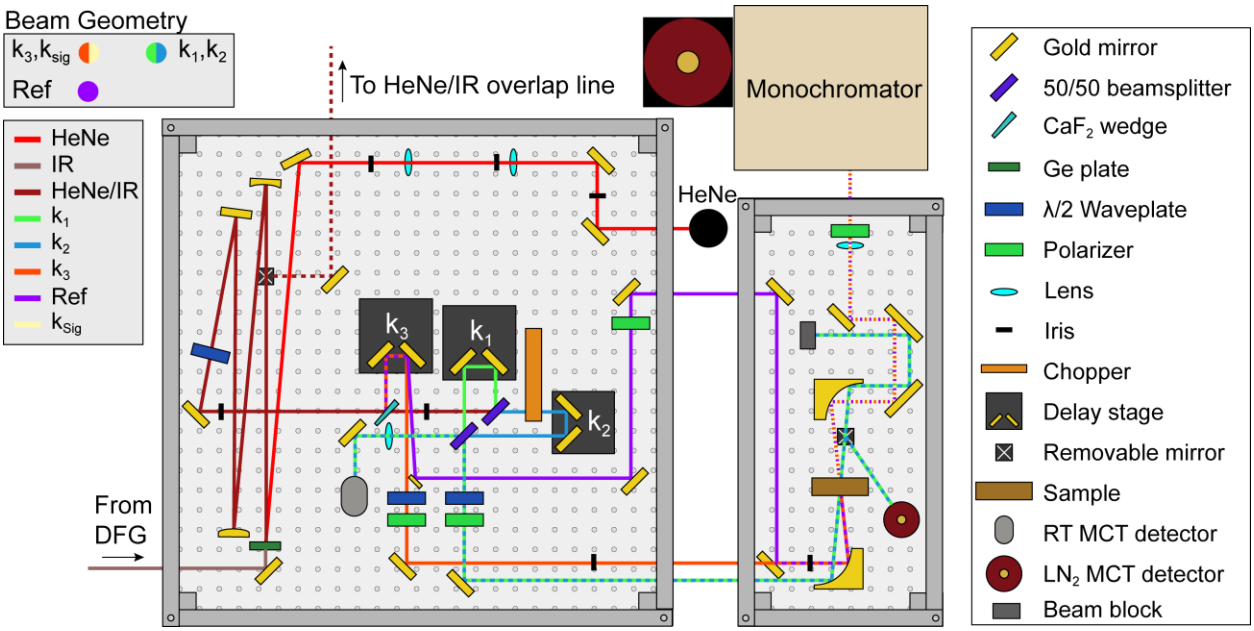

Figure 3.2: Schematic of the homebuilt interferometer and sample area of the compact 2D IR spectrometer. The optics and beams are color-coded according to the key in the figure.

Before aligning the interferometer, the HeNe/IR overlap is achieved as discussed in the previous section. The Ge window (5 mm AR coated, 7-12 μm, Thorlabs) that transmits mid-IR and reflects the tracer is near the input from the DFG. A reflective telescope with curved gold mirrors is used to collimate the IR and set the beam size to 6 mm. Before routing the beam into the interferometer the horizontally polarized IR light from the DFG is rotated 90° by a λ/2 waveplate (0.5 mm MgF_2, 6 μm, Karl Lambrecht Corporation) since the 50/50 beam splitters in the interferometer (1 mm CaF_2, 2-8 μm, Edmund Optics) are designed for vertically polarized

light. Two irises placed along the approach to the interferometer are used to align the beam. A CaF_2 wedge (1 mm, 0.5° pitch, Altechna) reflects ~5% of the IR off of the front face for use as the $\mathbf{k}_3$ probe pulse. A wedged optic is preferable to a flat optic in this case to eliminate the possibility of a ghost pulse collinear with $\mathbf{k}_3$ resulting from reflection off of the back face of the optic. Furthermore, since the wedge conveniently separates the two reflections, the back face reflection is later picked off with a D-shaped mirror and measured as a reference to correct for shot to shot fluctuations in laser power. The $\mathbf{k}_3$ beam is sent into a retroreflector mounted on a servo controlled delay stage (ANT95L, Aerotech) that has 1 nm (0.0067 fs) resolution, 75 nm (0.50 fs) repeatability, and 250 nm (1.66 fs) accuracy across 25 mm (167 ps) of travel. This delay line sets the waiting time τ_2 in the 2D IR experiment. The reflection off of the back face of the wedge is not sufficiently separated from $\mathbf{k}_3$ prior to the delay stage so the reference is picked off after the retroreflector. In principle this arrangement means that the alignment of the reference line depends on τ_2, but in practice this effect is negligible over the range of pump-probe delays and waiting times sampled in experiment.

The majority of light is transmitted through the CaF_2 wedge. Two 50/50 beam splitters in the Mach-Zehnder interferometer are used to split and then recombine the $\mathbf{k}_1$ and $\mathbf{k}_2$ beams after reflecting $\mathbf{k}_1$ off of a retroreflector mounted on a second servo controlled delay stage (ANT95L, Aerotech) and $\mathbf{k}_2$ off of a stationary retroreflector. The $\mathbf{k}_2$ beam is chopped at 500 Hz by a phase locked optical chopper (3501, Newport) triggered off of the 1 kHz signal generated by the timing delay electronics for the regenerative amplifier. Scanning the $\mathbf{k}_1$ delay sets the coherence time τ_1 in the 2D IR measurement and the Fourier transform of this time axis results in the ω_1 excitation frequency axis.

There are two outputs of the interferometer. In the first output, $\mathbf{k}_1$ is reflected off of the first beam splitter and transmitted through the second while the opposite is true for $\mathbf{k}_2$. Therefore each beam is transmitted once and reflected once, and the dispersion is matched for both pump beams. In the second output, $\mathbf{k}_1$ is reflected twice and $\mathbf{k}_2$ is transmitted twice, meaning that $\mathbf{k}_2$ passes through an additional 2 mm of CaF_2 relative to $\mathbf{k}_1$. The first output in which both beams travel through the same amount of material is therefore used for generating the 2D IR signal. The second output is focused onto a room temperature mercury cadmium telluride (MCT) detector (MCT5-020-H, Electro-Optical Systems Inc.) that measures an interferogram between $\mathbf{k}_1$ and $\mathbf{k}_2$ and is used for automatically correcting errors in the τ_1 time axis (phasing the spectrum) using the Mertz method described in Section 3.2.6.

Both the $\mathbf{k}_3$ probe and collinear $\mathbf{k}_1/\mathbf{k}_2$ pump beams are sent through waveplate (0.5 mm MgF_2, 6 μm, Karl Lambrecht Corporation) polarizer (ZnSe, 2-35 μm, Specac) (WP/P) combinations in order to independently control the polarization of the pumps and probe. The $\mathbf{k}_3$ polarization is stationary and set parallel to the vertical polarization of the interferometer. The waveplate and polarizer for the pumps are mounted in motorized precision rotation mounts (PRM1Z8, Thorlabs) allowing for the pump polarization to be switched programmatically between parallel and perpendicular. After the interferometer, the $\mathbf{k}_3$ and $\mathbf{k}_1/\mathbf{k}_2$ beams are arranged horizontally and offset by 3/4". Both are focused by a gold 90° off axis parabolic mirror (50338AU, Newport, 4" effective focal length) into the sample, which is typically held between two 1 mm thick CaF_2 windows separated by a 50 μm spacer. The windows are embedded in a brass block that can be temperature controlled using a programmable recirculating chiller (Ministat 125, Huber). The reference beam picked off by the D-mirror after the τ_2 delay stage is routed

independently to the gold parabolic mirror and is offset vertically below the $\mathbf{k}_3$ beam causing the reference to pass through the top of the sample.

The generated 2D IR signal is emitted in the same phase-matched direction as the $\mathbf{k}_3$ probe pulse. A second identical parabolic mirror collimates the beams and the signal. The residual pump light is dumped into a beam block. The signal, $\mathbf{k}_3$, and the vertically displaced reference beam are focused by a lens through the slit of a monochromator (Triax 190, Horiba) that disperses the light onto a 2 x 64 element MCT array connected to a boxcar integrator (IR-0144, Infrared Systems Development). A polarizer oriented parallel to the signal polarization is installed between the lens and slit to serve as a filter of residual pump light and scatter. The reference intensity is measured on the top stripe of the array. To ensure that the reference does not saturate the detector a polarizer installed in the reference line is used to attenuate as necessary. The 2D IR signal automatically interferes with and is heterodyned by the $\mathbf{k}_3$ probe and is detected on the lower stripe of the MCT pixel array. The waveplate in the $\mathbf{k}_3$ line is used to attenuate the k_3 probe so that it does not saturate the detector. Dispersion by the grating in the monochromator produces the detection frequency axis, ω_3.

3.2.4 Overlapping Pulses in Space and Time

Two important aspects of the spectrometer alignment that were not touched upon in the previous section are the spatial overlap of the signal generating fields at the focus of the parabolic mirror and determination of the relative timings between each of the pulses so that the coherence and waiting times in the 2D IR measurement are well defined. As long as the retroreflectors on the delay stages and the parabolic mirrors are well aligned, changing the pulse delay will not influence the spatial alignment of the beams. The $\mathbf{k}_3$ beam is used as a reference point for achieving overlap

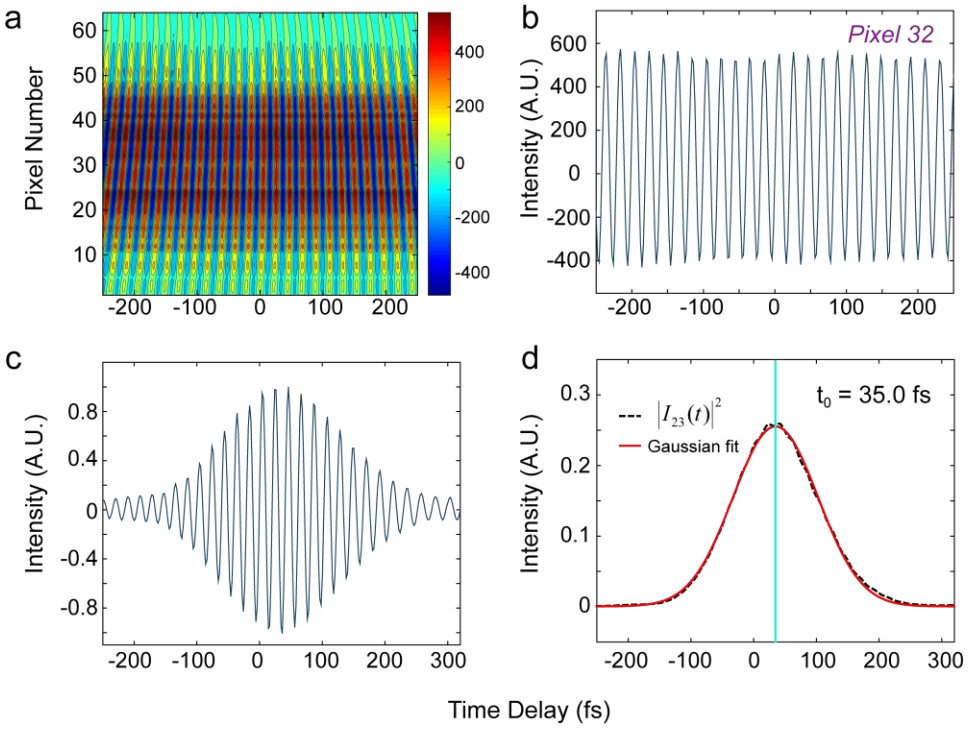

Figure 3.3: (a) Interference measured across the 64 pixels of the signal stripe. (b) The interference measured at pixel 32. (c) Integrating across all of the pixels produces an interferogram. (d) To determine t_0, a Gaussian is fit to the squared absolute value of the interferogram.

at the sample focus and two irises installed before the parabolic mirror are used to align $\mathbf{k}_3$ by adjusting the position of the retroreflector mounted on the τ_2 stage orthogonal to the direction of stage travel and a gold mirror positioned after the $\mathbf{k}_3$ WP/P combination. A 50 μm pinhole is positioned at the focus and adjusted to maximize the $\mathbf{k}_3$ light onto the lower pixel stripe of the MCT detector with all of the remaining beams blocked. Once set to the position of maximum $\mathbf{k}_3$ throughput, $\mathbf{k}_3$ is blocked and $\mathbf{k}_1$ is unblocked. Since $\mathbf{k}_1$ is not aligned into the monochromator, a curved gold focusing mirror installed temporarily behind the pinhole is used to detect the $\mathbf{k}_1$ throughput through the pinhole on a single channel MCT (MCT12-010, Electro-Optical Systems Inc.). A gold mirror positioned after the $\mathbf{k}_1/\mathbf{k}_2$ WP/P combination is used to maximize the $\mathbf{k}_1$ intensity through the pinhole onto the single channel detector. If $\mathbf{k}_1$ and $\mathbf{k}_2$ are recombined well on

the second beam splitter of the interferometer, then blocking $\mathbf{k}_1$ and unblocking $\mathbf{k}_2$ should result in equal pinhole throughput.

Once all three of the beams are spatially coincident at the focus of the parabolic they must also be overlapped in time. To determine the relative delay between two pulses, the modulation of the scattered intensity off of the pinhole is monitored as a function of the delay between the pulses. Scanning $\mathbf{k}_3$ against $\mathbf{k}_2$ establishes the τ_2 time axis. Since $\mathbf{k}_2$ is chopped, subtracting blocked shots from their adjacent unblocked shots removes the large background due to $\mathbf{k}_3$ intensity on the detector. A typical timing scan is stepped from -1000 fs to 1000 fs in 4 fs steps such that the full interferogram is collected and the oscillating signal is sufficiently sampled. Fig. 3.3a shows the scatter intensity measured across the 64 pixels of the signal stripe by scanning $\mathbf{k}_3$ against $\mathbf{k}_2$. Fig. 3.3b plots a slice at a single pixel near the center of the array. Integrating across the pixels results in the interferogram plotted in Fig. 3.3c. The scatter intensity is maximized when the two pulses are time coincident at the pinhole, corresponding to a delay of $\tau_2 = 0$ fs. In practice, the maximum is determined by fitting a Gaussian to the squared absolute value of the interferogram $I_{23}(t)$ after Fourier filtering, as shown in Fig. 3.3d. In this example, there is an offset of 35.0 fs relative to the previous time axis that must be corrected by redefining the stage position corresponding to time zero, t_0.

The relative delay between $\mathbf{k}_1$ and $\mathbf{k}_2$ sets the coherence time in the 2D IR measurement and it is this time axis that is scanned and Fourier transformed to produce the excitation axis, ω_1. As a result, the 2D IR spectrum is more sensitive to errors in τ_1 than in τ_2, since errors in the former will mix the absorptive and dispersive components in the Fourier transform.[2] Fortunately, the pump interferogram measured from the second output of the Mach-Zehnder interferometer provides a convenient and reliable method for determining t_0 for the τ_1 axis without the need to

install a pinhole at the sample focus and perform lengthy time scans. Typically scanning $\mathbf{k}_1$ against $\mathbf{k}_2$ from -150 to 150 fs in 2 fs steps produces an interferogram that is sufficient for setting t_0 by identifying the most intense central interference fringe. This timing scan can be easily rerun between 2D IR measurements to compensate for drift in t_0 due to position errors accumulated over the course of a data run. It is important to check and properly reset t_0 regularly since the phasing procedure applied in post processing is most reliable when timing errors are small and within a periodic cycle of the light, which for a wavelength of 6 μm corresponds to 20 fs.

3.2.5 Collecting and Processing 2D IR Spectra

Once the interferometer is aligned and the pulses are spatially overlapped and time coincident at the focus of the parabolic, a 2D IR spectrum can be measured. As an example, we will consider the 2D IR surface of the DNA oligonucleotide 5'-GATATATATC-3' acquired at a temperature of 80 °C. Appendix 3A describes how DNA samples are prepared for IR spectroscopy. The waiting time τ_2 is set to 150 fs before the 2D IR scan by sending the $\mathbf{k}_3$ delay stage to the appropriate offset. The coherence time τ_1 is stepped from -160 to 2500 fs in 4 fs steps. The scan is started at negative times ($\mathbf{k}_2$ arrives before $\mathbf{k}_1$) because the interferogram between the pump pulses recorded from the second output of the interferometer is used to phase the spectrum. A delay of -160 fs is around the minimum number of negative time points that fully captures the full pump interferogram. The τ_1 delay is stepped out to 2500 fs since this corresponds to ~2.5 times the vibrational lifetime of the in-plane nucleobase vibrations of DNA. The extent to which the τ_1 axis is scanned sets the resolution of the ω_1 axis, which for a delay of 2500 fs corresponds to 13.3 cm^{-1}. The resolution of the ω_3 axis is determined by the monochromator slit width as well as dispersion off the grating in the monochromator, and is 4 cm^{-1} for this example. The step size of 4 fs is

selected such that the ~6 μm signal oscillating with a ~20 fs optical period is sufficiently sampled above the Nyquist frequency.

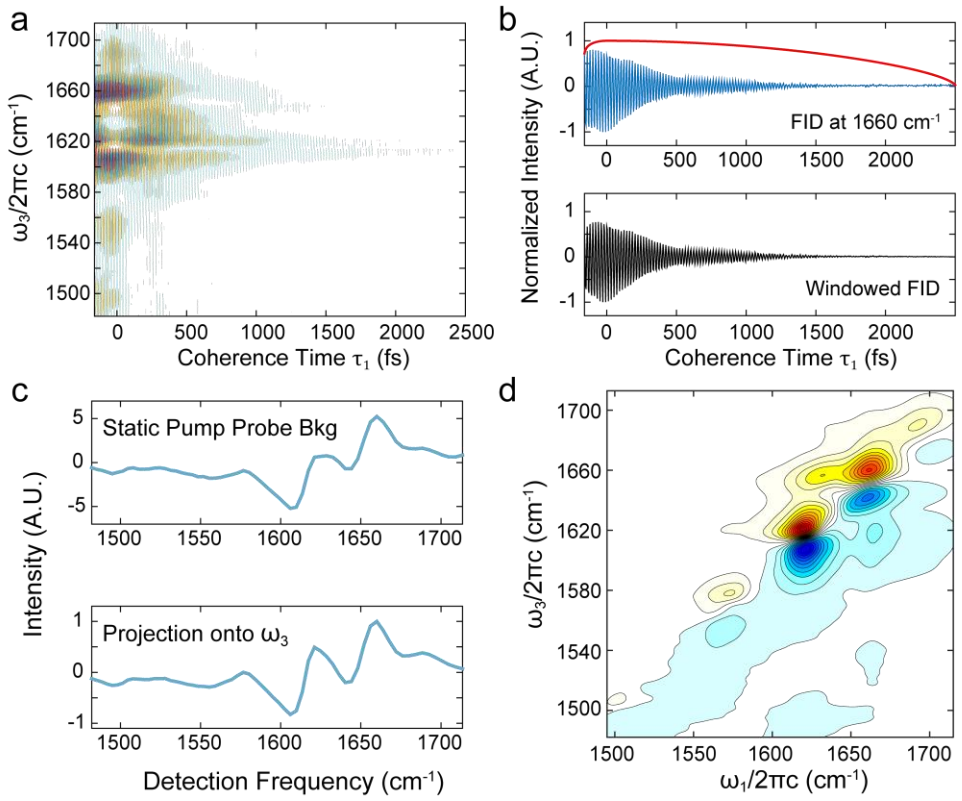

Figure 3.4: (a) Time-frequency representation of 2D IR data that reflects the raw signals recorded in the laboratory. The sample is 2 mM 5'-GATATATATC-3' at 80 °C. (b) Illustrative single pixel FID trace at 1660 cm^{-1}. Applying a Hann window with an exponent of 0.3 plotted in red results in the windowed FID plotted in black. (c) The static pump-probe background generated by k_2 and k_3 closely resembles the projection of the 2D IR surface onto the detection axis. (d) Frequency-frequency representation of the 2D IR data from panel a following windowing, zero padding, and phase correction.

In practice, data are collected in a mixed time-frequency representation as depicted in Fig. 3.4a. The detection axis is measured directly in the frequency domain by dispersion off of the monochromator grating. An oscillating free induction decay (FID) is recorded on each pixel of the

signal stripe of the array. Performing a fast Fourier transform (FFT) along each time trace results in a frequency-frequency representation that is far more intuitive to interpret than the mixed time-frequency representation of Fig. 3.4a. Several processing steps prepare the time domain data for FFT, but first some initial rearrangement of the raw intensities measured on the array is performed on the fly during data collection. Specifically, the signal is recorded in a referenced change in absorption (ΔOD) mode.

$$S_{2D}(\omega_3, \tau_2, \tau_1) = -\log\left(\frac{I_{open}}{R_{open}} \cdot \frac{R_{closed}}{I_{closed}}\right) \tag{3.1}$$

Here I_{open} refers to the signal intensity measured on the signal stripe of the array while R_{open} refers to the reference intensity measured on the reference stripe of the array when the chopper is not blocking $\mathbf{k}_2$. The I_{closed} and R_{closed} are the corresponding quantities measured when $\mathbf{k}_2$ is blocked by the chopper. Since the chopper is operating at 500 Hz, every other shot alternates between open and closed. Therefore the ΔOD quantity in eq 3.1 is calculated every two laser shots. As can be seen in Fig. 3.2, the reference beam is not chopped. Therefore R_{open} and R_{closed} only differ due to shot to shot fluctuations in laser power and their quotient should average to one. Dividing by the reference intensity on the fly therefore suppresses multiplicative noise due to fluctuations in laser power.

Chopping $\mathbf{k}_2$ and subtracting closed shots from open shots eliminates signals which do not have a contribution from $\mathbf{k}_2$, most notably the background intensity due to the heterodyning probe pulse. The stationary beam is chopped so that the $\mathbf{k}_1\mathbf{k}_3$ pump-probe signal that depends on τ_1 is removed on the fly. However, this still leaves a pump-probe signal generated by $\mathbf{k}_2\mathbf{k}_3$ on top of the 2D IR signal. Fortunately, this signal is static and amounts to an offset that can be subtracted off in post processing. Since the 2D IR signal oscillates about zero, taking the mean across τ_1 provides

a convenient method for determining the offset at each pixel. The static pump-probe spectrum determined in this way is plotted in Fig. 3.4c. Subtracting this spectrum off across the 64 pixels of the signal stripe effectively removes the static pump-probe background.

To avoid artifacts in the FFT, it is important that the signal decays to zero. In reality, electrical noise on the detector or laser intensity fluctuations can result in deviations from zero even long after the FID has dephased. Applying a window function ensures that the trace goes to zero, but at the possible risk of distorting the time domain signal. Throughout this thesis, an asymmetric Hann function raised to an exponent is used as the window function because this window provides a desirable balance between preserving resolution and reducing introduced artifacts.[3] Half of the symmetric Hann window, eq 3.2 is applied to the positive time signal.

$$w(t) = \left[\frac{1}{2}\left(1-\cos\left(\frac{2\pi t}{t_{\max}}\right)\right)\right]^{\alpha} \tag{3.2}$$

The time points are discrete and t_{max} represents the largest time point sampled, in this example 2500 fs. The exponent α determines how rapidly the window decays to zero. Increasing α windows out more long-time noise resulting in a smoother 2D IR surface, but it can also broaden line shapes. For data of sufficiently high quality, such as the example in Fig. 3.4, a small value of α less than one is adequate and does not significantly influence the line shapes. Fig. 3.4b shows a Hann window with α = 0.3 plotted over an illustrative FID trace at 1660 cm^{-1}. The windowed FID is plotted in the bottom panel of Fig. 3.4b.

It is convenient to interpolate the 2D IR surface to a finer frequency spacing along both frequency dimensions. Zero padding in the time domain is an efficient way to accomplish this interpolation along the ω_1 axis. The windowed time domain signal at each pixel is padded by appending 2 x 10^{14} zeros, which results in ~0.5 cm^{-1} spacing along the excitation axis. A uniformly

spaced ω_1 frequency axis can then be generated based off of the total number of time points and the 4 fs time step. After zero padding, a FFT is performed along each pixel of the array resulting in a frequency-frequency domain representation of the data. However, one crucial processing step remains. The 2D IR spectrum must be phase corrected to account for errors in the τ_1 time axis. This process is the topic of the next section.

The detection axis is interpolated to the same frequency spacing as the excitation axis directly in the frequency domain. Either a linear or cubic spline interpolant can be used. In the event that linewidths are comfortably wider than the pixel spacing in wavenumbers, as is typically the case for nucleic acid absorptions, a cubic spline is desirable over linear interpolation since it smooths the overall appearance of the spectrum.

3.2.6 Phasing 2D IR Spectra using the Mertz Correction

Perhaps the most pervasive challenge to reliably measuring and interpreting a purely absorptive 2D IR correlation surface regardless of considerations such as interferometer design or beam geometry is the problem of properly phasing the spectrum. The central issue is that the delay stages used to establish the coherence time axis in the 2D IR measurement are precise but not sufficiently accurate, as noted above. Position errors of even a few femtoseconds can lead to an ill-phased spectrum, that is, a mixing of the absorptive and dispersive features.[2] The result is a twisting and distortion of the line shapes and intensities. One can imagine that drawing conclusions from an ill-phased spectrum could likely lead to a significant misinterpretation of the system under study.

Fortunately the issue of phasing in time domain spectroscopy is well known and there are several established approaches in FTIR and 2D IR spectroscopy.[4-6] A considerable advantage of

2D IR in the pump-probe geometry is the ability to phase the 2D IR spectrum in much the same way as an FTIR spectrum. Specifically, the method used here is called the Mertz correction.[7] The Mertz method of phasing is robust, reliable, and easily automated but requires the simultaneous measurement of an independent pump interferogram during the collection of the 2D IR data. The second output of the Mach-Zehnder interferometer facilitates the collection of this necessary additional information. Fourier transformation of the pump interferogram yields the pump spectrum as well as a phase spectrum that can be used to correct for timing errors in the 2D IR measurement since the time point at which maximum constructive interference of the pump pulses occurs corresponds to the time point at which $\tau_1 = 0$ along the coherence time axis. We are explicitly assuming in this case that the phase error is dominated by a fixed offset in the τ_1 axis due to stage position error. In such case the coherence time axis is erroneously shifted by some amount $\tau_1 + \tau_{err}$ and the resulting phase error takes the linear form $\omega_1\tau_{err}$.

The pump interferogram collected during a 2D IR scan is sampled asymmetrically in time since the coherence time ranges from -160 to 2500 fs and the pump interferogram is centered around t_0. Therefore the first step when applying the Mertz correction is to truncate the pump interferogram such that it is symmetric about the maximum, as seen in Fig. 3.5a. Next the data are zero padded to five times the original length of the truncated symmetric interferogram. A Hann window is applied and the data are rotated to begin at the t_0 defined by the experimental (uncorrected) stage axis. The remaining task is to determine the frequency dependent phase offset $\omega_1\tau_{err}$ between the stage axis t_0 relative to the true t_0 defined by the maximum overlap of the pump pulses in time. The interferogram between $\mathbf{k}_1$ and $\mathbf{k}_2$ is Fourier transformed as in eq 3.3 and the phase spectrum is computed according to eq 3.4.

$$\tilde{I}_{12}(\omega_1) = \int_{-\infty}^{\infty} I_{12}(t) e^{i\omega t} dt \tag{3.3}$$

$$\phi(\omega_1) = \arctan\left(\frac{\mathrm{Im}(\tilde{I}_{12}(\omega_1))}{\mathrm{Re}(\tilde{I}_{12}(\omega_1))}\right) \tag{3.4}$$

The properly phased pump spectrum can then be obtained by applying the phase factor computed in eq 3.4.

$$S_{pump}(\omega_1) = \tilde{I}_{12}(\omega_1)e^{-i\phi(\omega_1)} \tag{3.5}$$

The normalized pump spectrum as well as the phase spectrum determined from the pump interferogram measured during the 2D IR scan from Fig. 3.4 are plotted in Fig. 3.5b. The phase correction is applied to the 2D IR spectrum in the frequency domain in exact analogy to eq 3.5, except that it is first necessary to interpolate the phase spectrum to match the frequency spacing along the ω_1 axis of the 2D IR surface.

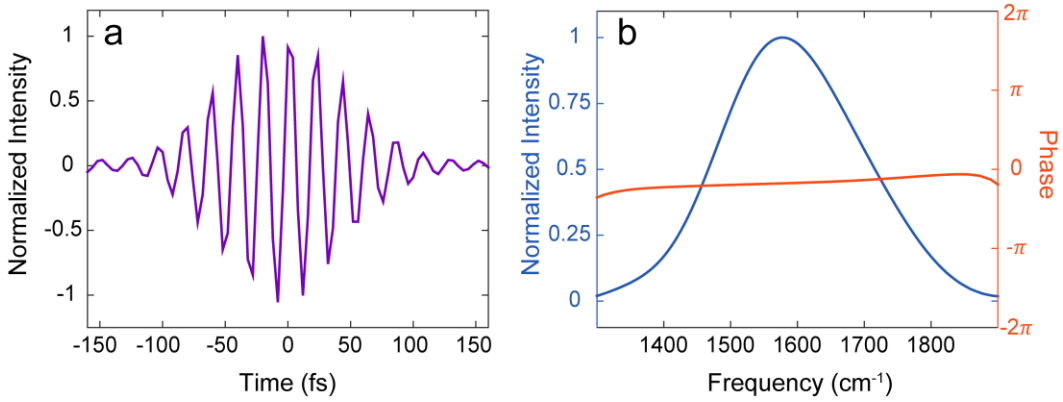

Figure 3.5: (a) The pump interferogram recorded in the time domain from the second output of the interferometer on a single channel MCT detector. The data have been truncated in time such that the trace is symmetric about t_0. (b) The phase corrected pump spectrum plotted in blue and the phase spectrum plotted in orange.

The pump interferogram is symmetric in time and therefore the positive and negative time data contain the same information. The 2D IR signal is inherently asymmetric in time, and the negative time data contain unwanted signal pathways in which both τ_1 and τ_2 are changing as a

function of the position of the $\mathbf{k}_1$ stage. Therefore after applying the phase correction factor to the 2D IR signal in the frequency domain, the data are inverse Fourier transformed back into the time domain and the negative time data are discarded. It is important to discard negative time data only after phase correction and the determination of the corrected t_0, otherwise one risks prematurely discarding pertinent signal information.

A final processing step involves correcting the 2D IR intensities for the finite bandwidth of the pumps and probe. Conveniently, the Mertz method measures the pump spectrum and dividing the 2D IR surface along the excitation axis by this spectrum corrects for intensity discrepancies due to the bandwidth limits of the pump. Since $\mathbf{k}_3$ is directly aligned into the monochromator, the probe spectrum can be measured directly in the frequency domain on the MCT pixel array and a similar correction can be applied along the detection axis.

3.3 Boxcar 2D IR Spectrometer for Transient Measurements

3.3.1 The Boxcar Temperature Jump Spectrometer

Although the compact 2D IR spectrometer is ideal for equilibrium measurements, many dynamic and kinetic processes in biology span decades in time past the vibrational lifetimes of the biomolecules involved. Introducing an electronically synchronized T-jump perturbation and tracking the response of the ensemble using a nonlinear IR probe extends the accessible window in time from picoseconds to milliseconds and almost everything in between. That being said, these measurements are not a trivial extension of the standard 2D IR instrument and experiment discussed in the previous section. The experimental design, data acquisition, and interpretation of results all increase in complexity considerably when a nanosecond temperature jump is introduced into the mix.

There are several advantages to performing transient experiments in the boxcar geometry as opposed to the simpler pump-probe geometry utilized in the compact 2D IR spectrometer. T-jump measurements are double difference measurements. The first difference is due to chopping and subtracting blocked from unblocked shots, as discussed above, while the second difference involves the subtraction of the equilibrium temperature spectrum from the spectrum sampled at some time after the rapid heating event to produce a time-resolved thermal difference spectrum. As a result, transient signals are routinely small in comparison to equilibrium measurements and enhanced sensitivity is required. For a given intensity of mid-IR light, the boxcar geometry offers the greatest possible nonlinear signal generation because the IR intensity can be equally distributed between all three of the signal generating fields with minimal waste. Recall that $\mathbf{k}_3$ in the pump-probe geometry was obtained from a small reflection off of a CaF_2 wedge and that this probe pulse had to be further attenuated so as not to saturate the detector. In the boxcar geometry there is no such restriction because the signal field is emitted in a unique direction and is externally heterodyned by a weak local oscillator (LO) field whose intensity can be tuned independently. Furthermore, in the Mach-Zehnder interferometer design half of the intensity of the probe pulses is lost to the second output of the interferometer. This light can be put to good use phasing the spectrum, as described in the previous section, but in reality 50% of the pump intensity is far more than is necessary to obtain a quality interferogram and much of this intensity would be put to better use generating 2D IR signal. The boxcar interferometer allows for the maximum and most evenly distributed intensity in each of the signal generating fields.

In contrast to the pump-probe geometry, the boxcar geometry is often described as background free since the third order signal of interest is emitted in a unique direction.[8] The ability to better control which signal pathways are measured is advantageous because it clarifies the

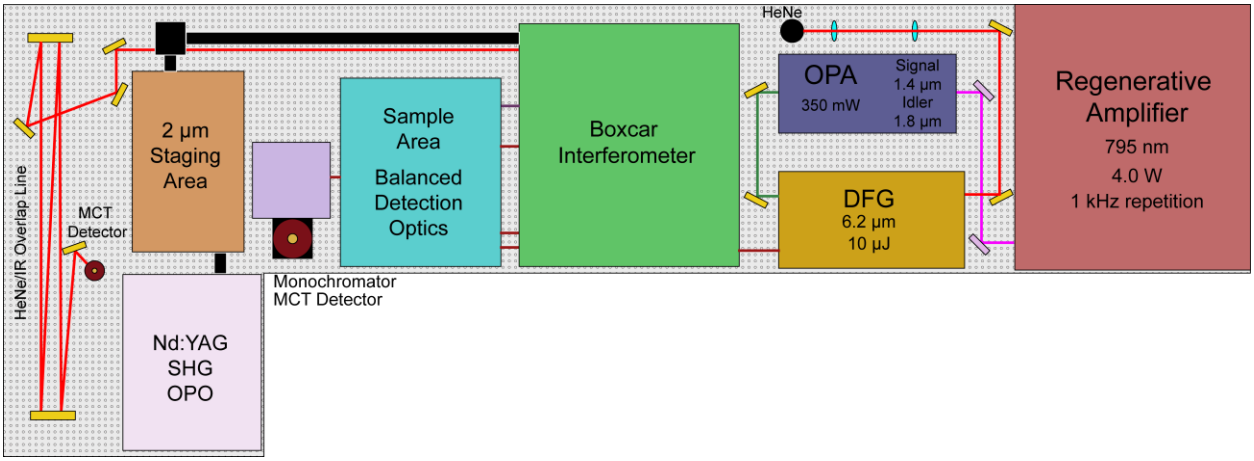

Figure 3.6: Diagram of the boxcar T-jump setup. A regenerative amplifier outputs 795 nm pulses which are sent into an OPA to generate a signal and idler wavelength. To obtain pulses in the mid-IR, a DFG process generates light at the difference frequency of the signal and idler. A boxcar interferometer produces the sequence of pulses and delays needed to perform a nonlinear measurement. The pulse sequence is focused into the sample in the sample area and the emitted signal is split and combined with an external LO on a beam splitter for balanced detection. The heterodyned transmitted and reflected signals are dispersed by a monochromator and each detected on a separate stripe of a dual stripe MCT array. An electronically synchronized Nd:YAG produces 1064 nm light which is then frequency doubled and used to pump an OPO to generate 2 μm light which is then routed to and focused into the sample to generate a nanosecond T-jump.

information content of complicated transient difference spectra. Employing a balanced detection scheme, described below, offers further advantages in a T-jump experiment. As we will see, balanced detection suppresses noise and removes unwanted background signals. Having an external LO that can be controlled without impacting signal generation and that is recorded on a shot by shot basis on the fly is also critical, since changes in LO transmission are used to both align the T-jump beam and to determine the magnitude of the temperature jump in post processing.

Generation of mid-IR light in the boxcar spectrometer is similar to the compact spectrometer. A few differences to note include the model of the regenerative amplifier (Libra, Coherent) and the OPA (TOPAS C, Light Conversion). The DFG is homebuilt and employs a collinear signal/idler geometry rather than the non-collinear geometry used in the commercial

NDFG. These differences are minor and ultimately the quantity of mid-IR light generated and sent into the interferometer is similar in both setups.

The details of the temperature jump spectrometer and experiment have been presented in detail previously.[9,10] The discussion here is therefore structured with an emphasis on updates to the experiment and on the practical aspects of acquiring and processing transient data. Fig. 3.6 shows the layout of the boxcar spectrometer. The arrangement of the mid-IR generating equipment is similar to the compact design in Fig. 3.1. The footprint of the boxcar interferometer and sample area are considerably larger in order to accommodate 2" optics, additional delay lines, and the optics required for balanced detection of the signal. The HeNe/IR overlap line is also several meters longer to account for the increased optical path through the boxcar setup.

The most obvious difference between the boxcar layout in Fig. 3.6 compared to the compact instrument in Fig. 3.1 is the addition of a second laser on the table. This Nd:YAG is used to generate the pulses that deliver the optical temperature jump. The 80 MHz repetition rate of the oscillator in the regenerative amplifier serves as the master clock for coordinating the arrival of the T-jump pulse relative to the probing mid-IR pulse sequence. Fig. 3.7 shows a diagram of the electronics and delay scheme that accomplishes the synchronization between the regenerative amplifier and the Nd:YAG. The electronics for controlling the regenerative amplifier (Synchronization and Delay Generator (SDG) Elite, Coherent) set the 1 kHz repetition rate (F1) by frequency dividing the 80 MHz output (RF) from the oscillator. The SDG is internally triggered to establish the (electronic) t_0 for coordinating the T-jump experiment.

The pump laser and Pockels cells (PC1 and PC2) in the regenerative amplifier are triggered and delayed off of the SDG at frequency F1. This arrangement differs from the triggering scheme used in the compact setup and from the scheme that is assumed in the Libra user manual. A more

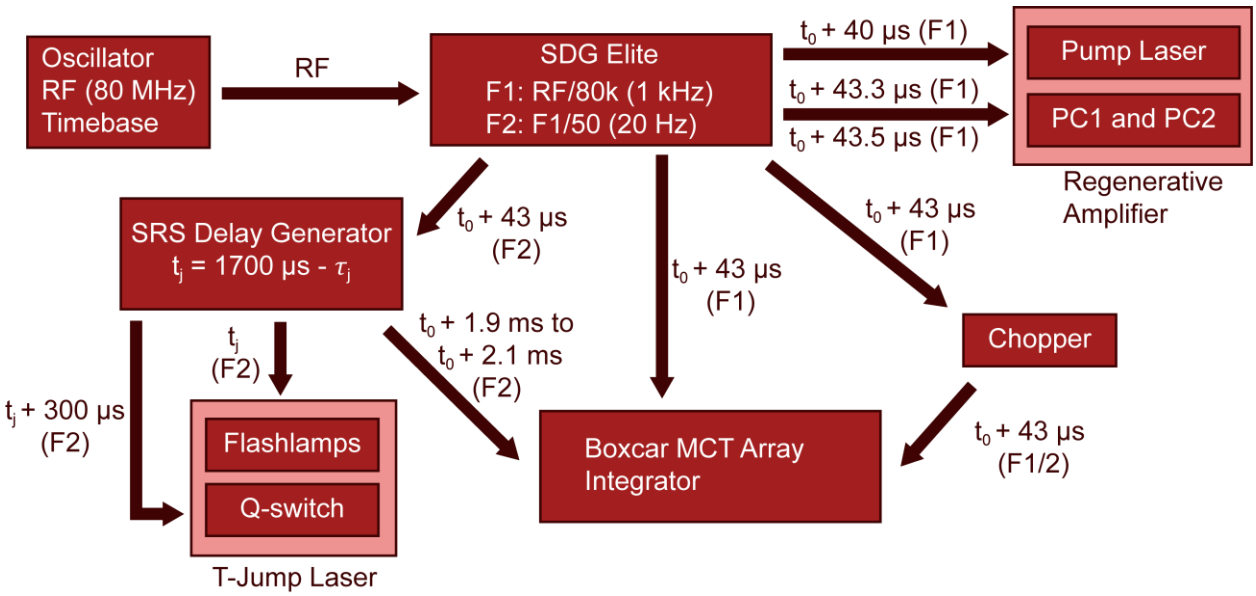

Figure 3.7: Schematic of the timing electronics and delays that run the T-jump experiment. The oscillator from the regenerative amplifier serves as the master clock (RF). The SDG Elite is internally triggered and frequency divides the RF to establish the 1 kHz (F1) and 20 Hz (F2) frequencies. The SDG coordinates the operation of the regenerative amplifier at a 1 kHz repetition rate, triggering the pump laser and the Pockels cells (PC1 and PC2). The chopper divides F1 in half and chops at 500 Hz. The chopper signal is passed to an external channel of the integrator for the MCT array so that blocked and unblocked shots can be sorted. The array acquisition itself is synched to F1. The delay passed to the flashlamps and the Q-switch in the Nd:YAG is generated by an SRS delay generator triggered at F2. The electronically controlled delay τ_j sets the interval between the arrival of the T-jump pulse and the first 6 µm pulse sequence.

typical arrangement involves internally triggering the pump laser Q-switch and synching the delay generator to the 1 kHz output. The chopper and integrator for the MCT array are also synchronized off of the SDG at F1. The chopper divides the 1 kHz input in half and chops at 500 Hz such that every other shot is blocked. This 500 Hz signal from the chopper is passed to an external channel of the array integrator so that blocked and unblocked shots can be sorted on the fly. The SDG Elite is capable of performing a single integer frequency division of F1. This divisor is set to 50 so that the second frequency output (F2) is 20 Hz, setting the repetition rate of the Nd:YAG laser.

3.3.2 Generation of 2 μm Pulses

A frequency divided output of the SDG is used to trigger an SRS delay generator (DG535, Stanford Research) at F2. This delay generator electronically sets the delay τ_j between the arrival of the T-jump pulse and the first 6 μm pulse sequence. The T-jump laser consists of a flashlamp pumped Q-switched Nd:YAG (YG981C, Quantel) that outputs 6-8 ns pulses at a 20 Hz repetition rate and with 1050 mJ of energy. The 1064 nm output is frequency doubled in a second harmonic generation (SHG) process within a temperature controlled deuterated potassium dihydrogen phosphate (KDP) crystal resulting in 520 mJ of 532 nm light. Residual 1064 nm light is discarded and the 532 nm pulses are used to pump a BBO-based optical parametric oscillator (OPO) (Opotek). The OPO process is similar to the OPA process discussed above, except the OPO is self-seeded rather than externally seeded. The OPO produces ~20 mJ of 728 nm signal and 1.98 μm idler. The idler is used to deliver the temperature jump in the T-jump experiment. The T-jump laser requires ~300 μs between the triggering of the flashlamps and the output of a 2 μm pulse while the regenerative amplifier outputs a pulse ~43 μs after the trigger. As a result, if the ND:YAG and regenerative amplifier are triggered at the same time, the T-jump pulse will always arrive after the 6 μm pulse. To overcome this limitation the previous 6 μm pulse must be used as the trigger for setting the T-jump delay, as illustrated in Fig. 3.7.

3.3.3 The Boxcar Interferometer and Balanced Detection

The separation of a mid-IR pulse into a series of copies that can be delayed relative to one another is accomplished in much the same way in the boxcar geometry as it is in the pump-probe geometry using beam splitters and retroreflectors mounted on delay stages. However, in the boxcar geometry no two beams are collinear and each can be controlled independently of the others.

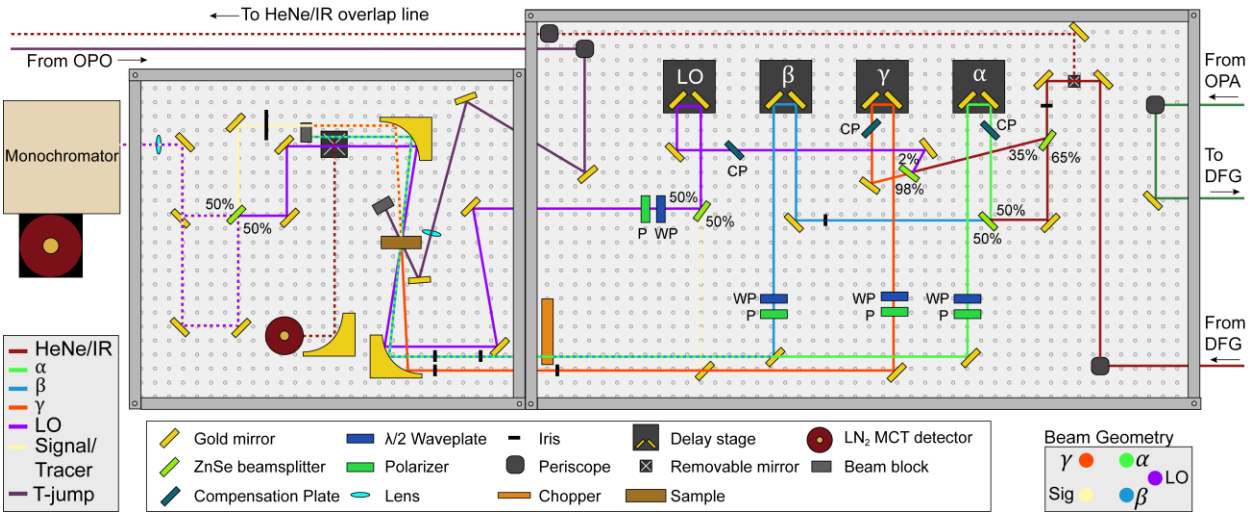

Figure 3.8: Schematic of the homebuilt interferometer and sample area of the boxcar temperature jump spectrometer. The optics and beams are color-coded according to the key in the figure.

Fig. 3.8 shows a diagram of the boxcar interferometer design as well as the sample area and balanced detection optics. Mid-IR light from the DFG is aligned through the interferometer using two gold mirrors to position the beam on an iris just prior to the interferometer and to an iris installed in the β delay line. The beams in the boxcar are labeled α, β, and γ rather than the $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$ labeling used for the compact spectrometer above because the pulse ordering is varied depending on whether the R or NR surface is being collected. This reversal of α and β is necessary for the R and NR signals to be emitted in the same direction in the boxcar geometry, which is desirable since otherwise one would need to realign to the desired signal or move the detector when switching between measurement of the rephasing and nonrephasing surfaces.

A series of ZnSe beam splitters are used to split off copies of the original mid-IR pulse. The first beam splitter reflects 35% and transmits 65%. The transmitted light is then split again on a 50/50 beam splitter to produce α and β, which are each sent to their own delay stages and through

waveplate/polarizers. This arrangement, in contrast to the pump-probe geometry, allows the polarization of the first two beams to be set independently. Returning to the 35% that was reflected off of the first beam splitter, 98% of this light is used as the γ beam, which serves as a stationary reference. The remaining 2% is reflected off of a beam splitter and is subsequently used as the LO and tracer. Note that each of the signal generating fields have approximately equal intensity and that together they account for ~99% of the total mid-IR light generated by the DFG. The LO and tracer each contain only ~0.5% but are still easily capable of saturating the detector. The β beam is chopped and subtracted on the fly such that only signals containing a contribution from β remain. Throughout the interferometer antireflective (AR) coated ZnSe compensation plates are installed to match the dispersion of each of the beams.

Each of the signal generating beams are arranged at three of the four corners of a 1" box and aligned into a gold parabolic mirror (10 cm effective focal length, 90° OAP, Janos Technology). Spatial overlap at the focus and relative pulse timings are achieved in much the same way as for the compact spectrometer using a 50 μm pinhole. The third order signal is emitted at the fourth corner of the box and the residual α, β, and γ beams are blocked. The tracer beam is aligned such that it is collinear with the signal at the fourth corner of the box, but it is only used for alignment purposes and is otherwise blocked. The LO is offset from the center of the box as shown in Fig. 3.8 and is also aligned through the pinhole. This design ensures that the LO passes through the same region of the sample as the remaining beams, but ~110 ps in advance so as not to interfere with signal generation. The LO path after the sample is longer than the signal path so that the signal catches up with the LO and can interfere at the detector.

The signal and LO are both directed to and recombined on a 50/50 ZnSe beam splitter AR coated on only one side. Due to the refractive index change between air and ZnSe ($n_1 < n_2$), the

signal reflecting off of the uncoated front face of the beam splitter accumulates a π phase shift while the LO internally reflected off of the back face does not. The reflected signal/LO are overlapped with the transmitted LO/signal on the beam splitter. In this way two π out-of-phase copies of the heterodyned signal are generated. These copies are vertically offset and focused through the 0.25 mm slit of the monochromator and dispersed onto the two stripes of the 2x64 MCT pixel array. The signal detected on the two stripes (S1 and S2) are given by

$$I_{S1}(\omega_3,\tau_2,\tau_1,\tau_{LO}) = \left|\tilde{E}_{sig}(\omega_3,\tau_2,\tau_1) + \tilde{E}_{LO}(\omega_3,\tau_{LO})\right|^2 \tag{3.6}$$

$$I_{S2}(\omega_3,\tau_2,\tau_1,\tau_{LO}) = \left|\tilde{E}_{sig}(\omega_3,\tau_2,\tau_1)e^{i\pi} + \tilde{E}_{LO}(\omega_3,\tau_{LO})\right|^2 \tag{3.7}$$

Assuming the electric fields can be expressed as plane waves,

$$\tilde{E}(\omega,\tau) = A(\omega,\tau)e^{i\varphi(\omega,\tau)} \tag{3.8}$$

The real part of eq 3.8 describes the real-valued electric field. For concision the argument of the exponential is expressed as a single phase term φ, which carries the dependence on the experimental delays and detection frequency. Terms related to the third order signal depend on ω_3, τ_2, and τ_1 while terms related to the LO depend on ω_3 and τ_{LO}. Expanding eq 3.6 and 3.7,

$$I_{S1}(\omega_3,\tau_2,\tau_1,\tau_{LO}) = A_{sig}{}^2 + A_{LO}{}^2 + 2A_{sig}A_{LO}\cos\left(\varphi_{sig} - \varphi_{LO}\right) \tag{3.9}$$

$$I_{S2}(\omega_3,\tau_2,\tau_1,\tau_{LO}) = A_{sig}{}^2 + A_{LO}{}^2 - 2A_{sig}A_{LO}\cos\left(\varphi_{sig} - \varphi_{LO}\right) \tag{3.10}$$

Acquiring the signal in this way and taking the difference between the two stripes on a shot-to-shot basis is called balanced detection. This acquisition scheme is advantageous in that the LO intensity and the homodyne term, $A_{sig}{}^2$ are canceled and correlated noise is suppressed. The $2\tilde{E}_{sig}\tilde{E}_{LO}$ term is the heterodyned signal of interest. The main downside of balanced detection is the added complexity of aligning the balanced detection optics, as it can be difficult to simultaneously achieve all of the necessary alignment objectives with the available degrees of freedom.

Appendix 3B contains a practical step by step guide for aligning the boxcar spectrometer as well as benchmarks for evaluating the quality of various stages of the alignment.

3.3.4 Overlapping the T-Jump Pulse and Determining the T-Jump Magnitude

Once the boxcar spectrometer is aligned with respect to all of the 6 μm beams, the task of overlapping the 2 μm and 6 μm pulses at the sample focus and electronically synchronizing the arrival of the T-jump pulse remains. The temperature dependent change in transmission of the low intensity solvent absorption due to the D_2O bend-libration combination band centered around 1550 cm^{-1} serves as a reference for achieving this overlap. The idler from the OPO is resonant with the OD stretch overtone of the D_2O solvent. The overtone is pumped rather than the fundamental because the ~10% absorption results in a more uniformly heated sample volume. Vibrational excitation of the solvent rapidly relaxes such that heating occurs within the nanosecond pulse envelope of the T-jump pulse.

While the 6 μm pulses are focused to a beam diameter of ~100 μm the 2 μm pulse is focused to a ~400 μm diameter spot. Therefore the heated volume is around four times larger than the sample volume being probed, which further ensures that the temperature change is as uniform as possible. Fig. 3.9a shows the LO spectrum recorded after a T-jump delay of τ_j = 5.6 μs in dark red with the initial temperature set to 30 °C. The 1 kHz repetition rate of the regenerative amplifier sets the arrival time of the subsequent LO pulses at 1 ms intervals. The next eleven LO spectra are plotted from red to blue for reference. Since the T-jump laser operates at 20 Hz, 50 mid-IR pulses arrive between each temperature jump event. The 50^{th} shot arrives 50 ms + τ_j ≈ 50 ms after the T-jump, and by this time the temperature has completely relaxed back to the initial equilibrium

temperature. Subtracting the equilibrium shot off from each of the shots to arrive earlier results in a t-LO thermal difference spectrum, plotted in Fig. 3.9b, that gives the change in LO transmission.

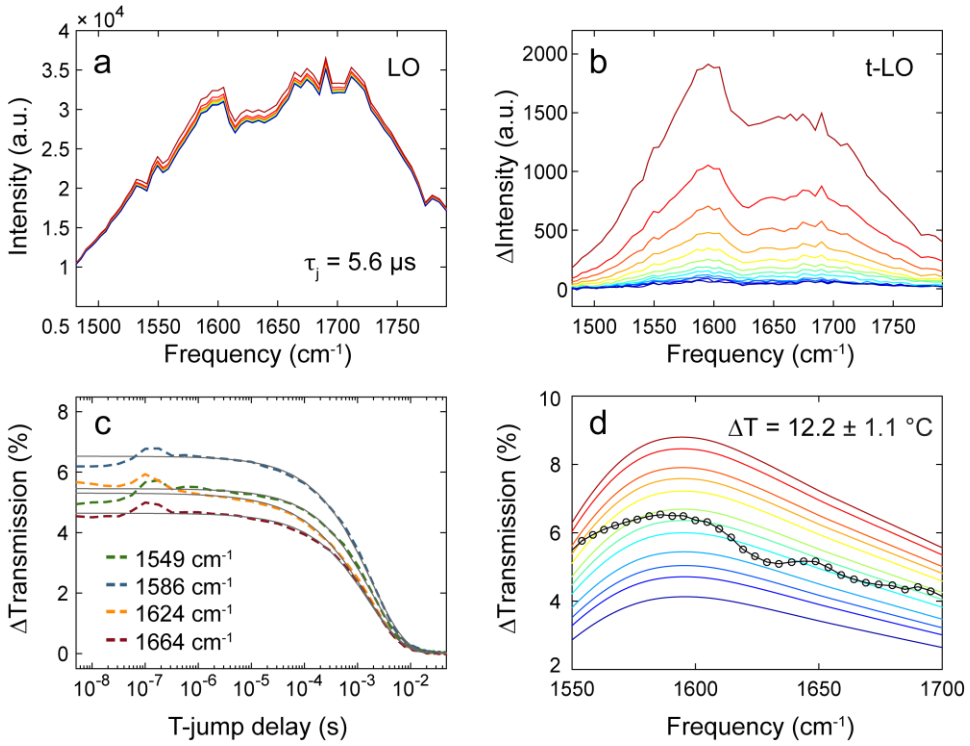

Figure 3.9: (a) Example of the LO spectrum recorded for a T-jump delay of τ_j = 5.6 μs followed by the LO spectrum sampled every ms for 11 ms (red to blue). The initial temperature is set to 30 °C. (b) Transient LO spectrum (t-LO) obtained by subtracting the equilibrium LO spectrum off from each of the spectra in panel a. (c) Select frequency traces showing the T-jump thermal profile. Gray lines are stretched exponential fits. (d) Checking the frequency dependent percent change in transmission of the LO (black points) against a reference set of %ΔTransmission curves obtained from a temperature dependent set of FTIR spectra of the D_2O bend-libration assigns the T-jump magnitude.

To determine the T-jump magnitude, the percent change in transmission of the LO is compared against a reference set of curves calculated from a series of temperature dependent FTIR spectra of the D_2O solvent sampled every ~1 °C. Fig. 3.9c shows four illustrative frequency slices through the full t-LO data set from which the spectra in Fig. 3.9b are drawn. These curves track

the frequency dependent change in solvent transmission and show the thermal profile of the T-jump, which is well fit by a stretched exponential. The first 200 ns are excluded from the fit because cavitation pressure waves propagating through the sample at the speed of sound can also cause changes in solvent transmission that do not report on temperature change. The value at which the fit plateaus for each frequency slice is taken as the percent change in transmission of the LO at that frequency. Since the initial temperature (T_i) of the sample can be measured accurately with a thermocouple, a series of reference curves for a spread of potential final temperatures (T_f) can be calculated from a set of FTIR absorbance spectra acquired at known temperatures, $A(T,\omega)$.

$$\Delta Trans(\%) = 100\left(\frac{10^{-A(T_f,\omega)} - 10^{-A(T_i,\omega)}}{10^{-A(T_i,\omega)}}\right) \tag{3.11}$$

Fig. 3.9d shows the percent change in transmission determined at each frequency of the t-LO spectrum from 1550-1700 cm^{-1} plotted in black checked against a set of reference curves calculated from FTIR. Each color coded curve corresponds to a single final temperature and reflects the frequency dependent change in transmission assuming an initial temperature of 30 °C. By matching the change in transmission of the LO at each frequency to the nearest reference curve, a T-jump magnitude (ΔT) can be assigned. Assigning the temperature change in this way results in a mean ΔT of 12.2 °C with a standard deviation of 1.1 °C across the frequency range.

Time zero ($\tau_j = 0$) for the simultaneous arrival of the 2 μm pulse and 6 μm pulse is defined by finding the flashlamp delay, t_j that corresponds to the point where the t-LO spectrum is at half of the maximum. This point reflects the midpoint in the rise of the thermal profile, as reported by the change in solvent transmission. Assuming that the temperature increase is delivered uniformly by the 2 μm pulse, then determining the delay at which the LO samples the midpoint of the rise should correspond to the point where the 2 μm pulse and 6 μm pulse are coincident in time.

3.3.5 Collection and Processing of Transient Data

Throughout this thesis the temperature jump data reported are transient heterodyne-detected dispersed vibrational echo (t-HDVE) spectra. Ultimately these signals are equivalent to pump-probe difference spectra, but acquired in an externally heterodyned boxcar geometry to maximize sensitivity and eliminate background signals. The Fourier transform spectral interferometry (FTSI) method discussed in Chapter 2 is used to separate the amplitude and phase information required to reconstruct the dispersed pump-probe signal. This section focuses on the practical aspects of how T-jump data are acquired and processed to arrive at time resolved thermal difference spectra and kinetic traces that relay the dynamic and kinetic information of interest.

The ~5 ns lower limit on the time resolution of the T-jump experiment is set by the width of the 2 μm pulse that delivers the temperature jump. In contrast the upper limit is not so simply defined. The 20 Hz repetition rate of the Nd:YAG sets a hard upper limit of 50 ms, since after 50 ms the next T-jump pulse arrives and reheats the sample all over again. Therefore the system must be re-equilibrated within the 50 ms window to avoid accumulated effects over the course of successive T-jump events. A more practical upper limit on the time resolution is dictated by thermal diffusion away from the heated sample volume, as reported through the change in LO transmission through the sample (Fig. 3.9c). The temperature (T_f) remains relatively constant for ~100 μs and then begins to relax back toward the initial temperature (T_i) with a stretched exponential profile that has a ~2 ms time constant. As a result, timescales which coincide with the thermal relaxation will no longer be measured at constant temperature, which complicates the task of interpreting the kinetics considerably even though the system response is tracked out to 50 ms.

Fig. 3.10 illustrates the nanosecond heating event and the resulting thermal profile delivered by the 2 μm pulse. In practice, the first time point sampled along the profile is determined

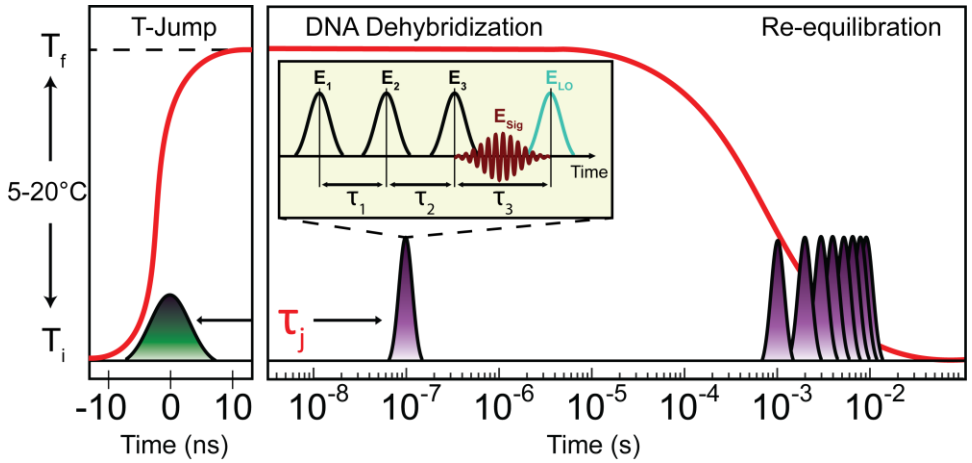

Figure 3.10: Nanosecond heating and resulting thermal profile imparted by the 2 μm pulse. The delay between the T-jump pulse and the first mid-IR pulse is set by the electronic delay τ_j while the subsequent pulses arrive in 1 ms intervals after τ_j due to the 1 kHz repetition rate of the regenerative amplifier.

by the delay between the center of the T-jump pulse and the first mid-IR pulse sequence set by the electronic delay τ_j as discussed above, but the remaining time points sampled along the thermal profile are fixed by the kHz repetition rate of the regenerative amplifier and occur at $\tau_j + n$ ms intervals up to $n = 50$, after which the next T-jump pulse arrives. As a consequence and in contrast to the equilibrium experiments described previously, successive 6 μm shots do not probe the same sample conditions and therefore cannot simply be averaged together. Further still, since every other shot is chopped only half of the shots along the thermal profile result in a detected third order signal and adjacent chopped shots cannot be subtracted. To account for these differences, T-jump data are treated as blocks of 50 mid-IR shots that each correspond to a unique T-jump delay τ_j and LO delay τ_{LO}. The T-jump delay τ_j is stepped through a set of values ranging from -5 ns to ~600 μs to sample the fastest range of accessible timescales while the shots arriving at 1 ms intervals after the first shot sample slower timescales. To record the chopped signal at all 50 shots along the thermal profile for a given τ_j and τ_{LO}, after collecting some number of specified shots, the chopper

phase is flipped by π and the same number of shots are collected. If the first, third, and so on odd shots were unblocked for the original chopper phase and the second, fourth, and so on even shots were chopped, then flipping the chopper phase by π reverses the situation such that the odd shots are chopped and the even shots generate signal. In this way both chopped and unchopped shots at every time point along the thermal profile are sampled and as before the blocked shots can be subtracted from the unblocked shots to eliminate unwanted contributions and background signals. The data are collected such that linear absorption along the detection axis is divided out on the fly.[10] Due to the relatively low 1-2 mM concentrations employed throughout this thesis, no further correction for linear absorption along the excitation axis is considered necessary and no such correction is applied.

Fig. 3.11 outlines how transient data are collected and organized prior to post-processing. Within a given average, the set of T-jump delays τ_j are stepped from smallest to largest delay. At each τ_j, the LO delay is stepped from τ_{LO} = -10, -5, 0, 5, and 10 fs to establish the phase contrast needed for the FTSI method. At each unique combination of τ_j and τ_{LO}, shots are acquired and the chopper phase is flipped as described above. The result is a 128 pixel by 50 shot matrix. The first 64 columns of the first row contain the HDVE spectrum recorded at τ_j after the T-jump, while rows 2-50 contain the shots arriving at 1 ms intervals after the first shot. The second set of 64 columns contains the corresponding LO spectrum recorded at each point along the thermal profile which is used as discussed in the previous section to determine the T-jump magnitude. After extracting phase and amplitude information, the known phase offset applied by τ_{LO} is shifted to τ_{LO} = 0, which should result in a spectrum that is equivalent to the dispersed pump-probe spectrum.[11] However, due to the accuracy limitations of the LO delay stage it is necessary to apply a further phase correction to account for any additional offset. Unfortunately the Mertz correction which was

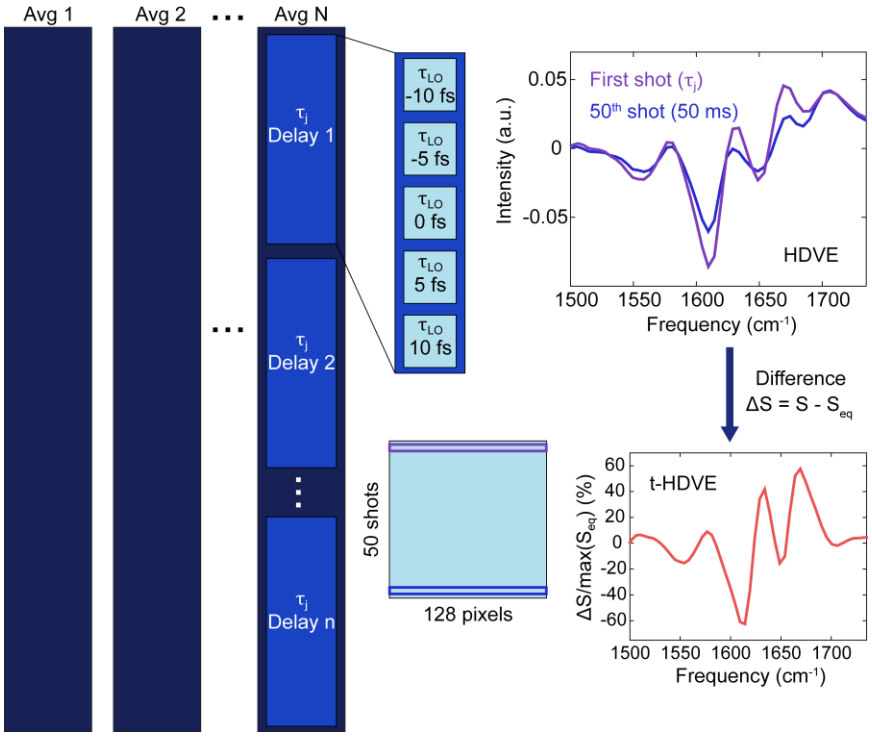

Figure 3.11: Diagram of how T-jump data are collected and organized in the laboratory and an example of the resulting HDVE spectra after phase correction and averaging in post processing. The lower right panel shows the final t-HDVE difference spectrum for an illustrative delay of τ_j = 100 μs.

robustly applied to automatically correct for phasing errors in the pump-probe geometry above is not easily implemented in the boxcar geometry. In fact the added challenge of addressing phase ambiguity is the greatest downside to this geometry.[8] A standard approach to phasing spectra in this case relies on the projection-slice theorem which states the equivalence of the projection of the 2D IR surface onto the detection axis with a pump-probe spectrum acquired at a delay equivalent to τ_2.[12] The R and NR surfaces are phased independently by fitting the projection of the respective surfaces to the dispersed pump-probe spectrum with respect to varying a phase correction term.

For t-HDVE timing errors in τ_1 are less of a concern since the coherence time is fixed at $\tau_1 = 0$. However, a similar approach is adopted for correcting errors in τ_{LO}. This requires the independent collection of a pump-probe spectrum at the initial temperature before the T-jump using the chopped β beam as the pump and the LO as the probe. The 49^{th} and 50^{th} shots, which are assumed to be at equilibrium at T_i, are fit against the pump-probe spectrum in the frequency domain with respect to a phase correction factor. The phase correction which results in the best fit for the equilibrium shots is then applied across the entire set of 50 spectra acquired along the thermal profile.

Once the phase is corrected such that all LO positions are shifted to $\tau_{LO} = 0$, the blocks of 50 shots corresponding to a given τ_j should be equivalent and can be averaged together. Transient data are presented as thermal difference spectra by subtracting off the equilibrium spectra contained in the 49^{th} and 50^{th} row from each of the other rows which correspond to the HDVE spectrum sampled along the thermal profile. Since τ_j is scanned and the remaining 49 shots in each set are fixed by the 1 kHz repetition rate, the set of transient data must be rearranged in order of arrival after the T-jump pulse. As part of this process, all of the *n* ms spectra collected at the same value of *n* after the first spectrum are averaged together since *n* ms + $\tau_j \approx$ *n* ms. The panel in the lower right corner of Fig. 3.11 shows an example of a t-HDVE spectrum for a melting DNA oligonucleotide duplex obtained through this process.

3.4 Acknowledgements

I thank Ann Fitzpatrick for her contributions building the compact 2D IR spectrometer and her easygoing mentorship. I thank Paul Stevenson for his efforts rebuilding the boxcar spectrometer

and for his helpful comments on Appendix 3b. I thank Nick Lewis, Ram Itani, and Brennan Ashwood for careful reading of this chapter.

3.5 References

1. DeFlores, L. P.; Nicodemus, R. A.; Tokmakoff, A., Two-dimensional Fourier transform spectroscopy in the pump-probe geometry. *Optics letters* **2007,** *32* (20), 2966-2968.

2. Gallagher Faeder, S. M.; Jonas, D. M., Two-dimensional electronic correlation and relaxation spectra: Theory and model calculations. *The Journal of Physical Chemistry A* **1999,** *103* (49), 10489-10505.

3. Harris, F. J., On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE* **1978,** *66* (1), 51-83.

4. Chase, D., Phase correction in FT-IR. *Applied Spectroscopy* **1982,** *36* (3), 240-244.

5. Backus, E. H.; Garrett-Roe, S.; Hamm, P., Phasing problem of heterodyne-detected two-dimensional infrared spectroscopy. *Optics letters* **2008,** *33* (22), 2665-2667.

6. Helbing, J.; Hamm, P., Compact implementation of Fourier transform two-dimensional IR spectroscopy without phase ambiguity. *JOSA B* **2011,** *28* (1), 171-178.

7. Mertz, L., Auxiliary computation for Fourier spectrometry. *Infrared Physics* **1967,** *7* (1), 17-23.

8. Johnson, P. J.; Koziol, K. L.; Hamm, P., Intrinsic phasing of heterodyne-detected multidimensional infrared spectra. *Optics Express* **2017,** *25* (3), 2928-2938.

9. Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A., Transient two-dimensional IR spectrometer for probing nanosecond temperature-jump kinetics. *Review of scientific instruments* **2007,** *78* (6), 063101.

10. Jones, K. C.; Ganim, Z.; Peng, C. S.; Tokmakoff, A., Transient two-dimensional spectroscopy with linear absorption corrections applied to temperature-jump two-dimensional infrared. *JOSA B* **2012,** *29* (1), 118-129.

11. Jones, K. C.; Ganim, Z.; Tokmakoff, A., Heterodyne-detected dispersed vibrational echo spectroscopy. *The Journal of Physical Chemistry A* **2009,** *113* (51), 14060-14066.

12. Jonas, D. M., Two-dimensional femtosecond spectroscopy. *Annual review of physical chemistry* **2003,** *54* (1), 425-463.

13. Krężel, A.; Bal, W., A formula for correlating pKa values determined in D2O and H2O. *Journal of inorganic biochemistry* **2004,** *98* (1), 161-166.

Appendix 3A: Preparation of DNA Samples for IR Spectroscopy

Modern oligonucleotide synthetic techniques have come a long way since the early days of DNA biophysics research in which samples were extracted and purified from natural sources such as calf thymus and salmon testes. This approach provided limited control over sequence composition and strand length. Today one can design and order any custom sequence of interest on the internet for only a few cents per base and have the sample arrive in the mail within a few days. However, oligonucleotide samples must still be prepared for IR spectroscopy by removing IR absorbing contaminants left over from synthesis and by exchanging labile protons in preparation for working in D_2O solutions. This appendix describes how to prepare nucleic acid samples for IR spectroscopy.

1. Typically DNA samples are purchased from IDT or Sigma at desalt grade purity and at 1 μmol synthesis scale. The minimum sample volume needed to run a single experiment is about 30 μL and the concentration ranges from 3-20 mg/mL.
2. Samples are prepared in D_2O solvent in order to remove interference from the H_2O bend at ~1650 cm^{-1}. This requires all labile protons in both the sample and the buffer to be HD exchanged prior to use. A good buffer for IR spectroscopy is sodium phosphate with NaCl or $MgCl_2$ added as needed.
3. Weigh up the buffer components required for the target pH according to the Henderson-Hasselbalch equation and the mono/divalent salts required for the target cation concentrations.
4. Dissolve the buffer in a sufficient volume of D_2O to create a solution that is no more concentrated than ~20 mg/mL. Solutions that are too concentrated tend not to lyophilize well.

5. Additional filtration is necessary to remove IR absorbing small molecules (such as carbonates) left over from nucleic acid synthesis. Add enough D_2O to the DNA sample to dissolve it entirely (for a typical sample, ~200 μL works well), vortex, and then transfer into a centrifugal filter tube (Amicon). Make sure the MW cutoff is below the MW of your sample. An oligonucleotide ten base pairs in length weighs around 3 kDa.
6. Spin through the filter into a collection tube on a benchtop centrifuge at a high speed (~13,000 rpm) for 10 minutes. Save the waste in order to check for possible sample loss through the filter.
7. Repeat step 6 using an additional 100 μL D_2O to rinse out the IDT vial before adding to the filter.
8. Add ~200 μL to the filter and invert into a clean collection tube. Spin for 2 minutes at 2,000 rpm.
9. Repeat step 8 into the same collection tube to maximize sample recovery.
10. After recovering the purified DNA sample off of the filter, make sure that the volume is such that the solution is sufficiently dilute for lyophilization (see step 4). If not, add additional D_2O.
11. Secure a kimwipe with a twist tie over the opening of the sample collection tube and the buffer Eppendorf tube to allow solvent to sublimate off in the lyophilizer.
12. Freeze both sample and buffer solutions by submerging the end of the tubes in liquid nitrogen.
13. Place the frozen tubes into a lyophilizer collection flask and seal it onto the lyophilizer. For typical sub mL volumes, around 24 hrs on the lyophilizer works well. Increase time with increasing sample volumes.

14. After lyophilization both sample and buffer tubes should contain a fluffy white powder that resembles a tiny cotton ball. If the sample appears gel like, redissolve in D_2O (perhaps more than before for a more dilute sample), and repeat the lyophilization process above. Pay close attention that the sample is completely frozen before sealing on the lyophilizer.
15. Once nicely HD exchanged and lyophilized, the buffer can be redissolved in the appropriate volume of D_2O. When checking the pH of the buffer with a standard glass electrode pH meter, remember to apply a correction for deuterated solutions.[13]
16. The exchanged DNA sample can now be weighed out and dissolved in the appropriate buffer volume for the target sample concentration. Remember when preparing a sample by weight that DNA is hygroscopic and it is impossible to remove all associated waters through lyophilization. In addition, lyophilized solid DNA is a salt with counter ions that contribute to the mass. Therefore weighing up double the target mass of oligonucleotide and then checking and adjusting the final concentration using the UV Nanodrop spectrometer works best.
17. For checking the concentration on the Nanodrop, dissolve 1 μL DNA solution into 99 μL H_2O (D_2O is not necessary for UV) and check the concentration using this 100x diluted sample. This dilution conserves DNA sample and is better suited for the Nanodrop's sensitivity range.
18. If the concentration is too high, dilute with additional buffer solution as needed. If too low, add a flake more of the lyophilized DNA sample. Check the concentration again on the Nanodrop and continue to adjust until the target concentration is achieved. Typically a concentration of within ±0.5 mg/mL of the target is acceptable.

19. If the sample is not going to be used immediately, it is important to store it properly to minimize HD exchange with ambient water. For shorter storage periods, parafilm around the edge of the sample tube and store in the fridge. For longer term storage, freeze instead. The best long term strategy is to store samples as lyophilized powders in the freezer. It is a good idea to also parafilm before annealing samples in a hot water bath to minimize HD exchange here as well.

Appendix 3B: Notes on Aligning the Boxcar Spectrometer

This appendix contains notes and tips compiled over several years of working with the boxcar spectrometer as well as a step by step alignment guide intended to serve as an introduction for beginners. The procedure outlined here is by no means definitive. The homebuilt spectrometers in the lab are the result of a collective and decentralized effort spanning several generations of graduate students and postdoctoral researchers. Much is owed to these past efforts, but one should always keep in mind that these homebuilt instruments (in contrast to commercial spectrometers) have not been engineered to be foolproof or even necessarily user-friendly. As a consequence there is no rote procedure that can be applied that will reliably result in a good alignment. Instead, the most productive approach is to treat the series of steps presented here as a conceptual framework rather than a strict formula and to think carefully about the purpose and objective of each step. Whenever possible, a brief motivation for each step as well as a metric for evaluating the quality of the outcome is included. Labels in this appendix are consistent with those in the diagram shown in Fig. 3.8.

DFG/HeNe-IR Overlap:

1. Align the collinear Signal/Idler (S/I) from the OPA to the two irises on either side of the DFG crystal. Use the bright visible spots. This is a white-ish color before the AGS (likely white light), and a red-pink color after (likely mixing of 6 μm with residual 800 nm).
2. Adjust the timing of the S/I first. This should be extremely sensitive. If not, the power you are seeing is likely spontaneous IR generation from the Idler.
3. When the timings are set, optimize the S/I overlap on power. Switching steps 2 and 3 will result in sub-optimal power and stability, as changing the overlap also changes the timings.
4. Center the HeNe spot on both mirrors of the IR telescope inside the DFG box. Make sure the IR is at least hitting both mirrors using thermal paper.
5. HeNe-IR Overlap: Near point is measured on the power meter placed at the output of the DFG box. Far point is the single channel MCT detector at the top of the table. Install the magnetic mirror at the entrance to the interferometer to send the beam down the overlap path and align to the detector. Be sure the HeNe is not clipping at any point.
6. Use the last two mirrors before the Ge plate to overlap the IR and HeNe, first mirror for the near point, second for the far point. Iterate near/far until no more improvement in overlap can be achieved. The far field detector should easily saturate when no lens paper/filter/irises are used to attenuate.
7. Once overlapped with the HeNe, remove magnetic mirror to send the beam into the interferometer.

Interferometer/Spatial Overlap/Timings:

8. Use the first two mirrors before the interferometer to align onto the two irises in the interferometer. Iterate until well aligned at both points.
9. Align α, β, γ, and the tracer into the parabolic using the irises mounted along the approach as a reference point. Pay close attention to which of the many HeNe reflections is overlapped with the IR. Close down the first iris in the interferometer in order to better see the multiple HeNe spots.
10. For the near iris, use the micrometers on the stage-mounted retroreflectors. Be sure not to adjust the micrometer which moves in the direction of stage motion, as this will create more work when finding timings later. For the far iris, use the mirrors after the WP/P combinations. Iterate until well aligned at both points.
11. Install the large magnetic mirror to send the beams after the parabolics into the single channel MCT inside of the balanced detection box. Block all beams except for γ and then use the magnetic mirror to align γ into the single channel detector. Do not cool down the detector or it will saturate.
12. Install a 50 μm pinhole onto the sample stage. The Z-position (direction of beam propagation) for pinhole alignment is near 22 mm. Check where the beams focus to determine this position if necessary by finding the Z-position that corresponds to the maximum throughput for each beam independently. Once the optimal Z-position is determined, set the Z micrometer and adjust the X and Y micrometer positions to optimize γ pinhole throughput. One should be able to measure around 8 V on the oscilloscope. (How well purged the boxes are will dictate throughput. 8 V for a well purged box, but only ~6 V if boxes have been open.)
13. The pinhole is now set, do not touch it going forward for best spatial overlap.

14. Next block γ, open α, and with the magnetic mirror translate the horizontal to send k1 into the single channel detector. If there is a bit of HeNe making it through the pinhole, use that as a guide. If unable to find α through the pinhole, remove the pinhole, optimize α pointing into the detector, then place pinhole back, but now it is necessary to revisit γ and make sure it is still optimized after finishing α.
15. Throughout the pinhole alignment procedure, use the large magnetic mirror to revisit the pointing into the single channel MCT (after switching between beams and when changing pointing through the pinhole) in order to ensure the beam is well centered on the detector element.
16. Use the last mirror before the parabolic on the α path to direct the α beam through the pinhole. Revisit the pre-parabolic iris positions and make sure the beam is still aligned on the irises. If not, use the retroreflector position to align through the irises and the last mirror to optimize pinhole throughput. For α, one should measure about 10 V throughput on the scope.
17. Block α and unblock β. Repeat process above (steps 14-16) but now in the β arm to optimize β throughput. The β beam is vertically displaced from α, so adjust the vertical alignment into the detector when switching between beams. For β, should measure about 9 V throughput on the scope.
18. Block β and unblock the tracer. Block the LO path near the WP/P combo. Repeat pinhole alignment (steps 14-16) above on the tracer arm. Adjust the horizontal alignment into the detector when switching between beams. The tracer throughput is weak, so cooling the detector may be necessary. However, do not add more than a few mL of liquid N_2, otherwise the detector will saturate.

19. Once α, β, γ, and the tracer have been maximized through the pinhole, block these beams and align the LO through the pinhole as well. There are two mirrors in the sample box to accomplish this. Use the first to align the LO onto the far edge of the second, which is an elliptical mirror.
20. Use the elliptical mirror to maximize the LO pinhole throughput. Once a small amount of tracer makes it through, use this to optimize the pointing into the single channel MCT using the large magnetic mirror.
21. All beams are now spatially overlapped at the sample focus and must now be overlapped in time. Begin by blocking all beams except β and γ. The γ arm is stationary, so we will time β up with respect to it first. Block the LO path in the balanced detection box and make sure the spatial filter is well set to block α, β, and γ after the sample.
22. If necessary, use the final gold mirror in the detection area (the one immediately before the lens and monochromator) to optimize scatter on the array. Close down the iris in the signal path of the detection or the slit of the monochromator until there is only a single contribution to the scatter on the detector. Multiple contributions will result in weird interference patterns with multiple maxima and minima instead of a single maximum corresponding to t_0.
23. Check for β-γ scatter on the array. If it is not centered on the MCT stripes, shift the monochromator grating using MainSpecControl.vi to center the scatter on the pixel array. The usual position is ~1070 nm on the "Go To" setting. (The grating calibration is off, but positions are reproducible!)
24. Move the Y-stage (β arm) to roughly find t_0 by finding the point of maximum constructive interference between the beams and set this position using the stage multiclient vi.

25. Using the "Timings" tab on the boxcar vi, select timings between β-γ and scan the Y-stage from -1000 to 1000 fs in 4 fs steps, averaging 200 shots per step. Run the boxcar_timing.m script in Matlab to determine t_0. Check both stripes of the array. They should agree to within a few fs. If not, improve how the scatter is centered on the array and repeat. Set the new t_0 using the stage multiclient vi.
26. Send β to t_0, block γ, and unblock α. Repeat steps 21-25 but now for α-β scatter scanning the X-stage (α arm). Timing up between α-β instead of α-γ reduces timing errors in τ_1.
27. Next time up the LO and β repeating steps 21-25 using the Z-stage (LO arm), but with a few changes: Move the chopper to chop the LO rather than β and block the sample path instead of the LO path in the balanced detection box.
28. Once done with timings, move the chopper back to the usual spot chopping β.

Homodyne Signal:

29. Now all beams are overlapped in space and time at the pinhole. Prepare a sample of 10 mg/mL diglycine (GG) and insert at the sample position. Send all beams to t_0, block the LO and the LO path in the sample box, and send α, β, and γ into the sample to generate homodyne signal. Adjust the sample stage Z position (direction of propagation of the beams) to find the optimum position. This will change sample to sample – different sample cells, spacers etc. will give rise to different optimum micrometer positions.
30. The goal for homodyne signal is ~1300 counts on the array (chopped mode) with the red and white stripes well balanced. Several factors influence the magnitude of the homodyne signal, including how well purged the box is, where the DFG wavelength is centered, and how much mid-IR is being generated, but ~1300 counts is a rough reference for 10 mg/mL GG.

31. If there is some signal on the array, proceed to step 32. If there is no signal, use the tracer to align onto the two stripes of the array after the balanced detection BS, then check for homodyne signal. The tracer should easily saturate the entire range of the array.
32. Set the X- and Y-stages both to 150 fs. To peak up the homodyne signal and balance the stripes, use the mirrors after the balanced detection BS. There are two mirrors that direct the transmitted beam. Use these mirrors to maximize the red stripe.
33. Use the BS and next mirror to maximize the reflected beam (white stripe). If just adjusting the horizontal/vertical does not result in sufficient improvement, a strategy here is to use the mirrors to walk the beams. In the horizontal, use the mirror furthest from the detector to move off of the (local) maximum slightly in a fixed direction, then use the mirror closest to the detector to try to peak past the previous maximum. Keep track of which direction you are moving on the furthest mirror. Check one direction, if this does not improve things, try the other. Repeat for the vertical and for the red stripe. Iterate until there is no more gain to be had and until both the stripes are well balanced/overlaid on the screen. This is much easier said than done!
34. Throughout this process, adjust the final mirror before the lens that focuses the two signal beams into the monochromator to maximize the homodyne signal.
35. Check that the signal beams are well centered on the lens. Use a card to make sure they die off similarly when blocked and when moved off the array using the final mirror before the detector (in both horizontal and vertical direction).

Heterodyne Signal:

36. Send the Z stage to around -110,000 fs. This number can change significantly (±20 ps) depending on alignment. Unblock all beams except for the tracer by placing a card on the back side of the BS that picks off the LO. Unblock the LO path in the balanced detection box.
37. Using the Reference mode in the software, optimize the LO intensity on the screen using the two mirrors unique to the LO path. Do not use any mirrors in the sample path, since this will misalign the balanced homodyne signal.
38. Run the boxcar vi in chopped mode and use the Aerotech software directly to jog the Z-stage backwards and forwards. Pay attention to how the fringes grow in and out as the LO interferes against the homodyne signal. The number of fringes will increase around either side of BSt_0. (Note BSt_0 here is referring to the timing at the balanced detection BS such that the homodyne and LO interfere maximally at the detector. We have already set the LO t_0 at the sample using pinhole scatter, and it is this t_0 that the boxcar vi recognizes as zero for the Z-stage).
39. Move the Z-stage backwards past BSt_0 until there are ~10 fringes on each stripe and use the LO path mirror furthest from the BS to maximize the fringes.
40. Jog the Z-stage to the other side of BSt_0 until there are ~10 fringes and use the LO path mirror closest to the BS to maximize the fringes.
41. Repeat steps 39-40 until the interference on either side of T0 is maximized simultaneously and nothing more can be gained on either side. If unable to accomplish this, switch which mirror is adjusted on either side of BSt_0 and repeat the process.
42. For array calibration, use the Timings tab to scan the LO against the homodyne signal ±2 ps from BSt_0 and generate the ω_3 axis using calibrate_array.m.

43. To set BSt_0 more precisely, first take a pump-probe (PP) by blocking all beams except for β and the LO. Send β to 150 fs and the LO to t_0 (Z-stage 0 fs). Remember to also take a PP background by setting the delay to 10,000 fs.
44. On the Phasing tab, load in the PP and the PP-bkg. Send the LO back towards BSt_0 (around -110,000 fs), change the acquisition mode to balanced detection, and adjust the LO timing to match the HDVE signal to the PP. Record this LO position.
45. When the alignment is complete, there should be over 6,000 counts of HDVE signal for 10 mg/mL GG in balanced detection mode at a waiting time of 150 fs.
46. Always collect and inspect a 2D IR surface of 10 mg/mL GG before moving on since the 2D IR spectrum is the most informative and sensitive reporter of misalignment.

Additional Notes and Tips:

47. Stage motion in the LabVIEW program is slow by default. This is important to accurately collect interferograms in a step-scan modality. However, for some applications, such as moving the LO between t_0 and BSt_0, there is a fast move built in. To access this, first move the stage to a position 1 fs away from the current position, then move to the far away position. The software switches to a fast move mode if the difference between sequential positions is >1000fs.
48. For very small signals, it may be advisable to collect data in a ΔOD mode. This mode can be accessed in the Run 2D tab of the software, and changing the collection mode from "Balanced Detection" to "Delta OD". Delta OD divides through by the LO on a shot-to-shot basis, which can improve the signal to noise ratio, however features at the edge of the LO spectrum will contain artifacts since you are dividing through by a small number.

49. Scatter can impede collecting a useful pump-probe spectrum. This can be alleviated for mild scatter by using the "Oscillate" option in the Pump-Probe tab in the software. This oscillates the LO stage at 13Hz (i.e. not a harmonic of the system) which suppresses scatter to a degree.
50. In 2D spectra, scatter shows up as a string of peaks along the diagonal. The first recourse here should always be sample preparation. Centrifugation of the sample before 2D experiments should be standard practice, as should careful window cleaning and sample preparation. If scatter persists, there are irises which may be carefully installed in the sample box to limit scatter that is not co-propagating with the signal. Finally, a polarizer may be installed before the monochromator and the polarization switched to ZZYY. However, scattered light does not necessarily preserve the polarization of the incident field, so even ZZYY spectra are not guaranteed to be scatter-free.

A Few Final Notes on "Phasing" in the Boxcar Geometry:

Phasing in this context refers to adjusting the delay between the LO and the signal generating fields to counter any phase acquired by the signal due to timing errors between α and β. This is accomplished by comparing the HDVE signal to a pump-probe taken at the same waiting time. Phase ambiguity is the single biggest source of uncertainty surrounding spectra acquired in the boxcar geometry. Errors in phasing can lead to the following artifacts:

- Suppression of cross peaks
- Artifacts which appear to be cross peaks, but are not real
- Line shape distortion – particularly phase twist
- Amplitude distortions

In addition to the phasing procedure applied in post processing discussed above, there is also a clear protocol for setting the initial LO phase. The first step is to acquire a pump-probe spectrum. With the PP as a reference, the BSt_0 is adjusted such that the HDVE on screen matches the PP spectrum using the Phasing tab of the vi. The phase error from mis-timing is proportional to $\omega\Delta t$. Therefore errors in BSt_0 will be most obvious when comparing peaks at well separated frequencies. Generally it is best to make sure that the highest frequency and lowest frequency doublets align with their corresponding peaks in the pump-probe.

Nucleic acids have many distinct peaks, and so it is comparatively easy to determine whether the spectrum is phased correctly. Proteins on the other hand, generally have broad, overlapping peaks, which do not readily lend themselves to a simple determination of the correct BSt_0. The match between the HDVE and PP can appear acceptable across several different optical cycles. In such cases it is often advisable to use GG as a reference before the protein sample is measured. The set BSt_0 phase will drift over time and must be reset often. For each sample, and even for the same sample over the course of a data run, the phasing should be checked and reset before each spectrum is collected.

Chapter 4

Extended Analysis Methods for Steady-State and Transient Infrared Spectroscopy

4.1 Introduction

In addition to the data processing steps discussed in Chapter 3, there are a number of additional processing and analysis methods used throughout this thesis that have been instrumental in bridging the gap between the raw signals measured in the laboratory and decipherable data that reveal the underlying story those signals tell. This chapter contains an assortment of such methods which, for lack of a better label, we will call extended analysis methods since they lie outside of the standard toolbox for analyzing steady-state or transient infrared data. None of the approaches discussed in this chapter are new, but rather ideas borrowed from signal processing and information theory are adapted and applied to IR spectroscopy. In section 4.2, a method for reducing the contribution due to additive noise in 2D IR data is presented. Section 4.3 describes a wavelet transform method for simultaneously removing background signals and noise in FTIR measurements. Section 4.4 outlines a maximum entropy method for reconstructing pure component spectra from the spectrum of a mixture. Lastly, section 4.5 presents a maximum entropy guided inverse Laplace transform method for transforming kinetic data into a corresponding spectrum of rates.

4.2 Noise Reduction in 2D IR through Spectral Subtraction

Noise is an inescapable reality when working with experimentally measured signals. Mechanical vibrations, air currents, fluctuations in laser power, and electrical interference picked up by the detector can all introduce noise and reduce data quality. Steps can be taken in the design of experimental instruments to reduce certain sources of noise. For example, floated optical tables dampen vibrations and enclosed beam paths reduce air currents. Dividing signals by a reference spectrum acquired on the fly reduces the detected shot to shot variation. However, despite our best efforts sources of noise can never be eliminated entirely and it is often necessary to take additional steps in post-processing to reduce noise contributions to data. An example of this approach discussed in the previous chapter is the windowing of time domain 2D IR data.

In this section, we apply spectral subtraction to reduce the noise features in 2D IR spectra. The concept underlying this approach is simple. If one can obtain a suitable estimate for the average noise spectrum in the absence of the desired signal, then the noise spectrum can be subtracted off from the measured spectrum to reduce its overall contribution.[1] This approach addresses only a certain type of noise. It must be additive and it must be stationary or slowly varying. In this case the measured signal $y(t)$ can be expressed

$$y(t) = x(t) + n(t) \tag{4.1}$$

where $x(t)$ is the signal of interest and $n(t)$ is the noise. Assuming that we are able to reasonably estimate or measure $n(t)$, the signal of interest can be isolated by rearranging eq 4.1 and introducing a parameter, α that scales $n(t)$.

$$x(t) = y(t) - \alpha n(t) \tag{4.2}$$

The value of α should reflect the level of confidence in the estimation of $n(t)$, with values less than one reflecting less confidence. Subtraction of the noise estimate can be performed in either the time or frequency domain.

In the context of 2D IR spectroscopy, this approach will only address a specific subset of the noise contributions to the overall spectrum. Specifically, variation in signal intensity over a hundreds of millisecond to second timescale due to slow fluctuations in laser power or detector operation. Assuming a step scan modality, this will amount to an essentially uniform offset across the 64 x 2 pixels of the mercury cadmium telluride (MCT) array for the signal averaged over the number of laser shots acquired at each time step. This type of noise at each pixel is characterized by the normalized mean square error relative to the mean intensity, $\langle I(pix, \tau_1)\rangle$ averaged across the total number of N shots acquired at each time step in τ_1.

$$\chi^2(pix,\tau_1) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left[I_i(pix,\tau_1) - \langle I(pix,\tau_1)\rangle\right]^2}{\sigma^2(pix,\tau_1)} \tag{4.3}$$

Where σ^2 is the noise variance. Such slowly varying noise averaged over successive laser shots is not addressed by the shot to shot referencing discussed in Chapter 3, since both the signal and reference will vary slowly together, resulting in a different offset at each time step relative to the other time steps sampled along the free induction decay (FID). When choosing whether or not to apply spectral subtraction to 2D IR data, it is important to properly characterize the noise and to evaluate whether or not this approach is appropriate. The mean square error in eq 4.3 approaches zero as the slowly varying noise approaches zero. Therefore large values of $\chi^2(pix,\tau_1)$ that are greater than unity suggest spectral subtraction may be applicable.

Fig. 4.1 shows an example of 2D IR data exhibiting noise characteristics that can be addressed by spectral subtraction. The sample is one of the DNA oligonucleotides discussed in Chapter 8 prepared at 2 mM concentration. The spectrum was recorded at 80 °C, well above the

melting temperature. The coherence time was stepped from -160 to 2500 fs in 4 fs steps. In the time domain each point along the FID measured at each pixel has a uniform offset, reflecting a slow variation in laser intensity relative to the timescale required to step the stage and average the 250 laser shots recorded at each time step. The FID tracked at 1662 cm^{-1} plotted in Fig. 4.1a illustrates how this slowly varying noise manifests in the time domain. An asymmetric Hann window with an exponent of 0.3 was applied to the data, as discussed in Chapter 3.

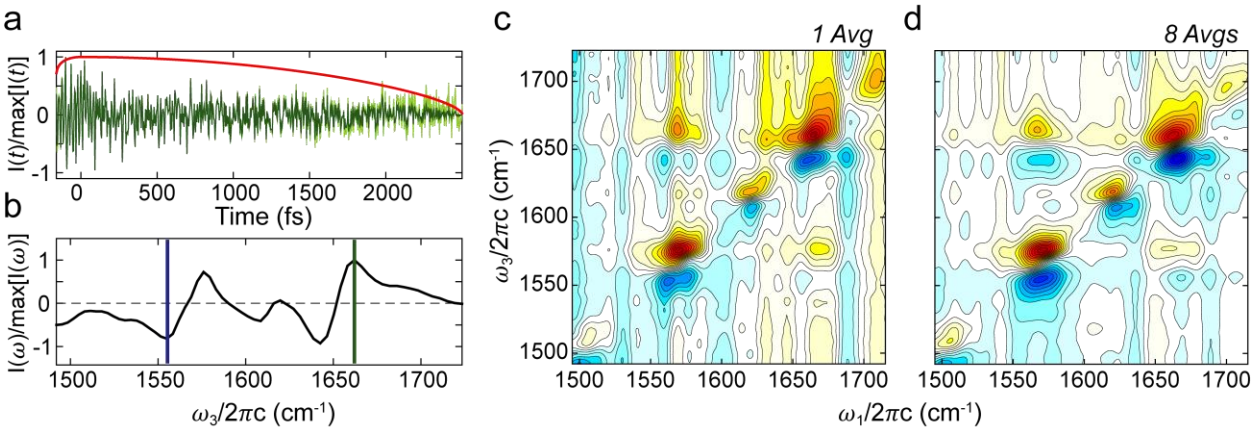

Figure 4.1: (a) The FID tracked at 1662 cm^{-1} plotted in light green. The Hanning window with an exponent of 0.3 is overlaid in red. The light green FID represents the data prior to windowing. (b) Projection of the 2D IR surface onto the detection axis. The green slice through the data corresponds to 1662 cm^{-1} while the blue slice corresponds to 1555 cm^{-1}. The corresponding FID's at these frequencies are plotted in Fig. 4.2. (c) A single average demonstrating the bands of noise parallel to the detection axis characteristic of slow variations in signal intensity. (d) The spectrum after averaging eight 2D IR surfaces. Although the data quality is improved relative to one average, there is a diminishing return with additional averaging and noise persists.

In the frequency domain, the random offsets due to slow intensity variation observed in the FID manifest as bands of noise at fixed ω_1 frequencies that run parallel to the detection axis. The single average of a 2D IR surface shown in Fig. 4.1c serves as an illustrative example. Assuming the source cannot be reasonably addressed, the standard approach when faced with this sort of

noise is to increase the number of shots acquired at each time step and/or to increase the number of 2D IR averages collected. In practice, in order for increasing the number of shots acquired at the kHz repetition rate of the regenerative amplifier to improve the data quality the noise variation must be rapid enough to be averaged out on a several ms timescale. Often this is not the case and increasing the number of shots acquired at each time step does not significantly reduce the contribution due to slowly varying noise. Averaging multiple 2D IR surfaces collected independently is usually more effective, as seen for the eight averaged surfaces in Fig. 4.1d. Although the averaged 2D IR surface is improved, there is a diminishing return as the number of averages is increased and the bands of noise discussed for the single average in Fig. 4.1c are still apparent. Even with eight averages, the level of noise remains problematic since the intensity and line shapes are still obscured to some degree.

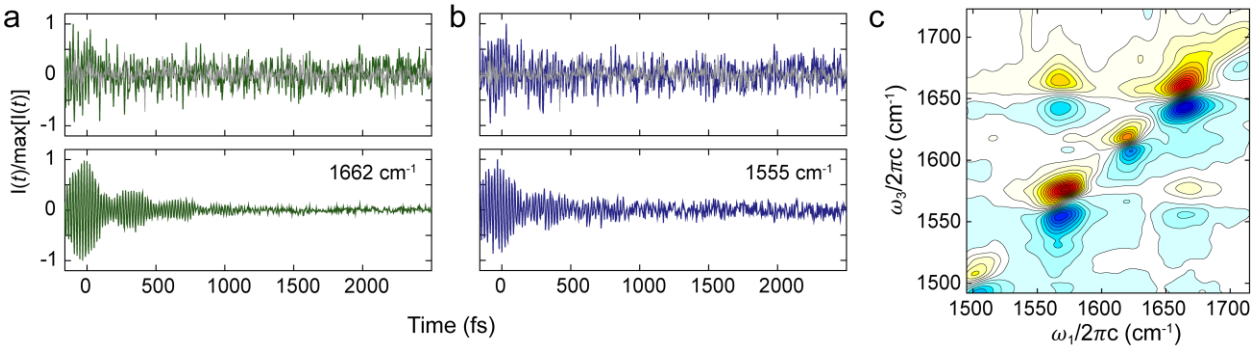

Figure 4.2: (a) The FID tracked at 1662 cm^{-1} plotted in green. The top panel shows the normalized FID prior to spectral subtraction. The estimated noise spectrum is overlaid in gray. The bottom panel contains the FID after subtraction of the noise spectrum. (b) The same quantities as plotted in panel a tracked at 1555 cm^{-1}. (c) The averaged 2D IR surface from Fig. 4.1d after spectral subtraction with $\alpha = 1$.

To apply spectral subtraction to the noisy data in Fig. 4.1, we must first obtain an estimate for the noise spectrum. Commonly the noise spectrum is measured on a separate channel

simultaneously with the signal. For a 2D IR measurement this approach is possible when a reference spectrum averaged across the number of shots collected at each time step is recorded. In this case the fluctuation of the average reference spectrum at each time step relative to the mean across the entire FID provides an estimate for the slowly varying noise spectrum. However, in many cases it is possible to obtain a reasonable estimate of the noise spectrum from the 2D IR data set directly. For a purely harmonic system, the stimulated emission/ground state bleach (GSB) and the excited state absorption (ESA) features in a 2D IR spectrum cancel and no signal is observed.[2] Therefore for a weakly anharmonic system, the projection of the 2D IR surface onto the excitation axis should essentially cancel the nonlinear signal, and any nonzero offset remaining will reflect the slowly varying noise.

This approach to estimating the noise spectrum is applied in Fig. 4.2. The FID tracked at two well-separated frequencies is plotted in Fig. 4.2a,b for 1662 cm^{-1} and 1555 cm^{-1}, respectively. The top panel shows the FID prior to spectral subtraction. Comparing the same time points across the two traces, particularly at longer times after the vibrational signal has relaxed, it is clear that there is a pattern to the noise that is conserved across the detection frequency range. This observation is consistent with a uniform offset across the pixel array at each time point sampled along the FID, identified above as characteristic of the slowly varying noise we are attempting to address. The noise spectrum estimated by taking the average across all pixels of the array at each time point is overlaid in gray and is normalized to 0.5 for illustration. After subtracting the noise spectrum from the time domain signal the FID is cleaned up considerably, as seen in the lower panel of Fig. 4.2a,b. For this example, $\alpha = 1$. The 2D IR surface is likewise improved. As seen in Fig. 4.2c, the bands of noise parallel to the detection axis have been suppressed.

For the example presented here, spectral subtraction effectively reduced contributions to the 2D IR spectrum due to slowly varying additive noise. However, as with any signal processing method employed to suppress noise, one should be mindful of potential influences on the signal of interest. For example, if the estimate for the noise spectrum is poor, subtraction could introduce artifacts or distort the desired signal. When applying the method for estimating the noise spectrum discussed above, one should always inspect the estimated noise spectrum to ensure that there is no residual FID character. There is of course no substitute for a stable instrument and a well aligned spectrometer. Every effort should be made to collect the highest quality data directly in the laboratory. But sometimes factors outside of one's control, such as a lemon of a regenerative amplifier or a failing detector, can introduce sources of noise that cannot be easily addressed. In such circumstances, spectral subtraction is one strategy for suppressing a particular subset of experimental noise.

4.3 Wavelet Transform Method for Simultaneous Background Removal and Noise Filtering in FTIR Spectroscopy

4.3.1 Introduction to the Discrete Wavelet Transform

Much of what can be learned from the infrared spectra of biomolecules can be derived, at least in part, from the linear spectrum. A Fourier transform infrared (FTIR) spectrum can be collected quickly and easily compared to a nonlinear measurement and on a commercially available instrument. Often insight derived from linear spectra guide the design of more labor and time intensive experiments. In fact, every project discussed in this thesis began with a careful set of FTIR spectra. The ability to collect, process, and interpret a quality linear spectrum should therefore not be taken for granted.

Ideally the solvent would not absorb any light in the frequency range of interest for the biomolecule. In this thesis, the focus is primarily on the 1500-1750 cm^{-1} range that contains the in-plane vibrations of the DNA nucleobases. Samples are therefore dissolved in D_2O to avoid absorption from the H_2O bend at 1645 cm^{-1}. Unfortunately, the D_2O spectrum is not completely flat in this frequency range either, but contains a low intensity absorption from the 1555 cm^{-1} D_2O bend-libration combination band. This solvent background must be subtracted off in post processing in order to isolate the absorption due to DNA. A common approach is to acquire a spectrum of the pure solvent and then to subtract this background spectrum off from the sample spectrum. In practice, it can be difficult to cleanly accomplish this subtraction since the path length varies sample to sample and the D_2O background varies with temperature. As a result a background spectrum must be acquired at each temperature and then scaled to match the sample spectrum. The baseline is shifted to zero and any residual slant in the baseline is corrected by fitting and subtracting a linear slope. Obtaining a set of quality FTIR spectra in this manner can be labor intensive, especially when sampling a large range of temperatures, pH's, or sample concentrations. A considerable level of user oversight is needed and the process requires a subjective assessment of the subtraction quality.

Ideally, this process could be automated in order to eliminate the need for collecting background spectra and to minimize user subjectivity. This section presents a method for automating the processing of FTIR spectra using a discrete wavelet transform (DWT) method. The basic approach presented here is adapted from methods developed for Raman spectroscopy, where simultaneous removal of the broad background due to the intrinsic fluorescence of the sample and the finely varying noise on top of the Raman signal is desirable.[3,4] A detailed discussion of wavelets and the wavelet transform (WT) are beyond the scope of this chapter. Several tutorials

are available in the literature.[5-7] In brief, the WT is analogous to the Fourier transform (FT) except wavelet functions are used as basis functions to decompose a signal rather than the sine and cosine functions used in the FT. The advantage of the WT is that wavelets are localized functions in both the time and frequency domain and it is therefore possible to isolate intervals in time that contain a localized characteristic frequency distribution. In contrast, the sine and cosine basis functions of the FT are delocalized and periodic in the time domain, and therefore the frequency content of a signal is assumed to be present at all points in time.

For the specific application discussed here, it is more useful to think of the WT as the projection of a signal onto a set of wavelet basis functions rather than as a transform between domain representations. Basis functions are obtained by scaling and shifting a prototype wavelet, which itself can be thought of as a bandpass filter. The basis functions therefore also behave like bandpass filters acting at different frequency scales. As a result the concept of scale is introduced in place of frequency when applying the WT.[6] In our case the frequency domain FTIR absorption spectrum will serve as the signal of interest. The features in the spectrum can be grouped into three categories according to their scale: highly localized variation due to noise, medium features corresponding to sample absorptions, and broad delocalized features originating from background solvent absorption. Applying the WT, the contributions to the FTIR spectrum across these scales can be separated, enabling a targeted and simultaneous removal of noise and background features from the spectrum.

4.3.2 Outline of the Algorithm and Selecting a Wavelet Basis

The most frequently used DWT algorithm is called the Mallat algorithm.[8] A selected prototype wavelet is used to generate low-pass and high-pass filters which are applied successively

at varying levels of scale. The number of filtering steps is referred to as the decomposition level. At the first level, a low-pass and high-pass filter are applied to the signal, resulting in a component containing smooth features and a component containing finer details, referred to as the approximation (cA) and detail coefficients (cD), respectively. At each successive decomposition level, a new pair of low-pass and high-pass filters at a coarser scale is applied to the cA from the last level of decomposition. In this way after N rounds of decomposition there are N detail coefficients but only a single approximation coefficient. After collecting the detail coefficients from each level of decomposition as well as the final approximation coefficient, the original signal can be reconstructed via an inverse DWT. The Mallat algorithm is outlined in Fig. 4.3.

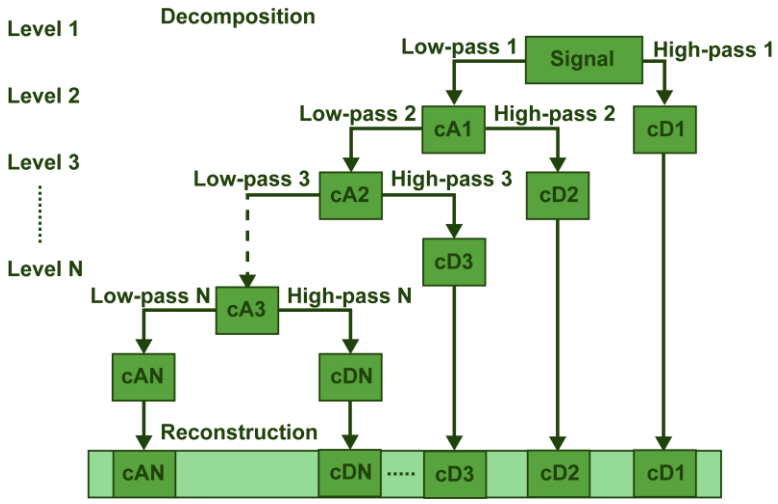

Figure 4.3: Diagram of the Mallat algorithm for performing the DWT. The approximation coefficient associated with decomposition level N is labeled cAN while the corresponding detail coefficient is labeled cDN.

When decomposing a signal in terms of a wavelet basis, the number of detail coefficients that are needed to faithfully reconstruct the original signal will be determined by how similar the prototype wavelet is to the original signal. In other words, the Mallat algorithm will converge more

quickly, requiring fewer levels of decomposition to represent the signal. Therefore one should select the prototype wavelet that best resembles the signal of interest. Fig. 4.4a shows an example of wavelets from four different families. Clearly some wavelets, such as the Haar and Daubechies examples plotted in the top two panels, bear little resemblance to FTIR signals and would be a poor choice for our purposes. In practice we and others have found that the best suited wavelet family for representing IR spectra in the liquid phase is the Symmlet family.[5] Several examples of wavelets from this family are plotted in Fig. 4.4b.

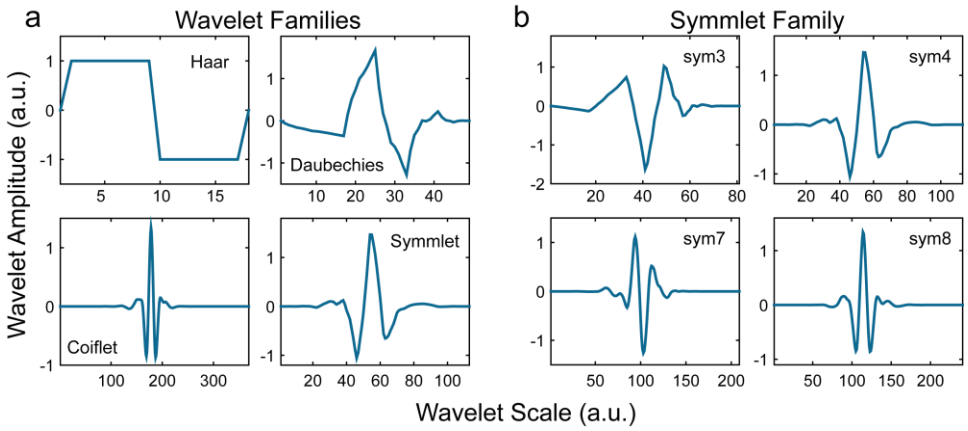

Figure 4.4: (a) Examples selected from several different wavelet families. (b) Various prototype wavelets from the Symmlet family, which is well suited for representing FTIR spectra in the liquid phase.

Ideally at this point one could select the appropriate prototype wavelet and decomposition level, apply the Mallat algorithm to the FTIR signal, and obtain a set of cD which contain the sample absorptions of interest plus the noise and a cA that contains the background. Those cD which correspond to noise could then be set to zero and the cA background could be subtracted off to remove solvent absorption. In reality this separation is not so straightforward and additional constraints are required in order to achieve a reasonable isolation of the background from the

sample contributions in the spectrum. One approach is to iteratively apply the DWT to a refined signal such that cA converges towards the background spectrum.[4] This refinement is achieved by setting everything in the spectrum that lies above cA after each iteration to the value of cA at those frequency points. The resulting "threshold spectrum" is then fed back into the DWT in place of the previous spectrum, and the process continues until all traces of sample absorption are removed from the cA.

An example of this approach is demonstrated in Fig. 4.5 for a DNA oligonucleotide prepared at 2 mM concentration and at 80 °C. The prototype wavelet is Symmlet 8 and six levels of decomposition are performed at each iteration. Fig. 4.5a shows the FTIR spectrum prior to any processing. Notice that the DNA spectrum in the 1500-1750 cm^{-1} range is riding on top of a broad and featureless absorption due to the D_2O bend-libration combination band. After one iteration of the DWT, the cA plotted in Fig. 4.5a results. Character from DNA absorption is clearly still present in the approximation coefficient, and as a consequence the cA runs through the middle of the original spectrum. Applying the thresholding method discussed above, all intensity that lies above cA is set to the value of cA at those frequency points, as shown in Fig. 4.5b. After feeding this threshold spectrum back into the DWT and repeating this process for 40 iterations, the approximation coefficient closely resembles the D_2O background. Subtracting off this extracted background spectrum from the original spectrum plotted in Fig. 4.5c results in the background corrected spectrum plotted in Fig. 4.5d. In addition to thresholding the spectrum at each DWT iteration, explicitly identifying frequency ranges which lack any absorbance due to DNA and excluding these from the analysis by resetting values to the original spectrum after each iteration can speed up the process and improve the quality of the extracted background spectrum.

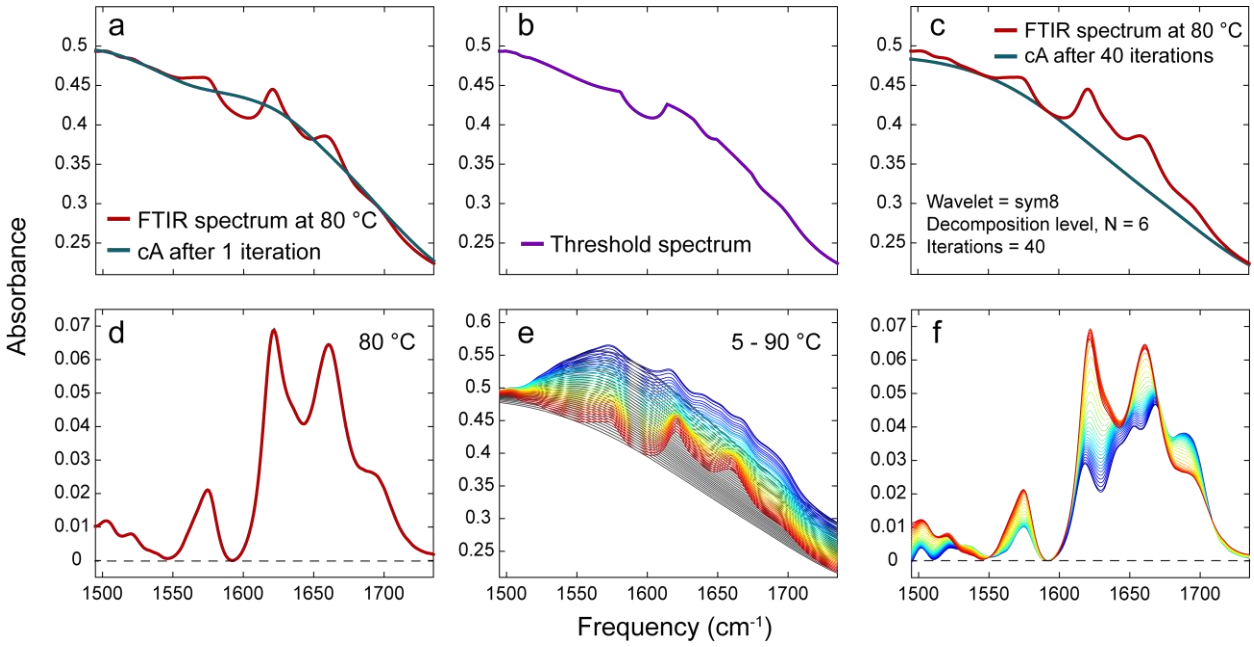

Figure 4.5: (a) FTIR spectrum of 2 mM DNA oligonucleotide at 80 °C in red and the approximation coefficient after one iteration of DWT in blue. (b) The threshold spectrum obtained by setting all points in the spectrum in panel a that lie above cA to the value of cA at that point. (c) The cA after 40 iterations of DWT closely resembles the D_2O background. (d) The processed spectrum at 80 °C resulting from subtracting the extracted background from the original spectrum. (e) Applying the same parameters across a set of FTIR spectra acquired from 5-90 °C (blue to red) resulting in the extracted background spectra plotted in gray. (f) Full set of background corrected spectra corresponding to the unprocessed data in panel e.

Some initial user oversight is still required in selecting a prototype wavelet, decomposition level, number of thresholding iterations, and possible background regions to be excluded from thresholding, but we have found that once a set of reasonable parameters is determined they apply equally well for related samples. The processing of large data sets can therefore be automated. For example, the same parameters used for the 80 °C spectrum in Fig. 4.5a apply equally well across a set of temperature dependent FTIR spectra acquired from 5-90 °C. The spectra before any further processing are shown in Fig. 4.5e running from low temperature in blue to high temperature in red along with the background spectrum extracted at each temperature in gray. The resulting set of

background corrected spectra is shown in Fig. 4.5f, indicating that the method is robust across a wide temperature range and insensitive to considerable reshaping of the spectrum as the oligonucleotide duplex melts. Therefore the DWT based method discussed here can be applied across large sets of related data without the need for user input.

In the present example, the signal to noise ratio is high enough to assume that all of the cD are dominated by DNA absorption features. For the FTIR data presented in this thesis there is generally no need for any additional noise removal and the DWT method is used only for background extraction. However, an example of simultaneous noise filtering and background removal is discussed in Section 4.3.4 since this property is a powerful added bonus of the method, particularly when working with problematic samples such as protein films or lipid vesicles.

4.3.3 Extracting the Solvent Background for use as a Thermometer

In addition to removing solvent absorption, the extracted background spectrum can also be used as an intrinsic probe to characterize the sample conditions assuming that the dependence of the solvent spectrum on the experimental variable is known. For example, the temperature dependence of the D_2O bend-libration combination band is shown in Fig. 4.6a, where a spectrum of pure D_2O has been acquired in ~1 °C steps from 4-90 °C. The peak frequency, line width, and intensity all change with temperature. In the laboratory a series of temperature dependent FTIR spectra are collected by stepping the temperature of a recirculating chiller that is connected to a brass jacket surrounding the sample cell. The bath temperature of the chiller does not match the temperature of the sample due to heat exchange with the environment. To measure the sample temperature as a function of the bath set point, a thermocouple is affixed to the center of the sample

window. This temperature calibration must be performed independently of measuring the spectrum.

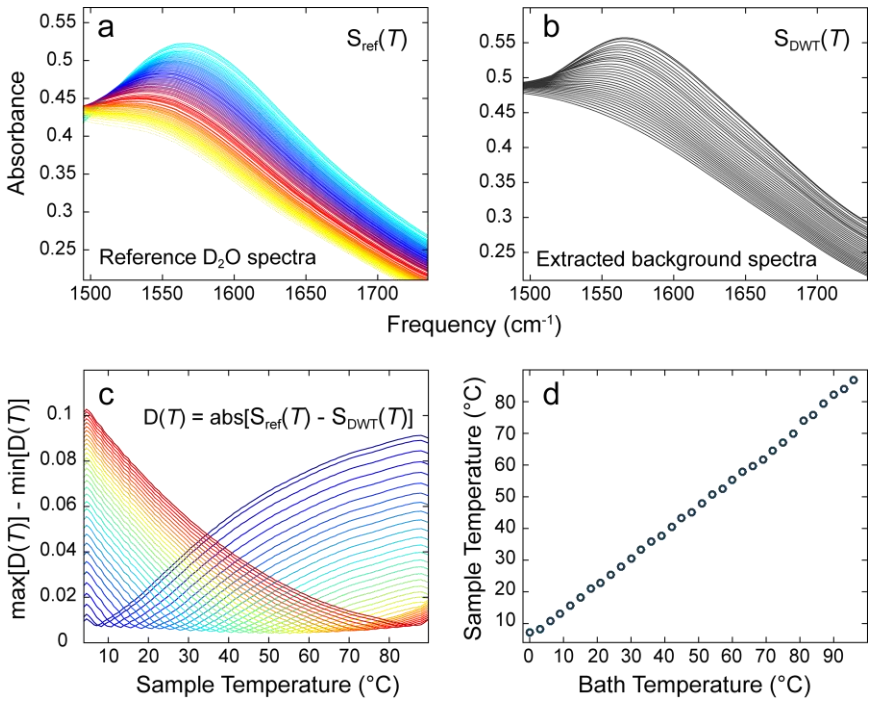

Figure 4.6: (a) The reference set of D_2O spectra spaced ~1 °C apart from 4-90 °C running from cyan to yellow with increasing temperature. (b) The extracted background spectra from the example in the previous section. (c) The difference between the maximum and minimum point in $D(T)$ plotted as a function of reference temperature. Each color coded curve corresponds to this quantity calculated for a single extracted background spectrum. The minimum point in each curve gives the temperature of the corresponding sample. (d) The mapping between the bath set point and the sample temperature obtained by checking the extracted background spectra against a known reference set of D_2O spectra.

Alternatively, the extracted background spectrum from the DWT method can be checked against a set of reference D_2O spectra at known temperatures. The set of D_2O spectra in Fig. 4.6a will serve as the reference set. Fig. 4.6b shows the extracted background spectra from the example in the previous section. In principle, the extracted background should be scaled to account for path length variation and fit against the reference set to determine the best match. However, we have

found that a much faster and simpler approach works reasonably well. To assign the temperature at which a spectrum was acquired, the extracted background is subtracted off from each reference spectrum. This produces a difference spectrum, $D(T)$ for each reference temperature. The best temperature match between an extracted background and the reference set should result in the flattest difference spectrum.

In this case a convenient one-dimensional proxy for the flatness of each $D(T)$ is the difference between the maximum and the minimum of the absolute value of $D(T)$. This quantity is plotted as a function of reference sample temperature in Fig. 4.6c for each extracted background spectrum. The color coding in the figure corresponds to the processed spectra in Fig. 4.5f. The minimum point in each curve indicates the temperature at which the corresponding sample spectrum was acquired. Fig. 4.6d shows how the bath set points map to the sample temperatures determined by checking the extracted background spectra against a set of reference D_2O spectra in this way.

4.3.4 Simultaneous Noise Filtering

The ability to simultaneously remove noise in addition to background features is a major advantage of DWT based processing methods. As an example of simultaneous background and noise removal, we consider the temperature dependent FTIR spectra of the antibiotic polypeptide gramicidin D in a dipalmitoylphosphatidylcholine (DPPC) lipid bilayer between 10-70 °C. Samples containing extended structures such as lipid vesicles or protein films are often subject to multiple internal reflections within the medium that can interfere and produce ripples on the spectrum. Fig. 4.7a shows the set of temperature dependent FTIR spectra before processing plus the extracted background spectra from the DWT method.

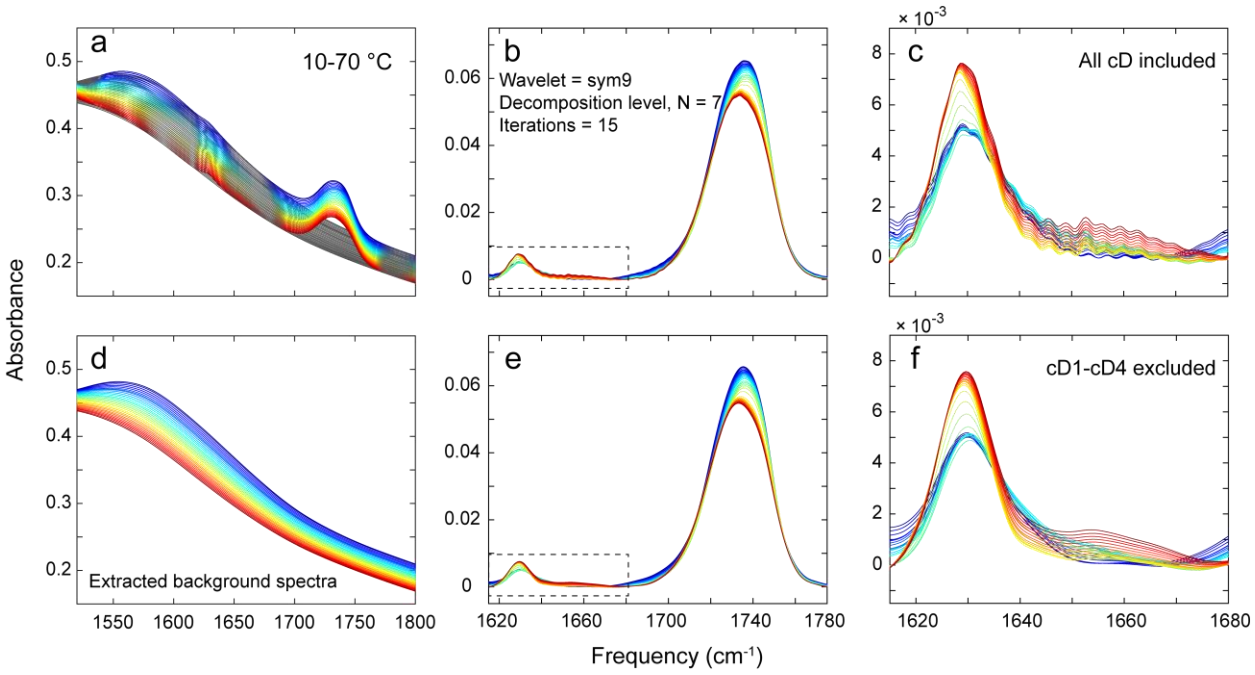

Figure 4.7: (a) Temperature series of gramicidin D in DPPC lipid bilayers ramped from 10-70 °C in ~2 °C. The DWT extracted background spectra are plotted in gray. (b) The processed spectra with all detail coefficients included. (c) Zoomed in view of the gramicidin D features showing interference ripples on top of the spectrum. (d) The set of extracted background spectra from panel a. (e) The processed spectra with the four highest frequency detail coefficients set to zero. (f) Zoomed in view analogous to panel c for the smooth spectra highlighting the removal of noise.

The optimal parameters for processing this set of spectra, listed in Fig. 4.7b, are different from the DNA oligonucleotide discussed above. For the purposes of this example we are not concerned with the assignment of the features in the spectrum, but this has been discussed in detail elsewhere.[9] Zooming in on the low intensity gramicidin D features in Fig. 4.7c highlights the considerable noise that is riding on top of the spectrum. Relative to the peptide absorptions, these ripples are sharply varying along the frequency axis and comparatively localized. This separation in scale suggests that the DWT will place these features into distinct detail coefficients. In that case those cD which contain noise can be filtered out before performing an inverse DWT to reconstruct the spectrum. Fig. 4.7e shows the spectra that result when the four highest frequency

detail coefficients are excluded. Once again zooming in on the low intensity gramicidin D features, Fig. 7f shows the considerable improvement in the quality of the spectra after removing the high frequency detail coefficients containing noise. Furthermore the original line widths of the peptide absorptions are well preserved, which is an advantage of the DWT method over other smoothing or filtering approaches which can broaden line shapes.

4.4 Spectral Component Reconstruction through Entropy Maximization

4.4.1 The Maximum Entropy Principle

The remaining two sections of this chapter concern maximum entropy (MaxEnt) methods for resolving pure components in the spectra of mixtures and for providing an additional constraint on the otherwise ill-posed inverse Laplace transform linking the time and rate domains. Although these applications are quite different, they are united by their reliance on the principle of maximum entropy. Stated simply, the MaxEnt principle dictates that when one is making inferences based on incomplete information, one should draw from the probability distribution that corresponds to the maximum entropy permitted by the information which is available.[10] The entropy in question is the Shannon-Jaynes or information entropy, of which the concept of thermodynamic entropy familiar to chemists has been argued a subset.[11] The Shannon-Jaynes entropy is a direct reflection of how much information is contained in each event, observation, character, or whatever signal is received, x_i out of all those possible, X and takes the form of eq 4.4.

$$H(X) = -\sum_i p(x_i) \ln p(x_i) \tag{4.4}$$

Where $p(x_i)$ is the probability associated with observing outcome x_i. As a simple example, consider a standard six sided die. The information entropy is maximized when the probability of rolling any side is equal, or 1/6. As soon as a bias is introduced, for example weighting the die such that rolling

a one is more probable, the entropy is reduced. To illustrate this, Fig. 4.8 plots the information entropy calculated as a function of an increasingly biased die. Initially all six outcomes are equally likely, as seen in the first inset showing a histogram of $1x10^6$ rolls of the die.

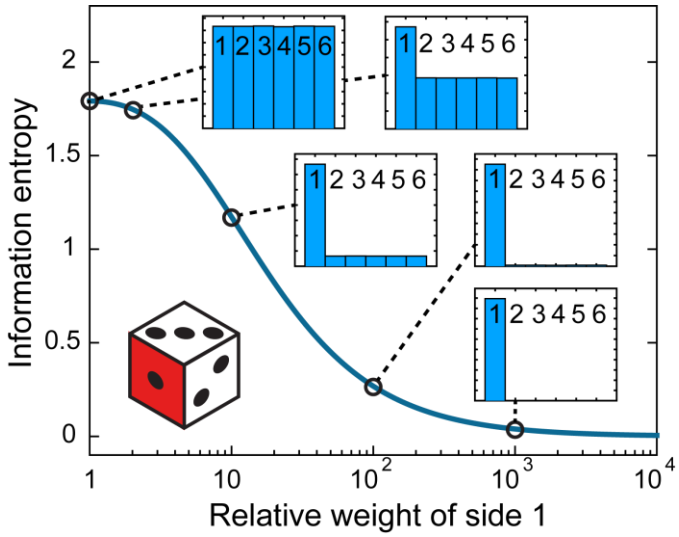

Figure 4.8: The information entropy calculated as a function of an increasingly biased die. Insets show the histogram of outcomes after $1x10^6$ rolls at each of the corresponding points along the entropy curve.

Biasing the die such that rolling a one has twice the weight of any other side results in a small reduction in entropy and after $1x10^6$ rolls the second histogram along the curve results, indicating that a one was rolled approximately twice as many times relative to any of the other possible outcomes. Following along the entropy curve, the die is further biased such that rolling a one is weighted ten times and then one hundred times over any other individual outcome, as reflected in the corresponding histograms and falling information entropy. Arriving at the last histogram where a roll of one is weighted a thousand to one over any other roll, the information entropy is reduced to essentially zero. At this point there is not much point in rolling the die since the act of doing so conveys no new information. The outcome is known with near certainty beforehand.

The point of this simple example is to illustrate the general principle that the information entropy is always maximized for the flattest probability distribution, or in other words the one in which all possible outcomes are equally probable. In the absence of any other information about the underlying probability distribution, this is therefore the most reasonable assumption since to assume anything otherwise would amount to artificially imposing bias. For the methods presented in the following sections, experimental observations provide prior knowledge. Generally we will see that introducing and maximizing an information entropy term has the effect of smoothing and localizing the features in the spectrum, population distribution, rate distribution, etc. as dictated and constrained by the information that we do know, that is, by the experimental data.

4.4.2 Construction of the Objective Function

The first maximum entropy method we will discuss was initially developed for the purposes of reconstructing the spectrum associated with each pure component present in a mixture that cannot otherwise be separated by physical or chemical means.[12] This approach relies on a combination of singular value decomposition, entropy maximization, and simulated annealing to arrive at a set of pure component spectra and their corresponding population distributions as a function of some experimental variable, such as reactant concentration, temperature, pH, voltage, or time. As we will see, the method can be generalized in the sense that it is not limited to separating chemical species in mixed solutions, but can resolve any measurably distinct subcomponent of the system, whether these are the various protonation states of a titratable group or the distribution of folded states in an ensemble of biomolecules. As long as the selected experimental observable is sufficiently sensitive to the presence or absence of some constituent of

the system, its contribution to the overall signal can be distinguished and characterized as a function of the experimental variable.

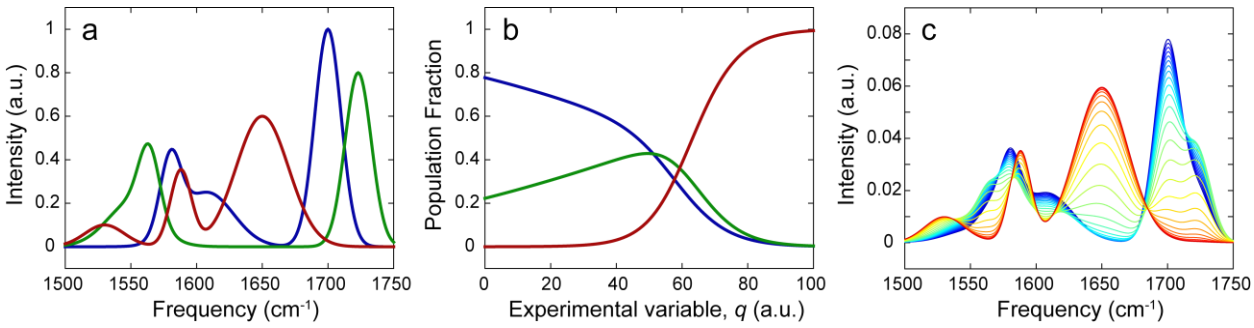

Figure 4.9: (a) The starting pure component spectra for the modeled spectra. (b) Corresponding population fractions as a function of a general variable, *q*. (c) Simulated set of mixture spectra as a function of *q* obtained through a linear combination of the component spectra weighted by the population profiles. The spectra run from low *q* in blue to high *q* in red.

For the purposes of illustrating the method, a set of model spectra $A(q)$ that vary as a function of some general variable q across the 1500-1750 cm^{-1} range will be considered. The three distinct spectroscopic contributions are shown in Fig. 4.9a and their corresponding population fractions as a function of the experimental variable, q are shown in Fig. 4.9b. It is important to note that the component spectra and population profiles do not necessarily lend themselves directly to a straightforward physical interpretation, but are simply a reflection of the resolvable information content contained within the measured signal. The challenge of interpreting the spectrum with regard to an underlying distribution of structures or states remains, but determining the number of resolvable component basis spectra needed to represent the data, their peak patterns and intensities, and how they are weighted as a function of an experimental observable can often provide valuable clues. The component spectra in this example were specifically designed to be challenging to separate in the sense that there are several highly congested frequency ranges with

overlapping absorptions. A simulated spectrum as a function of q is obtained by a linear combination of the component spectra weighted by their corresponding population at each value of q. As can be seen in Fig. 4.9c, this results in a fairly complicated dependence on q, with several overlapping features, subtle intensity changes, and frequency shifts. The simulated spectra run from low q in blue to high q in red. Elsewhere in this thesis, q is either temperature or pH and the measured signal is the FTIR spectrum of a DNA nucleotide or oligonucleotide in the 1500-1750 cm^{-1} frequency range.

To illustrate the ability of the MaxEnt method to resolve component basis spectra and population profiles, the set of spectra depicted in Fig. 4.9c is fed into the method. Since we know by design the underlying spectroscopic components and populations used to produce the simulated data set, evaluating the quality of the output is straightforward in this model case. The first step is to decompose the data through a singular value decomposition (SVD). Ref. 13 is a good introduction to the application of SVD to the analysis of multicomponent spectra. If the data in Fig. 4.9c are arranged into an $\omega \times q$ matrix such that each column contains the corresponding spectrum at each value of q sampled, then any single column of this matrix contains a spectrum depending on only a single value of q. Conversely, each row contains the behavior of a single frequency across the entire range of q. In this way, we can define a frequency space corresponding to the columns and a q space corresponding to the rows of the data matrix. Applying SVD to this matrix results in

$$A_{\omega \times q} + E = U_{\omega \times \omega} S_{\omega \times q} V^{T}_{q \times q} \tag{4.5}$$

where the left hand side has been expressed as the sum of a matrix $A_{\omega x q}$ containing the measured signal and a matrix E containing the noise. The right hand side of eq 4.5 is the result after SVD, which factorizes the data matrix into the product of three matrices U, S, and V. The columns of the

U and *V* matrices consist of an orthonormal set of frequency and *q* space vectors, respectively. The *S* matrix is a diagonal matrix whose nonzero elements are called the singular values. The singular values dictate how the vectors contained in *U* and *V* should be weighted to reconstruct the original data. Fig. 4.10 shows the first five SVD vectors obtained by decomposing the model data.

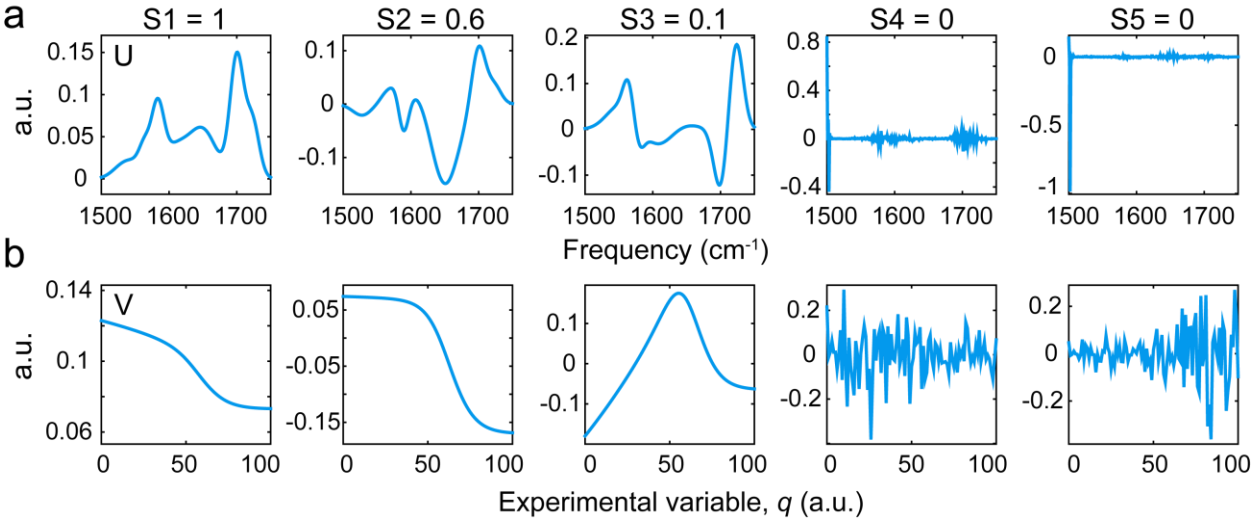

Figure 4.10: The first five vectors in (a) *U* and (b) *V*. The corresponding normalized singular value is indicated above each panel. The singular values quickly drop off, with the SVD vectors beyond the third component containing noise. Above S3 the S values are miniscule and round to zero.

The majority of the pertinent information is contained within the first few SVD vectors, with the remaining vectors containing noise. Therefore we can define a threshold *s*, and truncate the factorized matrices such that only those vectors which contain signal information are retained. From inspection of Fig. 4.10, $s = 3$ for the modeled set of spectra, as one would expect. After truncation eq 4.5 becomes

$$A_{\omega \times q} = U_{\omega \times s} S_{s \times s} V^{T}_{s \times q} \tag{4.6}$$

So far the data have been factorized into frequency and q space matrices and some of the noise has been discarded, but in general the raw SVD components cannot be interpreted physically, as seen by the seemingly arbitrary sign and intensity in Fig. 4.10. However, the separation of the data into spectroscopic components that share a similar q dependence serves as an excellent starting point for the MaxEnt method. A transformation matrix $T_{s\times s}$ is introduced that will remix the SVD vectors into a matrix of component spectra $a_{\omega\times s}$ and concentration profiles $c_{s\times q}$ according to

$$A_{\omega\times q} = U_{\omega\times s} S_{s\times s} T^{-1}_{s\times s} T_{s\times s} V^{T}_{s\times q} \tag{4.7}$$

$$a_{\omega\times s} = U_{\omega\times s} S_{s\times s} T^{-1}_{s\times s} \tag{4.8}$$

$$c_{s\times q} = T_{s\times s} V^{T}_{s\times q} \tag{4.9}$$

Going forward we construct a physically motivated objective function, F that will be minimized with respect to varying the elements of the transformation matrix $T_{s\times s}$. The objective function is composed of three terms that will be motivated in turn. The first is an information entropy term

$$H = -\sum_{\omega}\sum_{s} h_{\omega s} \ln(h_{\omega s}) \tag{4.10}$$

$$h_{\omega s} = \frac{|a'_{\omega\times s}|}{\sum_{\omega} |a'_{\omega\times s}|} \tag{4.11}$$

where $h_{\omega s}$ is defined in terms of the normalized absolute value of the derivative of the estimated spectrum with respect to frequency at each iteration of the optimization. The first derivative with respect to frequency is denoted by a prime in eq 4.11. As defined, the effect of minimizing H will smooth and localize the features in the derivative of the estimated spectra, which is equivalent to maximizing the information entropy since introducing any feature to the component spectra unsupported by the original data is disfavored. In the limit of zero prior information, meaning nothing is known regarding how much absorbance should be measured at a given frequency,

minimizing H will result in a set of completely flat component spectra, as one would expect for the case of maximum information entropy.

The second term in the objective function is a penalty term, P that biases the optimization towards both positive absorbance and population, which is what one would expect for physically meaningful results. The penalty function takes the form

$$P = \sum_{\omega}\sum_{s} W(a_{\omega\times s})a_{\omega\times s}^2 + \sum_{s}\sum_{q} W(c_{s\times q})c_{s\times q}^2 \tag{4.12}$$

where W is a weighting function that has the behavior

$$W(y) = \begin{cases} 0 & \text{if} \quad y \geq 0 \\ 1 & \text{if} \quad y < 0 \end{cases} \tag{4.13}$$

Therefore positive or zero values do not receive any penalty, but negative values are penalized by the square of their value, meaning that the severity of the penalty depends quadratically on the magnitude of the negative amplitude.

The final term is a dissimilarity term, D that serves as an added heuristic constraint rooted in the assumption that component spectra should be distinct from one another. Minimizing the inner product between the remixed frequency space vectors in $a_{\omega\times s}$ biases the optimization towards dissimilar component spectra. This term can be defined with respect to all of the components or with respect to a selected subset. The indices i and j in eq 4.14 can therefore refer to any two column vectors in $a_{\omega\times s}$ and any number of similar products between the vectors in $a_{\omega\times s}$ can be added to the objective function at the user's discretion.

$$D = \sum_{\omega} \hat{a}_i \cdot \hat{a}_j \tag{4.14}$$

Unlike the other terms in the objective function, the application of D requires oversight and should therefore be treated more carefully. The decision to include D and the form that it takes should reflect the researcher's knowledge of the system and level of experience. For instance, perhaps

two components in a mixture are chemically similar and one would expect their spectra to be similar as a result while a third component is chemically distinct. In this case, it would not make sense to impose a dissimilarity constraint between the first two components, but including the inner product between the third component and the first two vectors in the objective function could be supported. Overweighting or defining too robust of a dissimilarity constraint will lead to artifacts and over-resolution of the data. When in doubt or when not supported by some external prior knowledge or reasonable assumption about the system under study, *D* should be excluded from the objective function altogether.

4.4.3 Simulated Annealing Optimization

The objective function, *F* is the sum of the entropy, penalty, and dissimilarity terms described above. The relative contribution due to the penalty and dissimilarity terms are each independently set by the weighting parameters λ and γ, as indicated in eq 4.15.

$$F = H + \lambda P + \gamma D \tag{4.15}$$

In practice, we have found that setting λ = 100 and setting γ = 0 to exclude the dissimilarity term altogether are reasonable weightings when working with FTIR spectra of DNA. The task of minimizing *F* with respect to the elements of the transformation matrix T_{sxs} is complicated by the fact that the parameter space is large and there are many local minima in which the optimization can become trapped. Therefore an inbuilt MATLAB simulated annealing (SA) algorithm is used to minimize *F* rather than an algorithm relying on steepest descent. SA algorithms mimic the process of slowly cooling a heated metal to decrease defects, or, for that matter, the process of slowly cooling a heated solution containing complementary DNA strands such that the lowest energy, most highly hybridized, ensemble results.

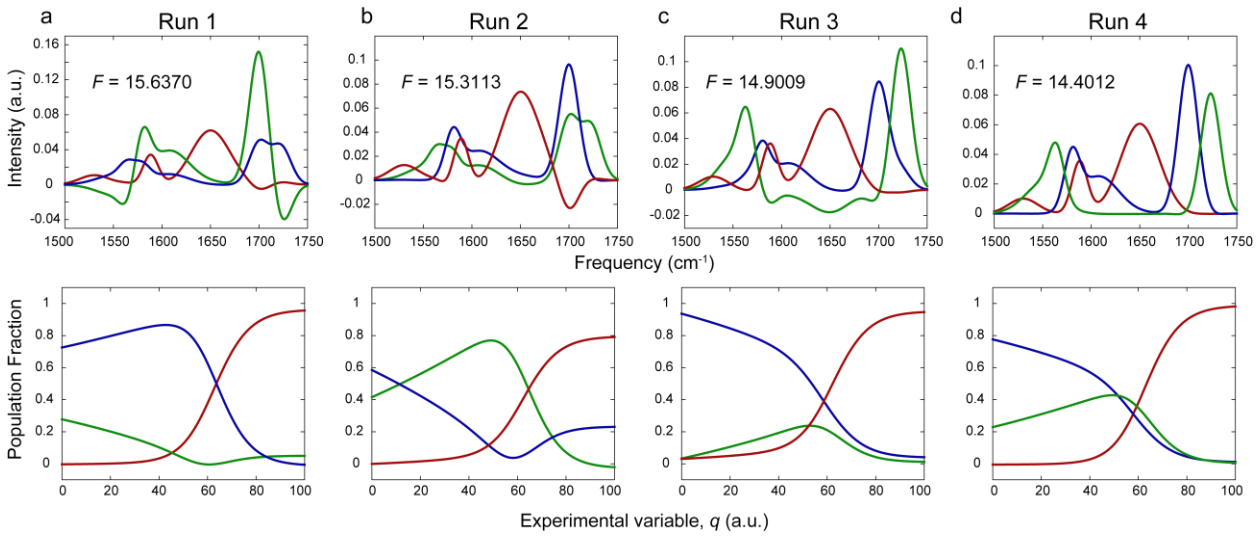

Figure 4.11: (a-d) The results of four separate minimizations of *F* for the model set of data in Fig. 4.9c arranged in order of descending values of *F*. The outcome with the smallest value of *F* in panel d is an accurate recovery of the original basis spectra and population profiles plotted in Fig. 4.9a,b.

In the SA algorithm, a "thermostat" is defined that starts at some high temperature. Each iteration generates a new set of points in parameter space that are selected randomly from a probability distribution that is proportional to the current temperature. All points that lower *F* are accepted, but there is also a finite probability of accepting points that raise *F* as well. This potential for accepting points that raise *F* makes the SA algorithm more robust and less likely to become trapped in local minima, especially at early iterations. With each iteration, the thermostat is lowered and the optimization converges towards a solution. It is important to rerun the minimization enough times to sufficiently explore the parameter space and to gain a sense for the value of the global minimum. Fig. 4.11 shows the outcome from four separate runs of the MaxEnt method applied to the model set of data in Fig. 4.9c. The value of $\lambda = 100$ and the value of $\gamma = 0$ such that no dissimilarity term is included in *F*. The initial guesses were held constant across the four runs. The outcomes in Fig. 4.11a,b show varying degrees of negative absorbance and both the

component spectra and population profiles are clearly nonphysical. As the value of *F* decreases, the results improve. For the lowest value of *F* in Fig. 4.11d the reconstructed component spectra and population profiles are an accurate recovery of the original basis spectra and population weights used to create the modeled data set in the first place, as can be seen by comparison to Fig. 4.9. The four outcomes plotted here were observed after rerunning the optimization many times and identifying common solutions. The highest fidelity reconstruction should correspond to the global minimum in *F*.

4.5 Maximum Entropy Guided Inverse-Laplace Transform

4.5.1 Rate Domain Representation of Kinetic Data

A common approach when analyzing kinetic data of the sort acquired in the temperature jump (T-jump) experiments discussed in this thesis is to fit time traces to a sum of discrete exponentials. The number of exponentials and the magnitude of their time constants can then be related to an appropriate model to extract rate constants and to describe the kinetics of the system.[14] Often when studying the unfolding of biomolecules, the observed kinetics are not so well behaved. Stretched exponential and other non-exponential behavior can result. In such cases, selecting the most appropriate and physically meaningful functional form with which to fit the data is a challenge. One approach is to obtain a representation of the data in a space of decay rates, thereby circumventing the need to fit time traces to an assumed functional form. Transformation between the rate and time domains is achieved by the Laplace transform.

$$I(t) = \int_0^\infty g(\lambda) e^{-\lambda t} d\lambda \tag{4.16}$$

where $g(\lambda)$ is the distribution of rate constants corresponding to the process $I(t)$. In practice data are collected in the time domain, so one would like to determine the distribution of rates which

satisfy eq 4.16 and describe the experimental data $I(t)$, which amounts to a numerical inverse Laplace transform (iLT) of the measured signal. Unfortunately this is known to be an ill-conditioned problem when applied to a signal containing non-zero noise.[15] As a result a large number of rate distributions will describe the data equally well and without some additional constraint it is impossible to arrive at a unique rate domain representation of the measured signal. The maximum entropy method (MEM) is one strategy for applying an additional constraint on the transform. A Shannon-Jaynes entropy term is added to the objective function such that a minimal number of additional assumptions are imposed on the data.

4.5.2 Minimization of the Objective Function

The MEM-iLT used in this thesis is an adaptation of the method outlined in Ref. 16 and we adopt similar notation here. Since the experimental observation window in T-jump IR measurements spans many decades in time, it is convenient to work in a logarithmic rate space. A discrete representation of eq 4.16 is given by

$$I(t) = \sum_{j=1}^{N} f_j e^{-\lambda_j t} \Delta(\log \lambda_j) \tag{4.17}$$

where N corresponds to the total number of time points sampled. The spacing $\Delta(\log\lambda_j)$ is set by uniformly dividing the interval defined by the logarithm of the inverse of the smallest and largest experimentally sampled time points into N equal bins. The set of f_j's constitute the rate distribution of interest. As discussed above, an additional constraint on the discrete numerical iLT is required to achieve a unique solution. Therefore an information entropy term is introduced.

$$H = -\sum_{j=1}^{N} f_j \left[\ln\left(\frac{f_j}{F_j} \right) - 1 \right] \tag{4.18}$$

The set of F_j's define the prior distribution which could in principle be used to incorporate any previous knowledge of the rate distribution. Throughout this thesis we will always assume no additional prior information is known about the rate spectrum and the F_j are set to a uniform low amplitude on the order of 10^{-4}. The set of F_j will also serve as the initial guess for the set of f_j. The form of the entropy function in eq 4.18 requires that the rate amplitude is positive, but in many instances measured signals can contain both positive and negative amplitude, such as for the T-jump measurements of interest here. Therefore the rate amplitude f_j is expressed as the difference between two positive rate amplitudes f_j^p and f_j^n, taking the form $f_j = f_j^p - f_j^n$. This difference is substituted for f_j in eq 4.17, but two separate entropy terms for the positive and negative rate amplitudes are defined in analogy to eq 4.18. The Laplace transform is incorporated into the objective function through the normalized mean square error between the experimental and the model trace, given by

$$\chi^2 = \frac{1}{N}\sum_{k=1}^{N}\frac{\left[I_f(t_k) - I_e(t_k)\right]^2}{\sigma_k^2} \tag{4.19}$$

where the subscripts *e* and *f* denote the experimental and fit values of $I(t)$, respectively and σ_k^2 is the noise variance associated with the k^{th} data point. When applied to the T-jump data in this thesis, an estimate for σ_k is obtained by assuming that a noise-free kinetic trace should vary smoothly and determining the standard deviation of the residual of a smoothing spline fit to the time trace at each frequency slice. The objective function takes the form

$$Q = H - \eta\chi^2 \tag{4.20}$$

where η is a Lagrange multiplier selected to satisfy the constraint that $\chi^2 = 1$. In practice the method is implemented in MATLAB (R2014b) and the function $-Q = \eta\chi^2 - H$ is minimized using the built in unconstrained multivariable function minimization routine with the trust-region algorithm. The

analytical gradient and Hessian of $-Q$ is supplied explicitly to the solver in order to improve optimization speed and reliability. The initial value of η is set to the mean of σ_k and is iteratively increased by a small amount after the objective function is minimized at each successive step. The optimization terminates once χ^2 approaches one and the resulting final set of f_j gives the rate domain representation of the data. When applying the MEM-iLT to multiple kinetic traces, such as the time traces recorded at each frequency slice through a T-jump data set, the rate spectrum at a given frequency does not depend on the others and the method is therefore amenable to parallelization, which cuts down on the computation time considerably.

4.5.3 Resolution of the MEM-iLT with Increasing Noise

To illustrate the application of the MEM-iLT method and to demonstrate the sensitivity of the rate spectrum to the form of the kinetic trace as well as the signal to noise ratio, the method is applied to modeled kinetic data. The modeled data is limited to single frequency time traces to simplify the discussion, but it is straightforward to generalize the conclusions drawn from these 1D examples to the 2D rate maps presented elsewhere in this thesis. The first set of model data considers the influence of the noise variance on the rate spectrum when there are multiple processes with varying amplitude and varying separation in timescales. The simplest scenario would consider an isolated single exponential process. The corresponding rate distribution after MEM-iLT is a Gaussian profile whose width is dictated by the noise variance in eq 4.19. In the limit that σ_k^2 approaches zero, the width of the rate distribution approaches zero and the rate domain representation of a single exponential process approaches a delta function centered at the inverse time constant of the exponential, as one would expect. In real kinetic measurements

multiple processes are often present. It is therefore important to establish the limitations of the MEM-iLT with respect to resolving multiple processes at different levels of noise.

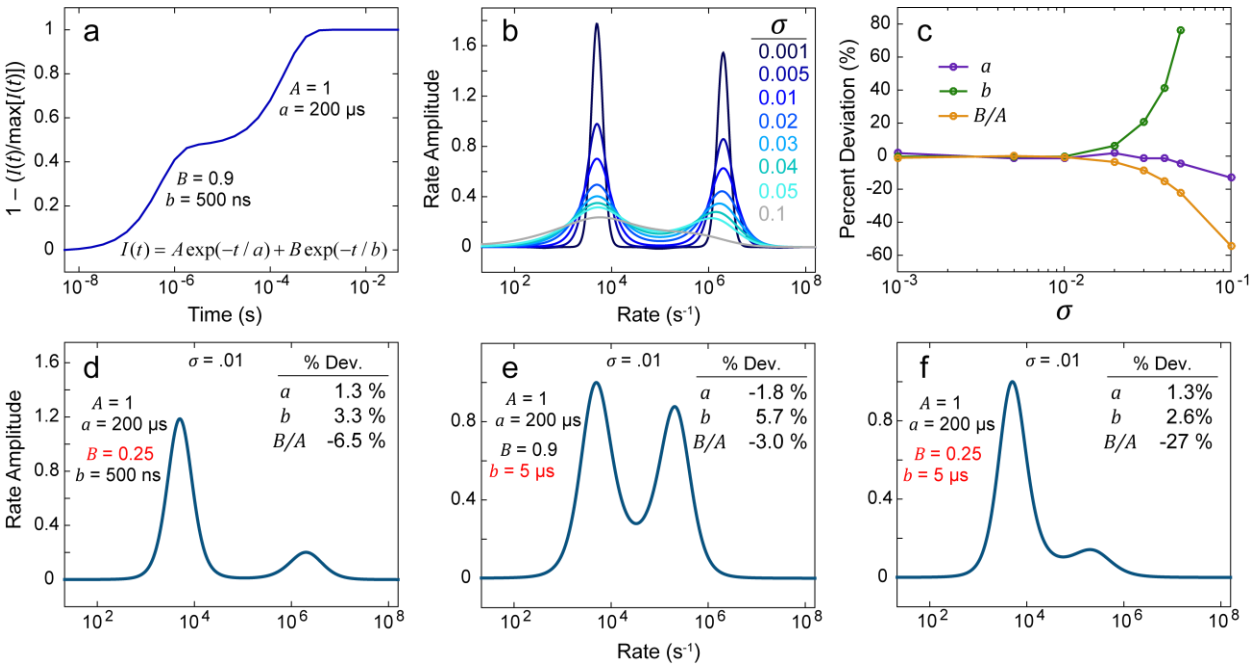

Figure 4.12: (a) Model biexponential kinetic trace in the time domain representing two well separated processes with similar amplitude. (b) The rate spectrum after MEM-iLT of the trace in panel a as a function of the standard deviation of the noise, σ. (c) Percent deviation as a function of σ between the rate constants and amplitude ratio after MEM-iLT vs the values initially supplied in the model time trace. (d-f) Testing the effects of reducing the amplitude ratio and separation of timescales between the two exponentials. The value of $\sigma = 10^{-2}$. Altered parameters are highlighted in red. The resulting percent deviation in each of the rate constants and the ratio of the amplitudes is indicated.

A biexponential kinetic trace in which the amplitudes of the two processes are similar and the time constants are well separated serves as an idealized scenario for establishing an upper bound on the permissible noise variance. Fig. 4.12a shows such a kinetic trace, where the amplitudes of the two exponentials, *A* and *B*, are within 10% of each other and the time constants of a = 200 µs and b = 500 ns are well separated along the time axis. The biexponential decay is

normalized and inverted to conform with our convention that a time trace that rises away from zero corresponds to positive rate amplitude while a trace relaxing towards zero corresponds to negative rate amplitude.

The value of σ was varied from 10^{-3}-10^{-1} resulting in the set of rate spectra plotted in Fig. 4.12b. Qualitatively, the trend with increasing σ shows a broadening of the rate distributions and, above a threshold of 10^{-2}, a merging of the two distributions. To quantify this loss of fidelity with increasing noise, the percent deviation of the inverse rate constants and amplitude ratio extracted from the rate spectrum versus the time constants and amplitude ratio set in the model trace is plotted in Fig. 4.12c. The rate constant for each process is taken to be the rate corresponding to the maximum in each distribution while the amplitude ratio is computed by comparing the integrated area of the distributions. Up to $\sigma = 10^{-2}$ the percent deviation is essentially zero for *a*, *b*, and the ratio *B*/*A*, indicating that the time constants and amplitude ratio are recovered with high accuracy. Above this point, the increasingly positive percent deviation in *b* and negative percent deviation in *a* indicate that the two time constants are tending towards each other. In fact for the largest value of $\sigma = 10^{-1}$ there is only one discernable maximum in the rate spectrum, with the lower amplitude rate distribution appearing as a shoulder on the larger amplitude distribution. A value for *b* in this case cannot be resolved. The negative percent deviation in the amplitude ratio is also consistent with merging distributions, since overlapping distributions will distort the ratio of the integrated areas.

For the biexponential plotted in Fig. 4.12a, $\sigma = 10^{-2}$ appears to be the largest permissible standard deviation of the noise. However, this is for an ideal case in which the two processes are well separated in time and have similar amplitudes. To test the effect of a larger discrepancy in amplitude, Fig. 4.12d shows the rate distribution corresponding to a time trace where the value of

B is lowered to 0.25 and $\sigma = 10^{-2}$. Regardless, the time constants are still recovered to within a few percent deviation. Some of this discrepancy (up to ± 2%) arises from the bin width of the discretized rate axis. The amplitude ratio deviates by -6.5%, which is likely due to the small amplitude of *B* relative to *A*. We can conclude that recovery of the time constants is not significantly affected by a mismatch in amplitudes as long as the processes are well separated in time, but that the recovery of the amplitude ratio is slightly impacted by the comparatively low intensity of *B*.

To test the ability to resolve more closely spaced processes in time, the value of *b* was set to 5 μs with all other parameters equal to those in Fig. 4.12a and $\sigma = 10^{-2}$. Fig. 4.12e shows the resulting rate spectrum after MEM-iLT. Clearly the two rate distributions overlap, but the percent deviation in the amplitude ratio is still only -3.0%. The percent deviation in time constant *b* is larger than in the previous case, but not significantly at 5.7%. In contrast time constant *a* is unaffected by the reduction in separation between the two processes. Fig. 4.12f shows the rate spectrum after setting $B = 0.25$ and $b = 5$ μs such that both the separation in time constants and amplitude ratio are reduced. Unsurprisingly, the percent deviation for the amplitude ratio is large at -27% since the rate distribution associated with *B* is both low amplitude and overlapping with the distribution associated with *A*. However, both time constants are still recovered with reasonable accuracy.

For all of the scenarios tested here with $\sigma = 10^{-2}$ the rate constants extracted from the transformed data were reasonably close to the initial values, with no deviation larger than ± 6%. The recovered ratio of the amplitudes proved more sensitive to both the separation of the time constants and the initial ratio of *B*/*A*. Although we can gain a sense for a practical estimate of the resolution of the method with respect to the noise variance, the insight derived from these model

scenarios should be regarded as only rough guidelines. The effectiveness of the MEM-iLT at resolving multiple processes will vary based on the particular circumstances of each application, but one can build an intuition by working with simulated kinetic data. For example, if two features in the rate domain overlap considerably, as seen in Fig. 4.12b for $\sigma > 10^{-2}$, then the time constants extracted from the rate spectrum are likely shifted towards the mean of the two time constants since the processes are not well resolved. For experimental T-jump data where σ is estimated by taking the standard deviation of the residual of a smoothing spline fit through the time trace, typical values range between $0.3\text{-}5 \times 10^{-3}$, which are at the lower end of the values tested here.

4.5.4 Example with Stretched Exponential Kinetics

The ability to analyze kinetic data without having to resort to fitting an assumed functional form in the time domain is an advantage of applying the MEM-iLT. The previous examples only considered exponential kinetics, but as noted above non-exponential processes are often observed when studying the unfolding of biomolecules. For the DNA oligonucleotide samples studied in this thesis, the most common non-exponential behavior takes the form of a stretched exponential in the time domain. There is no universal molecular explanation for stretched exponential kinetics, but a common interpretation rationalizes an observed stretched exponential in terms of the global relaxation of a system containing many closely spaced exponential processes each relaxing with their own characteristic time constant.[17,18] Stated most generally, stretched exponential behavior in this view is a reflection of heterogeneity in the kinetics of the system. The specific origin of this heterogeneity is not immediately apparent from kinetic measurements alone and can vary from system to system. Therefore the physical interpretation of observed stretched exponential kinetics must be treated on a case by case basis. Two possible limiting cases relevant to the unfolding of

nucleic acids involve a heterogeneous ensemble of folded structures which each unfold at different rates or conversely a highly homogeneous folded ensemble that is linked to the unfolded state by many possible kinetic pathways.

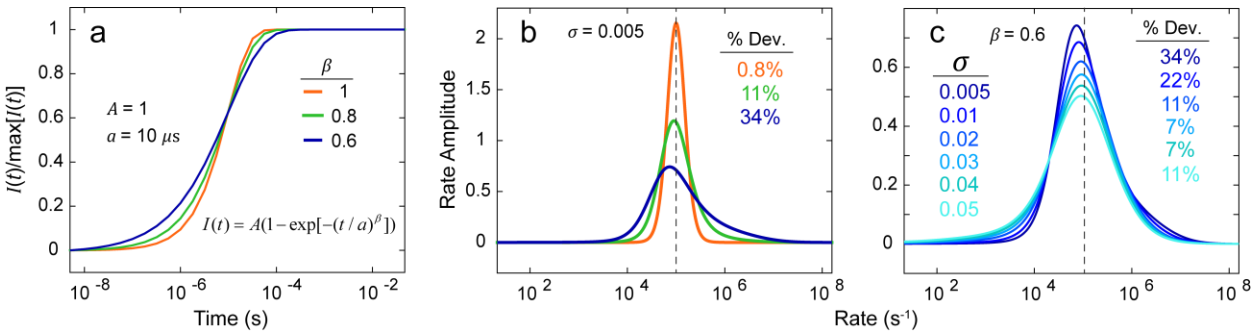

Figure 4.13: (a) Stretched exponential time traces with varying values of β but fixed amplitude and time constant. The parameters associated with the traces are indicated in the figure. (b) Rate domain representation of the time traces in panel a where the standard deviation of the noise was set to 0.005. The percent deviation of the inverse of the rate associated with the maximum amplitude versus the initial time constant of 10 μs increases as β increases. (c) The influence of increasing the noise variance on the β = 0.6 distribution. As σ increases the distribution symmetrizes and the percent deviation with the initial time constant decreases up until σ = 0.04, at which point the distribution begins to increasingly skew towards slower rates.

Fig. 4.13a plots three examples of stretched exponential time traces where the stretching parameter is set to β = 1, 0.8, or 0.6, which covers the range we observe for the unfolding of DNA oligonucleotides. Applying the MEM-iLT results in the rate representations shown in Fig. 4.13b. The standard deviation of the noise was set to $\sigma = 5 \times 10^{-3}$ to correspond to the upper bound of a typical estimate from experiment. In the case where β = 1, a single exponential trace is recovered and, as expected, the rate representation is a symmetric Gaussian profile centered at the inverse of the time constant initially supplied in the model time trace. As β is decreased and the exponential stretches, the corresponding rate distribution becomes increasingly asymmetric and skewed, with

a tail stretching out towards faster rate. The inverse of the rate at the maximum no longer matches the time constant initially supplied in the model, and this discrepancy, as quantified by the percent deviation, becomes increasingly large as β is decreased. However, a line drawn through the distribution at the inverse of the original time constant bisects the distribution into two parts with equal area, as indicated in Fig. 4.13b.

As the noise variance is increased, the stretched exponential character in the rate distribution discussed above is increasingly lost. Fig. 4.13c shows an example where σ is increased from 5×10^{-3} to 5×10^{-2} for the $\beta = 0.6$ time trace. Up until $\sigma = 3 \times 10^{-2}$, the distribution symmetrizes and the maximum shifts back toward the inverse of the time constant originally set in the time domain. Above $\sigma = 3 \times 10^{-2}$, the distribution increasingly distorts towards slower, rather than faster rate. Since a skew of the distribution towards slower rate is unsupported by the original time domain data, one should conclude that this is an artifact of the method and that the rate distribution is poorly resolved at these levels of noise.

Considering both the shift in the maximum of the rate distribution as an exponential is increasingly stretched and the tendency of noise to distort the distribution back towards exponential behavior highlights the difficulty of rigorously characterizing stretched exponential kinetics. Determining the point along the rate axis which bisects the area under the rate distribution into two equal parts effectively recovers the correct time constant for the model data discussed here. However, this method applies only for an isolated stretched exponential since the presence of additional kinetic processes will result in overlapping rate distributions. For the range of β values considered here, the maximum of the rate distribution provides a reasonable estimate for the characteristic timescale of the system, since even the worst discrepancy seen for the $\beta = 0.6$ case amounted to an error of a few microseconds (10 μs vs 13.4 μs). All things considered, this

approach is often the least ambiguous and most robust method for determining a characteristic rate, although one should not lose sight of the potential accumulation of error as β decreases.

4.5.5 Conclusion

It is important to always keep in mind that transformation to the rate domain cannot reveal any new information that is not already present in the time domain representation of the data. Employing the MEM-iLT simply results in a different, and often more intuitive, way of visualizing the experimental results. If a feature appears in the rate spectrum which is not supported by the original time domain measurement, it is an artifact of the transform method. For example, a common pitfall involves underestimating the noise variance. In this case deviations in the time trace due to noise can be misinterpreted as distinct kinetic processes and every bump and wiggle in the time domain can potentially be resolved as a peak in the rate domain. One can imagine that, depending on the severity of the underestimation of σ, the presence of these spurious features could result in any number of additional "processes" in the rate domain and that such noise artifacts are best distinguished in the raw time trace. When in doubt, it is better to err on the side of overestimating rather than underestimating the noise variance so as not to over-resolve the experimental data.

4.6 Acknowledgements

I thank William Memo Carpenter, Lukas Whaley-Mayda, and Chi-Jui Feng for careful reading of this chapter. I thank Paul Stevenson for sharing the peptide-lipid FTIR data.

4.7 References

1. Vaseghi, S. V., *Advanced digital signal processing and noise reduction*. John Wiley & Sons: 2008.

2. Khalil, M.; Demirdöven, N.; Tokmakoff, A., Coherent 2D IR spectroscopy: molecular structure and dynamics in solution. *The Journal of Physical Chemistry A* **2003,** *107* (27), 5258-5279.

3. Ramos, P. M.; Ruisánchez, I., Noise and background removal in Raman spectra of ancient pigments using wavelet transform. *Journal of Raman Spectroscopy* **2005,** *36* (9), 848-856.

4. Galloway, C.; Ru, E. L.; Etchegoin, P., An iterative algorithm for background removal in spectroscopy by wavelet transforms. *Applied spectroscopy* **2009,** *63* (12), 1370-1376.

5. Alsberg, B. K.; Woodward, A. M.; Kell, D. B., An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics and Intelligent Laboratory Systems* **1997,** *37* (2), 215-239.

6. Rioul, O.; Vetterli, M., Wavelets and signal processing. *IEEE signal processing magazine* **1991,** *8* (4), 14-38.

7. Walczak, B.; Massart, D., Noise suppression and signal compression using the wavelet packet transform. *Chemometrics and Intelligent Laboratory Systems* **1997,** *36* (2), 81-94.

8. Mallat, S. G., A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence* **1989,** *11* (7), 674-693.

9. Stevenson, P.; Tokmakoff, A., Distinguishing gramicidin D conformers through two-dimensional infrared spectroscopy of vibrational excitons. *The Journal of chemical physics* **2015,** *142* (21), 212424.

10. Jaynes, E. T., On the rationale of maximum-entropy methods. *Proceedings of the IEEE* **1982,** *70* (9), 939-952.

11. Jaynes, E. T., Information theory and statistical mechanics. *Physical review* **1957,** *106* (4), 620-630.

12. Widjaja, E.; Garland, M., Pure component spectral reconstruction from mixture data using SVD, global entropy minimization, and simulated annealing. Numerical investigations of admissible objective functions using a synthetic 7-species data set. *Journal of computational chemistry* **2002,** *23* (9), 911-919.

13. Hendler, R. W.; Shrager, R. I., Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. *Journal of biochemical and biophysical methods* **1994,** *28* (1), 1-33.

14. Nölting, B., *Protein folding kinetics: biophysical methods*. Springer: Berlin, 2006.

15. McWhirter, J.; Pike, E. R., On the numerical inversion of the Laplace transform and similar Fredholm integral equations of the first kind. *Journal of Physics A: Mathematical and General* **1978,** *11* (9), 1729-1745.

16. Kumar, A. T.; Zhu, L.; Christian, J.; 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. *The Journal of Physical Chemistry B* **2001,** *105* (32), 7847-7856.

17. Johnston, D., Stretched exponential relaxation arising from a continuous sum of exponential decays. *Physical Review B* **2006,** *74* (18), 184430.

18. Sabelko, J.; Ervin, J.; Gruebele, M., Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences* **1999,** *96* (11), 6031-6036.

Chapter 5

A Simple Oligonucleotide Lattice Model for Informing the Interpretation of Infrared Spectroscopy Experiments

5.1 Introduction

Infrared (IR) spectroscopy is a valuable experimental technique in the study of nucleic acid structure,[1,2] hydration,[3,4] folding dynamics,[5,6] and energy relaxation.[7,8] The toolbox of IR methods offers label-free structural information and time-resolved IR spectroscopies can access dynamics down to the picosecond timescale. Across the mid-IR, characteristic frequency ranges report on the sugar conformation, phosphate backbone orientation, glycosidic bond geometry, and the vibrations of the nucleobases themselves.[1] The 1500-1800 cm^{-1} range contains in-plane nucleobase modes and carbonyl stretches that are particularly useful for studying nucleic acid folding since the same physical interactions which mediate hybridization, namely hydrogen bonding and base stacking, are also largely responsible for shaping the infrared spectrum in this region. As a result the line shapes, intensities, and frequencies of these IR transitions are sensitive reporters of nucleic acid structure and, when tracked in time, base pairing dynamics.

Extracting the complete information content of an IR measurement requires a rigorous mapping between the conformation of the system and the spectrum corresponding to a particular distribution of structures. Unfortunately, no simple structure-spectrum correlation exists for nucleic acids due to the highly delocalized nature of the nucleobase vibrations.[9] In general the IR

spectrum of DNA cannot be modeled using a basis of local vibrational modes such has been done for protein amide I spectroscopy, where the spectrum can be effectively described in terms of coupled vibrations of the polypeptide backbone.[10,11] Theoretical work on identifying suitable basis modes and coupling parameters for simulating the IR spectroscopy of adenine-thymine (AT) and guanine-cytosine (GC) duplexes has been established,[12] but as of yet there is no general quantitative model for interpreting infrared spectra in terms of nucleic acid structure.

Turning to simplified models based on empirical results offers one route forward. For example, the ultraviolet (UV) spectrum of both single and double stranded DNA in the 215-310 nm range can be accurately predicted from the nucleobase sequence through a composition weighted sum over extinction coefficients experimentally derived for each of the ten dinucleotide nearest neighbors.[13] Ultraviolet spectroscopy is commonly employed to determine nucleic acid concentration, base pair composition, and duplex fraction due to the high sensitivity and low sample volume requirements of the measurement. The ability to accurately predict the shape of the UV spectrum, peak wavelength, and degree of hyperchromicity upon melting adds considerably to the utility of this technique. A similar approach for determining dinucleotide extinction coefficients in the mid-IR is possible in principle, but the comparatively significant influence of sequence composition, base pair formation, and hydration on IR spectra would likely necessitate a prohibitively large set of experimental conditions.

Recently we have found that a simple DNA lattice model can offer insight into the interpretation of both linear and two-dimensional IR experiments applied to study DNA oligonucleotide dehybridization.[14] Lattice models of varying complexity have been used in the past to successfully describe the melting transition and folding dynamics of both DNA[15,16] and RNA,[17] and have been used extensively to simulate various aspects of protein folding.[18,19] More

generally, these descriptions fall into a broad class of statistical thermodynamic models rooted in formulating a partition function on the basis of structural contacts, such as the Muñoz-Eaton[20] and Gō models.[21] Our motivation for developing a lattice model as opposed to a more complicated representation is to determine the simplest description consistent with our experimental results, thereby identifying the most essential degrees of freedom through a computationally cheap and intuitive approach. Furthermore the spectrum in the 1500-1800 cm^{-1} range primarily reports on the extent of paired nucleobase contacts, suggesting a simple model that describes hybridization in these terms could be a natural tool for interpreting IR experiments.

Here we seek to present our model in detail and to validate its utility in interpreting infrared experiments in terms of the predicted conformational ensemble. At its core, the model is a statistical extension of the nearest-neighbor (NN) model commonly used to predict DNA melting temperatures (T_m).[22] The NN model decomposes the thermodynamics of the helix to coil transition into additive contributions from the dinucleotide steps that make up a given oligonucleotide sequence. Applying the latest set of dinucleotide parameters and empirically derived corrections for monovalent and divalent cation concentrations, the NN model is capable of predicting T_m to within an average error of only 0.5 °C.[23] This approach assumes an all-or-nothing description of base pairing in which all possible base pairs are either fully broken or fully intact. Although this assumption proves remarkably successful at predicting hybridization thermodynamics for most nucleobase sequences, it neglects potential dimer heterogeneity that can play an important role in hybridization. Previous models which have incorporated the NN parameters into statistical treatments have been successful at simulating nucleic acid ensembles and thermodynamics as well as providing a unified description of the thermal denaturation of both oligomers and polymers.[16,24]

The basic organization of our model is as follows. At the finest level, discretized base pairing is enumerated such that all possible combinations of broken and intact contacts are generated and the enthalpy of a particular base pair configuration is set using the NN parameters. Configurational entropy is determined by self-avoiding random walks (SAWs) of beaded polymer chains on a 3D cubic lattice of nucleotide sized sites. Finally, concentration effects associated with the gain in translational entropy upon dimer dissociation are simulated on a larger 3D lattice of monomer sized sites. Note that we use the term "monomer" in this chapter to refer to a single-stranded oligonucleotide while the individual subunits that make up the strand will be referred to as nucleotides or beads. In the first half of the chapter, we present the details of the model and identify useful connections to experiment. In the second half, we will demonstrate the utility of the model in interpreting IR experiments on the melting of DNA oligonucleotides of varying sequence composition and length.

5.2 Details of the Model

5.2.1 Statistical Thermodynamics of the Dimer to Monomer Transition

A brief outline of the thermodynamics of a general dimer to monomer dissociation reaction is included in Chapter 1, Section 1.4. This descriptions lays the foundation for our lattice model and provides a connection to the quantities most often measured in experiment through the thermal melting curve. This chapter will connect the thermodynamic picture to the microscopic picture by building up the lattice model partition function. As before, we will restrict ourselves to the dissociation of self-complementary homodimers, but the model is easily extended to consider heterodimers as well.

We will assume that the dimer and monomer degrees of freedom are independent of one another and that the lattice model partition function, Q can be separated into the product of partition functions corresponding to each of the subcomponents of the system, as in eq 5.1.

$$Q = \prod_i Q_i \tag{5.1}$$

The partition function Q is related to the free energy G through eq 5.2. We will construct the lattice model partition function in detail below, but for now Q is defined generically in eq 5.3 as the sum over all microstates of the ensemble Boltzmann weights set by the free energy of a given microstate, G_α. The partition function is related to the Gibbs free energy G rather than the Helmholtz free energy F typical of an *NVT* ensemble because the experimentally derived parameters from the NN model that are used to set the energy of a microstate are determined with respect to Gibbs free energy. Within the context of the simple model, pressure is not accounted for explicitly and the difference between F and G is not considered.

$$G = -k_B T \ln Q \tag{5.2}$$

$$Q = \sum_\alpha e^{-G_\alpha / k_B T} \tag{5.3}$$

The chemical potential associated with component i in terms of the associated partition function is then given by

$$\mu_i = -k_B T \left(\frac{\partial \ln Q_i}{\partial N_i} \right)_{p,T,\{N_j, j \neq i\}} \tag{5.4}$$

where N_i is the number of molecules of component i.

We will assume that the N_i molecules of component i are indistinguishable and non-interacting. Therefore the partition function for component i can be expressed in terms of a molecular partition function, q_i associated with component i.

$$Q_i = \frac{q_i^{N_i}}{N_i!} \tag{5.5}$$

Using eq 5.2 and Stirling's approximation, we obtain an expression for the free energy in terms of the molecular partition function.

$$G_i = -N_i k_B T \ln \frac{q_i}{N_i} - N_i k_B T \tag{5.6}$$

In analogy to eq 1.3, differentiating with respect to N_i results in an expression for the chemical potential.

$$\mu_i = -k_B T \ln \frac{q_i}{N_i} \tag{5.7}$$

At equilibrium, the free energy change is zero and therefore $2\mu_M = \mu_D$. Using this relation and eq 5.7 we obtain an expression for the dissociation constant

$$K_d = \frac{N_M^2}{N_D} = \frac{q_M^2}{q_D} \tag{5.8}$$

which is related to eq 1.2 by $(\mathcal{N}_A V_{Tot})^{-1}$ where V_{Tot} is the total system volume. Eq 5.8 is a useful result because it relates the dissociation constant for the overall reaction to the monomer and dimer molecular partition functions.

5.2.2 Constructing the Lattice Model Partition Function

We construct the lattice model partition function starting with the assumption discussed above for eq 5.1 and expressing the system partition function as the product of a monomer and dimer partition function.

$$Q = Q_M Q_D \tag{5.9}$$

We also assume that the translational and internal degrees of freedom for each species are separable and further decompose the partition function into the product

$$Q = Q_{M,trans} Q_{M,int} Q_{D,trans} Q_{D,int} \tag{5.10}$$

In our model, each of these degrees of freedom is coarse grained onto a lattice.

5.2.3 The Translational Partition Function

We will begin by describing the translational partition function, which should reflect the excess translational entropy available to the monomer over the dimer as well as the decreasing free energy of the dimer as the total concentration of oligonucleotides is increased. The translational partition function is calculated by considering the number of possible monomer and dimer arrangements on a 3D lattice of monomer sized sites. Assuming the total oligonucleotide concentration is uniformly distributed across the entire system volume, the volume per oligonucleotide V is

$$V = \left(\mathcal{N}_A [C_{Tot}]\right)^{-1} \tag{5.11}$$

We discretize this volume into a 3D lattice of monomer sized sites by dividing V by the average volume of a single oligonucleotide strand. The monomer volume is calculated by assuming a spherical shape and determining the average radius of gyration given the number of nucleotides in the oligonucleotide chain, L_{Tot} and the length of a single nucleotide unit, l.[25]

$$\left\langle R_G^2 \right\rangle = \frac{l^2 L_{Tot}(L_{Tot}+2)}{6(L_{Tot}+1)} \tag{5.12}$$

The number of lattice sites per oligonucleotide strand, X is determined by

$$X = V\left(\frac{4}{3}\pi \left\langle R_G \right\rangle^3\right)^{-1} \tag{5.13}$$

The total number of sites in the entire system volume is then given by $N_{Tot}X$ where N_{Tot} is the total number of oligonucleotides. Assuming that the system is in the dilute limit where one can neglect the volume excluded by placing an oligonucleotide on the lattice, the number of ways of placing N_M monomers on a lattice of $N_{Tot}X$ sites is

$$\Omega_M = \frac{(N_{Tot}X)!}{N_M!(N_{Tot}X - N_M)!} \tag{5.14}$$

However, under dilute conditions when $N_M << N_{Tot}X$, this expression can be approximated by eq 5.15.

$$\Omega_M \approx \frac{(N_{Tot}X)^{N_M}}{N_M!} \tag{5.15}$$

In the context of the translational partition function, we define a dimer as any monomer with a neighbor occupying any one of b adjacent lattice sites, with $b = 6$ in the case of a 3D cubic lattice. The number of ways of arranging $2N_D$ oligonucleotides on the lattice such that they are in a dimer configuration is therefore equivalent to arranging N_D oligonucleotides in any of the $N_{Tot}X$ sites on the lattice and restricting N_D oligonucleotides to the $N_{Tot}b$ number of possible monomer-adjacent sites on the lattice. The expression for the dimer translational partition function in analogy to eq 5.15 is then

$$\Omega_D \approx \frac{(N_{Tot}X)^{N_D}(N_{Tot}b)^{N_D}}{N_D!} \tag{5.16}$$

We will assume that these configurational degeneracy counts Ω_i on a 3D cubic lattice representation of the total system volume are a sufficient representation of the translational partition function for the monomer and dimer.

Returning to the partition function Q_i associated with component i, incorporating the translational partition function calculated above, and treating the internal partition function in the

same manner as eq 5.5 where we assume that there are N_i indistinguishable copies of component i, we obtain

$$Q_i \approx \Omega_i q_{i,int}^{N_i} \tag{5.17}$$

The indistinguishability of dimers or monomers is accounted for here in the translational term of the partition function. Following the same procedure as before we arrive at expressions for the monomer and dimer chemical potentials, eq 5.18 and 5.19.

$$\mu_M = -k_B T \ln\left(\frac{N_{Tot} X q_{M,int}}{N_M}\right) \tag{5.18}$$

$$\mu_D = -k_B T \ln\left(\frac{N_{Tot}^2 X b q_{D,int}}{N_D}\right) \tag{5.19}$$

As above, at equilibrium $2\mu_M = \mu_D$ and we obtain an expression for the dissociation constant

$$K_d = \frac{N_M^2}{N_D} = \frac{X q_{M,int}^2}{b q_{D,int}} \tag{5.20}$$

This result provides an expression for the overall dimer to monomer equilibrium constant that is independent of the total system size and that can be calculated in terms of X, b, and the internal degrees of freedom modeled by the molecular partition functions associated with the monomer and dimer species. In principle, the $q_{i,int}$ include contributions from the entire phase space of the system, but for the purposes of the lattice model we focus solely on computing a coarse grained representation of the conformational degrees of freedom and neglect the rest.

5.2.4 The Internal Molecular Partition Function

The internal molecular partition function for component i, $q_{i,int}$ is determined by taking the sum of Boltzmann weights across all possible conformational microstates. The weight associated

with a particular microstate α is determined by the free energy of that microstate G_α, as in eq 5.3. For dimer states, the lattice model microstate free energies are set by considering the interplay between an enthalpic stabilization due to base pair formation and an increase in conformational entropy due to an unpaired nucleotide. In the simplified context of the lattice model, we will consider the free energy of the microstates to be defined solely by these terms. Initially we will only consider those dimer microstates with in-register base pairing, which one would expect to predominate for short and heterogeneous sequences. This assumption seems to be reasonable for the sequences studied in this thesis. Since we would like to consider sequence effects explicitly, every possible in-register base pairing scenario is modeled explicitly rather than in an averaged way. We therefore represent an oligonucleotide sequence as a 1D string of L_{Tot} sites, where L_{Tot} is the total number of nucleotides in the strand. A given site can exist in either an open or a closed state, with an open/closed site corresponding to a broken/intact base pair. All $2^{L_{Tot}}$ possible combinations of open and closed sites for a lattice of L_{Tot} sites are generated. The enthalpy of a given microstate, H_α is assigned by taking the sum of dinucleotide enthalpies from the NN model across the intact dinucleotide subunits of the microstate base pairing configuration, as shown in Fig. 5.1a,c for two illustrative microstates.

$$H_\alpha = \sum_{i=1}^{L_{Tot}-1} B_\alpha(i) B_\alpha(i+1) H_{NN}(i, i+1) \tag{5.21}$$

$$B(i) = \begin{cases} 0 & \text{if site } i \text{ is unpaired} \\ 1 & \text{if site } i \text{ is paired} \end{cases} \tag{5.22}$$

The experimentally derived dinucleotide enthalpies take into account sequence specific hydrogen bonding and base stacking effects.[22] For isolated pairing contacts that have a hydrogen bonding contribution in the absence of a stacking neighbor, we assign an additional enthalpic contribution of -2.1 kJ/mol for an adenine-thymine (AT) pair and -3.1 kJ/mol for a guanine-cytosine (GC) pair

based on an experimentally derived estimate for the stabilization due solely to hydrogen bonding in isolation.[26]

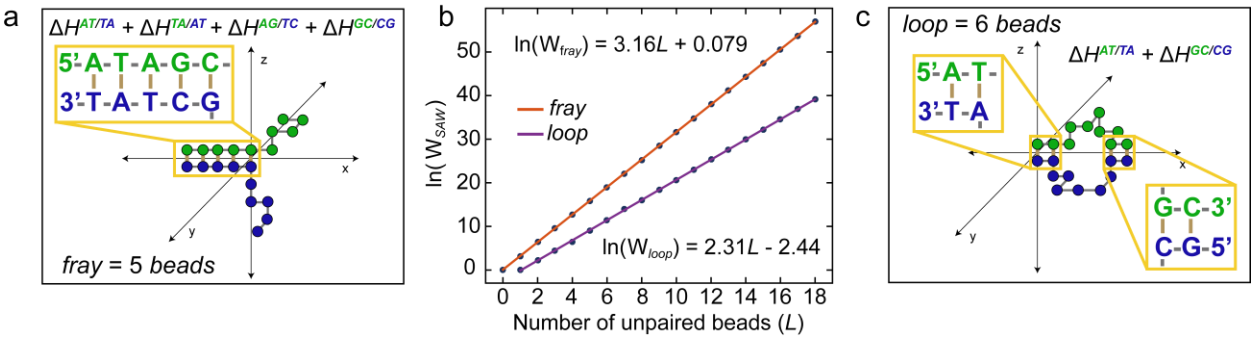

Figure 5.1: (a) Example of a microstate for an oligonucleotide where L_{Tot} = 10 with five intact base pairs and a frayed end of five nucleotide beads. The microstate enthalpy is assigned by taking the sum across the four intact dinucleotide subunits indicated in the panel. (b) The configurational entropy in units of k_B for frayed ends and internal loops of L = 0-18 nucleotide beads out of the L_{Tot} total beads in the strand based on self-avoiding random walks on a 3D cubic lattice. Linear fits with respect to L are indicated in the figure. (c) Example of a microstate with an internal loop of six nucleotide beads for the same sequence depicted in panel a.

The configurational entropy assigned to a particular arrangement of open and closed base pair sites is generated by allowing the unpaired bases to explore all possible arrangements on a 3D cubic lattice of nucleotide-sized sites through self-avoiding random walks. Each nucleotide in the sequence is represented as a bead on this 3D cubic lattice. Assuming in-register base pairing, a broken base pair or stretch of broken base pairs can exist in only one of two scenarios. A frayed end refers to a stretch of L = 1 to L = (L_{Tot} – 1) broken base pairs that includes a broken terminal pair while an internal loop refers to a stretch of L = 1 to L = (L_{Tot} – 2) broken base pairs flanked by at least one intact base pair on either side. An example of a microstate with a frayed end of L = 5 nucleotide beads for an L_{Tot} = 10 dimer is depicted in Fig. 5.1a and an example of a microstate with an internal loop of six beads is depicted in Fig. 5.1c for the same sequence. For simplicity,

frayed ends and internal loops, although self-avoiding, are not restricted to avoid clashing with the complementary strand. While frayed end configurations are allowed in all octants of the 3D lattice, each half of a looped configuration is restricted to those quadrants which lie above and below the plane that bisects the duplex parallel to the helical axis. For example, in Fig. 5.1c the green beads in the loop would be restricted to SAWs above the xy-plane while the blue beads would be restricted to SAWs below the xy-plane. This restriction has the effect of decreasing the number of possible loop configurations and is included as a reflection of the reduced flexibility of an internal loop as compared to a frayed end. The configurational entropy takes the form of a Boltzmann entropy, eq 5.23, where W_α is a count over all of the possible configurations of the free nucleotide beads for microstate α.

$$S_\alpha = k_B \ln(W_\alpha) \tag{5.23}$$

Fig. 5.1b plots the configurational entropy for frayed ends and internal loops of 0-18 nucleotide beads in length. For a stretch of broken base pairs of a given length, a frayed end provides a greater increase in entropy as compared to an internal loop and this disparity increases with increasing length. Now that we have defined the enthalpic and entropic contributions to the microstate free energies, the internal molecular partition function for component *i* can be expressed as

$$q_{i,int} = \sum_\alpha e^{-(H_\alpha - TS_\alpha)/k_B T} = \sum_\alpha W_\alpha e^{-H_\alpha/k_B T} \tag{5.24}$$

For the monomer internal molecular partition function, $q_{M,int}$, the microstate free energies are considered purely entropic and are generated by SAWs for a beaded chain equal in length to the total oligonucleotide length, $L_{Tot.}$ It should be noted that the model does not include any parameters that attempt to model cooperativity effects explicitly, but that these effects enter through the NN parameters and thereby emerge naturally.

5.2.5 Connecting the Model to Experimental Observables

Spectroscopic thermal melting curves are one of the most common experimental approaches to studying the DNA dimer to monomer transition. In this context, the extent of dehybridization is commonly expressed in terms of the fraction of intact base pairs as a function of temperature $\theta(T)$ under the assumption that changes in absorbance with temperature can be mapped to the number of broken base pairs in the DNA ensemble.[27] The overall contact fraction can be expressed as the product of an internal and external fraction, eq 5.25.

$$\theta(T) = \theta_{int}(T)\theta_{ext}(T) \tag{5.25}$$

Here $\theta_{int}(T)$ is the average fraction of contacts among those DNA with at least one intact base pair while $\theta_{ext}(T)$ is the fraction of DNA strands with at least one intact base pair out of the total number of DNA strands. The second term is essential for modeling the dimer to monomer transition associated with dehybridization. With the framework of the lattice model in place, we can calculate the internal fraction of intact base pairs in terms of the internal dimer partition function $q_{D,int}$.

$$\theta_{int}(T) = \frac{1}{L_{Tot}} \sum_{\delta} \frac{p e^{-G_{\delta}/k_B T}}{q_{D,int}} \tag{5.26}$$

This expression computes the average fraction of intact base pairs across the dimer microstates where the sum over δ is restricted to those microstates which have at least one intact base pair and p corresponds to the number of intact base pairs in microstate δ. For $\theta_{ext}(T)$ we use the dissociation constant, K_d computed above and the total concentration of DNA, $[C_{Tot}]$ to arrive at an expression for the fraction of oligonucleotide strands out of the total number of DNA strands that have at least one intact base pair.

$$\theta_{ext}(T) = 1 + \frac{K_d}{4[C_{Tot}]}\left(1 - \sqrt{1 + \frac{8[C_{Tot}]}{K_d}}\right) \tag{5.27}$$

This expression is equivalent to the dimer fraction $\theta_D(T)$ commonly applied when assuming a two state all-or-nothing dimer to monomer transition, in which case $\theta_{int}(T)$ must be 1. The product of eq 5.26 and eq 5.27, $\theta(T)$ is assumed to model the melting curve. The melting temperature, T_m is often defined as the temperature where $\theta(T) = 0.5$ and is routinely reported as a convenient proxy for DNA dimer stability.

5.2.6 Parameterization of the Model

In reality, the conformation of a single nucleotide unit in a DNA strand is determined by seven continuous dihedral angles. Reducing each nucleotide to a single bead fixed on a 3D cubic lattice will clearly undercount the conformational entropy severely, leading to nonphysical results. For example, consider the melting curve predicted by the lattice model for the sequence 5'-GATATATATC-3' in Fig. 5.2a. The overall sigmoidal shape of the curve is reasonable, but the T_m, defined as the temperature at which $\theta(T) = 0.5$, is predicted to be 591 °C! Furthermore the melting transition spans over 200 °C between the upper and lower baseline of the melting curve. These clearly nonphysical results motivate an additional correction to the model or parameterization against experiment such that more reasonable predictions are possible. The simplest correction would seek to account for the missing degrees of conformational freedom in an average way and one would expect that the magnitude of this entropy correction should grow exponentially with the number of free beads in a microstate configuration.[17]

We therefore introduce a base pair excess entropy parameter γ and the internal molecular partition function for component *i* becomes,

$$q_{i,int} = \sum_{\alpha} \gamma^{L} W_{\alpha} e^{-H_{\alpha}/k_B T} \tag{5.28}$$

where γ is a single parameter raised to the number of free nucleotides L in microstate α. The value of γ is set such that the lattice melting curve is shifted to a temperature range that agrees with experiment. As seen in Fig. 5.2b, applying this entropy correction results in good agreement with the experimental melting curve for this sequence when γ = 27.

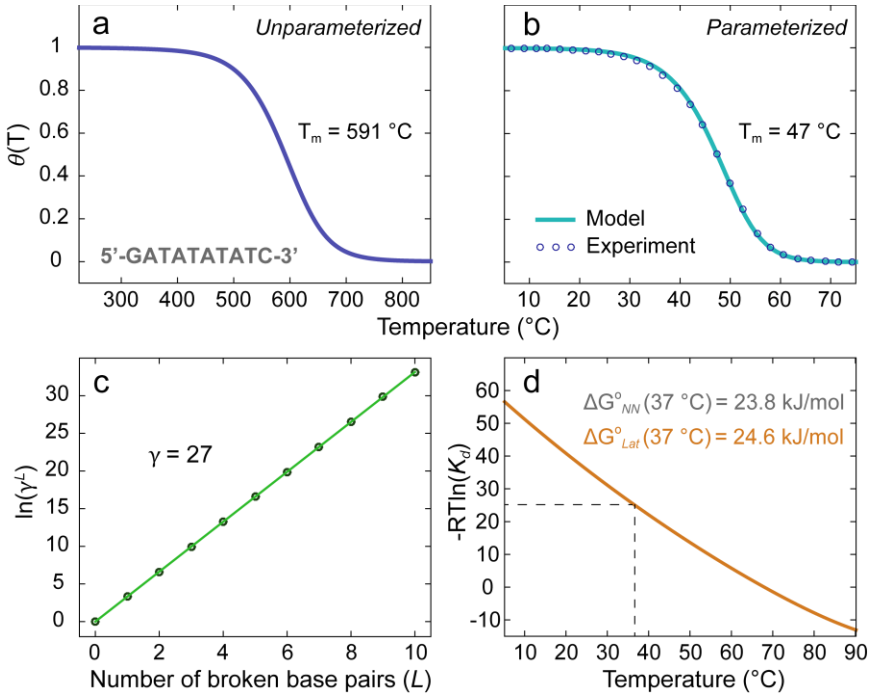

Figure 5.2: (a) The melting curve predicted for the sequence 5'-GATATATATC-3' without any further parameterization of the model results in an unphysically high T_m of 591 °C. (b) Melting curve predicted for the same sequence with a single entropic scaling parameter, γ raised to the number of broken bases in a microstate, L. The value of γ is parameterized against experiment. Points indicate the FTIR melting curve while the solid line plots the model prediction. (c) The entropic correction term for this sequence plotted as a function of L in units of k_B. (d) Free energy for the dimer dissociation reaction predicted by the parameterized lattice model. The lattice model prediction is compared against the NN model ΔG° at 37 °C, demonstrating reasonable agreement.

The length and sequence dependence of this parameter is discussed in the next section. The Fourier transform infrared (FTIR) melting curve was measured as described below. The entropy correction

term for this sequence is plotted as a function of the number of free bases, L in Fig. 5.2c. In the absence of experimental data, the NN model T_m can just as easily be used to set the value of γ.

The standard free energy of dimer dissociation as a function of temperature is plotted in Fig. 5.2d for the parameterized lattice model. As a further check on the influence of γ, we compare the predicted ΔG^o at 37 °C against the NN model at 37 °C since the latest salt-corrected version of the NN model computes this value using a set of experimentally derived correction parameters to account for the effects of both monovalent and divalent cations. The agreement between the two models is reasonable despite the fact that the lattice model does not include any explicit correction for cation concentration. However, cation effects on DNA hybridization are primarily entropic, so it is likely that the salt concentration also influences the value of γ when parameterizing against experiment.[22,23]

5.3 Validation of the Model

5.3.1 Validation with Respect to Nucleobase Sequence

Determining the role of sequence composition and nucleobase ordering in DNA hybridization is a question which is nearly as old as the discovery of the DNA double helix itself and which has been investigated in some detail.[28-30] We have studied the sequence-dependent mechanism of DNA dehybridization through both steady-state and time-resolved IR spectroscopy, finding that the placement of GC pairs in an otherwise AT sequence shapes the dehybridization mechanism.[14,31] We used our DNA lattice model to assist in the interpretation of experimental results, but we have not previously presented the model in detail. Here, we return to our well-studied set of DNA oligonucleotides in order to validate that the lattice model effectively captures the sequence dependent effects observed in IR experiments. The self-complementary

oligonucleotides are ten base pairs in length and include the sequences 5'-ATATGCATAT-3' (GC-core), 5'-ATGATATCAT-3' (GC-mix), and 5'-GATATATATC-3' (GC-ends). We will refer to each of the sequences by the name in parentheses for convenience. Fig. 5.3a shows an illustrative set of temperature dependent FTIR acquired between 5-90 °C for the GC-core sequence. The sample conditions were 2 mM oligonucleotide in 50 mM deuterated pD 7.2 sodium phosphate buffer, 240 mM NaCl, and 18 mM $MgCl_2$.

We will discuss the spectral features indicative of DNA hybridization and melting curve analysis in detail in Chapter 6. For the purposes of validating the lattice model, we are interested in how well the shape of the melting curve is reproduced as a function of nucleobase sequence. Experimental melting curves that reflect the global changes in the 1500-1750 cm^{-1} range were determined by taking the normalized second singular value decomposition (SVD) component and subtracting off linear fits to the sloping baselines. Fig. 5.3d shows the second SVD vector measured for the GC-core sequence with linear fits to the upper and lower baselines indicated. After a similar FTIR temperature series, SVD analysis, and baseline correction, the experimental melting curves in Fig. 5.3b were determined for each of the sequences. These melting curves report primarily on the fraction of intact base pairing in the oligonucleotide ensemble and should therefore relate directly to the $\theta(T)$ computed by the lattice model. Fig. 5.3c shows the corresponding $\theta(T)$ calculated for each of the sequences after parameterizing γ against experiment. The magnitude of the entropy scaling parameter does not vary significantly across the oligonucleotide set, as seen in Fig. 5.3f, but it appears to track the trend in melting temperatures measured by experiment.

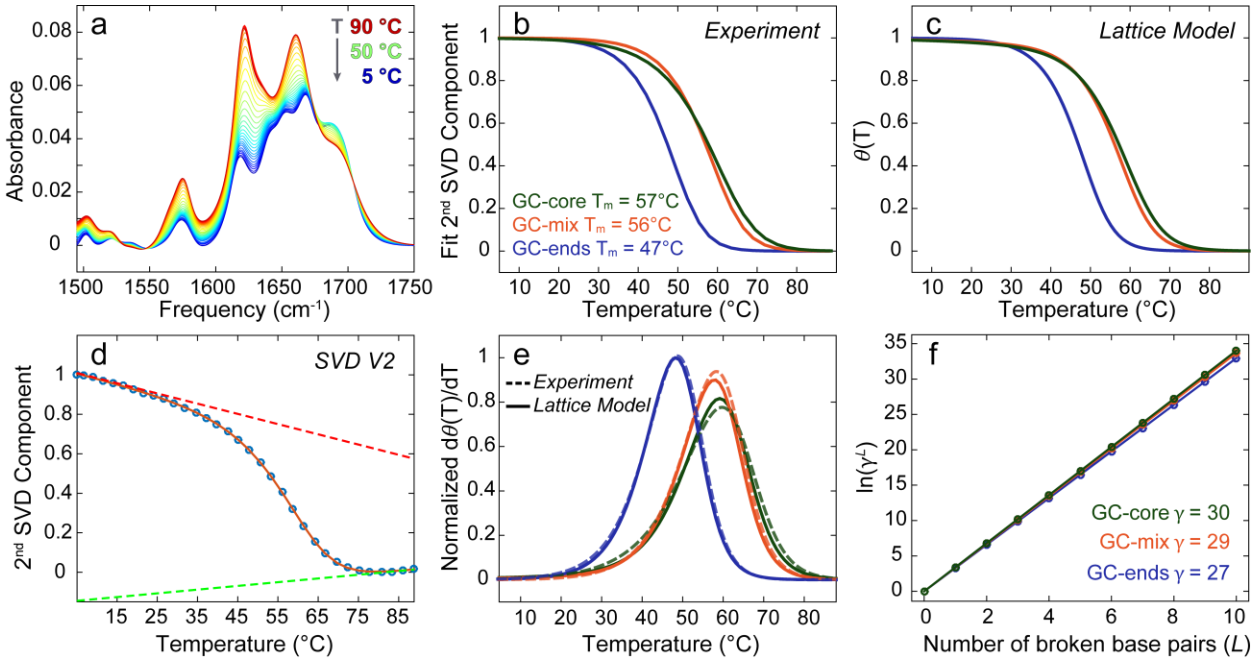

Figure 5.3: (a) Temperature dependent FTIR spectra collected between 5-90 °C in the 1500-1750 cm^{-1} range for the GC-core sequence. (b) Baseline corrected experimental melting curves obtained from the second SVD component for each of the sequences. (c) Lattice model melting curves corresponding to the experimental data in panel b. (d) Example of the second SVD component extracted from the FTIR temperature series in panel a. Linear baseline fits are indicated. (e) First derivatives of the experimental and lattice model melting curves normalized to the maximum magnitude of the GC-ends differential melting curve. (f) Entropy scaling parameter for each of the sequences.

For the simplest description of a dimer to monomer transition, the so-called all-or-nothing picture, all possible base pairs are either fully intact (dimer) or fully broken (monomer). From the perspective of the lattice model this corresponds to a scenario in which the value of $\theta_{int}(T)$ can only assume a value of 1. The melting curve in this case is a direct reporter of the dimer fraction and takes the form of a sharply transitioning sigmoid perfectly symmetric about the inflection point, which in this limiting case also corresponds to T_m. The GC-ends melting curve best exemplifies such an all-or-nothing dehybridization. For the other sequences a more diverse dimer ensemble, characterized by a loss of base pairing prior to duplex dissociation, is observed. Since

these partially melted dimer configurations are most prevalent at temperatures below T_m, this results in a distortion of the low temperature side of the melting curve away from a symmetric sigmoid along with a more gradual dimer to monomer transition, as best exemplified by the GC-core sequence. In the context of the lattice model, this corresponds to a scenario in which $\theta_{int}(T)$ can adopt values below 1 with increasing temperature and it is this term that results in distortions to the shape and symmetry of the melting curve. The GC-mix sequence represents an intermediate case between the GC-ends and GC-core sequences.

Comparing Fig. 5.3b and 5.3c, it appears that the lattice model reasonably reproduces the trend in T_m and the shape of the melting curves across the set, but comparing subtle shifts in the shape of melting curves can be challenging. Plotting the first derivative with respect to temperature offers a clearer means of comparison since any changes in the shape of the curve are magnified in the derivative. Fig. 5.3e shows the differentiated melting curves normalized to the maximum magnitude of the derivative of the GC-ends melting curve for both the experimental and lattice model curves. In the first derivative it is easier to visualize asymmetry about the inflection point and the width of the differentiated curve is a direct reflection of the sharpness of the dimer to monomer transition. Here too we see reasonable agreement between the model and experiment.

So far we have compared experimental melting curves that reflect the global changes to the FTIR spectrum in the 1500-1750 cm^{-1} frequency range against modeled melting curves that consider the full oligonucleotide ensemble. The agreement between the simulated melting curves and experiment suggests that the model reasonably captures configurational heterogeneity among dimers as well as the overall dimer/monomer equilibrium as a function of temperature. However, we would also like to evaluate if the model correctly predicts the composition of the melting dimer ensemble. Infrared spectroscopy provides a degree of nucleobase specific insight through the

unique vibrational spectrum of each base and it is therefore possible to monitor the melting of AT and GC base pair domains independently. For instance between 1500-1590 cm^{-1} the spectrum is dominated by GC features and tracking thermal changes to the spectrum in this frequency range produces a melting curve that reports primarily on the dehybridization of the GC regions of the duplex. Above 1590 cm^{-1} the spectrum becomes somewhat congested with overlapping absorptions from all four nucleobases. We have previously identified distinct AT and GC cross-peak regions in two dimensional infrared (2D IR) spectra that alleviate this congestion.[14] However, because of the 80% AT content in these sequences as well as the large increase in intensity of the 1622 cm^{-1} A ring mode upon the loss of base pairing, it is reasonable in this case to assume that the 1590-1630 cm^{-1} frequency range tracks AT base pairing.

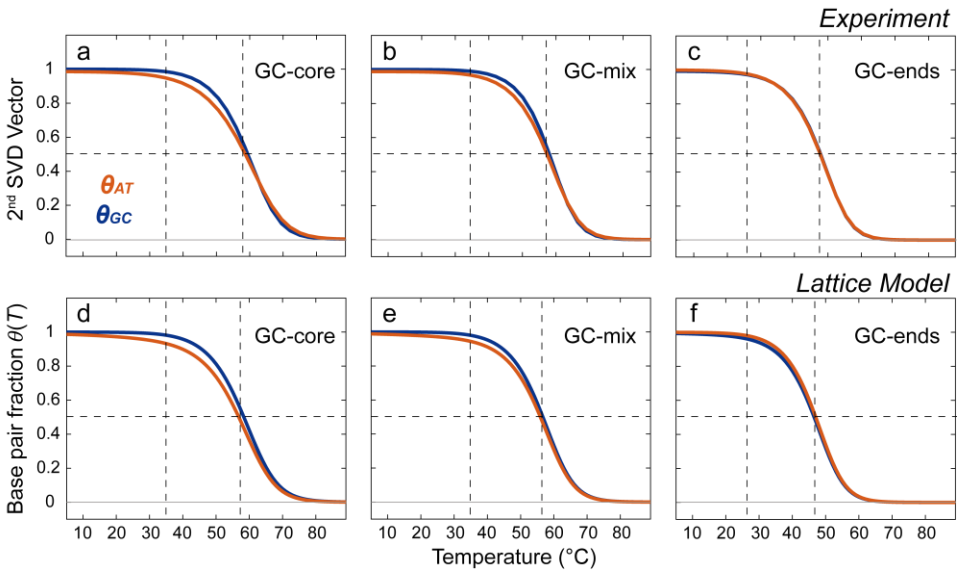

Figure 5.4: AT and GC specific experimental melting curves obtained from an SVD analysis restricted to the 1500-1590 cm^{-1} (GC) and 1590-1630 cm^{-1} (AT) frequency ranges for the (a) GC-core, (b) GC-mix, and (c) GC-ends sequences. (d-f) The corresponding AT and GC specific melting curves simulated with the lattice model by calculating the average fraction of intact AT and GC base pairs among dimers at a given temperature and then multiplying by $\theta_{ext}(T)$.

Fig. 5.4a-c shows the GC and AT specific melting curves measured by restricting the SVD analysis to the respective GC and AT frequency ranges discussed above. Comparing the AT and GC melting curves for the GC-core sequence suggests that the AT termini of this sequence begin to melt at lower temperature than the GC center of the duplex, as evidenced by the relative drop in the AT curve at lower temperature and increased asymmetry compared to the GC curve. The GC-mix sequence exhibits similar behavior, but the difference between the AT and GC melting curves is less pronounced. For the GC-ends sequence the AT and GC melting curves are essentially overlaid, suggesting that both the center and termini of the duplex dehybridize at the same temperature. To verify whether or not this sequence-dependent melting behavior measured in IR experiments is captured by the lattice model, we calculate $\theta_{AT}(T)$ and $\theta_{GC}(T)$ base pair fractions in analogy to the simulated melting curves $\theta(T)$ above, but instead of including all base pairs when computing $\theta_{int}(T)$, the AT and GC base pairs are considered independently. Fig. 5.4d-f shows the modeled AT and GC melting curves for each of the sequences. The trends across the oligonucleotide set discussed for the experimental curves are reproduced. Reasonable agreement suggests our simple model not only captures shifts in the overall dimer ensemble, but also correctly predicts which base pair contacts dehybridize first with increasing temperature.

One benefit of the relative simplicity of our lattice model is that there are a tractable number of microstates. One can visualize such a coarse grained ensemble directly in order to build an intuition for how the simulated ensemble relates to experimental signals. For example, Fig. 5.5a plots a population profile for the GC-core sequence across 5-90 °C where microstates have been grouped according to their total number of intact base pairs. Taking a single temperature slice through this profile provides insight into the heterogeneity of the oligonucleotide ensemble at a given temperature point. Plotting the sum over the population of all dimers with at least one broken

base pair (dashed green line) demonstrates how the population of partially dehybridized dimers gradually accumulates with increasing temperature, peaks around 7 °C below T_m, and then sharply drops off as the temperature is further increased.

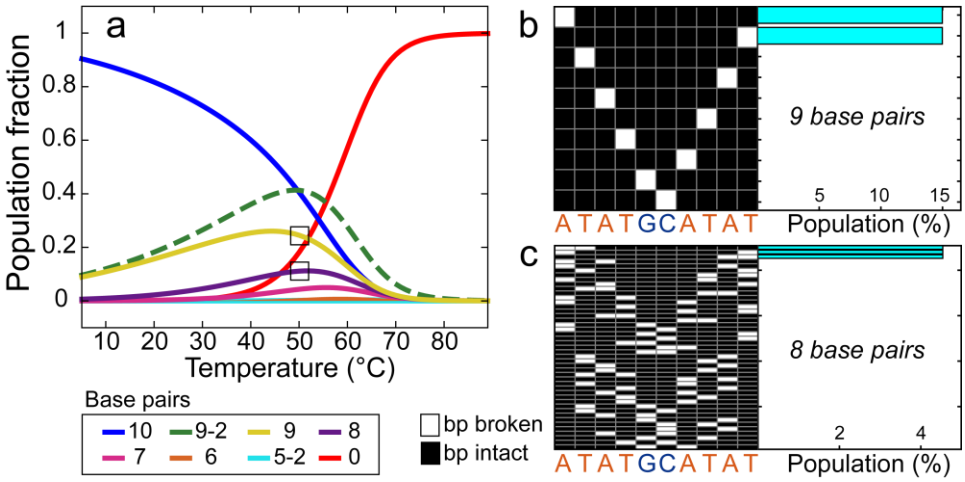

Figure 5.5: (a) Population profile of microstates grouped according to their total number of intact base pairs for the GC-core sequence. Contact plots showing all possible base pairing configurations at 50 °C for sequences with (b) 9 intact base pairs and (c) 8 intact base pairs. Each row represents a unique base pairing configuration. A white box indicates a broken base pair while a black box indicates an intact base pair at a particular site along the strand. The population percentage associated with each configuration is indicated by the horizontal bar graph to the left.

At a finer level of detail, the base pairing configurations themselves can be visualized directly through contact plots such as those in Fig. 5.5b,c which represent all of the possible base pairing arrangements for the GC-core sequence with nine and eight intact base pairs. Each row represents a unique base pairing configuration, with the oligonucleotide sequence indicated along the horizontal axis. A white box represents a broken base pair while a black box represents an intact base pair at a given site along the strand. The configurations are arranged top to bottom in order of decreasing contribution to the overall dimer population. The horizontal bar graph on the

right side of the contact plot indicates the percentage of the total population represented by each base pairing configuration at 50 °C. From these plots, it is clear that the dimer configurations with frayed ends are significantly more populated than dimers with internal loops, suggesting that the initial loss of AT base pairing observed in the experimental melting curve originates from the terminal base pairs. This example serves to illustrate how the lattice model can help rationalize experimental observations and build intuition for the interpretation of IR experiments.

5.3.2 Validation of the Model with Respect to Oligonucleotide Length

Sequence dependent effects are built into the lattice model through explicitly accounting for all possible base pairing configurations and then assigning enthalpies based on the NN dinucleotide parameters. The ability of the model to capture experimentally observed sequence effects is therefore primarily related to the enthalpic contribution to the partition function. With the aim of evaluating the entropic contributions to the model, we designed a length series of oligonucleotides with the sequence 5'-C$(AT)_n$G-3' where n = 3, 4, 5, 6, 7 or 8. As discussed above, there are two entropic terms in the partition function and both are sensitive to oligonucleotide length. The first is determined by enumerating the conformational degrees of freedom through SAWs on a 3D cubic lattice. The second is the single free parameter in the model, γ. Fig. 5.6a shows the parameterized $\theta(T)$ calculated for each of the sequences in the length series while Fig. 5.6b shows the corresponding experimental melting curves. Sample conditions were identical to those used for the sequence dependence study above. Fig. 5.6c plots the value of the γ parameter as a function of oligonucleotide length. Since γ is set against experiment to reproduce the measured T_m for each sequence, we must first evaluate the trend in melting temperatures predicted by the unparameterized model to assess if the SAWs qualitatively capture the correct length dependence.

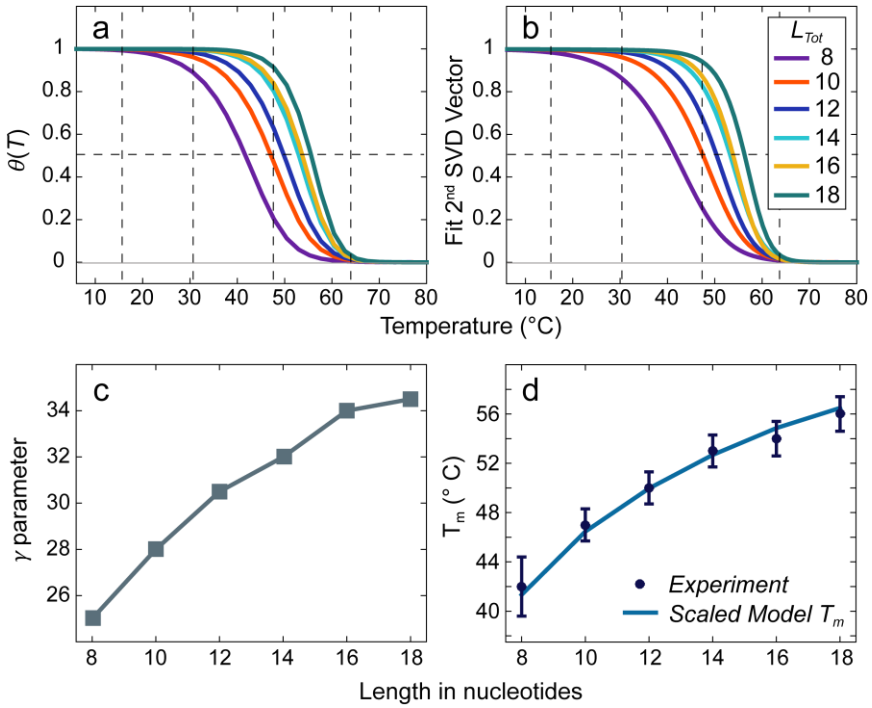

Figure 5.6: (a) The base pairing fraction $\theta(T)$ calculated across the length series using the parameterized lattice model. (b) The corresponding experimental melting curves for the length series. (c) The scaling parameter γ as a function of oligonucleotide length. (d) The experimental trend in T_m vs length plotted in dark blue points. The T_m trend predicted by the unparameterized lattice model scaled by a factor of 12.87 to shift the curve to the same temperature range as experiment is indicated by the light blue line.

Fig. 5.6d plots the experimentally measured trend in T_m as a function of length as the dark blue points. The light blue line represents the trend in T_m predicted by the unparameterized model. For the sake of comparison the model trend is scaled down by a factor of 12.87 to shift the curve into the same temperature range as the experimental points. The lattice model without any further parameterization reproduces the qualitative T_m trend, suggesting that the SAWs successfully model the scaling of the conformational entropy as a function of oligonucleotide length. The melting temperature appears to increase rapidly at first with the addition of nucleotides to the DNA strand,

but then begins to level off at longer lengths. A similar trend is observed for the γ parameter in Fig. 5.6c, suggesting that the entropic contribution due to SAWs of free nucleotide beads and the contribution due to γ share a similar length dependence. This result supports our physical interpretation of γ as largely accounting for missed degrees of conformational freedom coarse grained out by reduction of the system onto a 3D cubic lattice. Comparing the simulated and experimental melting curves directly, the model reasonably reproduces the measured trend. The most pronounced deviation between the model and experiment is for the shortest sequence (n = 3). The experimental melting curve displays a more steeply sloping transition than the simulated melting curve. This discrepancy is likely due to the comparatively low T_m of this sequence shifting the melting curve to a temperature range in which the low temperature baseline cannot be adequately sampled in experiment. As a result there are insufficient points to accurately fit the baseline slope and applying a baseline correction can distort the shape of the melting curve to some extent by imparting artificial linear character on low temperature points which do not belong on the baseline. Nevertheless the melting temperature agrees reasonably well with the T_m = 44 °C predicted by the NN model and the deviation in the slope of the transition is relatively minor between model and experiment.

5.3.3 Modeling the FTIR Spectrum using the Lattice Model Population Distributions

Since our lattice model is designed to inform the interpretation of IR experiments, it would be ideal to have a direct means of generating simulated spectra in terms of the model. Among empirically based methods, the most comprehensive and robust approach would involve the experimental determination of a set of dinucleotide component spectra for both the paired and unpaired state. One could then use the Boltzmann weights of the lattice model partition function

to generate a spectrum of the ensemble predicted by the model based on the intact/broken dinucleotides in each microstate. However, determining dinucleotide component spectra requires simultaneously fitting a massive amount of experimental data. For example, extracting the dinucleotide extinction coefficients for the UV spectrum in the 215-310 nm range required a data set consisting of over 200 single stranded oligonucleotides and over 80 unique duplex sequences.[13] A similar approach in the mid-IR would likely require a far larger data set since the FTIR spectrum in this frequency range is more sensitive to sequence composition, base pairing, and DNA hydration than the comparatively broad and featureless UV spectrum.

A rougher approach involves determining suitable component spectra for collections of microstates rather than for each dinucleotide subunit. A spectrum can then be calculated by using the sum over the appropriately grouped microstate populations to weight the contribution from the corresponding component spectrum, assuming that the experimentally observed spectrum can be expressed as a concentration weighted linear combination of these basis spectra. To illustrate this approach, we return to the population profile for the GC-core sequence in Fig. 5.5a. Microstates are grouped into three categories: dimers with all ten possible base pairs intact (p1), dimers with at least one broken base pair (p2), and monomer strands that have zero intact base pairs (p3). Fig. 5.7a plots these three populations as a function of temperature. Component spectra corresponding to each of these populations were determined using a maximum entropy method described in detail in Chapter 4. In brief, this method determines the minimum number of pure component spectra needed to represent the experimental FTIR temperature series by determining reweighting parameters for the SVD vectors subject to several physically motivated constraints, including a Shannon entropy term and a penalty term for negative absorbance/population values.

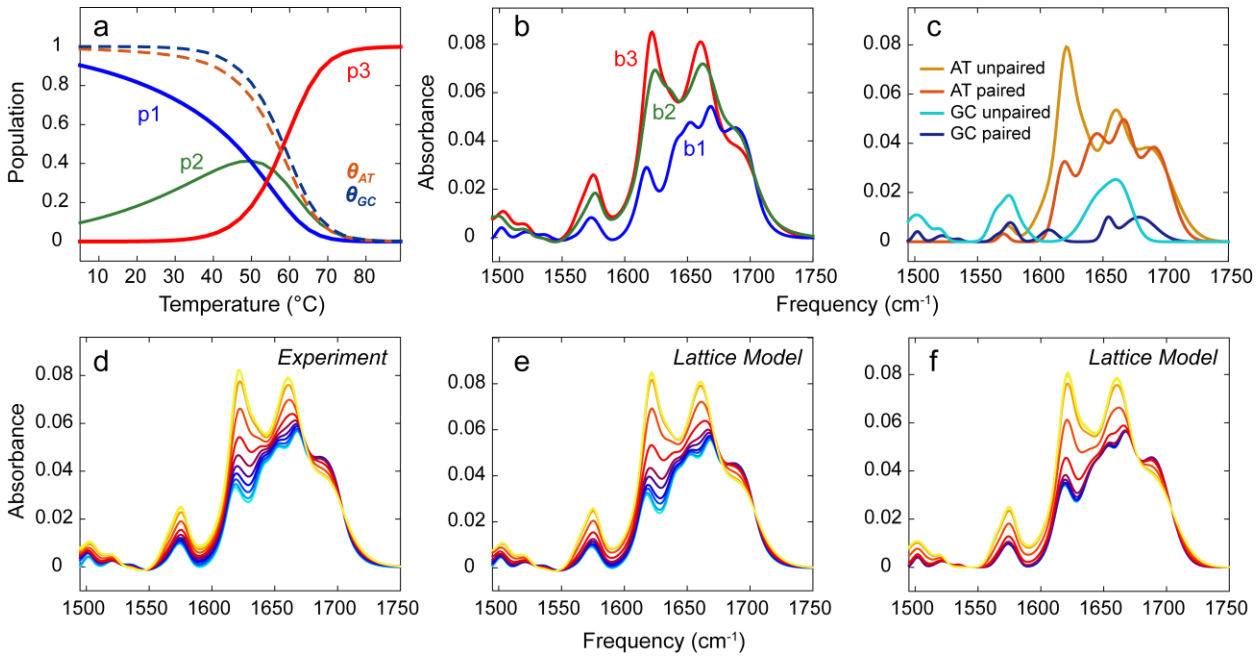

Figure 5.7: (a) Population profile for the GC-core sequence. Microstates are grouped into three populations: p1 fully paired dimers (blue), p2 partially dehybridized dimers (green), and p3 monomers (red). The fraction of intact AT and GC base pairs as a function of temperature are plotted as the dashed orange and blue curves. (b) Basis spectra extracted from a maximum entropy method applied to the GC-core FTIR temperature series. (c) AT/GC unpaired/paired basis spectra determined by fitting sums of Gaussians to the highest/lowest temperature experimental spectra. (d) Experimental FTIR spectra for the GC-core sequence across 5-90 °C in ~8 °C steps. (e) Corresponding simulated FTIR spectra determined using the lattice model populations in panel a to weight the basis spectra in panel b. (f) Simulated FTIR spectra determined by weighting the AT and GC component spectra in panel c by $\theta_{AT}(T)$ and $\theta_{GC}(T)$ plotted in panel a.

For the GC-core sequence, the method returns the three component basis spectra in Fig. 5.7b. These spectra are a reflection of the minimum amount of information needed to represent the experimental data and therefore each pure component spectrum does not necessarily relate to a simple physical interpretation. However, it is clear that the basis spectrum b1 plotted in blue and the basis spectrum b3 plotted in red in Fig. 5.7b closely resemble the lowest and highest temperature spectra, respectively, suggesting that b1 contains the component spectrum corresponding to the highly paired dimer and b3 contains the component spectrum corresponding

to the monomer state. The remaining basis spectrum b2 falls in between b1 and b3, suggesting that it is the contribution to the experimentally observed spectrum due to broken base pairs within dimers. A simulated FTIR temperature series $S_{Lat}(T)$ can be calculated through the matrix multiplication

$$S_{Lat}(T) = B_{\omega \times b} P_{p \times T} \tag{5.29}$$

where the columns of B contain the basis spectra and the rows of P contain the lattice model populations. Fig. 5.7e shows the FTIR spectrum for the GC-core sequence computed in this way at temperature points spaced ~8 °C apart from 5-90 °C. The corresponding experimental spectra are shown in Fig. 5.7d. The FTIR spectra generated by the lattice model in this way show reasonable agreement with experiment, suggesting that our assignment of the basis spectra in terms of the grouped microstates is plausible and that the microstate populations from the lattice model capture the correct behavior.

Spectral fitting of the experimental data is another possible route for obtaining basis spectra. As an example of this approach, we fit the lowest temperature and highest temperature experimental spectra to a sum of fourteen Gaussian functions. The initial guesses and bounds for the peak position, widths, and amplitudes of each Gaussian are informed by the well-characterized experimental spectra of the free nucleotides as well as the spectroscopic signatures of base pairing.[1,9] Nevertheless, this approach overfits the spectrum and even within the experimentally informed constraints there are many closely related degenerate fits to the data. Furthermore, overlapping Gaussians are not necessarily the best choice of line shape functions for the most accurate description of the spectrum, but for the sake of illustration we adopt this simplified approach. We group the spectral fits according to AT and GC base pairing. Fig. 5.7c shows the unpaired and paired basis spectra associated with each type of base pair obtained by fitting the

highest and lowest temperature experimental spectra, respectively. To model the FTIR spectrum in this case, the average fraction of intact AT and GC base pairs from the lattice model, plotted as the dashed curves in Fig. 5.7a, are used to weight a sum over the AT and GC basis spectra. For example $\theta_{AT}(T)$ weights the contribution from the paired AT basis spectrum at a given temperature while 1 - $\theta_{AT}(T)$ weights the contribution from the unpaired AT basis spectrum. Fig. 5.7f shows the simulated FTIR spectrum computed in this way. Comparison against the experimental spectra in Fig. 5.7d suggests that this method is less accurate than the maximum entropy based approach. Qualitative trends in intensity, line shape, and peak position are reproduced, with particularly good agreement with experiment at high temperature. However, the low temperature behavior of the simulated spectra suggest a much sharper transition than observed in experiment, with the first six temperature points exhibiting little change.

5.4 Conclusion

We have developed a simple oligonucleotide lattice model for informing the interpretation of IR spectroscopy experiments. The model is based on a statistical extension of the NN model where degrees of freedom at each level of detail are coarse grained onto a discrete lattice. Introducing a single free parameter results in reasonable agreement with experiment. We have validated that the model effectively captures both sequence and length dependent effects for model oligonucleotide sequences. By simulating melting curves and FTIR spectra based on a combination of suitable basis spectra weighted by the lattice model population distributions, the model can be directly related to IR experiments. The primary utility of our lattice model is its relative simplicity. Although all possible base pairing configurations are considered explicitly, there are still few enough microstates that they can be visualized directly, allowing one to build an intuition for how

experimental observables relate to the underlying oligonucleotide ensemble. For sequences up to about 20 nucleotides, the model runs in several seconds to minutes on a desktop computer and in any widely available software package, such as MATLAB. The current model only considers in-register base pairing for sequences up to 20 base pairs in length. For short sequences, this assumption is likely valid, but for longer strands, particularly if the base sequence is repetitive, shifted register dimers should be considered. In addition, the current model considers the free energy of monomer states as purely entropic. Again, this assumption is likely supported for short sequences, but for longer sequences the possibility of stable intramolecular folded states, such as hairpins, should be accounted for as this could shift the dimer-monomer equilibrium considerably. Although it would be fairly straightforward to incorporate these additional details in the future, at present the model is most applicable for short sequences. Even with this restriction we have found the insight provided by the lattice model indispensable in our studies of the biophysics of DNA dehybridization in the chapters to follow.

5.5 Acknowledgements

I thank Chi-Jui Feng and Brennan Ashwood for their careful reading of this chapter. I thank Ryan Menssen and Xinxing Zhang for collecting the length-dependent FTIR data.

5.6 References

1. Banyay, M.; Sarkar, M.; Gräslund, A., A library of IR bands of nucleic acids in solution. *Biophysical chemistry* **2003,** *104* (2), 477-488.

2. Krummel, A. T.; Zanni, M. T., DNA vibrational coupling revealed with two-dimensional infrared spectroscopy: insight into why vibrational spectroscopy is sensitive to DNA structure. *The Journal of Physical Chemistry B* **2006,** *110* (28), 13991-14000.

3. Khesbak, H.; Savchuk, O.; Tsushima, S.; Fahmy, K., The role of water H-bond imbalances in B-DNA substate transitions and peptide recognition revealed by time-resolved FTIR spectroscopy. *Journal of the American Chemical Society* **2011,** *133* (15), 5834-5842.

4. Yang, M.; Szyc, Ł.; Elsaesser, T., Decelerated water dynamics and vibrational couplings of hydrated DNA mapped by two-dimensional infrared spectroscopy. *The Journal of Physical Chemistry B* **2011,** *115* (44), 13093-13100.

5. Brauns, E. B.; Dyer, R. B., Time-resolved infrared spectroscopy of RNA folding. *Biophysical journal* **2005,** *89* (5), 3523-3530.

6. Stancik, A. L.; Brauns, E. B., Rearrangement of partially ordered stacked conformations contributes to the rugged energy landscape of a small RNA hairpin. *Biochemistry* **2008,** *47* (41), 10834-10840.

7. Hithell, G.; Shaw, D. J.; Donaldson, P. M.; Greetham, G. M.; Towrie, M.; Burley, G. A.; Parker, A. W.; Hunt, N. T., Long-range vibrational dynamics are directed by Watson–Crick base pairing in duplex DNA. *The Journal of Physical Chemistry B* **2016,** *120* (17), 4009-4018.

8. Szyc, Ł.; Yang, M.; Elsaesser, T., Ultrafast Energy Exchange via Water− Phosphate Interactions in Hydrated DNA. *The Journal of Physical Chemistry B* **2010,** *114* (23), 7951-7957.

9. Peng, C. S.; Jones, K. C.; Tokmakoff, A., Anharmonic vibrational modes of nucleic acid bases revealed by 2D IR spectroscopy. *Journal of the American Chemical Society* **2011,** *133* (39), 15650-15660.

10. Dijkstra, A. G.; Jansen, T. l. C.; Knoester, J., Modeling the vibrational dynamics and nonlinear infrared spectra of coupled amide I and II modes in peptides. *The Journal of Physical Chemistry B* **2011,** *115* (18), 5392-5401.

11. DeFlores, L. P.; Ganim, Z.; Nicodemus, R. A.; Tokmakoff, A., Amide I′− II′ 2D IR spectroscopy provides enhanced protein secondary structural sensitivity. *Journal of the American Chemical Society* **2009,** *131* (9), 3385-3391.

12. Lee, C.; Park, K.-H.; Cho, M., Vibrational dynamics of DNA. I. Vibrational basis modes and couplings. *The Journal of chemical physics* **2006,** *125* (11), 114508.

13. Tataurov, A. V.; You, Y.; Owczarzy, R., Predicting ultraviolet spectrum of single stranded and double stranded deoxyribonucleic acids. *Biophysical chemistry* **2008,** *133* (1-3), 66-70.

14. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A., Sequence-dependent mechanism of DNA oligonucleotide dehybridization resolved through infrared spectroscopy. *Journal of the American Chemical Society* **2016,** *138* (36), 11792-11801.

15. Araque, J. C.; Panagiotopoulos, A. Z.; Robert, M. A., Lattice model of oligonucleotide hybridization in solution. I. Model and thermodynamics. *The Journal of chemical physics* **2011,** *134* (16), 165103.

16. Everaers, R.; Kumar, S.; Simm, C., Unified description of poly-and oligonucleotide DNA melting: Nearest-neighbor, Poland-Sheraga, and lattice models. *Physical Review E* **2007,** *75* (4), 041918.

17. Chen, S.-J.; Dill, K. A., RNA folding energy landscapes. *Proceedings of the National Academy of Sciences* **2000,** *97* (2), 646-651.

18. Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I., Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **1994,** *33* (33), 10026-10036.

19. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G., Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics* **1995,** *21* (3), 167-195.

20. Muñoz, V.; Eaton, W. A., A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proceedings of the National Academy of Sciences* **1999,** *96* (20), 11311-11316.

21. Taketomi, H.; Ueda, Y.; Gō, N., Studies on protein folding, unfolding and fluctuations by computer simulation: I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *International journal of peptide and protein research* **1975,** *7* (6), 445-459.

22. SantaLucia, J., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences* **1998,** *95* (4), 1460-1465.

23. 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* (19), 5336-5353.

24. Markham, N. R.; Zuker, M., DINAMelt web server for nucleic acid melting prediction. *Nucleic acids research* **2005,** *33*, W577-W581.

25. Flory, P. J., *Statistical Mechanics of Chain molecules*. Hanser Publishers: New York, 1988.

26. Kool, E. T., Hydrogen bonding, base stacking, and steric effects in DNA replication. *Annual review of biophysics and biomolecular structure* **2001,** *30* (1), 1-22.

27. Wartell, R. M.; Benight, A. S., Thermal denaturation of DNA molecules: a comparison of theory with experiment. *Physics Reports* **1985,** *126* (2), 67-107.

28. Craig, M. E.; Crothers, D. M.; Doty, P., Relaxation kinetics of dimer formation by self complementary oligonucleotides. *Journal of molecular biology* **1971,** *62* (2), 383-401.

29. Pörschke, D.; Uhlenbeck, O.; Martin, F., Thermodynamics and kinetics of the helix-coil transition of oligomers containing GC base pairs. *Biopolymers* **1973,** *12* (6), 1313-1335.

30. Petersheim, M.; Turner, D. H., Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry* **1983,** *22* (2), 256-263.

31. Sanstead, P. J.; Tokmakoff, A., Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *The Journal of Physical Chemistry B* **2018**.

Chapter 6

Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy

The work presented in this chapter has been published and is reprinted with permission from:
Sanstead, P. J.; Stevenson, P.; Tokmakoff, A., Sequence-dependent mechanism of DNA oligonucleotide dehybridization resolved through infrared spectroscopy. *Journal of the American Chemical Society* **2016,** *138* (36), 11792-11801.

6.1 Abstract

Despite its important role in biology and nanotechnology, many questions remain regarding the molecular mechanism and dynamics by which oligonucleotides recognize and hybridize to their complementary sequence. The thermodynamics and kinetics of DNA oligonucleotide hybridization and dehybridization are often assumed to proceed through a two-state, all-or-nothing, dissociation pathway, but deviations from this behavior can be considerable even for short sequences. We introduce a new strategy to characterize the base pair specific thermal dissociation mechanism of DNA oligonucleotides through steady-state and time-resolved infrared (IR) spectroscopy. Experiments are interpreted with a lattice model to provide a structure specific interpretation. This method is applied to a model set of self-complementary 10 base pair sequences in which the placement of GC base pairs is varied in an otherwise AT strand. Through a combination of Fourier transform infrared (FTIR) and two dimensional infrared (2D IR) spectroscopy, experiments reveal varying degrees of deviation from simple two-state behavior. As

the temperature is increased, duplexes dissociate through a path in which the terminal bases fray, without any significant contribution from loop configurations. Transient temperature jump experiments reveal timescales of 70-100 ns for fraying and 10-30 μs for complete dissociation near the melting temperature. Whether or not frayed states are meta-stable intermediates or short-lived configurations during the full dissociation of the duplex is dictated by the nucleobase sequence.

6.2 Introduction

The hybridization and dissociation of nucleic acids are central events in numerous biological processes ranging from replication[1,2] to gene regulation[3,4], as well as technological applications ranging from bioassays[5,6] to DNA nanotechnology[7,8]. Technological applications depend on the remarkable property of high fidelity base pairing that results in selective binding between complementary sequences, and from the growing predictive power in designing sequences for self-assembly. It is well established that many factors influence the thermodynamics and kinetics of DNA hybridization, including strand length,[9,10] strand concentration,[11,12] base composition,[13,14] and the concentration of monovalent and divalent cations.[15]

Even though our knowledge of the thermodynamics of DNA hybridization is broad, our understanding of the kinetics, dynamics, and molecular mechanism by which single stranded oligonucleotides diffuse into contact, recognize, and hybridize to their complementary sequence remains limited. In addition to the atomistic insight provided by all-atom molecular dynamics simulations,[16-18] recently developed coarse-grained methods that represent DNA with a reduced number of interaction sites per nucleotide have simulated the hybridization mechanism in detail.[19,20] These models predict rich hybridization dynamics including initial nucleation of a few key contacts followed by zippering of the remaining base pairs (bps), shifted register slithering of

one strand along another, and various internal displacement schemes depending on the nucleobase sequence.[21,22] At this time additional experimental insight is required to directly investigate these or other potential mechanisms. Past studies relying on ultraviolet (UV) hyperchromicity lack base-specific structural sensitivity.[9,14,23] Förster resonance energy transfer (FRET) experiments offer localized insight into strand proximity,[24-26] but they do not directly reveal pairing between individual bases and raise questions regarding how the fluorescent tag may potentially alter the dissociation mechanism.

The experimental challenges are such that we are still limited in our ability to characterize structural variation in DNA oligonucleotide duplexes. Especially for short sequences, DNA melting is often assumed to proceed in an all-or-nothing two-state fashion in which all of the bps are either entirely intact or broken, but in reality dimers have the potential to adopt a variety of partially associated states. The two-state assumption greatly simplifies the analysis and proves to be an adequate description for many applications involving both oligonucleotide and polymeric DNA.[11,13,27] However evidence suggests that deviations from the two-state assumption are possible for short sequences and even probable with certain motifs such as terminal AT base pairs (bps), highly heterogeneous sequences, and hairpin structures.[28-33]

Experimentally, UV absorption spectroscopy is the most widely used method to track the thermal denaturation of DNA where the hyperchromicity at ~260 nm traces the melting curve. A van't Hoff analysis can be performed to extract the enthalpy, entropy, and free energy of dissociation assuming that a two-state picture is adequate. Comparing calculated van't Hoff thermodynamics to model-independent calorimetric results is a common check for multistate behavior, where a discrepancy suggests a failing of the two-state model.[12,31] The standard UV

method is rapid, convenient, and requires a small amount of sample, but offers little structural or mechanistic insight into dehybridization.

In order to describe DNA dehybridization in finer detail, a structurally sensitive experiment that can observe and quantify partially associated configurations of complementary strands on the time-scale of their interconversion is desirable. For kinetic studies, the ps-ms timescales relevant to nucleic acid structural changes necessitate techniques with sufficient temporal resolution, which have included time-resolved IR[34,35], UV[9,14,23,31], and fluorescence spectroscopies.[25,36] The more complex challenge involves untangling stable thermodynamic intermediates with varying degrees of disorder from transiently formed kinetic states.

To address this challenge, we have developed an IR spectroscopy based strategy to characterize the dehybridization of DNA oligonucleotides. The IR spectral region from 1450-1800 cm^{-1} contains in-plane base vibrations including carbonyl stretches and ring breathing modes that provide a unique fingerprint for each of the four DNA nucleobases.[37] Furthermore the spectrum in this region is shaped dramatically depending on whether or not a base is paired or free.[38] Through a combination of Fourier transform infrared (FTIR) and two dimensional infrared (2D IR) spectroscopy, the dehybridization of DNA strands can be tracked at a level of detail that distinguishes GC and AT bps. Equilibrium melting experiments are complemented with transient temperature jump (T-jump) experiments that track the ns-μs dissociation dynamics of DNA oligonucleotides in real-time. To provide a detailed interpretation of spectroscopic results, we use a lattice model that builds on the nearest-neighbor (NN) model[13] and allows the explicit consideration of partially melted duplex configurations. This model accurately captures the T_m as well as the shape of the melting curve, and provides an extra level of insight into the subpopulations of intact AT and GC bps for the sequences studied here. The details and validation of the lattice

model are the subject of Chapter 5. In this study, the method has been applied to a model set of self-complementary DNA 10-mers where the placement of GC base pairs is varied in an otherwise AT sequence. Pre-melting events such as fraying are distinguished from the loss of final contacts resulting in dissociation to the monomer state. We find the propensity to fray as well as the manner in which fraying proceeds is dictated by the placement of the GC pairs within the sequence.

6.3 Results

6.3.1 Temperature Ramp FTIR of Model Oligonucleotide Sequences

For the purposes of developing the methods and analysis required to characterize the dissociation of DNA oligonucleotides using IR spectroscopy, we designed a set of model sequences that have the same length and bp content, but varying base sequence. Since GC bps show greater stability than AT bps,[39] we expect that different regions of the oligomers will have contrasting thermodynamics, assuming they do not melt in a strictly two-state manner. Additionally the placement of GC pairs in an otherwise AT sequence provides a natural and noninvasive probe of local structure. An oligonucleotide set that fits these criteria is: 5'-ATATGCATAT-3' (GC-core), 5'-ATGATATCAT-3' (GC-mix), 5'-GATATATATC-3' (GC-ends), and 5'-ATATATATAT-3' (AT-all). We will refer to the sequences by the shorthand names in parentheses for convenience. The term "dimer" is used to represent any configuration of two associated strands, while "monomer" refers to a single strand without hydrogen bonds to another strand.

As an initial assessment, FTIR temperature ramp experiments were performed to track the helix-to-coil transition between 5-90 °C. We have previously assigned the vibrational modes for each of the nucleobases in the 1500-1750 cm^{-1} range, finding these vibrations to be highly coupled

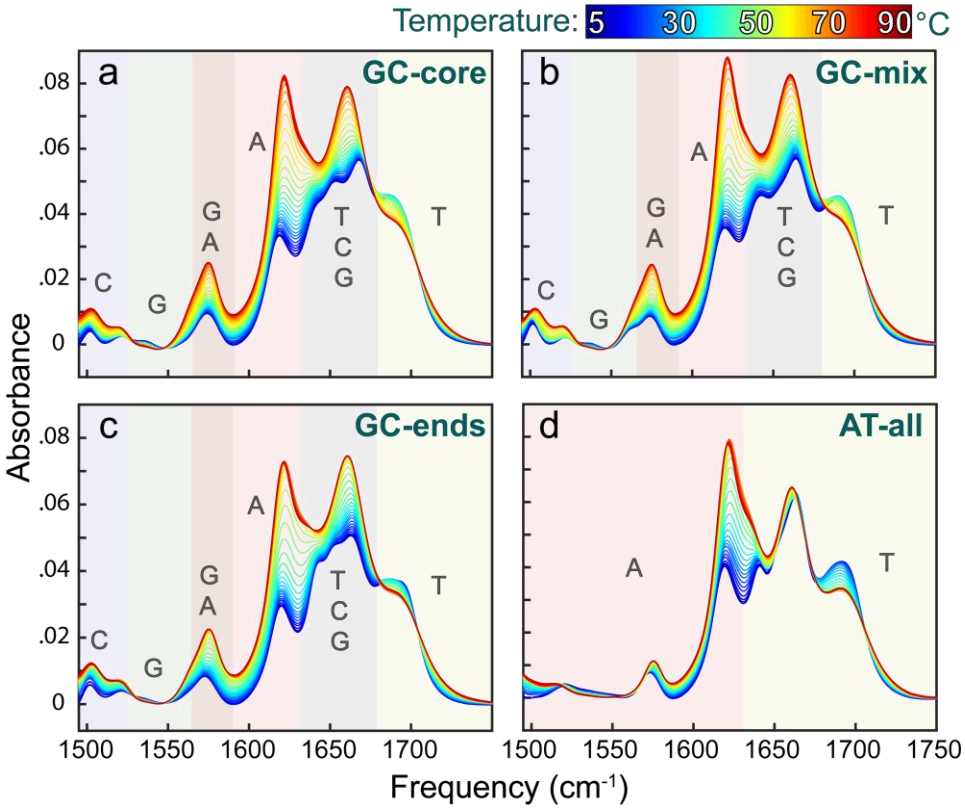

Figure 6.1: Temperature ramp FTIR spectra 5-90 °C with 2 mM oligonucleotide, 50 mM pD 7.2 sodium phosphate buffer, 240 mM NaCl, 18 mM $MgCl_2$ for (a) GC-core, (b) GC-mix, (c) GC-ends, and (d) AT-all. Shading indicates the contributions of the four nucleobases to the spectra.

and delocalized across the base.[37] In the thermally dissociated state the IR spectrum closely resembles a composition weighted sum of the individual free nucleotides that make up the sequence, but variations between the sequences, primarily small differences in ring mode intensity, reflect differing degrees of single strand stacking interactions in the monomer. In contrast the formation of a DNA duplex significantly influences and shapes the vibrational frequencies, intensities, and line shapes.[38,40] In addition, much of the character from the isolated bases is evident in the spectra, allowing meaningful and base-specific structural insight. Many changes to the spectrum are apparent with increasing temperature (Fig. 6.1) but the most prominent feature is the

doubling in intensity of the 1622 cm^{-1} A ring mode upon duplex melting. This peak reports primarily on the amount of paired A since the intensity is suppressed upon the formation of a stacked and hydrogen bonded Watson-Crick (WC) pair. Similar to the A ring mode, the C ring modes at 1503 cm^{-1} and 1520 cm^{-1} as well as the G ring modes at 1564 cm^{-1} and 1575 cm^{-1} are likewise suppressed upon duplex formation. These features report on the status of GC base pairs, although the presence of an additional A ring mode at 1574 cm^{-1} (Fig. 6.1d) that overlaps with the higher frequency G ring modes complicates things slightly. The shading in Fig. 6.1 highlights the contributions from each of the nucleobases across the 1500-1750 cm^{-1} frequency range.

Above 1630 cm^{-1} the spectrum becomes congested with multiple overlapping peaks. However some base-specific information is still discernable such as the T carbonyl mode at 1690 cm^{-1} that increases in intensity when T is paired. With increasing temperature the loss of fine structure and the large intensity growth around 1660 cm^{-1} are due to increasingly unpaired T, C, and G but the GC features in this frequency range are largely obscured beneath T contributions since the sequences have only 20% GC content.

6.3.2 Melting Curve Analysis

To describe the global spectral changes between 1495-1750 cm^{-1} upon dehybridization, we analyzed the IR spectra over the 5-90 °C temperature range using singular value decomposition (SVD). Temperature-dependent changes to the IR spectrum are obtained from the normalized 2^{nd} SVD component, shown in Fig. 6.2 as the blue data points. Melting curves are obtained from this data by subtracting the sloping baselines indicated by the dashed lines in Fig. 6.2 at high and low temperature. In practice, the melting curves (Fig. 6.3a) were extracted from a fit to a two-state thermodynamic model that allows for a temperature dependent enthalpy and entropy, which is

described in Chapter 1. Analyzing the full IR spectrum results in melting curves similar to those obtained by tracking the intensity of a single broad feature in the UV spectrum and therefore serves as the starting point of our analysis. We have previously compared the T_m trends obtained from FTIR and UV methods for DNA oligonucleotides, finding good agreement between the methods and with UV-based model predictions.[15,41] The melting curves are assumed to report on the fraction of intact base pairs (θ_{bp}) and T_m is defined as the temperature at which 50% of all bps are intact.[11,12]

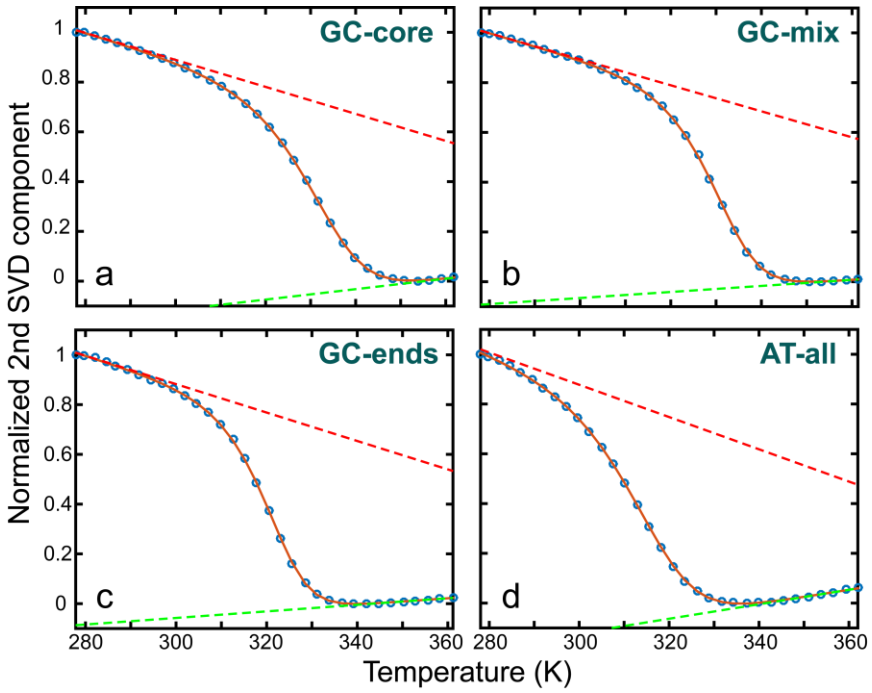

Figure 6.2: Normalized 2nd SVD component from temperature ramp FTIR (blue points) fit to the two-state model described in Chapter 1 (solid red line). Sloping baseline fits are plotted for illustration using dashed lines.

It is apparent that the base sequence and composition determine both the shape of the melting curve and the melting temperature. AT-all, lacking any GC content, has the lowest T_m at 40 °C. Despite their identical GC content, GC-core, GC-mix, and GC-ends show distinct melting curves due to the variable placement of the GC pairs in each sequence. The T_m for GC-ends is

47 °C, and the melting curve is a sharp sigmoid symmetric about the inflection point, as expected for two-state melting.[42] At 57 °C and 56 °C, both GC-core and GC-mix have a T_m that is shifted to higher temperature, but the shape of the melting curves differ. Qualitatively, the four curves also vary in their slopes at $\theta_{bp} = 0.5$ as well as the degree of symmetry about that point. For example, slight deviation from an idealized sigmoidal melting curve in the case of GC-core hints at a possible departure from two-state behavior; however additional information is required in order to understand the sequence effects on dehybridization in detail.

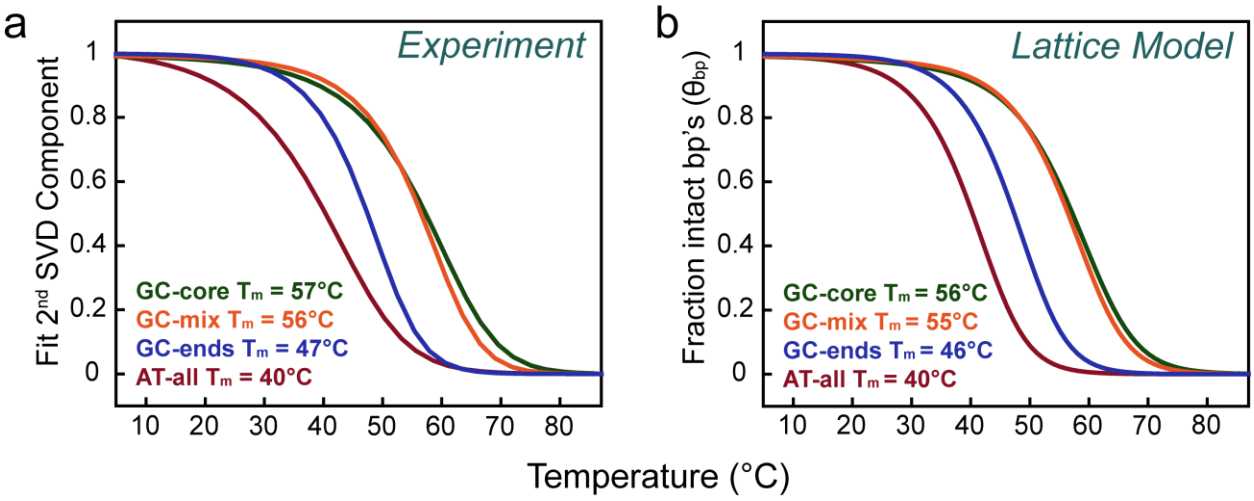

Figure 6.3: Comparison of melting curves (a) determined from a two-state analysis fit to the 2nd SVD component from FTIR temperature ramps, and (b) calculated from the fraction of intact base pairs in the lattice model.

As discussed in Chapter 1, the simplest thermodynamic interpretation of the melting curves invokes a two-state assumption and a van 't Hoff analysis to determine the free energy of dissociation for each of the sequences. This approach assumes that the change in specific heat ΔC_p is zero and that the enthalpy is temperature independent. A more detailed description still within the two-state all-or-nothing assumption of base pairing allows for a temperature-independent

non-zero ΔC_p and therefore allows for a temperature dependent ΔH and ΔS. To go beyond an all-or-nothing description of the thermodynamics of base pairing the statistical model described in Chapter 5 can be used to calculate the free energy of dehybridization. A comparison of these three thermodynamic treatments (Fig. 6.4) shows varying degrees of agreement across the sequences.

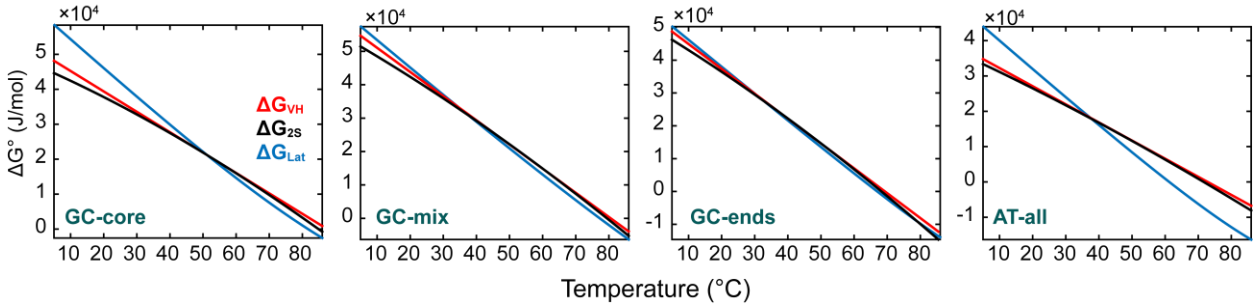

Figure 6.4: Comparison of the free energy of dissociation determined by van 't Hoff analysis, fitting the melting curves to a two-state model with a non-zero ΔC_p, and the statistical lattice model from Chapter 5.

Comparing the free energies of dissociation calculated from the van 't Hoff analysis, two-state thermodynamics with non-zero ΔC_p, and the structurally explicit lattice model offers further insight into possible deviations from all-or-nothing dissociation. Since the first two treatments invoke a two-state assumption explicitly, one would expect them to prove most successful when applied to those sequences for which the two-state assumption holds best and to break down for those sequences for which it does not. Once again GC-ends appears essentially two-state, with the ΔG^o predictions from all three models nearly overlaid across the entire temperature range. In contrast the $\Delta G°$ curves for GC-core suggest a breakdown of the two-state model for this sequence, with increasing discrepancy between the models with increasing distance from T_m. GC-mix once again appears to be an intermediate case between GC-ends and GC-core. The agreement between the lattice model and the experimentally derived van 't Hoff and two-state model thermodynamics

is worse for AT-all than for the remaining sequences, but this is not surprising given the deviation between the experimental and predicted melting curves for this sequence discussed below.

For all of the sequences all of the models prove consistent for some range around the melting point, but the temperature range over which this agreement persists appears to be dictated by the degree to which the two-state assumption holds. This result is not unexpected considering that both the van 't Hoff and two-state model rely on a set of reference thermodynamics determined at T_m, assume some dependence of the thermodynamic parameters on temperature, and then extrapolate to determine the dissociation free energy at any temperature of interest with respect to the reference. The melting temperature T_m is selected as the reference temperature because this corresponds to a point at which the data are steeply varying and we are therefore most sensitive and confident in the measurement at this point. It seems reasonable that the further away from the reference temperature one attempts to extrapolate the less accurate the result will be and that the rate at which this deviation occurs should reflect the accuracy of the underlying assumptions in the model with respect to the system under consideration, hence the better agreement across the entire temperature range for GC-ends where the two-state assumption appears accurate.

6.3.3 Temperature Ramp 2D IR

Despite the ability to clearly distinguish base-specific features in the FTIR, spectral congestion and overlap can pose a challenge when interpreting the spectra. As a step toward experimentally characterizing which bases are intact in partially melted dimer structures, we used 2D IR spectroscopy to report on base pairing of a single type, either AT or GC. 2D IR separates AT and GC reporter regions by spreading information out onto a second frequency axis. The peaks in a 2D IR spectrum consist of oppositely signed doublets (red above blue) with the on-diagonal

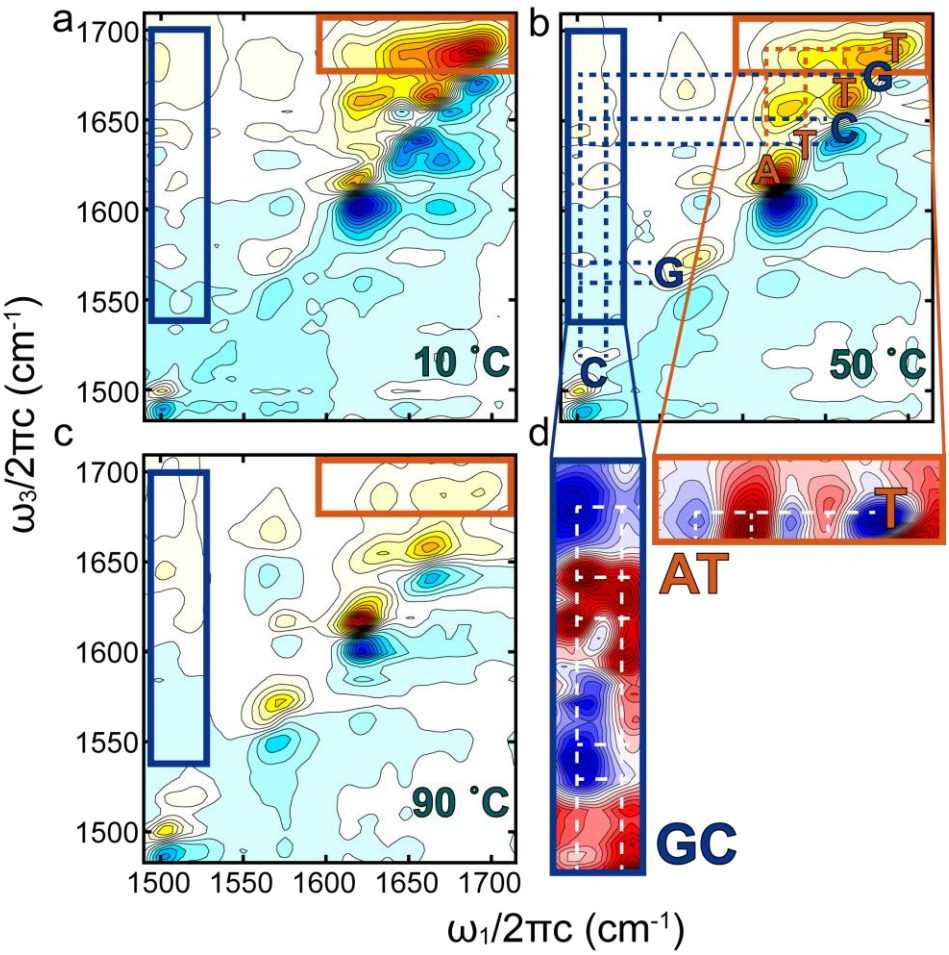

Figure 6.5: Representative 2D IR spectra at (a) 10, (b) 50, and (c) 90 °C from the GC-core temperature ramp. The GC and AT cross peak regions are indicated by the blue and orange boxes, respectively. Dashed lines serve as a guide for locating cross peaks. (d) The second SVD component spectra demonstrate the changes in the cross peak regions upon thermal dehybridization, with blue/red (red/blue) doublets representing loss (gain).

peaks mirroring the peaks in the linear spectrum and the off-diagonal cross peaks reporting on the coupling between the vibrational modes. In addition to the intensity changes and frequency shifts discussed above, the formation of a WC pair gives rise to additional intermolecular cross-peaks due to vibrational couplings across the hydrogen bonded bases. These cross-peaks offer a direct measure of the extent of intact WC pairs and therefore tracking these off diagonal features provides

a means of isolating the response from either the AT or GC bps. Temperature ramp 2D IR was used to monitor the helix-to-coil transition for each of the sequences. Three representative 2D IR spectra from the GC-core temperature ramp (Fig. 6.5) demonstrate the substantial spectral changes observed as the duplex melts and dissociates.

To separately quantify GC and AT base pairing, SVD analysis was performed on GC and AT specific cross-peak regions. The blue box in Fig. 6.5 defines the GC region and includes cross peaks between the 1500-1520 cm^{-1} C ring modes and the G ring modes in the 1540-1575 cm^{-1} range as well as to the G carbonyl mode at 1680 cm^{-1}. The orange box defines the AT region and contains the 1690 cm^{-1} T carbonyl peak as well as its cross peaks to the lower frequency T modes and the 1625 cm^{-1} A ring mode. Dashed lines in Fig. 6.5 illustrate the coupling of these features while the second SVD component spectra (Fig 6.5d) demonstrate how these cross-peaks change in response to duplex melting. A red above blue doublet represents intensity gain while a blue above red doublet represents intensity loss. However, these cross-peak regions are crowded with information and many of the features overlap. As a result much of the doublet structure is obscured and, roughly, blue features represent intensity loss while red features represent intensity gain. The loss of intermolecular cross-peaks is clearly visible as well as the gain of intramolecular cross-peaks upon duplex dissociation as the vibrations are localized onto the unpaired base.

Comparing the SVD vectors across the set of sequences provides insight into differences in the local base pairing environment. The SVD analysis restricted to the GC and AT cross-peak regions yields a set of frequency space (U) and a set of temperature space (V) vectors which when weighted by their corresponding singular values (S) return the original data through the matrix multiplication USV^T. The first vectors (U1, V1) typically resemble an average across the entire data set. The second vector U2 contains the primary spectral changes with increasing temperature

while V2 traces the temperature dependence of the features in U2. Therefore an effective melting curve for the AT (θ_{AT}) and GC (θ_{GC}) bps can be determined (Fig. 6.8a-d) from the second SVD components, since they track how the GC/AT features evolve with temperature.

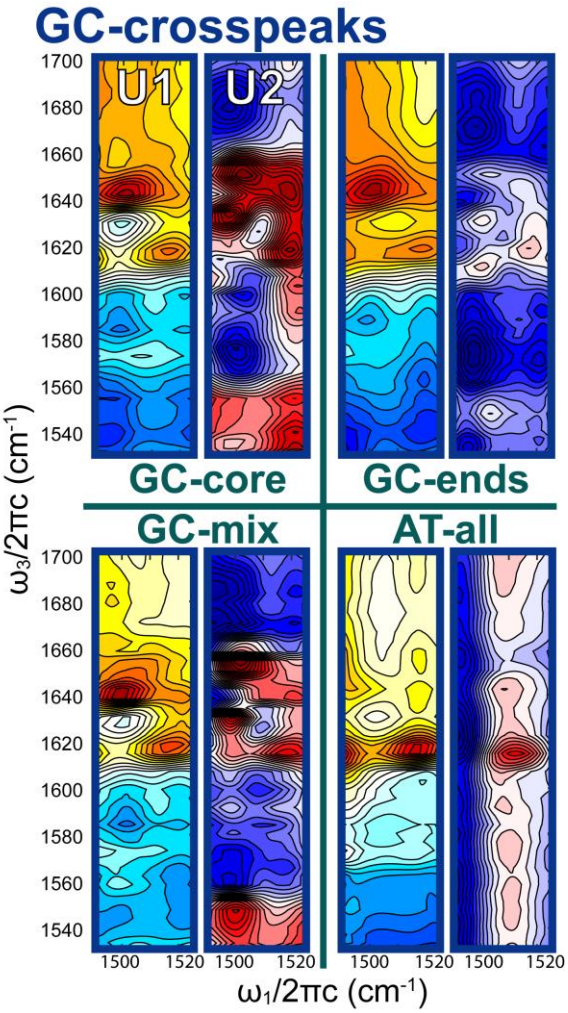

Figure 6.6: The first two frequency space vectors, U1 and U2, from the SVD analysis of the GC cross peak region defined in Fig. 6.5

A comparison of the U1 vectors (Fig. 6.6-6.7) against the temperature dependent 2D IR surfaces (Fig. 6.5) reveals that the U1 vectors closely resemble an average across the temperature

series. The U2 vectors for the GC (Fig. 6.6) and AT (Fig. 6.7) cross-peak regions convey the primary spectral changes in each of their respective regions. The sign of the SVD vectors is arbitrary, but comparison against the observable changes in the temperature ramp allows the assignment of loss and gain features. From the intensity loss of the 1690 cm^{-1} T carbonyl mode and the loss of the cross-peak to the 1622 cm^{-1} A ring mode in the AT region as well as the loss of the intermolecular cross peaks between the C ring modes at 1500-1525 cm^{-1} and the G modes at 1575 cm^{-1} and 1680 cm^{-1} in the GC region, we assign blue/red doublets as loss features and red/blue doublets as gain features. In general as the temperature increases the intermolecular cross-peaks disappear as base pairs break and the vibrational coupling across the WC pairs mediated by the hydrogen bonds and base stacks in the duplex are disrupted. A gain in intensity of the intramolecular cross peaks is observed in conjunction with the loss of the intermolecular cross-peaks, which can be explained by increasing localization of the vibrations onto the individual nucleobases as they dissociate and unstack.

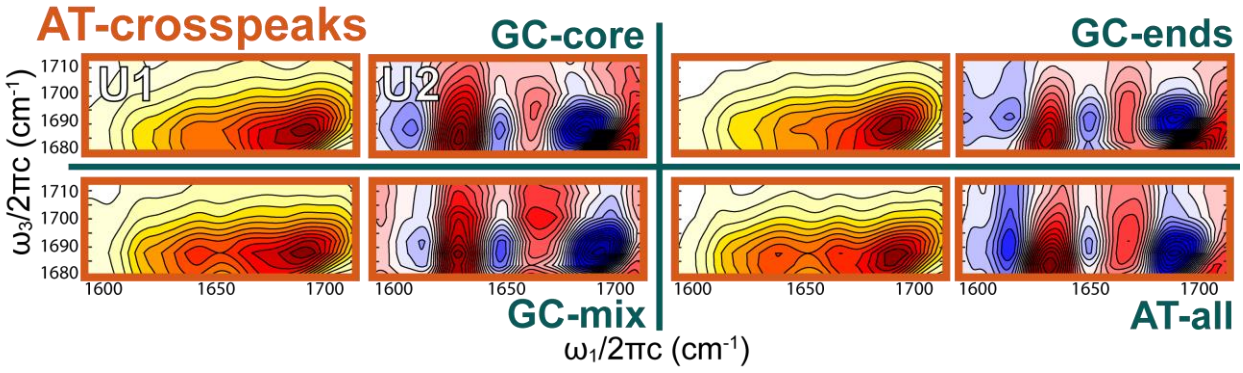

Figure 6.7: The first two frequency space vectors, U1 and U2, from the SVD analysis of the AT cross peak region defined in Fig. 6.5

It is interesting to note that the GC U2 vectors are distinct for each sequence while the AT U2 vectors are essentially the same regardless of sequence. The differences in the GC U2 vectors

are likely due to the fact that the GC base pairs in each sequence represent a unique local environment. For instance a GC pair at the center of the GC-core sequence has one AT and one GC neighboring pair while for GC-mix all GC pairs are flanked by AT pairs. The GC pairs in GC-ends cap the sequence and only have a single AT neighbor while the other side of the WC pair is solvent exposed. This less stacked, less intermolecularly coupled environment is reflected in the U2 vector, where the large increase in the intramolecular C cross peak intensity seen for GC-mix and especially GC-core as the four peak pattern centered around $\omega_1 = 1515$ cm^{-1} and $\omega_3 =$ 1640 cm^{-1} is greatly reduced in the GC-ends U2 vector. This reduction in intensity gain is consistent with the solvent exposed position that the GC pairs occupy in the GC-ends sequence, which bears more resemblance to the environment encountered by a solvent exposed free nucleotide. The GC region for the AT-all sequence, having no GC pairs, contains essentially noise.

The AT cross peak regions show far less variation across the sequences, but considering the local environment of the majority of the AT pairs in which the flanking base pairs are also AT pairs, this result is not unexpected. From the perspective of most AT pairs the sequence context is similar independent of which oligonucleotide in the set it belongs. As a consequence the changes observed in the AT cross peak region are similar for all of the sequences considered here.

6.3.4 Maximum Entropy Method to Reconstruct Spectral Component Amplitudes

To account for the possibility of spectroscopically resolvable structural heterogeneity in the dimer ensemble along the helix-to-coil transition, we employed a maximum entropy method to evaluate the minimum number of spectral components contributing to the temperature ramp FTIRs for each sequence.[43] The method is described in detail in Chapter 4. The reconstructed amplitudes are plotted in Fig. 6.8e-h. All of the sequences have a spectral component that dominates at low temperature (blue) and a second that dominates at high temperature (red),

suggesting these components correspond to the fully paired dimer and monomer, respectively. The 3rd component amplitude (green) is therefore assigned as a spectral contribution from partially melted dimers.

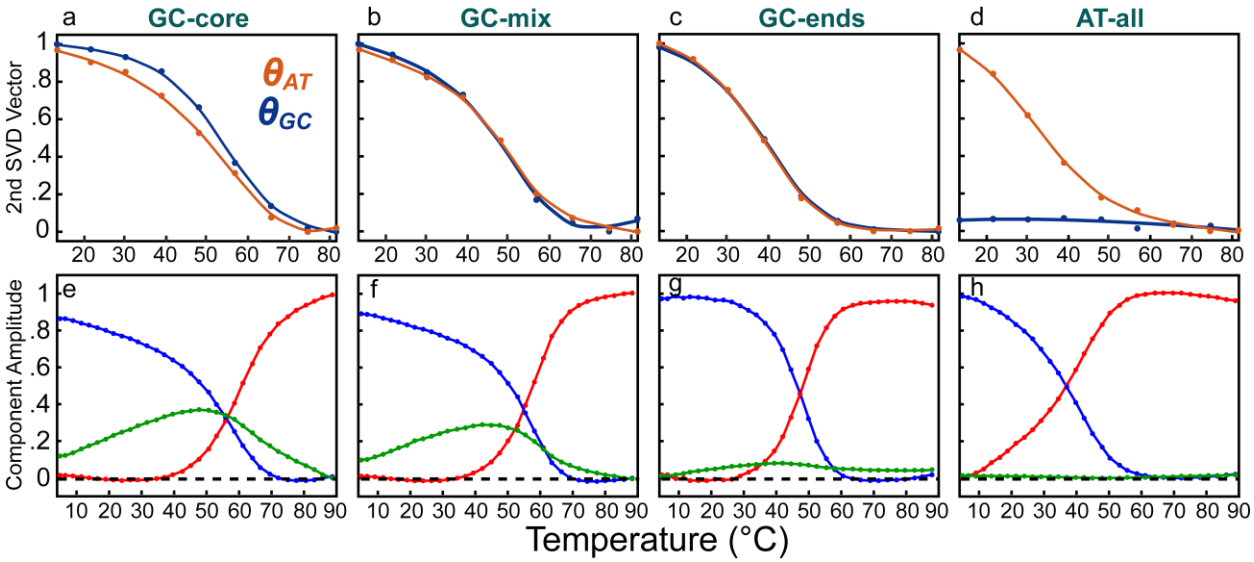

Figure 6.8: The AT and GC melting curves measured from temperature ramp 2D IR (top panel) and the maximum entropy reconstructed spectral amplitude profiles from the FTIR SVD analysis (bottom panel) for (a,e) GC-core, (b,f) GC-mix, (c,g) GC-ends, and (d,h) AT-all.

The reconstruction for GC-core suggests a sizeable partially melted population along the helix-to-coil transition represented by the considerable 3rd component amplitude peaking at nearly 40% ~10 °C below T_m. Insight from 2D IR reveals the nature of these partially melted configurations. The θ_{GC} transition is shifted to higher temperatures with respect to θ_{AT} and appears to be a sharper, more two-state like transition, while θ_{AT} appears to have a more gradual and sloped upper baseline indicating loss of AT contacts at temperatures well below T_m (Fig. 6.8a). These results suggest a dissociation pathway where the terminal AT base pairs begin to fray resulting in an ensemble of fully duplexed and partially melted dimers. However from the perspective of a GC

bp embedded within the core of the sequence dissociation is essentially two-state, consistent with the sharper transition observed for θ_{GC}. GC-mix shows less pronounced contrast between θ_{GC} and θ_{AT} as well as a smaller contribution from a 3rd spectral component amplitude, consistent with a reduced population of partially melted dimers with respect to GC-core (Fig. 6.8b). The maximum entropy 3rd component amplitude is virtually flat and the AT/GC bp fractions from 2D IR are essentially overlaid for GC-ends (Fig. 6.8c,g), as one would expect for two-state dissociation.

6.3.5 Temperature Jump Dissociation Kinetics

Changes in the IR spectrum were tracked in response to a five nanosecond 18 °C T-jump to provide a direct window into the dehybridization mechanism. The T-jump spectrometer and experiment have been described in detail previously[44,45] and in Chapter 3. Instead of collecting T-jump difference spectra for the full 2D IR surface, we collected transient heterodyned dispersed vibrational echo (t-HDVE) spectra, which are related to the projection of the 2D IR spectrum onto the ω_3 axis. Since the ω_1 axis is not resolved, the data acquisition time is significantly faster allowing finer sampling of kinetic traces. However, projection onto a single frequency axis can result in overlapping spectral features that must be accounted for when interpreting t-HDVE data. The spectral changes in the temperature ramp 2D IR spectra (Fig. 6.5) serve as a valuable guide when identifying features that track the GC and AT responses independently. The induced absorption of the A ring mode centered around 1595 cm^{-1} demonstrates substantial intensity change upon loss of AT pairing and when projected onto the ω_3 axis provides a clear marker for the AT features. Similarly the response of the induced absorption of the G ring modes centered around 1545 cm^{-1} serves as a reporter of GC base pairing.

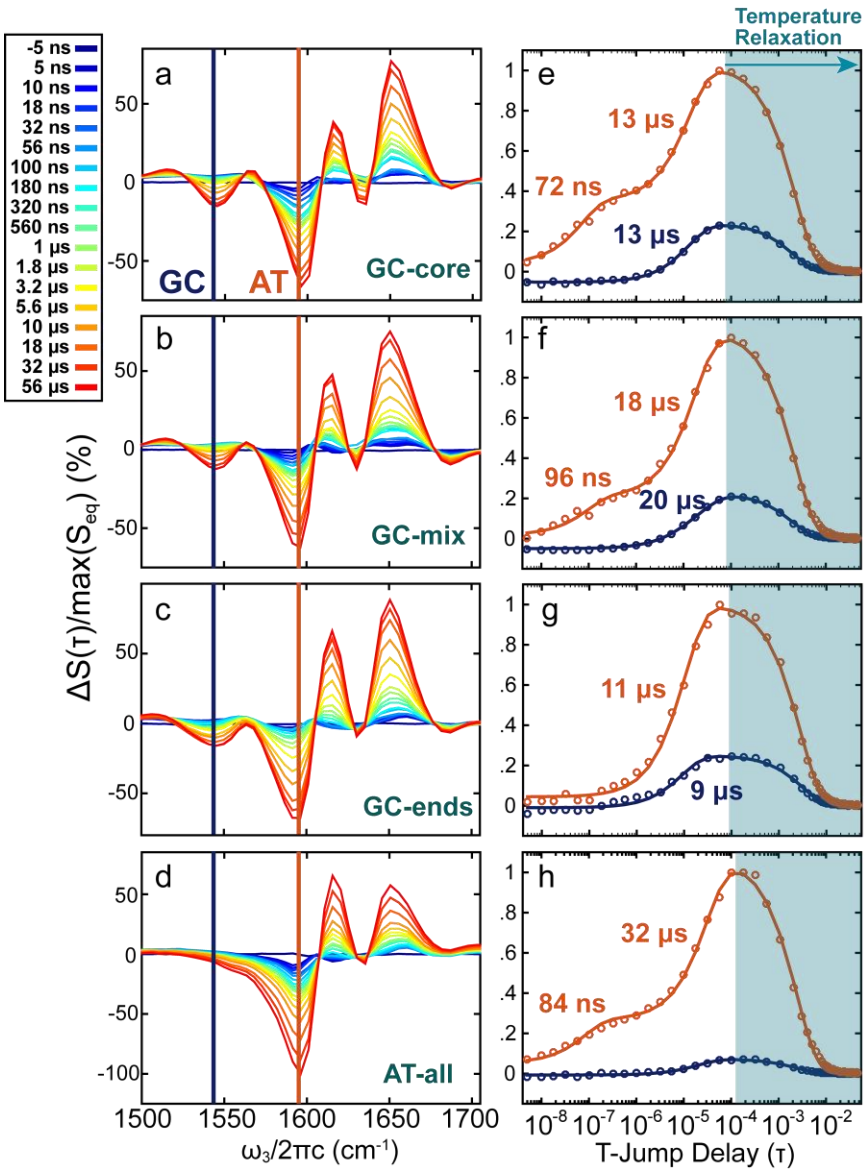

Figure 6.9: The t-HDVE spectra for the (a) GC-core, (b) GC-mix, (c) GC-ends, and (d) AT-all sequences. Kinetic traces tracked at the 1595 cm^{-1} AT (orange) and 1545 cm^{-1} GC (blue) frequencies normalized to the maximum of the AT trace. Displayed time constants are from exponential fits.

Temperature jumps of 18 °C were performed across the center of the reconstructed spectral amplitude distributions with the initial temperature set 5 °C below the maximum of the 3rd component amplitude plotted in green (Fig. 6.8). Transient HDVE spectra for each of the

sequences (Fig. 6.9) reveal significant spectral changes in the time-window 5 ns-100 μs. To distinguish changes specific to AT and GC pairs, the kinetics were tracked at 1595 cm^{-1} and 1545 cm^{-1}, respectively. All transient responses rise away from the baseline until a T-jump delay of ~100 μs, after which the signal drops as a result of the temperature re-equilibration of the sample. The clear differences in the AT and GC kinetics in the case of the GC-core and GC-mix sequences illustrate the nucleobase specificity of these experiments. Whereas all of the equilibrium melting curves can be evaluated in terms of approximate two-state melting, the T-jump experiments for three out of the four oligonucleotides show clear bimodal melting kinetics, which is inconsistent with the expectation for all-or-nothing unfolding where all of the bps break in concert and one would expect a single consistent exponential trace regardless of which base feature is tracked.

Fitting kinetic traces to exponentials to extract time constants, we find that both AT and GC traces for all of the sequences display a relaxation response between 10-30 μs, except the 1545 cm^{-1} trace for the AT-all sequence which is essentially flat, since there are no GC pairs in the sequence. This relaxation process is consistent with full dissociation into the monomer state since it is universal and concerted between the AT and GC bps.

A faster 70-100 ns response is observed only from the AT features for all of the sequences that have terminal AT pairs, but not for GC-ends, suggesting the more weakly paired AT's at the termini respond more-or-less independently of the GC's more central to the helix. Therefore, this early time response would appear to involve fraying of the terminal AT base-pairs. The magnitude of the fraying response is greatest for GC-core, consistent with a larger population of partially melted dimer configurations for this sequence, whereas GC-mix and AT-all show a reduced early time response. For GC-ends both the AT and GC traces are well fit by a single exponential rise

with a ~10 μs time constant consistent with a two-state dissociation mechanism. These observed timescales are in agreement with those reported previously for the unfolding and dissociation of small nucleic acid systems initiated by a rapid temperature jump.[14,35]

6.3.6 Comparison to the Lattice Model

To help interpret experimental results in terms of conformational variation within the melting duplex ensemble, we used a statistical lattice model for DNA hybridization similar to those used by others,[23,46,47] which we summarize here and discuss in more detail in Chapter 5. The model enumerates the microstates available to complementary DNA strands, broadly grouping them according to their configuration of intact base pairs. The base pairing scheme for a particular microstate is represented by a 1D lattice of intact and broken bp sites. The oligonucleotide conformation is reduced to a beaded chain, and the configurational degrees of freedom for unpaired bases are enumerated through self-avoiding random walks on a 3D cubic lattice. These bp configurations range from fully and partially melted "dimer" microstates, to fully dissociated "monomer" configurations. The statistical weight for a given configuration is assigned with a microstate enthalpy using the unified set of nearest-neighbor parameters developed by SantaLucia.[13] Additionally, translational degrees of freedom and concentration effects are generated through a separate 3D lattice where the formation of a dimer is defined by two molecules occupying adjacent lattice sites. For the purpose of comparing the lattice model to two-state thermodynamics, we calculate the dissociation constant for the dimer-monomer equilibrium K_d from monomer (fully dissociated strands) and dimer (all other configurations) partition functions.

The lattice model provides a detailed description of the temperature-dependent conformational variation within the ensemble of paired oligonucleotides. To demonstrate the

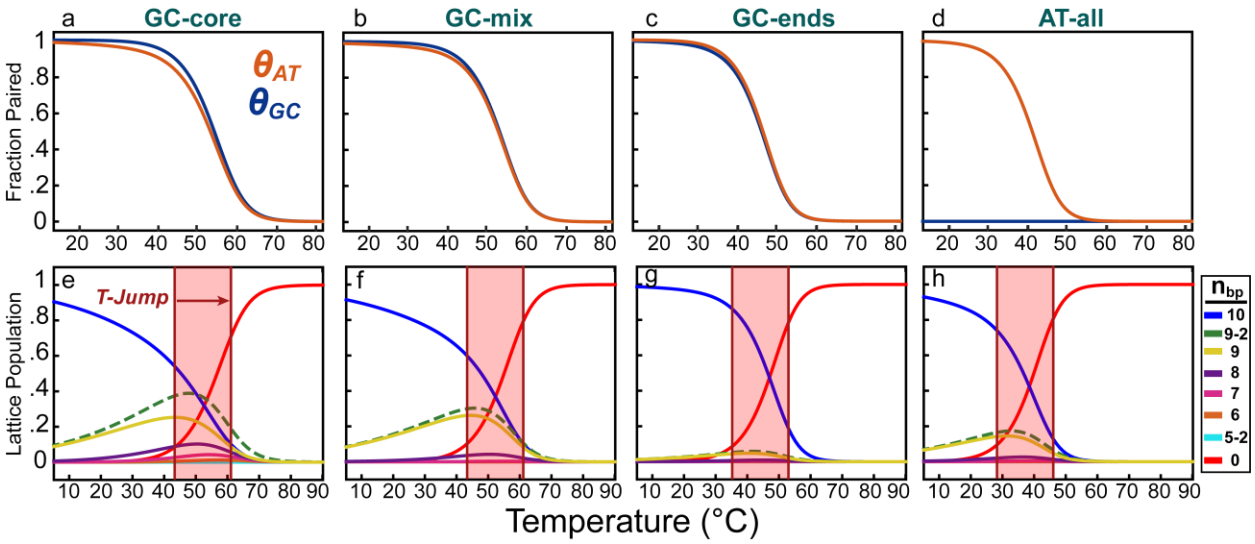

Figure 6.10: The lattice model AT and GC bp fractions (a-d) and population profiles (e-h) for (a,e) GC-core, (b,f) GC-mix, (c,g) GC-ends, and (d,h) AT-all. The population curves correspond to fully intact dimers (blue), frayed dimers (dashed green), and fully dissociated monomers (red). Otherwise the microstates are grouped according to their number of intact bps, n_{bp}. The red shading indicates the temperature jumps from the t-HDVE experiments in Fig. 6.9.

model's predictions, Fig. 6.10e-h illustrate the relative populations of conformers grouped by number of intact base pairs (n_{bp}) for each of the sequences. We observe that the fully base paired dimer and monomer account for the vast majority of the population at the temperature extremes. For both GC-core and GC-mix near T_m a significant fraction of the population exists as conformers with only 6-9 intact base pairs, but conformers with <6 intact base pairs (n_{bp} = 5-2) have negligibly low populations. In the case of GC-mix, 99% of the total dimer population with one or more broken base pairs can be accounted for by states with only 1-2 broken bps, while for GC-core 99% of the partially melted population can be accounted for by states with only 1-4 broken bps. Within these disordered configurations, those with frayed ends are heavily favored over loop microstates and we therefore call configurations with at least one broken and one intact bp "frayed dimers", represented by the dashed green profiles in Fig. 6.10e-h. To better visualize the composition of the

n_{bp} sub-ensembles, illustrative contact plots that explicitly show the microstates and their associated populations are displayed in Fig. 6.11 for three values of n_{bp}.

The model predicts that regardless of sequence, fraying initiates at the end of the helix, but how the accumulation of frayed dimer population proceeds as the temperature increases is dictated by the nucleobase sequence. GC-core frays more-or-less symmetrically around a stable core of intact bps while GC-ends preferentially frays from one end of the helix. GC-mix and AT-all fray more asymmetrically than GC-core, but still from both ends of the helix. A common trend is a bias towards the preservation of intact GC over AT bps. Therefore the placement of GC pairs within the sequence directs which configurations are populated along the helix-to-coil transition.

Furthermore the lattice model provides a window into the starting ensemble for T-jump experiments and offers a sense of how the population will shift in response to the temperature change. The temperature ranges corresponding to the T-jump experiments above are indicated by the red shading in Fig. 6.10. From the shifts in population, the model suggests that the majority of the ns fraying response observed for GC-mix corresponds to the loss of a single terminal AT bp. For GC-core more diverse configurations are possible with populated microstates tolerating up to four broken bps, but here still much of the ns response is likely accounted for by the loss of a single AT bp.

To make more direct connections with experiment, the lattice model was used to calculate the average fraction of intact base-pairs (θ_{bp}) for comparison against the FTIR melting curves for each of the sequences (Fig. 6.3), as well as the fraction of intact AT (θ_{AT}) and GC (θ_{GC}) base pairs across the transition (Fig. 6.10a-d) for comparison against the AT and GC melting curves determined by 2D IR (Fig. 6.8a-d). The variation in melting temperatures as well as the shape of the melting curves calculated from the model agrees well with experiment. However the agreement

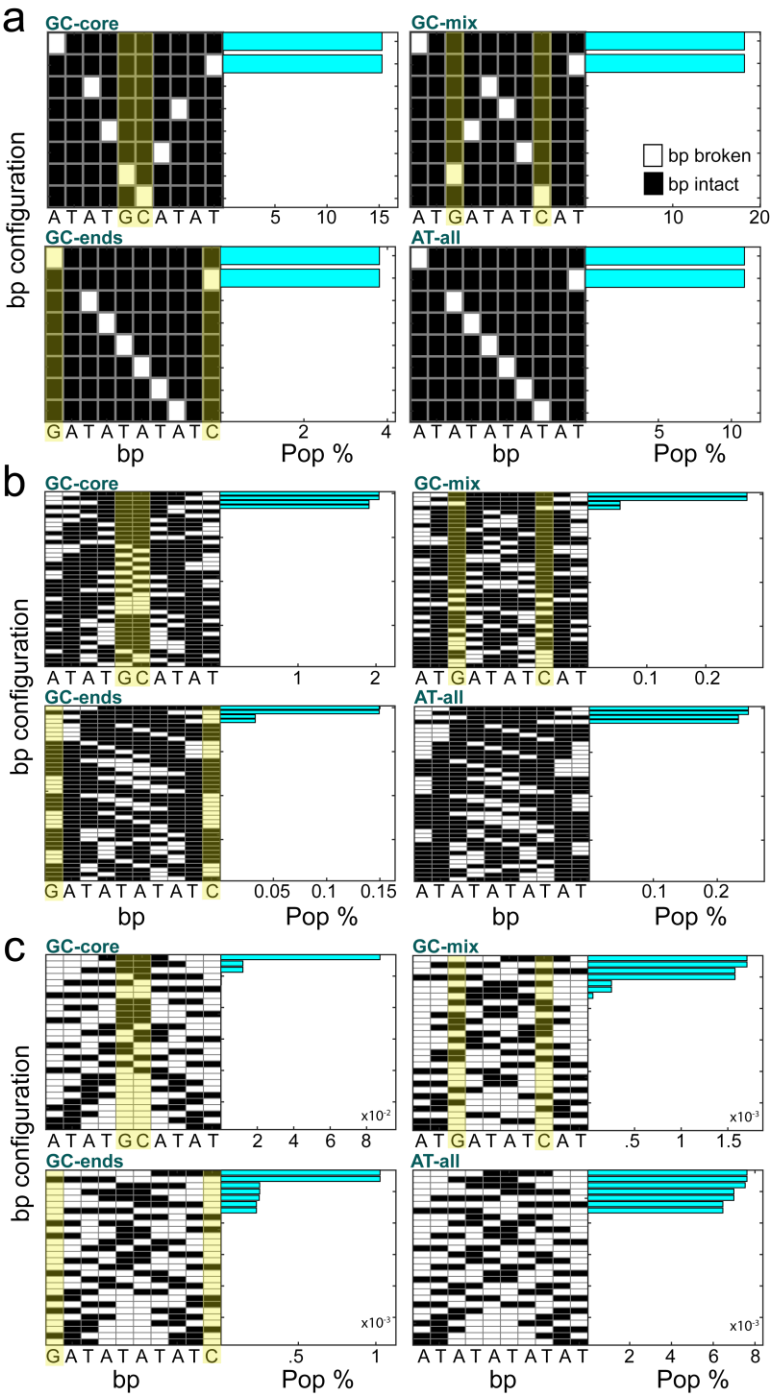

Figure 6.11: Contact plots generated by the lattice model at T_m for each of the sequences for (a) $n_{bp} = 9$, (b) $n_{bp} = 7$, and (c) $n_{bp} = 4$. GC base pairs are highlighted. The horizontal bar graphs indicate the percent of the population represented by a particular microstate.

for the AT-all sequence is comparatively weak, with the experimental melting transition appearing less sharp than the model prediction. We believe this discrepancy arises from a combination of factors. The comparatively low T_m of 40 °C makes it difficult to accurately subtract baselines from the melting curve (Fig. 6.2). In addition, the repetitive AT-all sequence could have a significant contribution from overhanging shifted-registry states which are currently not considered in the model.

It is interesting to note that θ_{GC} drops at a slightly lower temperature than θ_{AT} for GC-ends in contrast to the GC-core/GC-mix sequences where AT pairs are lost before GC. This result suggests that the GC bps at the end of the helix break at slightly lower temperatures than the AT core of the sequence, but that once the capping GC pairs go the rest of the duplex easily follows. This description is consistent with those states predicted by the model to be appreciably populated for GC-ends as the temperature increases (Fig. 6.11).

Returning to the spectral component amplitudes reconstructed from FTIR (Fig. 6.8e-h), there is a strong resemblance between the 3rd component amplitude plotted in green with the temperature dependent populations of frayed dimers in the lattice model. In Fig. 6.10e-h the lattice model population profiles are grouped into three fractions: dimers with all base pairs intact (n_{bp} = 10, blue), frayed dimers where the populations of dimers with at least one broken and one intact bp are summed together (n_{bp} = 9-2, dashed green), and monomer configurations lacking any intact bps (n_{bp} = 0, red). The similarity between the experimental spectral component amplitudes and the model predictions represented in this way is striking. It suggests that FTIR experiments are indeed sensitive to the heterogeneity of partially melted configurations in the dimer ensemble and that the population distributions of the lattice model are realistic.

It is important to note that the number of experimentally resolvable spectral components in equilibrium experiments does not necessarily correspond to the number of distinct thermodynamic states, as discussed in Chapter 4 when describing the maximum entropy method. Since the changes in the IR spectrum primarily reflect the number of intact vs broken bps at a given temperature, and the spectra at the temperature extremes provide limits on the true duplex vs the monomer spectrum, one would expect a resolvable spectral component between these extremes, if it exists, to correspond to those configurations which have at least some broken and some intact bps. The inability to resolve frayed dimers for AT-all (Fig. 6.8) despite the prediction of ~17% partially melted population near T_m is likely explained by the fact that the spectrum for this sequence (Fig. 6.1d), having the most homogeneous composition, contains lower information content compared to the mixed AT/GC sequences making it difficult to resolve a distinct 3rd spectroscopic component using the maximum entropy method.

6.4 Discussion

6.4.1 Evidence for Sequence-Dependent Heterogeneous Dehybridization

Our results indicate that IR spectroscopy can spectrally isolate different contacts within the DNA duplex, thereby giving base pair resolved information on melting transitions, conformational disorder, and melting kinetics. Using the AT and GC fractions measured from 2D IR spectroscopy in conjunction with the maximum entropy spectral component amplitudes, it is possible to experimentally resolve the heterogeneity of configurations in the dimer ensemble and to identify which bps are involved in the formation of frayed dimers. T-jump experiments reveal the kinetics involved in transitioning from a fully paired dimer through partially melted configurations into the random coil state.

Although the dynamics of dissociation vary with sequence, our results indicate that dissociation proceeds from the end of the helix for all of the sequences tested. This suggests a common dissociation pathway in which the terminal bases fray first, but where the stability of the resulting frayed configuration is dictated by the nucleobase sequence. Experimentally resolving the base-specific details of GC-ends dehybridization proves more challenging due to its effectively all-or-nothing behavior, and the evidence for terminal GC fraying should be regarded as suggestive rather than conclusive. However, in addition to the predicted offset between the θ_{AT}/θ_{GC} curves and the model prediction of a small frayed population consisting almost entirely of structures with a single broken terminal GC pair, it is worth noting the small deviation in the GC-ends 1545 cm^{-1} kinetic trace that peaks around 100 ns visible in Fig 6.9g. Although the amplitude of this feature is too small to justify fitting a biexponential, the slight deviation from a single exponential rise as well as the 2 μs shift in the GC time constant towards earlier time with respect to the AT time constant are both consistent with the presence of a small amplitude ns response corresponding to a minor amount of GC fraying for this sequence. As a result of the low magnitude of this early-time signal, we conclude that the GC-ends sequence represents an essentially classic two-state dissocation, but that dissociation still likely proceeds from the ends of the duplex in a mechanism consistent with the other sequences.

In addition to the T-jump evidence, the IR spectrum is significantly shaped depending on whether a base is Watson-Crick paired or free in solution, and we would expect bubbling bps to be just as evident in the steady-state experimental data as fraying bps, but we see no evidence for conformations in which internal AT bps bubble between intact capping GC pairs, suggesting a lower limit exists for the size of possible loop states and explaining why GC-ends dissociates in an apparent two-state manner. This lack of bubbling for DNA 10-mers is not only consistent with

our own lattice model predictions, but also with previous calculations that suggest bubbling is unfavorable for short oligonucleotides[48,49] as well as experimental evidence suggesting a minimum length of ~20 bps is necessary to nucleate bubble formation.[50] Due to the previously discussed uncertainty regarding AT-all's melting curve and the fact that we are unable to resolve a third spectroscopic component from the FTIR maximum entropy reconstruction we cannot confidently analyze this sequence in detail at this time, but some degree of deviation from a simple two-state dissociation appears likely in light of our results. We propose, in order of increasing deviation from two-state all-or-nothing dehybridization: GC-ends < AT-all < GC-mix < GC-core.

6.4.2 Interpretation of Experimental Observations with the Lattice Model

The lattice model allows direct consideration of the microstates populated along the helix-to-coil transition and can help inform the interpretation of the spectroscopic results. For GC-ends there is a lack of partially melted configurations and the population is dominated by a ratio of completely intact duplex to monomer across the entire temperature range as would be expected for a two-state dissociation (Fig. 6.10g). For both GC-core and GC-mix a sizeable frayed population is observed, with a nearly 40% and 30% contribution near T_m, respectively (Fig 6.10e-f). These stable partially melted configurations are characterized by the loss of terminal AT bps and the trends in θ_{AT} and θ_{GC} are well reproduced by the model (Fig. 6.10a-d). Even at the lowest temperature sampled ~5 °C only ~90% of the population is fully duplexed for both the GC-core and GC-mix sequences. The shape of the frayed dimer distribution indicates that as the temperature increases a few initial bp contacts break leading to partially melted stable configurations, the population of which accumulates gradually and peaks ~10 °C below T_m. But as the temperature

increases past T_m the ensemble of fully intact dimer and partially melted dimer states quickly drops off coincident with a sharp rise in the monomer population.

The population profile for the AT-all sequence predicts a comparatively small frayed dimer population, with 85% accounted for by states with only a single broken bp. Moreover, microstates with 9 AT bps and the fully intact duplex of 10 AT bps would yield similar IR spectra, shedding additional light on the difficulty of resolving a distinct spectroscopic component for partially melted configurations for this sequence using the maximum entropy method (Fig. 6.8h).

The lattice model also offers an opportunity to ask whether the frayed configurations, heterogeneous melting, and bimodal kinetics truly reflect transitions involving more than two thermodynamic states separated by barriers $>k_BT$. To test this, we used the lattice model to calculate a free energy landscape as a function of the number of intact base pairs for each oligonucleotide (n_{bp}) from the temperature-dependent probability distribution as $\Delta G(n_{bp}) = -k_BT\ln[P(n_{bp})]$, where the monomer state is taken as the reference state. These surfaces are plotted for the temperature extremes, T_m, and temperatures 20 °C above and below T_m (Fig. 6.12). Each sequence is predicted to have two stable minima separated by a barrier at $n_{bp} = 1$ at all temperatures, as expected for thermodynamically two-state melting. The primary difference between sequences is the shape and location of the dimer minimum on this landscape. Whereas the position of the two minima are at $n_{bp} = 0$ and $n_{bp} = 10$ for GC-ends, the position of the minimum in the dimer basin shifts from $n_{bp} = 10$ to $n_{bp} = 7$ between 5 and 90 °C for GC-core. GC-mix appears to be the sequence closest to a three-basin energy landscape, with a high temperature shoulder at $n_{bp} = 6$. This sequence dependent two minima structure of the free energy surfaces is consistent with the free energy surfaces measured directly by single molecule force clamp experiments for hairpins containing alternating AT and GC blocks of bps.[51]

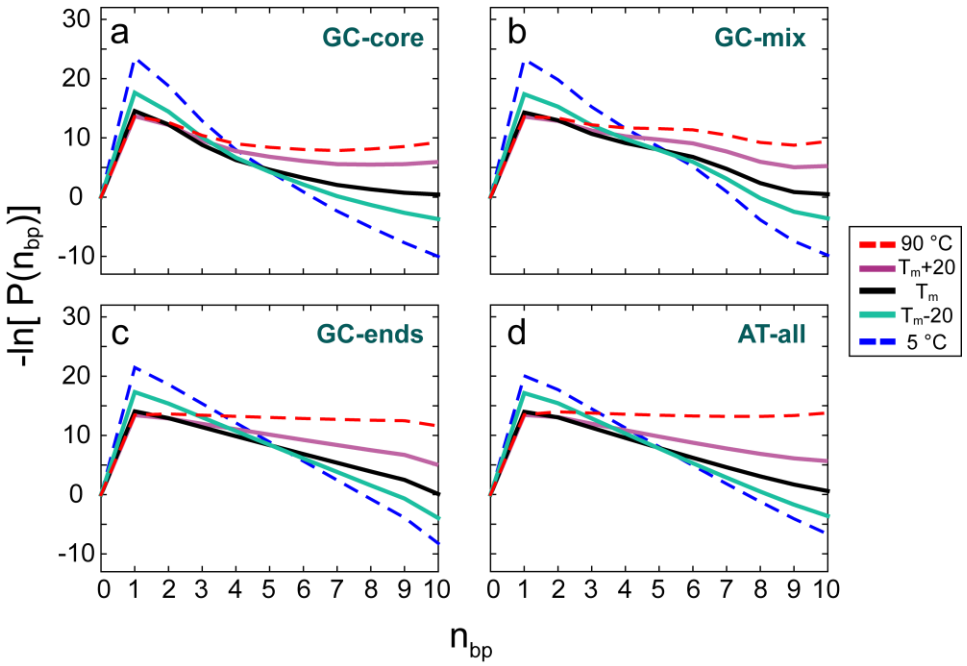

Figure 6.12: Free energy surfaces calculated using the lattice model for (a) GC-core, (b) GC-mix, (c) GC-ends, (d) AT-all at 5 °C, T_m - 20 °C, T_m, T_m + 20 °C, and 90 °C.

Although a previous report has suggested the feasibility of using the nearest-neighbor parameters to analyze hybridization kinetics,[52] interpretting dehybridization dynamics based on a free energy surface calculated from the equilibrium description of the lattice model should be done with caution. However, it does raise the question of how one would observe biexponential kinetics in a thermodynamically two-state system. Taking GC-core as an example, we see from Fig. 6.12a that our T-jump is far from perturbative, functionally reshaping the energy landscape along which the dissociating system evolves.[53] As a consequence the fast time-scale in the relaxation kinetics reflects re-equilibration within the dimer basin through rapid fraying of the termini, followed by full dissociation through an activated process. This result highlights an important distinction when defining what makes a system "two-state". Although we find that some sequences demonstrate a diversity of partially melted conformations near T_m as opposed to only monomers and fully paired

duplexes connected through the all-or-nothing loss of base pairing assumed by the standard two-state description of oligonucleotide dehybridization, we do not find definitive evidence for stable intermediate states in the thermodynamic sense of a local minimum along the dissociation pathway. While such low dimensional energy landscapes calculated from a simple model should be regarded as suggestive rather than definitive, the observed time-scales for fraying and diffusion on the free energy surface are consistent with this picture. It is possible that the "paired" state at higher temperatures reflects many frayed configurations that exchange relatively rapidly on a 10-100 ns time-scale, and that the full dissociation events are relatively rare events oberved on the 10-30 μs time scale. This description is evaluated in detail in Chapter 7 by characterizing the relaxation kinetics for multiple initial and final temperatures.

The hybridization mechanism suggested by our results is consistent with the zippering mechanism derived from both experiment[9,10,14] and simulation[21,22] for heterogeneous sequences in which an initial critical nucleus of stable bps forms as the rate limiting step followed by rapid "zippering" of the remaining base pairs into place. However the fast response we assign to duplex fraying has not been analyzed in detail in these past experiments due to the limits on the time resolution of capacitive discharge temperature jump techniques. Fraying has been studied in equimolar mixtures of A and U RNA oligonucleotides, but this report was restricted to temperatures below T_m due to the difficulty of isolating fraying from single strand destacking at higher temperatures.[54] Measuring the UV hyperchromicity reports primarily on the degree of global base stacking while IR experiments track the base-specific features necessary to distinguish which bases are invovled in pre-dissociation melting. We believe our results provide direct experimental confirmation of the zipper mechanism inferred from these early temperature jump

experiments as well as the sequence directed zippering predicted in coarse grained simulations of DNA hybridization.

6.5 Conclusions

The structural resolution afforded by linear and 2D IR spectroscopy along with the ns-ms time resolution of T-jump IR experiments presents an effective strategy for studying the dissociation of DNA oligonucleotides in detail. Combining these data with a simple lattice model allows explicit consideration of the microstates populated along the helix-to-coil transition. We have applied these methods to a model set of DNA 10-mers and find that the extent of deviation from the two-state assumption is dictated by the nucleobase sequence. Overall, the sequences tested can be arranged in order of increasing deviation from simple two-state behavior: GC-ends < AT-all < GC-mix < GC-core, where at either extreme GC-ends dissociates in an essentially all-or-nothing manner while GC-core passes through sizeable partially melted populations along the transition. The observed temperature trends and kinetics follow naturally when weighing the interplay of base pairing, where AT base pairs are generally weaker than GC base pairs, against the entropic drive to dissociate the duplex.

Despite the contrast in the amount of populated partially melted configurations for each of the sequences, our data suggest that all of these self-complementary sequences dissociate along a similar dehybridization pathway, with the ends of the helix fraying first. The stability of these frayed dimers, and ultimately the degree of deviation from an all-or-nothing dissociation, is dictated by the nucleobase sequence, as is the nature of the frayed dimer configurations themselves. A combination of infrared spectroscopies proves capable of resolving and characterizing stable partially melted configurations that give rise to heterogeneity in the folded ensemble, tracking and

identifying base specific features of the dissociation mechanism on a ns-ms timescale, and ultimately provides an experimentally consistent and noninvasive means to address the DNA dehybridization mechanism in detail beyond a simple two-state picture. We believe the strategy presented here will prove generally applicable to the detailed characterization of folding and hybridization in nucleic acid systems.

6.6 Acknowledgements

I thank Paul Stevenson for useful discussions and for his help on the boxcar 2D IR spectrometer.

6.7 References

1. Bell, S. P.; Dutta, A., DNA replication in eukaryotic cells. *Annual review of biochemistry* **2002,** *71* (1), 333-374.

2. Kunkel, T. A.; Bebenek, K., DNA Replication Fidelity*. *Annual review of biochemistry* **2000,** *69* (1), 497-529.

3. Mandal, M.; Breaker, R. R., Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology* **2004,** *5* (6), 451-463.

4. Sherwood, A. V.; Henkin, T. M., Riboswitch-Mediated Gene Regulation: Novel RNA Architectures Dictate Gene Expression Responses. *Annual Review of Microbiology* **2016,** *70* (1), 361-374.

5. Schulze, A.; Downward, J., Navigating gene expression using microarrays—a technology review. *Nature cell biology* **2001,** *3* (8), E190-E195.

6. Lin, X.; Sun, X.; Luo, S.; Liu, B.; Yang, C., Development of DNA-based signal amplification and microfluidic technology for protein assay: A review. *TrAC Trends in Analytical Chemistry* **2016,** *80*, 132-148.

7. Seeman, N. C., From genes to machines: DNA nanomechanical devices. *Trends in biochemical sciences* **2005,** *30* (3), 119-125.

8. Seeman, N. C., An overview of structural DNA nanotechnology. *Molecular Biotechnology* **2007,** *37* (3), 246-257.

9. Wetmur, J. G.; Davidson, N., Kinetics of renaturation of DNA. *Journal of molecular biology* **1968,** *31* (3), 349-370.

10. 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. *Journal of molecular biology* **1971,** *62* (2), 361-381.

11. Mergny, J.-L.; Lacroix, L., Analysis of thermal melting curves. *Oligonucleotides* **2003,** *13* (6), 515-537.

12. Marky, L. A.; Breslauer, K. J., Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers* **1987,** *26* (9), 1601-1620.

13. SantaLucia, J., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences* **1998,** *95* (4), 1460-1465.

14. Pörschke, D.; Uhlenbeck, O.; Martin, F., Thermodynamics and kinetics of the helix-coil transition of oligomers containing GC base pairs. *Biopolymers* **1973,** *12* (6), 1313-1335.

15. 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* (19), 5336-5353.

16. Hagan, M. F.; Dinner, A. R.; Chandler, D.; Chakraborty, A. K., Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA. *Proceedings of the National Academy of Sciences* **2003,** *100* (24), 13922-13927.

17. Maffeo, C.; Yoo, J.; Comer, J.; Wells, D.; Luan, B.; Aksimentiev, A., Close encounters with DNA. *Journal of Physics: Condensed Matter* **2014,** *26* (41), 413101.

18. Mentes, A.; Florescu, A. M.; Brunk, E.; Wereszczynski, J.; Joyeux, M.; Andricioaei, I., Free-Energy Landscape and Characteristic Forces for the Initiation of DNA Unzipping. *Biophysical journal* **2015,** *108* (7), 1727-1738.

19. Ouldridge, T. E.; Louis, A. A.; Doye, J. P., Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *The Journal of chemical physics* **2011,** *134* (8), 085101.

20. 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. *The Journal of chemical physics* **2013,** *139* (14), 144903.

21. Ouldridge, T. E.; Šulc, P.; Romano, F.; Doye, J. P.; Louis, A. A., DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic acids research* **2013,** *41* (19), 8886-8895.

22. Hinckley, D. M.; Lequieu, J. P.; de Pablo, J. J., Coarse-grained modeling of DNA oligomer hybridization: length, sequence, and salt effects. *The Journal of chemical physics* **2014,** *141* (3), 035102.

23. Ma, H.; Proctor, D. J.; Kierzek, E.; Kierzek, R.; Bevilacqua, P. C.; Gruebele, M., Exploring the energy landscape of a small RNA hairpin. *Journal of the American Chemical Society* **2006,** *128* (5), 1523-1530.

24. 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 research* **2007,** *35* (9), 2875-2884.

25. Ma, H.; Wan, C.; Wu, A.; Zewail, A. H., DNA folding and melting observed in real time redefine the energy landscape. *Proceedings of the National Academy of Sciences* **2007,** *104* (3), 712-716.

26. Morrison, L. E.; Stols, L. M., Sensitive fluorescence-based thermodynamic and kinetic measurements of DNA hybridization in solution. *Biochemistry* **1993,** *32* (12), 3095-3104.

27. Rauzan, B.; McMichael, E.; Cave, R.; Sevcik, L. R.; Ostrosky, K.; Whitman, E.; Stegemann, R.; Sinclair, A. L.; Serra, M. J.; Deckert, A. A., Kinetics and thermodynamics of DNA, RNA, and hybrid duplex formation. *Biochemistry* **2013,** *52* (5), 765-772.

28. SantaLucia, J.; Allawi, H. T.; Seneviratne, P. A., Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **1996,** *35* (11), 3555-3562.

29. Breslauer, K. J.; Frank, R.; Blöcker, H.; Marky, L. A., Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences* **1986,** *83* (11), 3746-3750.

30. Patel, D. J.; Hilbers, C., Proton nuclear magnetic resonance investigations of fraying in double-stranded d-ApTpGpCpApT in aqueous solution. *Biochemistry* **1975,** *14* (12), 2651-2656.

31. Breslauer, K. J.; Sturtevant, J. M.; Tinoco, I., Calorimetric and spectroscopic investigation of the helix-to-coil transition of a ribo-oligonucleotide: rA7U7. *Journal of molecular biology* **1975,** *99* (4), 549-565.

32. Yin, Y.; Zhao, X. S., Kinetics and dynamics of DNA hybridization. *Accounts of chemical research* **2011,** *44* (11), 1172-1181.

33. Nonin, S.; Leroy, J.-L.; Gueron, M., Terminal base pairs of oligodeoxynucleotides: imino proton exchange and fraying. *Biochemistry* **1995,** *34* (33), 10652-10659.

34. Brauns, E. B.; Dyer, R. B., Time-resolved infrared spectroscopy of RNA folding. *Biophysical journal* **2005,** *89* (5), 3523-3530.

35. Stancik, A. L.; Brauns, E. B., Rearrangement of partially ordered stacked conformations contributes to the rugged energy landscape of a small RNA hairpin. *Biochemistry* **2008,** *47* (41), 10834-10840.

36. Narayanan, R.; Zhu, L.; Velmurugu, Y.; Roca, J.; Kuznetsov, S. V.; Prehna, G.; Lapidus, L. J.; Ansari, A., Exploring the energy landscape of nucleic acid hairpins using laser temperature-jump and microfluidic mixing. *Journal of the American Chemical Society* **2012,** *134* (46), 18952-18963.

37. Peng, C. S.; Jones, K. C.; Tokmakoff, A., Anharmonic vibrational modes of nucleic acid bases revealed by 2D IR spectroscopy. *Journal of the American Chemical Society* **2011,** *133* (39), 15650-15660.

38. Banyay, M.; Sarkar, M.; Gräslund, A., A library of IR bands of nucleic acids in solution. *Biophysical chemistry* **2003,** *104* (2), 477-488.

39. Marmur, J.; Doty, P., Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *Journal of molecular biology* **1962,** *5* (1), 109-118.

40. Krummel, A. T.; Zanni, M. T., DNA vibrational coupling revealed with two-dimensional infrared spectroscopy: insight into why vibrational spectroscopy is sensitive to DNA structure. *The Journal of Physical Chemistry B* **2006,** *110* (28), 13991-14000.

41. 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 chemical biology* **2015,** *11* (2), 470-477.

42. Lumry, R.; Biltonen, R.; Brandts, J. F., Validity of the "two-state" hypothesis for conformational transitions of proteins. *Biopolymers* **1966,** *4* (8), 917-944.

43. Widjaja, E.; Garland, M., Pure component spectral reconstruction from mixture data using SVD, global entropy minimization, and simulated annealing. Numerical investigations of admissible objective functions using a synthetic 7-species data set. *Journal of computational chemistry* **2002,** *23* (9), 911-919.

44. Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A., Transient two-dimensional IR spectrometer for probing nanosecond temperature-jump kinetics. *The Review of scientific instruments* **2007,** *78* (6), 063101-063101.

45. Jones, K. C.; Ganim, Z.; Peng, C. S.; Tokmakoff, A., Transient two-dimensional spectroscopy with linear absorption corrections applied to temperature-jump two-dimensional infrared. *JOSA B* **2012,** *29* (1), 118-129.

46. Chen, S.-J.; Dill, K. A., RNA folding energy landscapes. *Proceedings of the National Academy of Sciences* **2000,** *97* (2), 646-651.

47. Everaers, R.; Kumar, S.; Simm, C., Unified description of poly- and oligonucleotide DNA melting: nearest-neighbor, Poland-Sheraga, and lattice models. *Phys Rev E Stat Nonlin Soft Matter Phys* **2007,** *75* (4 Pt 1), 041918.

48. Zimm, B. H., Theory of``melting"of the helical form in double chains of the DNA type. *The Journal of Chemical Physics* **1960,** *33* (5), 1349-1356.

49. Applequist, J.; Damle, V., Thermodynamics of the helix-coil equilibrium in oligoadenylic acid from hypochromicity studies. *Journal of the American Chemical Society* **1965,** *87* (7), 1450-1458.

50. Zeng, Y.; Montrichok, A.; Zocchi, G., Bubble nucleation and cooperativity in DNA melting. *Journal of molecular biology* **2004,** *339* (1), 67-75.

51. Woodside, M. T.; Anthony, P. C.; Behnke-Parks, W. M.; Larizadeh, K.; Herschlag, D.; Block, S. M., Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid. *Science* **2006,** *314* (5801), 1001-1004.

52. Ohmichi, T.; Nakamuta, H.; Yasuda, K.; Sugimoto, N., Kinetic property of bulged helix formation: analysis of kinetic behavior using nearest-neighbor parameters. *Journal of the American Chemical Society* **2000,** *122* (46), 11286-11294.

53. Chung, H. S.; Shandiz, A.; Sosnick, T. R.; Tokmakoff, A., Probing the Folding Transition State of Ubiquitin Mutants by Temperature-Jump-Induced Downhill Unfolding. *Biochemistry* **2008,** *47* (52), 13870-13877.

54. Pörschke, D., A direct measurement of the unzippering rate of a nucleic acid double helix. *Biophysical chemistry* **1974,** *2* (2), 97-101.

Chapter 7

Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization

The work presented in this chapter has been published and is reprinted with permission from:
Sanstead, P. J.; Tokmakoff, A., Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *The Journal of Physical Chemistry B* **2018,** *122* (12), 3088-3100.

7.1 Abstract

We have studied two model DNA oligonucleotide sequences which display contrasting degrees of heterogeneous melting using an optical temperature jump to trigger dehybridization and a nonlinear infrared (IR) spectroscopy probe to track the response of the helix ensemble. This approach offers base-sensitive structural insight through the unique vibrational fingerprint characteristic of each nucleobase as well as time resolution capable of following unfolding across nanoseconds to milliseconds. We observe pre-dissociation unzipping of the helical termini, loss of final dimer contacts, and rehybridization of the dissociated strands all in a single measurement. Complete dissociation of the dimer is found to be well described by Arrhenius kinetics for both sequences, with dissociation barriers in the range of 160-190 kJ/mol. A sequence with terminal adenine-thymine (AT) base pairs and a guanine-cytosine (GC) core returns a large amplitude fast response ranging from 70-170 ns originating only from the AT base pairs. Variable temperature jump (T-jump) experiments in which the final temperature (T_f) is fixed and the initial temperature

(T_i) is varied such that different starting ensembles all evolve on the same final free energy surface were employed to explore the features of the underlying potential that dictates hybridization. These experiments reveal that the unzipping of the AT termini is an essentially barrierless process and that both activated and downhill events are necessary to describe the dehybridization mechanism. While our results are largely consistent with the classic nucleation-zipper picture, new insights regarding the nature of base pair zippering refine the mechanistic details of the fastest DNA hybridization dynamics.

7.2 Introduction

Since the discovery of the double-helical structure of DNA in 1953[1] our understanding of nucleic acids has progressed far beyond a static archive for the storage of genetic information. Many of the most important functions performed by DNA are highly dynamic, whether in the biological roles of replication,[2,3] transcription,[4,5] and gene regulation[6,7] or in the expanding field of DNA-based nanotechnology, where the predictable yet reversible nature of base pairing enables the design of increasingly complex nanoscale structures,[8,9] molecular machines,[10,11] and computing devices.[12] Understanding the fundamental motions and structural changes underlying these processes is thus of central importance.

Much of the foundation for the current knowledge of the kinetics and dynamics of DNA hybridization were laid by early temperature jump (T-jump) experiments that used a capacitive discharge to rapidly heat a nucleic acid solution and ultraviolet (UV) absorbance to monitor the system's response.[13-16] These researchers were among the first to note the now well-established observation that dimer dissociation is effectively described by a two-state Arrhenius treatment, but that the association process follows anti-Arrhenius behavior.[13,15] A mechanism in which the

formation of a critical nucleation site of several in-register base pairs is the rate limiting step, followed by rapid "zippering" of the remaining pairs into place was proposed to account for the experimental observations. Due to limitations in these early experiments such as relatively slow microsecond heating and a lack of structural insight from the UV spectrum, this mechanism could not be observed directly but was rather inferred by fitting the available data to kinetic models which incorporate a nucleation barrier as well as a fundamental rate constant for adding a base pair to the growing helix.[13-16] This analysis necessarily relied on a many parameter fit to limited information. More recent work has contributed finer insight to this initial picture, such as fluorescence correlation spectroscopy measurements of local base dynamics[17,18] and single-molecule optical force pulling measurements to map out barriers and free energy landscapes,[19] to name a few, but the basic concept underpinning the so-called nucleation-zipper mechanism of DNA hybridization has thus far stood the test of time. Despite remarkable experimental progress, many aspects of the hybridization mechanism, particularly the transient interactions between strands that precede the critical nucleation event, possible rearrangement schemes following an out of register nucleation, and the rapid zippering of the remaining pairs into the helical spiral, have evaded direct experimental interrogation.

Although noted in early T-jump experiments,[13,15] sub-microsecond timescales were not fully resolved. Subsequent experimental work has attributed fast dynamics in oligonucleotide dissociation to end fraying effects, which are essentially the unzipping of the terminal base pairs. A time-resolved Stokes shift (TRSS) study employed a coumarin reporter variably placed at the center and near the end of the helix to explore potential ultrafast aspects of fraying across a 40 fs to 40 ns range, identifying a 5 ps process characteristic of the helical terminus.[20] The authors assign this timescale to increased dye solvation due to enhanced base mobility at the termini rather than

base pair opening, but these experiments establish a 40 ns lower bound on fraying dynamics. An upper bound of 1 ms has been proposed by imino proton exchange NMR measurements,[21,22] although one could argue that the ~10 μs time resolution of early capacitive discharge T-jump experiments could alternatively serve as the upper bound even though these authors may not implicate fraying directly to account for the presence of unresolved fast timescales.[13,15] Chapter 6 presented preliminary IR T-jump experiments where a 70-100 ns response near T_m was observed for oligonucleotides containing terminal AT base pairs. These timescales were assigned to unzipping of the termini since they are observed only for the AT features and correlate with predicted populations of frayed helices.[23] However, the previous chapter did not contain any further characterization of base zippering in detail beyond the proposed assignment of these additional nanosecond timescales.

To date many of the most detailed descriptions of the critical dynamic events in DNA function, from the encounter and recognition of the involved strands to their rearrangement into an operational structure or device, are restricted to simulation due to the relevant timescales and length scales spanning many orders of magnitude. Interestingly these models often predict rich dynamics even for short oligonucleotides.[24-26] Additional experimental insight is necessary to fully understand nucleic acid folding in real time. To this end, Chapter 6 detailed the development of a steady-state and time-resolved infrared (IR) spectroscopy based strategy for studying the dehybridization of DNA and demonstrated the utility of this approach by characterizing the sequence-dependent thermodynamics and kinetics of a model set of DNA oligonucleotides. The previous chapter focused primarily on equilibrium experiments.

The focus of this chapter is the kinetics and dynamics of DNA dehybridization studied through time-resolved T-jump IR spectroscopy. This approach offers several advantages in that IR

spectroscopy is label-free, minimally perturbative, and provides a degree of base-sensitive structural information.[27] Furthermore an optically triggered T-jump allows for heating on timescales of a few nanoseconds and our instrument[28] can track the response of the system from ns-ms, allowing the direct observation of pre-dissociation events, the loss of final dimer contacts, and rehybridization all in a single measurement. By varying the initial temperature at which the system is prepared but tuning the magnitude of the temperature jump such that the final temperature is fixed, features of the underlying free energy surface can be explored.[29] This strategy enables us to identify activated vs downhill events in DNA dehybridization. We conclude that both contribute to the folding mechanism, with nucleation being the primary barrier and the unzipping of the termini a seemingly barrierless process.

The goals of this study are twofold. First, we would like to directly measure and describe previously inaccessible fast dynamics in the dehybridization mechanism, such as the rate of unzipping of the termini, which precede the better-characterized dissociation process. Second and more broadly, we seek to develop a sequence-specific experimental approach to study the folding of nucleic acids in detail across the full range of relevant timescales. This approach provides a comprehensive experimental probe of the nucleation-zipper mechanism of DNA oligonucleotide hybridization and lays the foundation for a direct method of observing nucleic acid folding in real-time through multidimensional IR spectroscopy. Subsequent chapters will apply the approach developed in this chapter to study non-canonical DNA sequences containing modified cytosine bases relevant to epigenetic regulation in eukaryotes.

7.3 Results

7.3.1 Temperature Dependent FTIR

The two model DNA sequences from the previous chapter that displayed the largest contrast in both their thermodynamics and dissociation kinetics near their respective melting temperatures (T_m) were selected for a detailed characterization of DNA dehybridization via T-jump IR spectroscopy. Both self-complementary sequences are ten base pairs in length and have identical adenine-thymine (AT) and guanine-cytosine (GC) content, where the placement of two GC base pairs is varied in an otherwise AT sequence: 5'-GATATATATC-3' (GC-ends) and 5'-ATATGCATAT-3' (GC-core). We will refer to the sequences by the shorthand names in parentheses for convenience. Despite the identical length and base pair composition of these sequences, we noted in the previous chapter a ten degree offset in T_m, significant contrast in the diversity of structures present in the dimer ensemble, and an additional fast response in the near-T_m kinetics of the AT features of the GC-core sequence that is absent in the GC-ends sequence. Temperature dependent Fourier transform infrared (FTIR) spectra across 5-90 °C were used to initially characterize DNA duplex melting, or the "helix-to-coil" transition. The 1500-1750 cm^{-1} frequency range contains in-plane ring modes, carbonyl stretches, and ND_2 bending modes that are sensitive to the hydrogen bonding and base stacking interactions that facilitate base pairing.[27,30] A representative high and low temperature FTIR spectrum, taken at 90 and 5 °C respectively, illustrates the substantial changes to the IR spectrum in this frequency range upon hybridization (Fig. 7.1a,c). The bases that roughly contribute to each feature in the spectrum are indicated in the figure. Although we have discussed the temperature dependent FTIR spectra of these sequences in detail in Chapter 6, a brief description is included here in order to introduce the infrared spectra

of the oligonucleotides, highlight the spectroscopic changes indicative of melting, and to motivate the sampling range of the T-jump experiments.

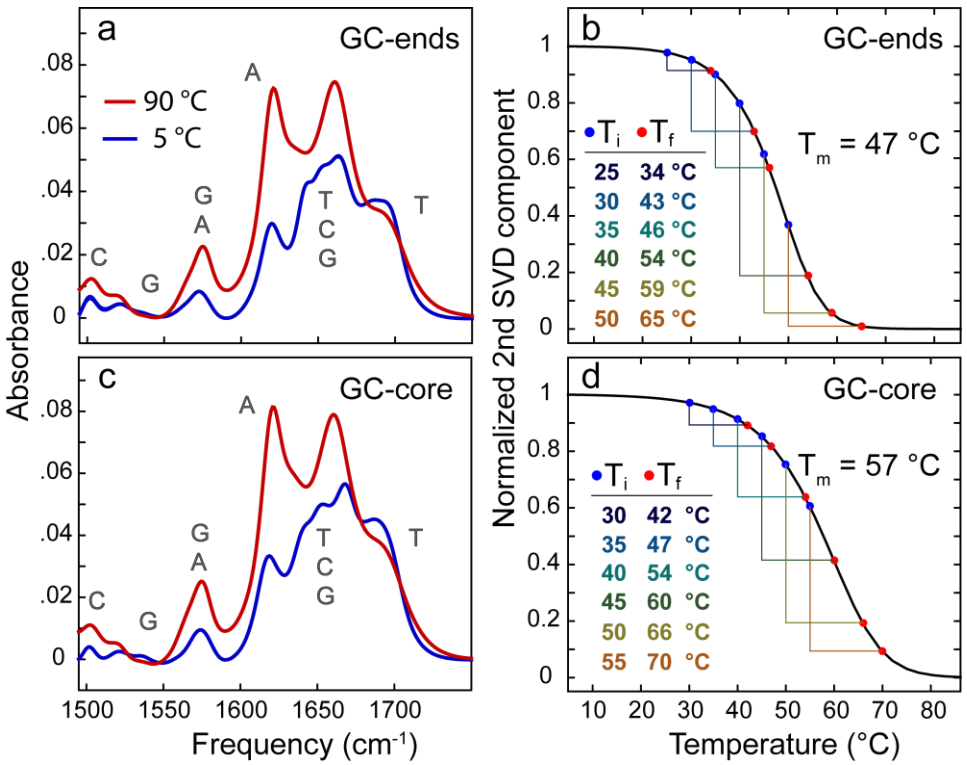

Figure 7.1: Representative high (90 °C) and low (5 °C) temperature FTIR spectra with 2 mM oligonucleotide, 50 mM pD 7.2 sodium phosphate buffer, 240 mM NaCl, 18 mM $MgCl_2$ for (a) GC-ends and (c) GC-core. Melting curves derived from the 2nd SVD component of the FTIR temperature ramp for (b) GC-ends and (d) GC-core. The temperature jumps sampled across the helix-to-coil transition for each sequence are indicated.

Of the many changes to the IR spectrum upon dehybridization, the most striking is the ~130% growth in intensity of the 1622 cm^{-1} A ring mode. This mode is a delocalized combination of C=C stretching, C-H as well as ND_2 bending, and in-plane pyrimidine vibrations.[30] Since the AT content in these sequences is large (80%) and this mode is highly sensitive to base pairing, the 1622 cm^{-1} feature is an excellent reporter on the extent of intact AT contacts. The low frequency

side of the spectrum below 1600 cm^{-1} is dominated by G and C features which are primarily characterized by delocalized in-plane ring motions.[30] These features are similarly suppressed upon base pair formation and therefore intensity growth in this frequency range is indicative of the loss of GC contacts. Above 1630 cm^{-1} the spectrum becomes comparatively congested with overlapping contributions from T, C, and G vibrations, but the high AT content suggests that T is primarily responsible for these features. However it is clear that the variable position of the GC pairs influences the shape of the spectrum in this range, particularly at low temperature.

To account for global changes to the spectrum in response to increasing temperature, melting curves were derived from a singular value decomposition (SVD) analysis through the normalized, baseline corrected, second SVD component, as discussed in the previous chapter. The melting curves serve as a guide for designing T-jump experiments that will map out the entire helix-to-coil transition. The curves are spanned from upper to lower baseline by six T-jumps ranging in magnitude from 9-16 °C. The initial and final temperatures for each of the T-jump measurements sampled across the melting transition are indicated in Fig. 7.1b and 7.1d for the GC-ends and GC-core sequences, respectively.

7.3.2 Assigning Features in the Transient Spectrum using 2D IR

Nonlinear IR spectroscopies have been used increasingly to study the biophysics of nucleic acids due to enhanced structural insight compared to linear IR as well as picosecond time resolution.[31-34] Two-dimensional infrared (2D IR) spectroscopy is particularly powerful since it can reveal vibrational couplings, quantify anharmonicities, and separate homogeneous and inhomogeneous contributions to vibrational line shapes, all of which can contribute additional structural information. The 2D IR spectra of the GC-ends and GC-core sequences are reported in

Chapter 6, where distinct cross-peak regions sensitive to AT and GC base pairing were identified that reveal varying degrees of heterogeneous melting.[23] For the purposes of studying dissociation kinetics in real time the acquisition of the full transient 2D IR surface is prohibitive, since it requires the scanning of an additional time delay to resolve the excitation axis which adds considerably to the time required for a single measurement. Recording the heterodyne detected vibrational echo (HDVE) spectrum instead offers a compromise between acquisition time and spectral information content.[35] The coherence time is not scanned in an HDVE measurement, allowing for more rapid data acquisition and finer sampling of the kinetics. However this also means that the excitation axis is not resolved and the resulting spectrum is related only to the projection of the full 2D IR surface onto the detection axis.

The T-jump spectrometer and experiment have been described in detail previously.[28,35] This approach allows the evolution of the HDVE spectrum in response to a 5 ns optical T-jump to be tracked across a ns-ms time range. Data are reported as t-HDVE spectra, which are obtained by subtracting off the equilibrium HDVE spectrum acquired at the initial temperature (T_i) from the HDVE spectrum acquired at each point along the thermal profile imparted by the T-jump pulse. Temperature dependent equilibrium 2D IR spectra as well as their difference spectra provide a valuable guide to understanding the features in the one-dimensional t-HDVE spectra. Illustrative 2D IR surfaces for the GC-core sequence at 80 and 10 °C are plotted in Fig. 7.2a,b as well as the difference spectrum between the high and low temperature surfaces in Fig. 7.2c. The projection of the 2D IR surface onto the detection axis (Fig. 7.2d) is a more direct comparison to the HDVE spectrum, while the difference spectrum between projections at different temperatures is a more direct comparison to the t-HDVE spectrum measured in our T-jump experiment.

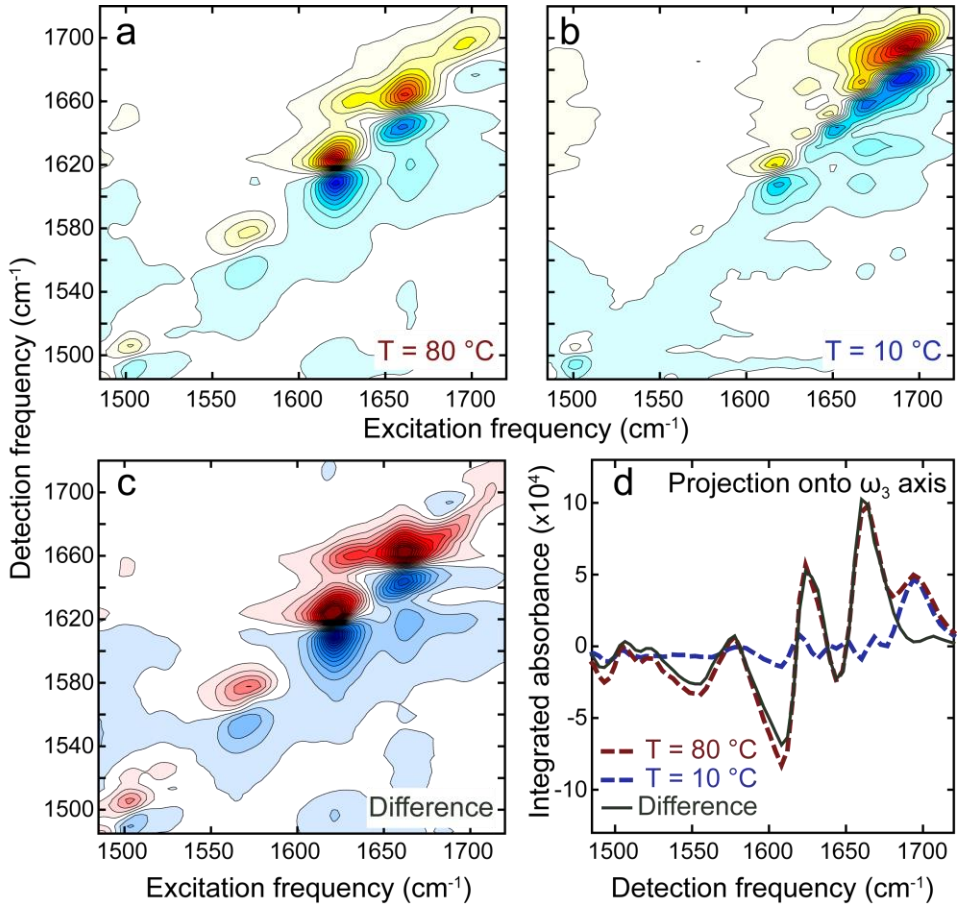

Figure 7.2: Illustrative (a) 80 °C and (b) 10 °C 2D IR surfaces of the GC-core sequence prepared with the same sample conditions reported in Fig. 7.1. (c) The difference spectrum between the high and low temperature surfaces. (d) Projection onto the detection axis at each temperature as well as the difference of the projections, which closely resembles the t-HDVE spectra in Fig. 7.3b.

The peaks along the diagonal of the 2D IR spectra can be mapped to the peaks in the linear high and low temperature spectra in Fig. 7.1c, but the peaks in a 2D IR spectrum are oppositely signed doublets corresponding to ground state bleach (0→1 transition, red/orange) and excited state absorption (1→2 transition, blue/cyan) shifted to lower detection frequency by the anharmonicity.[36] The intensity changes and frequency shifts observed in the linear spectrum with increasing temperature are also evident in the 2D IR spectrum, but with the added advantage of a

second frequency axis spreading out some of the overlapping features and the higher order dipole scaling sharpening linewidths. For example, the excited state absorption (ESA) of the 1622 cm^{-1} A ring mode and the ESA of the 1565-1575 cm^{-1} G ring modes are well isolated and it is clear from inspection of the 2D IR surfaces that projection onto the detection axis does not significantly congest these features with other absorptions. Additionally, the formation of Watson-Crick pairs gives rise to cross-peaks in the low temperature spectrum representative of inter-base vibrational couplings that are lost with increasing temperature, such as the cross-peaks observed between the 1622 cm^{-1} A mode and the 1690 cm^{-1} T carbonyl mode.

For the difference spectrum representation plotted in Fig. 7.2c, a red-over-blue doublet corresponds to a gain in both GSB and ESA intensity while an inverted blue-over-red doublet represents intensity loss. The difference spectrum is clearly dominated by gain features, consistent with the substantial intensity increases upon unfolding discussed for the linear spectrum above. Projecting the 2D IR difference spectrum onto the detection axis provides the best comparison to t-HDVE, and the features in this projection are in excellent agreement with t-HDVE spectra measured for the GC-core sequence (Fig. 7.3b). This connection facilitates assignment of the t-HDVE features using the more information rich 2D IR surfaces as a guide.

7.3.3 Sampling Across the Melting Transition with T-jump IR Spectroscopy

The temperature jump experiments indicated across the melting curves in Fig. 7.1b,d were measured using t-HDVE spectroscopy with the aim of mapping out the entire helix-to-coil transition and fully resolving the dissociation kinetics for each sequence. As discussed in Chapter 3, the thermal profile imparted by the T-jump pulse, as measured by the change in mid-IR transmission of the D_2O solvent, takes the form of a heated plateau (T_f) that remains relatively

constant for tens of microseconds followed by a stretched-exponential relaxation back toward T_i on a several millisecond timescale. As a consequence, the magnitude of the t-HDVE spectrum initially rises away from zero as the DNA ensemble prepared at T_i responds to the sudden jump to T_f and evolves towards the higher temperature ensemble. This signal rise can have either positive or negative sign depending on whether the feature is a GSB or ESA. Such a straightforward assignment of the peaks in information dense transient difference spectra should not be taken for granted, but the significant change in intensity between the dimer and monomer spectra first noted in the FTIR and best characterized by the 2D IR difference spectrum enables clear assignment of loss/gain t-HDVE features for these oligonucleotides. An example of a t-HDVE data set is given in Fig. 7.3c, corresponding to a 15 °C T-jump from T_i = 45 °C for the GC-core sequence.

As the thermal energy dissipates and the system begins to relax back toward T_i, the magnitude of the t-HDVE spectrum once again approaches zero. Six illustrative time points demonstrate how the t-HDVE spectrum responds to the T-jump (Fig. 7.3b). Color coded lines in Fig. 7.3c indicate the delay slices corresponding to the select t-HDVE spectra. As seen in the single frequency kinetic traces in Fig. 7.3a, the difference features of the t-HDVE spectrum increase up to 100 μs, but by 1 ms the spectrum tends back towards the zero baseline as the system approaches re-equilibration at 45 °C. The similarity of the t-HDVE spectrum to the detection axis projection of the 2D IR difference spectrum (Fig. 7.2d) strongly suggests a common origin for these spectroscopic features. Although the response of the t-HDVE spectrum in time contains all of the kinetic information of interest, this representation is not a particularly clear or intuitive way to extract kinetics. Transforming the data into a rate domain representation eliminates the need to assume and fit some functional form in the time domain and offers a more intuitive way to extract kinetic information directly from observed rates. Transformation from the time domain to the rate

domain is achieved by performing a numerical inverse Laplace transform (iLT). However, as discussed in Chapter 4, this is an ill-conditioned problem when applied to a signal with nonzero noise and a large number of distributions can fit the data equally well. The maximum entropy method (MEM) is one strategy to achieve a unique solution to this problem, and the approach detailed in Chapter 4 is applied to the t-HDVE data presented here.

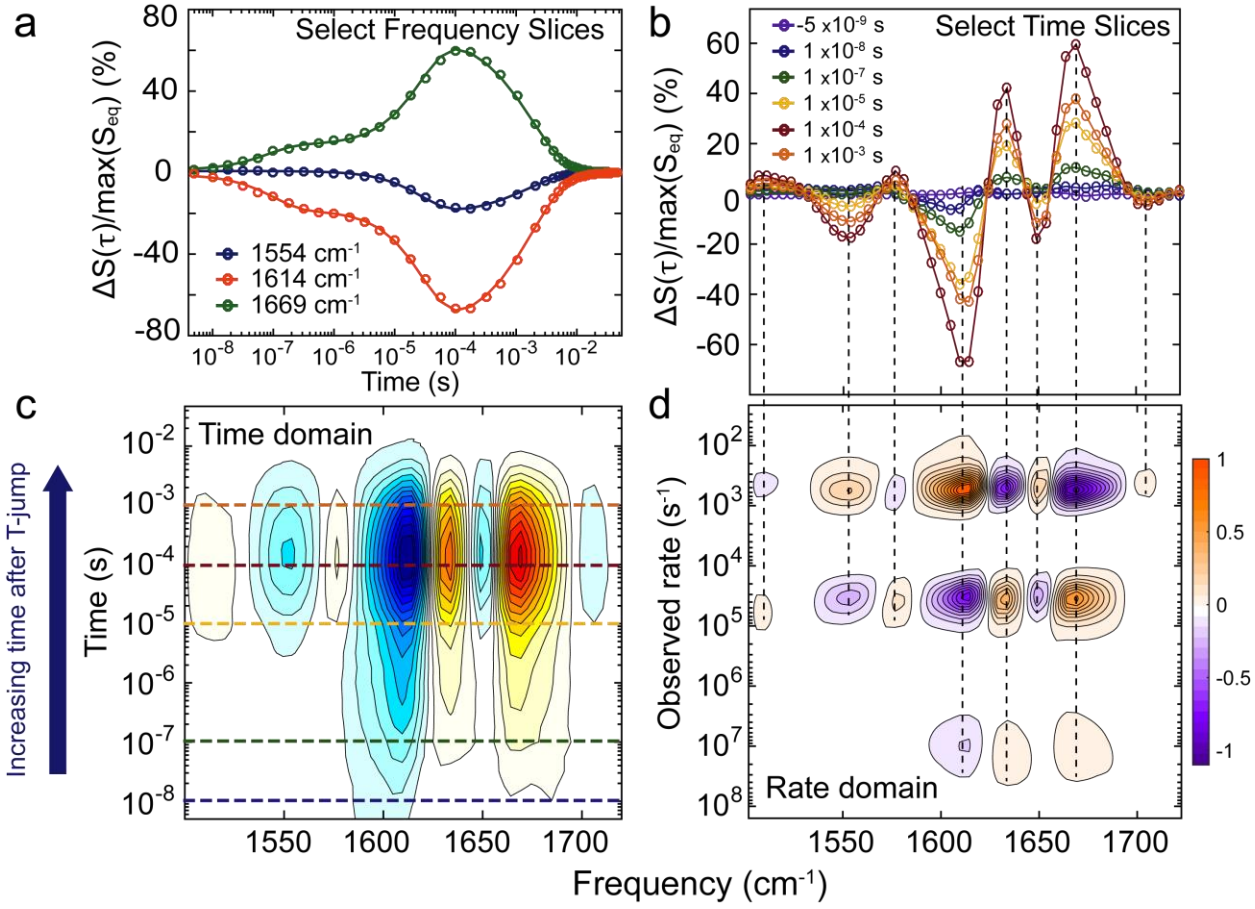

Figure 7.3: (a) Single frequency kinetic traces selected from the full data set in panel c (b) Representative t-HDVE spectra at select delays. Time slices are indicated by the color coded lines in panel c. (c) Time domain t-HDVE data set for the GC-core sequence with T_i = 45 °C and T_f = 60 °C. Sample conditions are identical to those reported in Fig. 1. (d) The rate spectrum representation of the T-jump data set in panel c. Positive rate amplitude is orange while negative amplitude is purple. Black dashed lines serve as a guide for relating how positive and negative features of the t-HDVE spectrum evolve in the rate domain.

As an example of the rate representation of t-HDVE data, the rate spectrum obtained by transforming the full data set in Fig. 7.3c is plotted in Fig. 7.3d. Orange features represent positive rate amplitude while purple features represent negative amplitude. Dashed lines serve as a guide for relating how GSB and ESA features in the t-HDVE spectrum evolve in the rate domain. Note that for a positively signed GSB, positive rate amplitude indicates a gain in magnitude for that feature (orange) while negative amplitude indicates intensity loss (purple). The inverse is true for the negatively signed ESA features, where negative amplitude indicates a gain feature and positive amplitude indicates a loss of magnitude. Taking a slice along the rate axis at a given frequency gives the rate spectrum at that frequency. The rate axis is arranged such that the corresponding time after arrival of the T-jump pulse is increasing along the y-axis (decreasing rate), as this is the most intuitive way to visualize the data in terms of the experiment.

The MEM-iLT was applied to the set of T-jumps indicated in Fig. 7.1b,d to obtain a set of rate spectra that track across the helix-to-coil transition. The full set of time domain t-HDVE spectra are plotted in Fig. 7.4 and the corresponding set of rate spectra for the GC-core and GC-ends sequences are plotted in Fig. 7.5. The rate spectrum of the GC-core sequence contains three bands of amplitude that are well separated along the rate axis. We label these three regimes the fast response at around 10^7 s^{-1} (red), the slow response at around 10^4-10^5 s^{-1} (green), and the relaxation response between 10^2-10^3 s^{-1} (blue). The slow response speeds up noticeably with increasing temperature, with a clear shift in these distributions towards larger rate as T_f increases. Dashed lines in Fig. 7.5 serve as a guide to visualize this shift. Furthermore, the slow response appears to be universal and concerted across the entire frequency range, with all features of the t-HDVE spectrum responding in tandem. In contrast the fast response is observed for the 1614 cm^{-1} A ring mode ESA as well as the more congested features at higher frequency, including the GSB

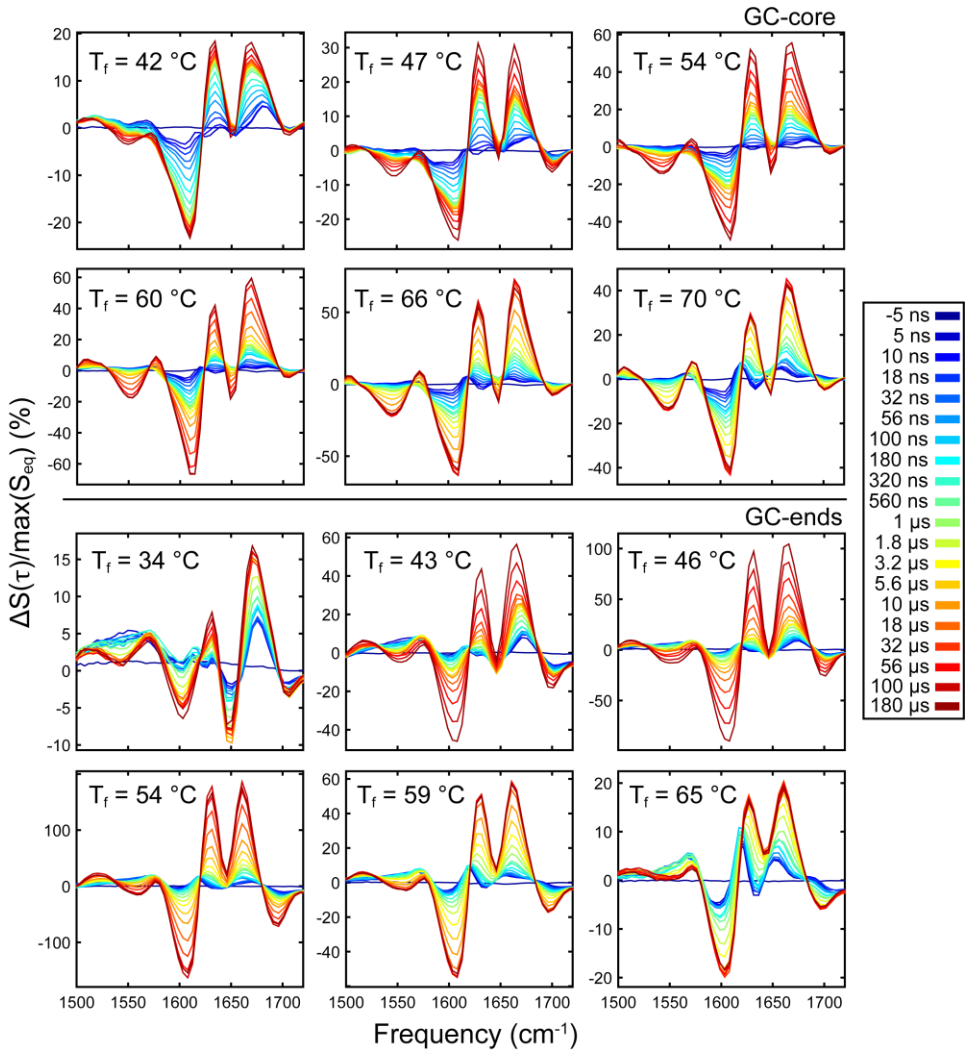

Figure 7.4: The t-HDVE spectra prior to transformation into the rate domain representations of Fig. 7.5. The final temperature is indicated in each panel and the initial temperature is indicated in Fig. 7.1b,d. The set of delays sampled is indicated in the color-coded legend. The relaxation of the signal back toward equilibrium at T_i is omitted for clarity.

of this A ring mode and T, G, and C contributions at 1630 cm^{-1} and 1665 cm^{-1}, as assigned through reference against the 2D IR difference spectrum in Fig. 7.2. The maximum amplitude rate at which the fast response is centered does not appear to depend as strongly on temperature compared to the shifting slow response. However the relative amplitude of the fast response does noticeably change in that it decreases with increasing T_f. The relaxation response, so named because it corresponds

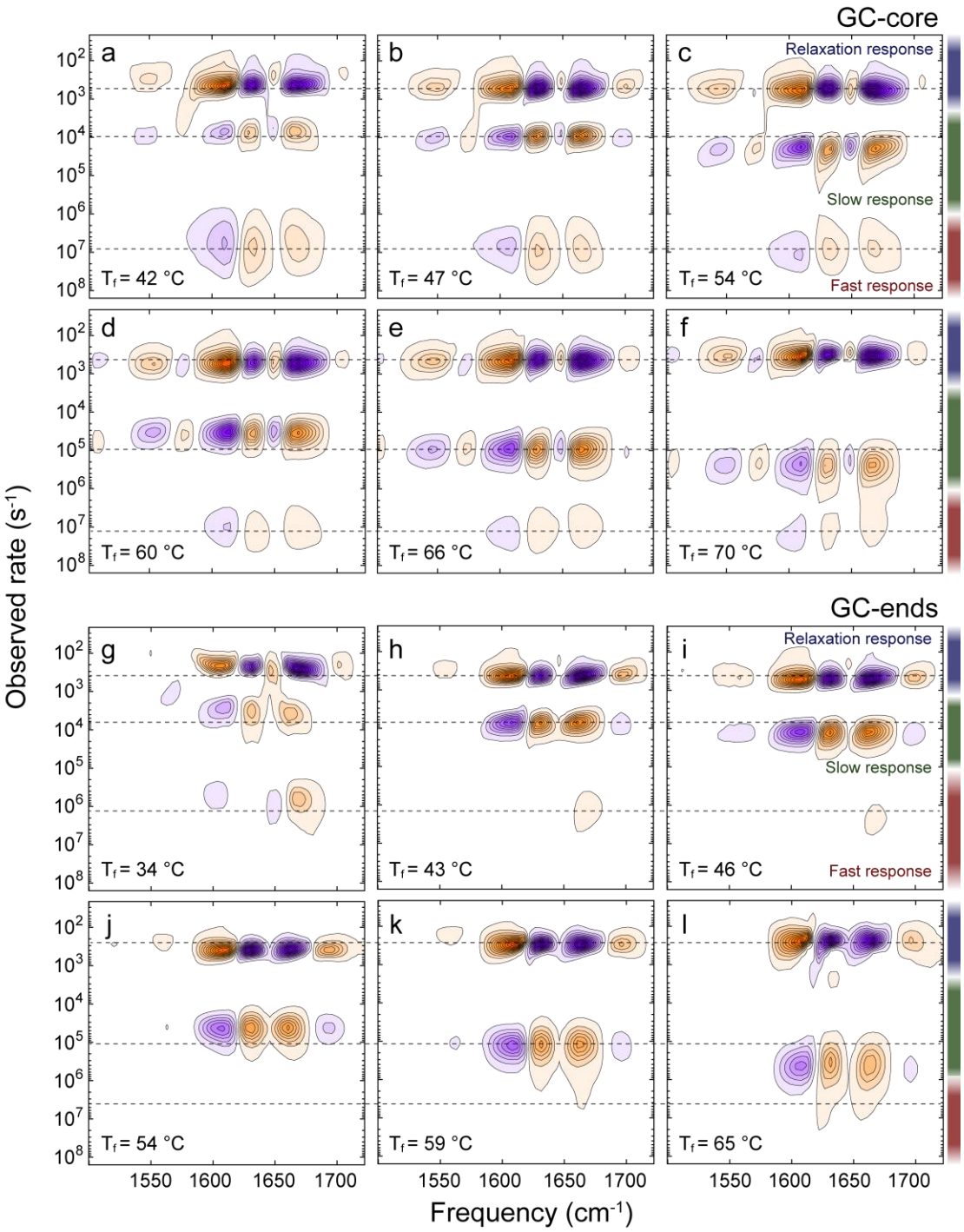

Figure 7.5: Rate spectrum representations of the t-HDVE data acquired across the (a-f) GC-core and (g-l) GC-ends melting curves. The T_i for each T-jump experiment is indicated in Fig. 7.1b,d above. Orange represents positive rate amplitude while purple represents negative amplitude. The approximate ranges of the fast, slow, and relaxation responses are indicated by the red, green, and blue color bars, respectively. Dashed lines serve as a guide for visualizing shifts along the rate axis.

with the time range over which the solvent thermal profile relaxes back towards equilibrium at T_i, appears to speed up from T_f = 42 °C through T_f = 54 °C, but then begins to slow down again above T_f = 54 °C.

The corresponding rate spectra sampled across the GC-ends melting curve behaves similarly in the evolution of the slow and relaxation responses with increasing T_f, but the fast response is markedly different. First, the rate amplitude in the fast range is greatly reduced, with only the lowest temperature spectrum showing any appreciable response. Second, the amplitude ratios in the GC-ends fast response show a different frequency pattern, with the most intensity belonging to the 1665 cm^{-1} feature corresponding to primarily T but also overlapping C and G contributions. This pattern is in contrast to the GC-core sequence, where the ESA and GSB of the 1622 cm^{-1} A ring mode are the most prominent features in the fast range. Third, what little fast response is measured for the GC-ends sequence is around two times slower than that observed for the GC-core sequence. In fact for T_f = 59 and 65 °C the fast response appears to merge with the slow response and it is difficult to cleanly separate the two ranges.

In order to better visualize the trends in observed rate across the helix-to-coil transition, the amplitude-weighted mean of the maximum amplitude rate across all frequencies was calculated at each T_f across each of the fast, slow, and relaxation regimes and plotted as a function of $\underline{T_f}$ in Fig. 7.6. The mean observed rate for the slow response is well fit by a single exponential with offset. The fits provide an empirical mapping between T_f and the average observed rate (λ) in the slow response range. Moreover an exponential dependence of λ on T_f suggests the slow response can be described by Arrhenius kinetics. For both sequences, the λ of the fast response appears linear in T_f, however the trend for the GC-ends sequence should be regarded with less confidence due to the reduced rate amplitude and ambiguity at high temperature for the fast response of this

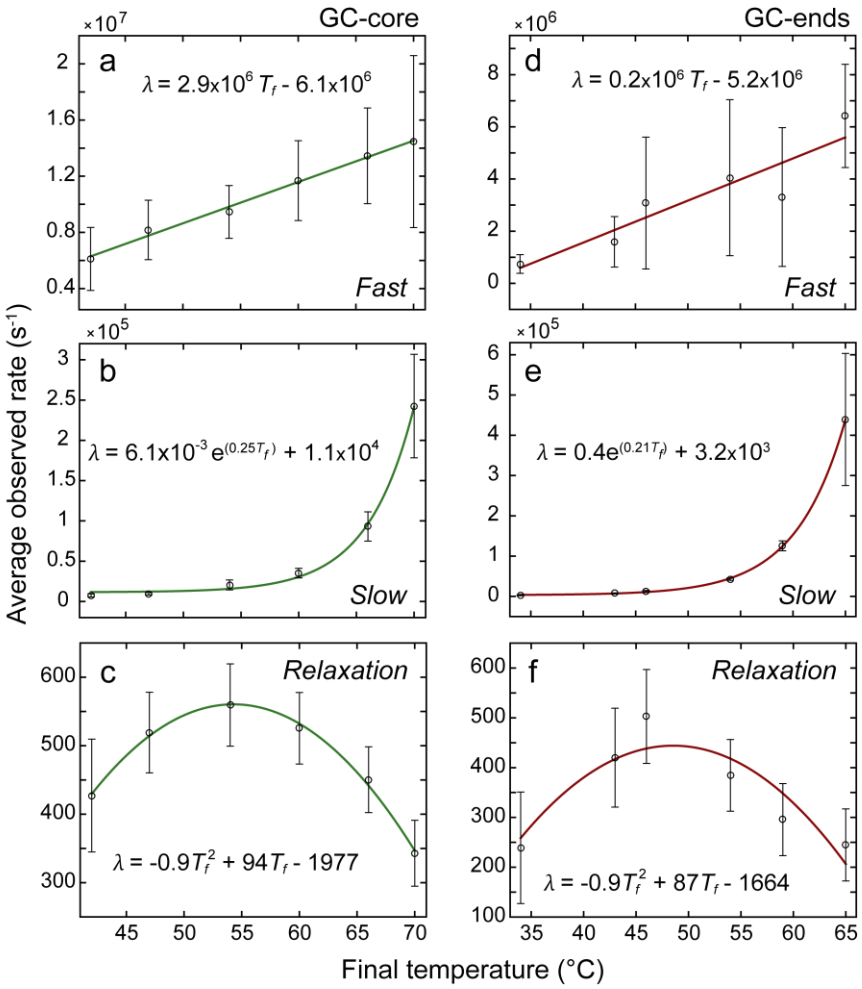

Figure 7.6: The average observed rate (λ) calculated from the amplitude weighted mean across the maximum of the (a,d) fast, (b,e) slow, and (c,f) relaxation regimes plotted vs T_f for the GC-core and GC-ends sequences. Error bars represent the amplitude weighted standard deviation of λ. An empirically motivated fit in each of the response ranges is indicated in the corresponding panel.

sequence discussed above. The T_f dependence of λ for the relaxation response is empirically described by a quadratic function. The GC-core sequence is particularly well fit by an inverted parabola. However, both sequences show similar behavior in that λ increases with increasing temperature up until ~T_m and then shows the opposite trend above T_m, with λ decreasing as the temperature is further increased.

7.3.4 Arrhenius Kinetics Describe the Slow Response

Since the slow response exhibits the exponential dependence on T_f expected for Arrhenius kinetics, we apply an Arrhenius treatment to extract the rate constants, activation energies, and prefactors for both DNA sequences. The range of rates over which the slow response occurs is consistent with previously reported timescales for dissociation of duplex DNA into monomer strands for oligonucleotides around ten base pairs in length[15,16] and is consistent with our previous assignment of this response.[23] We therefore assign the slow response to the dimer-to-monomer transition where the observed rate is related to the association (k_a) and dissociation (k_d) rate constants through eq 7.1.[37]

$$\lambda = k_d + 4[M_{eq}]k_a \tag{7.1}$$

Here $[M_{eq}]$ is the monomer concentration at equilibrium. We determine λ as discussed above, and determine $[M_{eq}]$ by assuming that the melting curves in Fig. 7.1b,d report on the fraction of intact dimers (θ_D) and that this is related to the equilibrium constant (K_D) through eq 2.

$$\theta_D = 1 + \frac{K_D}{4C_{tot}}\left(1 - \sqrt{1 + \frac{8C_{tot}}{K_D}}\right) \tag{7.2}$$

Given that the total concentration of DNA strands (C_{tot}) is known and that $K_D = [M]^2/[D] = k_d/k_a$, the dissociation and association rate constants can be determined from the measured λ and the equilibrium constant extracted from the melting curve.

Arrhenius plots for the GC-ends and GC-core sequences are plotted in Fig. 7.7a and b, respectively. The activation energy for the dissociation (E_d) and association (E_a) as well as the corresponding prefactors (A_d and A_a) are determined by fitting the logarithm of the Arrhenius equation to a plot of ln(k) vs 1/T. The values for the fit corresponding to each sequence are indicated in the appropriate figure panel. Both sequences exhibit linear Arrhenius plots and the

extracted activation energies of around 175 kJ/mol for dissociation and around -60 kJ/mol for association are consistent with previous reports for similar oligonucleotides.[13,15]

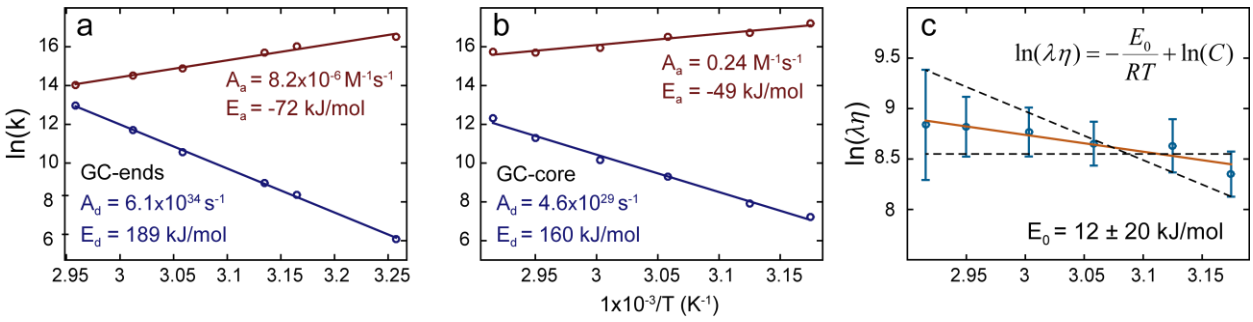

Figure 7.7: (a,b) Arrhenius plots derived from two-state analysis of the slow response of the GC-ends and GC-core sequences. (c) Fit of the GC-core viscosity-scaled fast response to Kramers' equation in the high friction limit. The orange line indicates the best fit while the dashed lines show the maximum and minimum slope lines through the data used to estimate the error in E_0.

It should be noted that the Arrhenius description as typically applied to the dimer-to-monomer transition of DNA oligonucleotides assumes a two-state all-or-nothing dissociation of the duplex in which all of the possible in-register base pairs must be either fully intact or fully broken. In light of the results discussed in Chapter 6 demonstrating a sizeable population of frayed dimer configurations for the GC-core sequence, we relax the all-or-nothing assumption in recognition that the activation energies determined here reflect an average over the heterogeneity of dimer species present. However, with regard to the two-state assumption, we found no evidence for stable intermediate states in the thermodynamic sense of additional minima in the free energy surface separated from the dimer and monomer wells by barriers larger than k_BT. We therefore assume the two-state assumption still holds in describing the slow response, but a possible source of error is introduced by imposing this assumption on the FTIR melting curve, which we have observed is more sensitive to the loss of base pairing rather than the dimer fraction.[23] In practice

deviations in the shape of the melting curve due to this effect are minor and we adopt the common assumption that the melting curve reports on the dimer fraction for the sake of comparing our results against the many past Arrhenius descriptions of DNA hybridization.

7.3.5 Variable T-Jump Experiments Suggest Zippering is Essentially Barrierless

The most striking differences between the GC-core and GC-ends rate spectra is the substantial amplitude in the sub-μs range observed for the GC-core sequence. The lack of amplitude in this range for the GC-ends sequence suggests an essentially all-or-nothing dissociation on the timescale of our measurement. For the GC-core sequence, the frequency pattern of the fast amplitude suggests fraying of the terminal AT base pairs, consistent with our previous assignment of this response.[23] The relative amplitude of the fraying response appears to decrease with increasing temperature, as the ratio of coil to helix increases (Fig. 7.5). This result is consistent with the fraying picture, as one would expect the relative contribution due to unzipping of the termini to predominate at lower temperatures where a diverse dimer ensemble is more favorable than at higher temperatures increasingly dominated by monomer strands. The fast timescales range from 70-170 ns across the 42-70 °C range sampled and display a linear temperature dependence (Fig. 7.6a).

Rapid zippering in the nucleation-zipper picture has been proposed to be a hydrodynamically limited process with a predicted activation energy of ~30 kJ/mol.[14] In order to account for a possible viscosity influence on the fast response, we follow the treatment of Ansari et al. and fit the observed rate to an expression derived from Kramers' equation in the high friction limit (eq 7.3).[38]

$$\lambda = \frac{C}{\eta}\exp\left(-\frac{E_0}{RT}\right) \tag{7.3}$$

In eq 7.3, C is an adjustable fit parameter that is related to the shape of the potential in the standard expression for a diffusive barrier crossing,[39] E_0 is the barrier height, and η is the solvent viscosity. This expression neglects an internal viscosity term applied for proteins, since we anticipate this factor is negligible for oligonucleotides.[40] To fit the experimental data, we take the logarithm of eq 7.3 and rearrange the expression such that the experimentally measured rate is scaled by the known viscosity of the D_2O solvent at T_f. The slope of a line fit to the logarithm of the viscosity scaled rates vs $1/T_f$ will be proportional to the barrier height. This analysis (Fig. 7.7c) returns a barrier of 12 ± 20 kJ/mol, where the uncertainty has been estimated from the maximum and minimum slope lines through the data indicated by the dashed lines in the figure. The value of E_0 is small and within the experimental uncertainty, suggesting the barrier to unzipping may be negligibly low. In such case the fraying observed in our experiment would be an essentially barrierless process dominated by diffusive dynamics, since the exponential in eq 7.3 approaches unity as the barrier height approaches zero. It is important to note that the barrier height from the best fit line of ~$4k_BT$ is near the limit where transition state theory is predicted to break down,[41] not to mention that eq 7.3 is derived under the assumption that the barrier height is substantially larger than k_BT and therefore this expression likely no longer adequately describes the fast response.[39] Further still, this approach does not cleanly separate the influence of temperature and viscosity on the zippering rate.

These concerns with regard to a Kramers description of rapid fraying clearly motivate a more detailed investigation of the fast dynamics that can confirm whether or not the zippering process is effectively barrierless and therefore diffusion limited. One possible strategy involves fixing the T-jump magnitude along with T_i and then increasingly adding viscogens to the solution, but this will introduce unwanted interactions with the oligonucleotides that can shift the melting

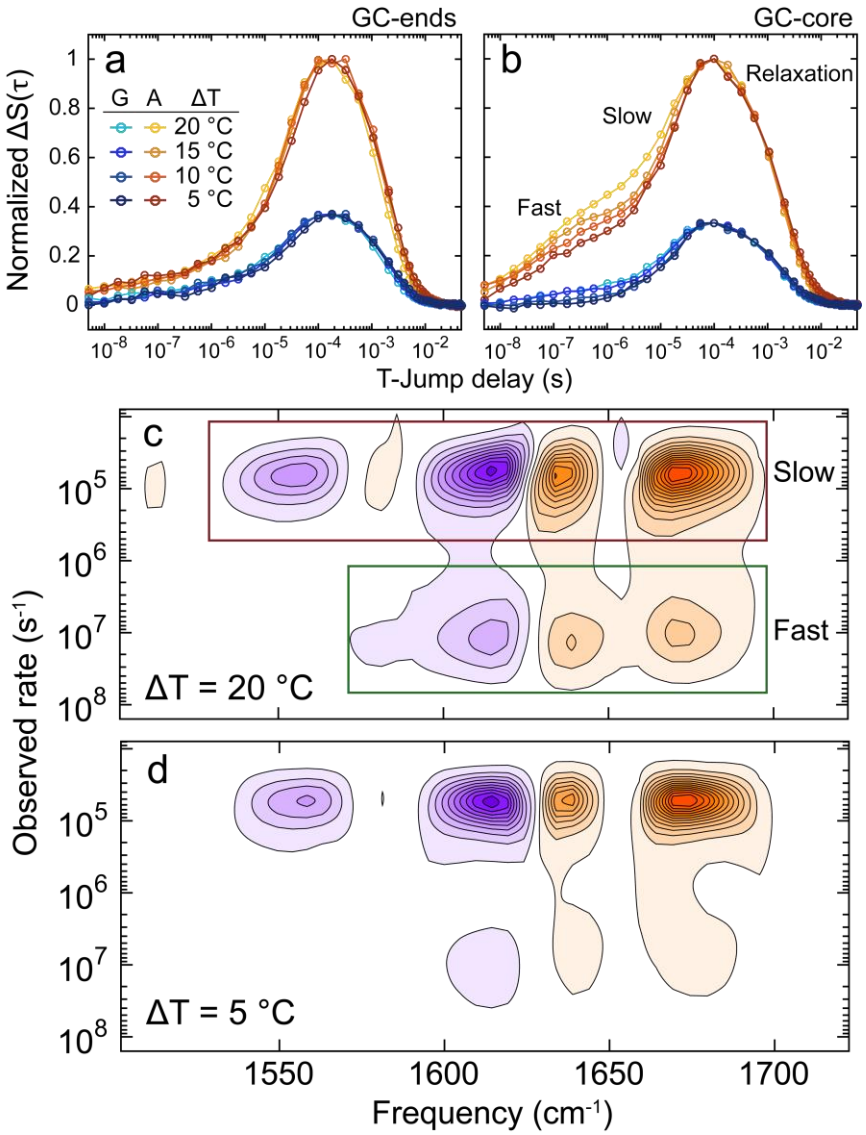

Figure 7.8: Variable T_i, fixed T_f time-domain t-HDVE kinetic traces tracked at the 1555 cm^{-1} G ring mode ESA (blue) and the 1614 cm^{-1} A ring mode ESA (orange) for the (a) GC-ends and (b) GC-core sequences. The jump magnitude ΔT associated with each trace is indicated in the figure. The rate spectrum corresponding to the (c) largest and (d) smallest T-jumps for the GC-core sequence across the full frequency range and across the slow and fast rate range.

point.[14] The most complete characterization would reveal the underlying free energy surface dictating hybridization. One approach to explore the nature of this potential at least in part is to fix T_f and vary T_i such that the different ensembles prepared at the set of initial temperatures all evolve

on the same free energy surface set by T_f. The goal of this set of measurements is to determine how drastically and in what manner the free energy surface is reshaped in response to the T-jump, thereby revealing the essential features of the surface. This variable initial potential, fixed final potential strategy was initially developed to distinguish between activated and downhill mechanisms in protein folding.[29] In order to resolve these contributions to the observed rates measured for DNA oligonucleotides, we designed a set of variable T-jump experiments in which T_f was fixed at 4 °C above T_m for each of the sequences. The magnitude of the T-jump (ΔT) was varied between 20, 15, 10, and 5 °C, as measured by the change in transmission of the D_2O solvent.[42] In order to fix T_f at 51 °C for the GC-ends sequence and 61 °C for the GC-core sequence, T_i was set such that ΔT would take the system to the target final temperature.

Select variable T-jump time traces obtained by taking single frequency t-HDVE slices at the maximum response of the ESA of the G ring modes around 1555 cm^{-1} and the ESA of the A ring mode at 1614 cm^{-1} are plotted in Fig. 7.8a and 7.8b for the GC-ends and GC-core sequences, respectively. These features were selected because they occur in relatively uncongested frequency ranges and offer insight into the dissociation of the GC and AT regions of the duplex. For the sake of comparison the traces are normalized to the maximum of the ΔT = 20 °C response of the A ring mode so as to account for the amplitude changes expected due simply to the fact that the amplitude of the slow response traces out the melting curve (Fig. 7.9). For the GC-ends sequence there is negligible amplitude in the fast response range independent of which base is probed and the normalized kinetic traces overlay. The average observed rate of the slow response across the variable T-jump set displays a consistent value that does not depend on ΔT. This observed rate as measured through the response of the A mode is $3.2 \pm 0.2 \times 10^4$ s^{-1} while the rate measured through the G modes is $3.4 \pm 0.3 \times 10^4$ s^{-1}, in reasonable agreement. Error bars represent the standard

deviation of the rates measured across the variable T-jump set. Both values are also consistent with the average observed rate of 3.2 ± 0.5 x10^4 s^{-1} obtained by averaging across the entirety of the slow response.

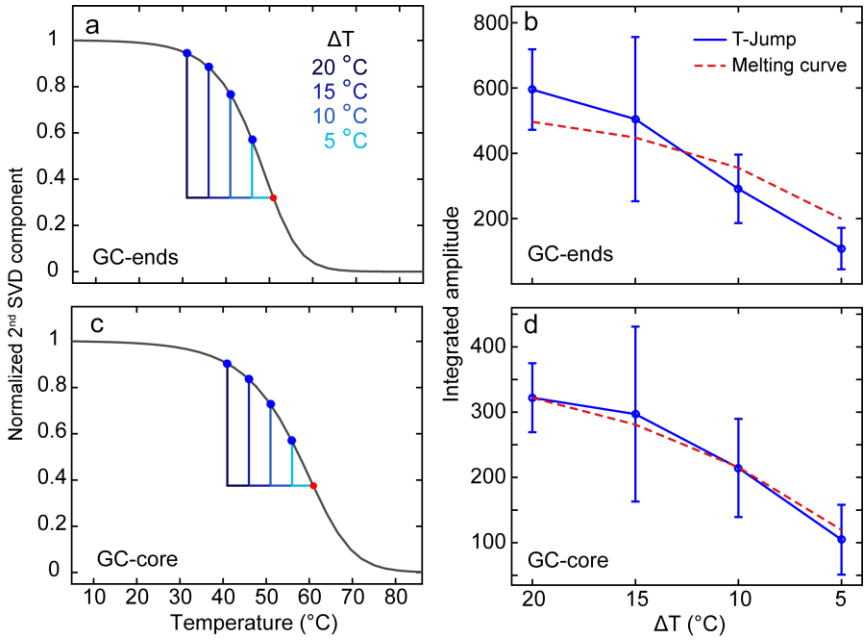

Figure 7.9: The amplitude of the slow response traces out the melting curve. The melting curve and temperature ranges of the variable T-jump experiments carried out for the (a) GC-ends and (c) GC-core sequences. Blue points represent the T_i while the red point represents the fixed T_f. The integrated absolute value of the slow response at each ΔT (blue line) compared to the change in amplitude from the melting curve for each temperature change (red dashed line) for (b) GC-ends and (d) GC-core.

For the GC-core sequence the dependence on T_i and the T-jump magnitude is less straightforward. As before, transformation to the rate domain offers a more intuitive means to extract rate and amplitude trends from the T-jump data. The rate spectra corresponding to the largest (ΔT = 20 °C) and smallest (ΔT = 5 °C) jumps for the GC-core variable T-jump experiments are shown in Fig. 7.8c,d. Consistent with the GC-ends sequence, the average observed rate associated with the slow response appears to be independent of T_i, yielding a consistent value of

$5.9 \pm 0.5 \times 10^4$ s^{-1} whether tracked at the A mode, G modes, or the average across the entire slow response range. In contrast, it is apparent that the fast response displays a T_i dependence. As observed in both the A mode kinetic traces (Fig. 7.8b) and the rate spectra (Fig. 7.8c,d), the relative amplitude of the fast response decreases with decreasing jump magnitude, dying off as ΔT approaches zero. Integrating across the absolute value of the amplitude in the fast response range, indicated by the green box in Fig. 7.8c, offers a convenient representation of how this amplitude tracks the T-jump magnitude (Fig. 7.10c). The average observed rate associated with the GC-core fast response also appears to depend on T_i. A plot of the amplitude weighted mean rate across the fast response range vs ΔT (Fig. 7.10d) illustrates that λ peaks around $1.3 \pm 0.2 \times 10^7$ s^{-1} when $\Delta T = 15$ °C, corresponding to a T_i of 46 °C, but falls off as T_i is decreased to 41 °C or increased to 51 and 56 °C. The smallest T-jump magnitude returns the slowest rate. When $\Delta T = 5$ °C, λ decreases to around $7.3 \pm 2.2 \times 10^6$ s^{-1}.

7.4 Discussion

7.4.1 Connection to Past Studies of DNA Dehybridization

For the slow response assigned to the dimer-to-monomer transition, the Arrhenius temperature dependence, extracted rate constants, magnitude of the activation energies, and the reasonable effectiveness of a two-state dimer/monomer model to account for the observed kinetics are all consistent with previous temperature jump studies of the dehybridization of DNA oligomers of similar length and composition.[13,15] The structural insight afforded by IR spectroscopy reveals that there is no discernable difference in the observed rate of the slow response depending on which part of the helix is probed, with both AT and GC features returning identical rates to within the error of the measurement. This result confirms that the final dissociation of the dimer into the

monomer state involves a concerted loss of the remaining intact base pairs. The negative activation energy measured for the association process necessarily cannot correspond to an elementary reaction step, but given the high dimensionality of the association problem, in which two strands must diffuse into proximity, establish initial contacts, and then orient such that hybridization can proceed, this result is not surprising. In fact anti-Arrhenius behavior for the association rate is well-documented across several different experimental techniques and has been attributed to the presence of a metastable intermediate in the rate-limiting step,[13] the free energy barrier arising from a significant entropic loss upon formation of the transition state,[43-45] and to increasing numbers of conformational configurations as temperature increases.[40] All of these explanations are complementary and consistent with the trends we observe in k_a as well as with what one would expect for a dimerization process initiated by a critical nucleation step preceded by a diffusive encounter plus reorientation. Finally, the behavior of the slow response observed in the variable T-jump experiments definitively supports the assignment of this timescale to a barrier crossing between the dimer and monomer ensembles. Since the barrier height of an activated process is set by the temperature at which the kinetics are observed, the rate measured for such a process should only depend on T_f, with T_i only influencing the amplitude of the response. Indeed the average observed rate of the slow response for both sequences is measured to be independent of T_i and is therefore consistent with an activated barrier crossing.

The relaxation response reports on the rehybridization of the dissociated monomer strands as the thermal energy dissipates and the temperature returns to T_i. As a result, it is more difficult to extract reliable kinetic information from this response since the temperature is no longer constant and the oligonucleotide response must be deconvolved from the thermal relaxation. Therefore we will not draw quantitative conclusions from this range of rates for now, but a

discussion of the qualitative trends is nevertheless informative. One should note that the T_f temperature axis for the relaxation response in Fig. 7.6c,f is no longer strictly correct due to the relaxing temperature of the system. Therefore there is uncertainty in which temperature corresponds to the maximum observed rate of rehybridization. However the temperature range sampled for the T-jump that exhibits a maximum in the measured relaxation response offers reasonable bounds. In this case the maximum rate would lie in a range from a few degrees below T_m to 11 or 17 °C below T_m, for the GC-ends and GC-core sequences respectively. This range for the maximum rate is in agreement with stop-flow UV experiments that measured recombination directly, where the hybridization rate is observed to reach a maximum at ~0.9T_m (in Kelvin), then drop off with increasing temperature[46] as well as Förster resonance energy transfer (FRET) rapid mixing experiments where the peak hybridization rate is measured ~10 °C below T_m.[43]

7.4.2 Modeling the Fast Response as Diffusion on a Reshaped Free Energy Surface

We have established that the Arrhenius description of the dimer-to-monomer transition as well as the behavior observed for the rehybridization rate align well with past studies of DNA hybridization and this agreement serves to validate our nonlinear IR spectroscopy approach to studying nucleic acid folding. However, we have also directly resolved fast fraying of the termini that, to our knowledge, has not been characterized in detail experimentally. The apparent breakdown in a Kramers description of zippering previously employed to successfully describe the role of solvent viscosity on protein conformational changes[38] provides an initial indication that the barrier to fraying may be negligibly low. Variable T-jump experiments in which T_f is fixed and T_i is varied provide a direct indication that the fraying response is a diffusion limited process, since the T_i dependence of the GC-core fast response is inconsistent with an activated process where one

would expect the barrier height, and therefore the rate, to be set by the free energy surface at T_f alone. Such a T_i dependent unfolding rate has been reported for the BBL protein, which is a well-studied downhill folder.[29] The trend reported for the fast rate, with a maximum at ΔT = 15 °C (Fig. 7.10d), as well as the shape of the amplitude trend (Fig. 7.10c) are at first glance puzzling.

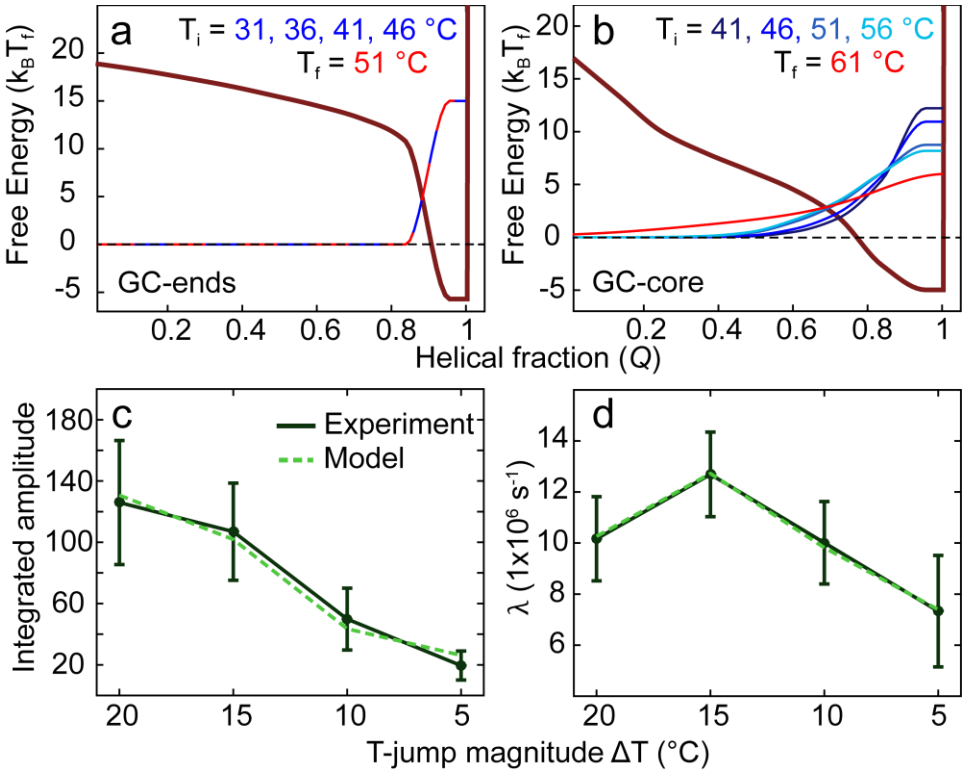

Figure 7.10: (a) The dimer basin free energy surface at T_f proposed for the GC-ends sequence. The helical fraction (Q) is defined in the main text. The normalized populations at the set of T_i (blue) and at T_f (red) are superimposed over the free energy surface for reference. (b) The dimer basin free energy surface proposed for the GC-core sequence used in the diffusion simulations to model the observed trends in the fast response as rapid re-equilibration within the frayed dimer ensemble. (c) The trend in the integrated GC-core fast response rate amplitude as a function of ΔT compared against the change in average helical fraction between T_i and T_f predicted by the model. (d) The trend in amplitude weighted average observed rate as a function of ΔT compared against the rate constant calculated for the decay of the helical fraction in the diffusion simulations for the GC-core sequence.

To explore the possible origins of these observations, we turn to simulations in which the fast response is modeled as rapid re-equilibration within the frayed dimer ensemble in response to a reshaped dimer basin following the temperature jump. Diffusion within a potential is described by the Smoluchowski equation (eq 7.4).

$$\frac{\partial P(Q,t)}{\partial t} = D_0 \frac{\partial}{\partial Q}\left[\frac{\partial P(Q,t)}{\partial Q} + \frac{1}{k_B T_f}\frac{\partial G(Q,T_f)}{\partial Q} P(Q,t)\right] \tag{7.4}$$

We introduce a reaction coordinate, *Q*, which we will assume tracks the structural changes to which our IR probe is primarily sensitive, namely the breaking of hydrogen bonds, the reduction of base stacking, and the unwinding of the helix towards free coils. We therefore define *Q* generically as the "helical fraction", which can range between a value of 1 corresponding to fully paired B-form DNA and a value of 0 corresponding to a complete loss of Watson-Crick character without considering the dimer-to-monomer transition. In eq 7.4, *P* is the time dependent dimer population, D_0 is the diffusion coefficient, which for simplicity is assumed to be independent of *Q*, and *G* is the one-dimensional free energy surface set by T_f on which the population initially set by T_i diffusively re-equilibrates. Caution should be taken when employing such a coarse model so as not to over-interpret the experimentally justifiable conclusions, but this simple description does appear to capture some of the salient characteristics of the free energy landscape and offers a plausible, albeit low-dimensional, explanation for all of the experimental results.

7.4.3 The GC-ends Sequence Dehybridizes with a Concerted Loss of All Base Pairs

Based on the observation that there is negligible amplitude in the fast response range for the GC-ends sequence and that the observed kinetics show no T_i dependence in the fixed T_f variable T-jump experiments, we propose the dimer basin free energy surface depicted in Fig. 7.10a. The normalized dimer population distributions at the set of T_i and at T_f are superimposed over the free

energy surface for reference. This surface has a sharp minimum near $Q = 1$ and regardless of T_i, the starting population occupies this low point. Normalizing the initial dimer population so as to only consider conformational shifts within the dimer ensemble, all of the initial distributions across the T_i set are similar in shape and overlay. Furthermore, since all of the starting population starts out in the deep minimum near $Q = 1$ and the walls of this minimum are steep, the population cannot diffuse outwards and no rapid re-equilibration within the dimer basin in response to the T-jump is observed. The dissociation mechanism dictated by such a free energy surface is essentially all-or-nothing in nature. Dimers with helical fraction lower than 1 are energetically unfavorable and therefore the only observed timescale corresponds to barrier crossing out of the sharp minimum in the dimer basin into the monomer basin, which necessarily exists on a reaction coordinate orthogonal to the dimer-specific helical fraction, Q. Therefore the proposed surface in Fig. 7.10a accounts for both the lack of amplitude in the rapid fraying response range as well as the behavior of the GC-ends sequence in the variable T-jump experiments and suggests an essentially all-or-nothing mechanism adequately applies in this case.

As can be seen in Fig. 7.5g-l, there is a small amplitude feature in the fast response range for the GC-ends sequence. This feature is most pronounced for the $T_f = 34$ °C surface in panel g, and is seen to disappear below the level of contouring presented in the figure as temperature increases until the fast response amplitude merges with the tail of the slow response at high temperature, as discussed above. Due to the low intensity of this feature and the ambiguity at high temperature, it is difficult to cleanly define this response. However, as can be seen in Fig. 7.6d, the mean rate across this range appears to follow a roughly linear trend consistent with the behavior observed for the GC-core sequence, but the rate is around a factor of two slower. In contrast to the GC-core sequence, the fast amplitude peaks at 1665 cm^{-1}. Unfortunately this frequency lies in the

most congested range sampled, with overlapping contributions from T, G, and C. This congestion in addition to the comparatively poor signal to noise for this feature makes it difficult to say anything definitive about the origin of this response. The 80% AT content would suggest that the signal at 1665 cm^{-1} should be dominated by T, and this is certainly true for the linear and 2D IR spectra, but one must keep in mind that the t-HDVE spectra are difference measurements. It is therefore possible that the early-time changes to the spectrum at this frequency are dominated by a loss of GC base pairing at the end of the duplex.

Upon base pairing, the overlapping G and C modes centered around 1665 cm^{-1}, which have primarily carbonyl stretching character, split as the G peak blue shifts to around 1675 cm^{-1}.[27] Therefore one would expect that the loss of GC contacts would correspond to a growth in intensity at 1665 cm^{-1}. The slower timescale measured for the GC-ends sequence fast response is consistent with the increased stability of GC base pairs. However, one would also expect an intensity increase from the G ring modes around 1565 cm^{-1}, but we do not measure an appreciable fast response in this frequency range. It is possible that the changes in the ring mode intensity are simply smaller than the changes to the spectrum due to the shifting G carbonyl mode and therefore fall below the noise. If what fast response can be measured for this sequence does correspond to the loss of the GC cap at the terminus, it is consistent with the ends-initiated unzipping mechanism suggested above for the GC-core sequence. However, there is simply not enough confidence in the small signal measured to conclude this definitively. Within the resolution of our experiment, GC-ends appears to undergo an essentially concerted all-or-nothing loss of base pairing. Future experiments conducted at low temperature where the signal contribution from dimer dissociation is reduced and the spectrum is dominated by pre-dissociation effects could provide additional insight into the origins of the GC-ends fast response.

7.4.4 Fraying in the GC-core Sequence Appears to be Diffusion Limited

The free energy surface proposed for the GC-core dimer basin does not have as sharp of a minimum near $Q = 1$ and also shows a more diverse distribution of frayed dimers at each T_i (Fig. 7.10b). The design of both the shape of the dimer basin potential and the starting population distribution at each T_i as well as the assignment of dimer heterogeneity to frayed configurations are informed by our previous equilibrium experiments and modeling for these sequences discussed in Chapter 6. In contrast to the GC-ends sequence, the shape of the starting population distribution shows a dependence on T_i, with the dimer distribution spreading out towards lower helical fraction as T_i is raised. The shape of the dimer basin at T_f is also broader and flatter. Taken together, these initial conditions result in the diffusive spread of the initial population out into the dimer basin resulting in the T_f population distribution in Fig. 7.10b.

The diffusion simulations numerically propagate the Smoluchowski equation, eq 7.4, starting with the initial populations and the potential shown in Fig. 7.10b. The shape of this potential is a refined version of the GC-core free energy surface from Chapter 6. The surface is plotted with respect to the more general reaction coordinate, Q, and the increased steepness of the potential near high helical fraction is found to be essential for reproducing the experimental results. Normalized dimer starting populations at each of the initial temperatures are calculated from the lattice model discussed in Chapter 5. Using a finite difference method, the discretized evolution of the dimer population, P, is tracked by stepping eq 7.5 by a small time increment h.

$$P(Q,t+h) = P(Q,t) + hD\left[\frac{\partial^2 P(Q,t)}{\partial Q^2} + \frac{1}{k_B T_f}\left(\frac{\partial G(Q,T_f)}{\partial Q}\frac{\partial P(Q,t)}{\partial Q} + \frac{\partial G^2(Q,T_f)}{\partial Q^2}P(Q,t)\right)\right] \quad (7.5)$$

The variables in eq 7.5 are defined as for eq 7.4. Since T_f is fixed across the set of T-jumps, the free energy surface, G, is fixed and we further assume that G is time independent. After each step

in time the derivatives of *P* with respect to the reaction coordinate *Q* are updated. The diffusive spread of the dimer population initially prepared at each T_i is shown in Fig. 7.11a. In the figure, the time evolution is indicated by the color of the distribution, which runs from blue at t = 0 to red at the final simulation step. The simulation time was propagated for ~10,000 steps, resulting in a final population distribution across the T-jump set that varied by less than 1%. After 10,000 steps, the dimer population remained normalized to within 5%. Renormalizing the population at each step did not noticeably influence the results.

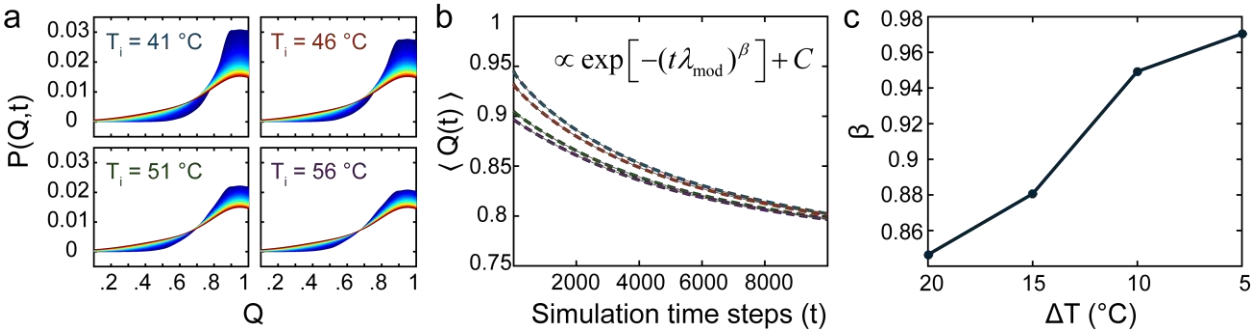

Figure 7.11: (a) Diffusion of the dimer population prepared at each T_i out into the dimer basin. Simulation time runs from blue to red. (b) The decay of the ensemble average helical fraction, *Q*, across 10,000 simulation steps (color coded dashed lines). The populations were propagated until the final populations agreed to within <1%. The decays are best fit by a stretched exponential decay with offset (gray solid lines). (c) The stretching parameter β as a function of ΔT from the fits in panel b. The decay of $\langle Q(t) \rangle$ approaches single exponential kinetics (β = 1) as ΔT goes to zero.

The quantities from the model which are compared against experiment rely on the ensemble averaged value of *Q* as a function of time. Fig. 7.11b shows the decay of this quantity for each of the simulated jumps. The amplitude trend from experiment is compared against the change in $\langle Q(t) \rangle$ between T_i and the final population distribution (Fig. 7.10c). To extract a characteristic time constant for each of the decays in Fig. 7.11b, each trace was fit to a stretched exponential with offset since the decays are best fit by this function. The model rates (λ_{mod})

compared against experiment in Fig. 7.10d are the inverse of the time constant from these fits. The calculated values are scaled to match experiment by the Q diffusion coefficient, $D_0 = 3.3 \times 10^{12}\ Q^2/s^{-1}$.

It is interesting to note that stretched exponential behavior naturally arises out of the model and that the degree to which the decay is stretched depends on the T-jump magnitude ΔT. The stretching parameter, β, is plotted as a function of ΔT in Fig. 7.11c. As can be seen in the figure, β approaches 1 as ΔT approaches zero, indicating that the re-equilibration approaches exponential behavior as the T-jump becomes negligibly small. This result suggests that the diffusion-limited re-equilibration within the dimer ensemble follows a reshaping of the free energy surface that is increasingly far from perturbative for growing values of ΔT. Consequently increasingly distinct dimer configurations lying further apart in Q from the final distribution are accessible resulting in the effect of stretching out the observed kinetic trace. If this description is correct, it would suggest that the rapid zippering reaction is essentially continuous with respect to temperature. Unfortunately, the differences in the stretching parameter suggested by the model over the range of ΔT sampled are quite small and we are unable to confidently resolve such small changes in β with stretched exponential fits to the data.

It is possible that the rate spectrum representation could be useful in quantifying stretched exponential kinetics in the future without having to rely on fitting low-amplitude signals in the time domain. One would expect a stretched exponential time trace to result in an asymmetric rate distribution, with the skew of the distribution related to the stretching parameter β, as demonstrated in Chapter 4. At this time we have not done a careful characterization of how stretched exponential time domain t-HDVE data manifests in the rate domain. As a further complication, the noise variance estimate in the MEM-iLT can influence the width of the rate distribution and any analysis

relying on subtle shifts in the shape of the peaks in the rate domain would require a careful characterization of the influence of this effect as well.

The agreement between the model and experiment is not surprising given that the shape of the potential and starting ensembles were designed to account for the experimental observations. However, neither were drawn arbitrarily. Both the general shape of the dimer basin and the trend in starting populations with T_i are consistent with the characterization of these sequences in the previous chapter. Connecting the shifts in dimer population to a molecular picture, the drive towards lower values of Q in response to the T-jump corresponds to transitions from more highly helical to more highly frayed configurations through rapid unzipping of the termini in a diffusion limited re-equilibration of the dimer ensemble that precedes barrier crossing to the monomer basin. The simulations reveal an interplay between the diffusive spread of the initial population away from $Q = 1$ and the opposing slope of the dimer basin potential. These competing factors offer an explanation for the maximum in the average observed rate at $\Delta T = 15$ °C. For a purely diffusive re-equilibration, one would expect the fastest rate to be associated with the most highly frayed starting ensemble, since this distribution would most closely resemble the final distribution. However, unzipping of the termini results in dimer configurations that lie higher in energy than the $Q = 1$ dimer, and therefore the slope of the potential increasingly opposes the spread of the frayed ensemble with decreasing helical fraction. The starting population at $T_i = 46$ °C re-equilibrates most rapidly following the T-jump because it strikes a balance between the initial width of the distribution and the amount of starting population on steeper regions of the potential. This explanation relies on a simplified picture that neglects possibly important factors such as higher-dimensional reaction coordinates[47] or the influence of the roughness of the free energy

landscape.[44,48] Nevertheless, the model offers a self-consistent explanation for all of the experimental observations.

7.5 Conclusions

This chapter detailed the study of the dehybridization of two contrasting model DNA oligonucleotides using an optically triggered temperature jump to induce unfolding followed by a nonlinear IR spectroscopy probe. This approach proves powerful in that it offers base specific insight, as encoded through the distinct pattern of vibrational absorptions unique to each of the nucleobases, as well as insight into secondary structure through the substantial reshaping of these line shapes and intensities in response to hybridization. Time resolution spanning ns-ms allows kinetics to be tracked across many decades in time, and for the sequences studied here we observe in a single measurement unzipping of the termini, the loss of final dimer contacts, and ultimately rehybridization as the ensemble relaxes back towards the initial conditions. By monitoring dimer dissociation we measure the activation energy for dimer separation and find this process to be well described by Arrhenius kinetics, consistent with many previous studies. However, in addition to this activated process, we also observe essentially barrierless fraying of the helical termini in the GC-core sequence. This new insight regarding the nature of base pair zippering refines the mechanistic details of the fastest hybridization dynamics following helix nucleation. As a result, our findings suggest that the kinetic models routinely employed in the past that impose a fundamental zippering rate constant on each pairing step post nucleation result in a somewhat artificial description of hybridization. Our data support a picture in which the dimer ensemble can exist in a broad basin where the populated configurations are not separated by appreciable barriers, suggesting rapid interconversion can occur between all of the accessible frayed structures within

the well. In contrast, placement of the GC pairs at the helical termini results in an essentially all-or-nothing hybridization mechanism in which only the activated process is observed, as measured for the GC-ends sequence. Going forward we believe the approach presented here provides a route to study the structural dynamics of nucleic acids in detail across a wide range of timescales and can offer unique insight into the features of the underlying folding free energy surface. The next two chapters will apply the approach developed in Chapters 6 and 7 to study non-canonical DNA sequences containing modified cytosine bases relevant to epigenetic regulation in eukaryotes.

7.6 Acknowledgements

I thank Paul Stevenson for careful reading of the manuscript that covers much of the material in this chapter.

7.7 References

1. Watson, J. D.; Crick, F. H., Molecular structure of nucleic acids. *Nature* **1953,** *171* (4356), 737-738.

2. Costa, A.; Hood, I. V.; Berger, J. M., Mechanisms for initiating cellular DNA replication. *Annual review of biochemistry* **2013,** *82*, 25-54.

3. Bramhill, D.; Kornberg, A., Duplex opening by dnaA protein at novel sequences in initiation of replication at the origin of the E. coli chromosome. *Cell* **1988,** *52* (5), 743-755.

4. Choi, C. H.; Kalosakas, G.; Rasmussen, K. Ø.; Hiromura, M.; Bishop, A. R.; Usheva, A., DNA dynamically directs its own transcription initiation. *Nucleic Acids Research* **2004,** *32* (4), 1584-1590.

5. Cordes, T.; Santoso, Y.; Tomescu, A. I.; Gryte, K.; Hwang, L. C.; Camará, B.; Wigneshweraraj, S.; Kapanidis, A. N., Sensing DNA opening in transcription using quenchable Forster resonance energy transfer. *Biochemistry* **2010,** *49* (43), 9171-9180.

6. Polach, K.; Widom, J., Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *Journal of molecular biology* **1995,** *254* (2), 130-149.

7. Ziller, M. J.; Gu, H.; Müller, F.; Donaghey, J.; Tsai, L. T.-Y.; Kohlbacher, O.; De Jager, P. L.; Rosen, E. D.; Bennett, D. A.; Bernstein, B. E., Charting a dynamic DNA methylation landscape of the human genome. *Nature* **2013,** *500* (7463), 477.

8. Jones, M. R.; Seeman, N. C.; Mirkin, C. A., Programmable materials and the nature of the DNA bond. *Science* **2015,** *347* (6224), 1260901.

9. Douglas, S. M.; Dietz, H.; Liedl, T.; Hogberg, B.; Graf, F.; Shih, W. M., Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* **2009,** *459* (7245), 414-418.

10. Teller, C.; Willner, I., Functional nucleic acid nanostructures and DNA machines. *Current opinion in biotechnology* **2010,** *21* (4), 376-391.

11. Bath, J.; Turberfield, A. J., DNA nanomachines. *Nature nanotechnology* **2007,** *2* (5), 275-284.

12. Sakamoto, K.; Gouzu, H.; Komiya, K.; Kiga, D.; Yokoyama, S.; Yokomori, T.; Hagiya, M., Molecular computation by DNA hairpin formation. *Science* **2000,** *288* (5469), 1223-1226.

13. Craig, M. E.; Crothers, D. M.; Doty, P., Relaxation kinetics of dimer formation by self complementary oligonucleotides. *Journal of molecular biology* **1971,** *62* (2), 383-401.

14. Wetmur, J. G.; Davidson, N., Kinetics of renaturation of DNA. *Journal of molecular biology* **1968,** *31* (3), 349-370.

15. 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. *Journal of molecular biology* **1971,** *62* (2), 361-381.

16. Pörschke, D.; Uhlenbeck, O.; Martin, F., Thermodynamics and kinetics of the helix-coil transition of oligomers containing GC base pairs. *Biopolymers* **1973,** *12* (6), 1313-1335.

17. Chen, X.; Zhou, Y.; Qu, P.; Zhao, X. S., Base-by-base dynamics in DNA hybridization probed by fluorescence correlation spectroscopy. *Journal of the American Chemical Society* **2008,** *130* (50), 16947-16952.

18. Klostermeier, D.; Millar, D. P., Time-resolved fluorescence resonance energy transfer: A versatile tool for the analysis of nucleic acids. *Biopolymers* **2002,** *61* (3), 159-179.

19. Woodside, M. T.; Anthony, P. C.; Behnke-Parks, W. M.; Larizadeh, K.; Herschlag, D.; Block, S. M., Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid. *Science* **2006,** *314* (5801), 1001-1004.

20. Andreatta, D.; Sen, S.; Pérez Lustres, J. L.; Kovalenko, S. A.; Ernsting, N. P.; Murphy, C. J.; Coleman, R. S.; Berg, M. A., Ultrafast Dynamics in DNA:"Fraying" at the End of the Helix. *Journal of the American Chemical Society* **2006,** *128* (21), 6885-6892.

21. Leroy, J. L.; Kochoyan, M.; Huynh-Dinh, T.; Guéron, M., Characterization of base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. *Journal of molecular biology* **1988,** *200* (2), 223-238.

22. Nonin, S.; Leroy, J.-L.; Gueron, M., Terminal base pairs of oligodeoxynucleotides: imino proton exchange and fraying. *Biochemistry* **1995,** *34* (33), 10652-10659.

23. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A., Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *Journal of the American Chemical Society* **2016,** *138* (36), 11792-11801.

24. Ouldridge, T. E.; Šulc, P.; Romano, F.; Doye, J. P.; Louis, A. A., DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic acids research* **2013,** *41* (19), 8886-8895.

25. Hinckley, D. M.; Lequieu, J. P.; de Pablo, J. J., Coarse-grained modeling of DNA oligomer hybridization: length, sequence, and salt effects. *The Journal of chemical physics* **2014,** *141* (3), 035102.

26. Zgarbová, M.; Otyepka, M.; Šponer, J. í.; Lankaš, F.; Jurečka, P., Base pair fraying in molecular dynamics simulations of DNA and RNA. *Journal of chemical theory and computation* **2014,** *10* (8), 3177-3189.

27. Banyay, M.; Sarkar, M.; Gräslund, A., A library of IR bands of nucleic acids in solution. *Biophysical chemistry* **2003,** *104* (2), 477-488.

28. Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A., Transient two-dimensional IR spectrometer for probing nanosecond temperature-jump kinetics. *Review of scientific instruments* **2007,** *78* (6), 063101.

29. Lin, C.-W.; Culik, R. M.; Gai, F., Using VIP T-jump to distinguish between different folding mechanisms: application to BBL and a Trpzip. *Journal of the American Chemical Society* **2013,** *135* (20), 7668-7673.

30. Peng, C. S.; Jones, K. C.; Tokmakoff, A., Anharmonic vibrational modes of nucleic acid bases revealed by 2D IR spectroscopy. *Journal of the American Chemical Society* **2011,** *133* (39), 15650-15660.

31. Krummel, A. T.; Zanni, M. T., DNA vibrational coupling revealed with two-dimensional infrared spectroscopy: insight into why vibrational spectroscopy is sensitive to DNA structure. *The Journal of Physical Chemistry B* **2006,** *110* (28), 13991-14000.

32. Yang, M.; Szyc, Ł.; Elsaesser, T., Femtosecond two-dimensional infrared spectroscopy of adenine-thymine base pairs in DNA oligomers. *The Journal of Physical Chemistry B* **2011,** *115* (5), 1262-1267.

33. Hithell, G.; González-Jiménez, M.; Greetham, G. M.; Donaldson, P. M.; Towrie, M.; Parker, A. W.; Burley, G. A.; Wynne, K.; Hunt, N. T., Ultrafast 2D-IR and optical Kerr effect

spectroscopy reveal the impact of duplex melting on the structural dynamics of DNA. *Physical Chemistry Chemical Physics* **2017,** *19* (16), 10333-10342.

34. Hithell, G.; Ramakers, L. A.; Burley, G. A.; Hunt, N. T., Applications of 2D-IR Spectroscopy to Probe the Structural Dynamics of DNA. In *Frontiers and Advances in Molecular Spectroscopy*, Elsevier: 2018; pp 77-100.

35. Jones, K. C.; Ganim, Z.; Tokmakoff, A., Heterodyne-detected dispersed vibrational echo spectroscopy. *The Journal of Physical Chemistry A* **2009,** *113* (51), 14060-14066.

36. Khalil, M.; Demirdöven, N.; Tokmakoff, A., Coherent 2D IR spectroscopy: molecular structure and dynamics in solution. *The Journal of Physical Chemistry A* **2003,** *107* (27), 5258-5279.

37. Nölting, B., *Protein folding kinetics: biophysical methods*. Springer: Berlin, 2006.

38. Ansari, A.; Jones, C. M.; Henry, E. R.; Hofrichter, J.; Eaton, W. A., The role of solvent viscosity in the dynamics of protein conformational changes. *Science* **1992,** *256* (5065), 1796-1798.

39. Nitzan, A., *Chemical Dynamics in Condensed Phases*. Oxford University Press: Oxford, 2006.

40. Wallace, M. I.; Ying, L.; Balasubramanian, S.; Klenerman, D., Non-Arrhenius kinetics for the loop closure of a DNA hairpin. *Proceedings of the National Academy of Sciences* **2001,** *98* (10), 5584-5589.

41. Ma, H.; Gruebele, M., Low barrier kinetics: dependence on observables and free energy surface. *Journal of computational chemistry* **2006,** *27* (2), 125-134.

42. Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B., Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry* **1996,** *35* (3), 691-697.

43. 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 research* **2007,** *35* (9), 2875-2884.

44. Ansari, A.; Kuznetsov, S. V.; Shen, Y., Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proceedings of the National Academy of Sciences* **2001,** *98* (14), 7771-7776.

45. Sorgenfrei, S.; Chiu, C.-y.; Gonzalez Jr, R. L.; Yu, Y.-J.; Kim, P.; Nuckolls, C.; Shepard, K. L., Label-free single-molecule detection of DNA-hybridization kinetics with a carbon nanotube field-effect transistor. *Nature nanotechnology* **2011,** *6* (2), 126-132.

46. Ross, P. D.; Sturtevant, J. M., On the kinetics and mechanism of helix formation: The two stranded poly (A+U) complex from polyriboadenylic acid and polyribouridylic acid. *Journal of the American Chemical Society* **1962,** *84* (23), 4503-4507.

47. Chung, H. S.; Khalil, M.; Smith, A. W.; Ganim, Z.; Tokmakoff, A., Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (3), 612-617.

48. Zwanzig, R., Diffusion in a rough potential. *Proceedings of the National Academy of Sciences* **1988,** *85* (7), 2029-2030.

Chapter 8

Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability

The work presented in this chapter has been published and is reprinted with permission from:
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 Chemical Biology* **2015,** *11* (2), 470-477.

8.1 Abstract

In the active cytosine demethylation pathway, 5-methylcytosine (mC) is oxidized sequentially to 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC). Thymine DNA glycosylase (TDG) selectively excises fC and caC but not cytosine (C), mC, and hmC. We propose that the electron-withdrawing properties of -CHO and -COOH in fC and caC increase N3 acidity, leading to weakened hydrogen bonding and reduced base pair stability relative to C, mC, and hmC, thereby facilitating the selective recognition of fC and caC by TDG. Through ^{13}C NMR we measured the pK_a at N3 of fC as 2.4 and the two pK_a of caC as 2.1 and 4.2. We used isotope-edited IR spectroscopy coupled with density functional theory (DFT) calculations to site-specifically assign the more acidic pK_a of caC to protonation at N3, indicating that N3 acidity is increased in fC and caC relative to C. IR and UV melting studies of self-complementary DNA oligomers confirm reduced stability for fC-G and caC-G base pairs. Furthermore, while the fC-G base pair stability is insensitive to pH, the caC-G stability is reduced as pH decreases and the

carboxyl group is increasingly protonated. Despite suggestions that fC and caC may exist in rare tautomeric structures which form wobble GC base pairs, our two-dimensional infrared (2D IR) spectroscopy of fC and caC free nucleosides confirm that both bases are predominantly in the canonical amino-keto form. Taken together, these findings support our model that weakened base pairing ability for fC and caC in dsDNA contributes to their selective recognition by TDG.

8.2 Introduction

Thymine DNA glycosylase (TDG) is known to recognize and excise the thymine moieties from G-T mismatches in double-stranded DNA (dsDNA) by N-glycosidic bond hydrolysis, and to initiate base replacement through the DNA base-excision repair (BER) pathway.[1,2] It can also remove uracil and 5-hydroxymethyluracil (hmU) from mismatches with guanine.[3,4] This enzyme plays a central role in cellular defense against genetic mutation caused by the spontaneous deamination of cytosine (C) and 5-methylcytosine (mC), and thereby helps maintain genome integrity.[5]

Recently another major role of TDG has been recognized in epigenetic regulation through an active mC demethylation pathway.[6,7] Methylation and demethylation at the C5 position of cytosine are critical for transcriptional regulation and genome reprogramming in eukaryotes.[6-8] Unlike the well-known methylation pathway, the active demethylation pathway was poorly understood until the recent discovery of sequential oxidation steps by the ten-eleven translocation (TET) family of enzymes.[9] TET enzymes can oxidize mC to 5-hydroxymethylcytosine (hmC),[10,11] oxidize hmC to 5-formylcytosine (fC), and then oxidize fC to 5-carboxylcytosine (caC) in a stepwise manner.[6,9,12,13] TDG can excise fC and caC from dsDNA to give an abasic site both *in vitro*[14] and in mammalian cells,[15-18] and fC showed greater activity than caC under physiological

conditions.[14] The abasic site can be replaced by cytosine through downstream BER, completing the active demethylation pathway.[7] It has been shown through binding affinity studies that TDG preferentially binds fC-G and caC-G over mismatched T-G and U-G base pairs in duplex DNA despite greater excision activity towards the latter pairs. This observation indicates preferential recognition of fC and caC by TDG and suggests the need for a more detailed understanding of the properties of fC, caC, and their influence on the DNA duplex.[19] We suspect that fC-G and caC-G base pairs have weakened stability due to their modification, which contributes to their selective flipping by TDG in the genome.

Hashimoto et al. proposed that fC/caC may favor an imino tautomeric state that forces a wobble structure when paired across from G (similar to G-T and G-U mispairs) and thereby facilitates the flipping of fC/caC into the active site of TDG.[20] The observation that the TDG catalytic domain binds significantly more weakly to C, mC, and hmC than to fC and caC supports the existence of a discrimination step before stable complex formation.[19,20] Alternatively Maiti et al. provided an explanation attributing the TDG specificity to the N-glycosidic bond stability,[21] as estimated by the electronic substituent constant (σ_m) of the C5 substituent[22] or the N1 pK_a value of the pyrimidine base.[23] The observation that the TDG catalytic domain has higher activity towards G-caC base pairs at pH 5.5 compared to pH 7.5 and 8.0 is consistent with this picture of N-glycosidic bond stability as the origin of TDG activity, but that alone cannot fully account for the activity at neutral or higher pH. In addition, this study does not address whether TDG can flip C, mC, or hmC into the active site. If TDG does flip each base into the active site and selectivity is due only to N-glycosidic bond stability, then such a recognition process must be inefficient considering the ~3 billion base pair human genome.

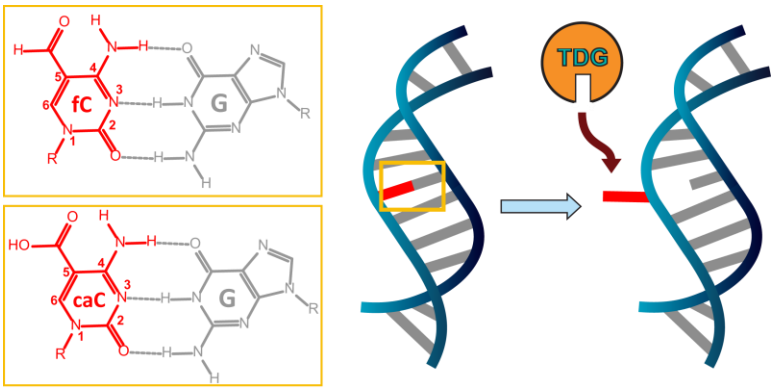

Figure 8.1: Structure of fC and caC base pairs and schematic of the extrahelical flip and recognition by TDG.

We propose that the electron-withdrawing -CHO and -COOH substituents at C5 in fC and caC not only decrease the pK_a of N1 and weaken the N-glycosidic bond, but also decrease the electron density at N3 (and thus the pK_a). This would result in weakened hydrogen bonding of the G-fC and G-caC base pair and thereby facilitate flipping of fC and caC for recognition by TDG, Fig. 8.1. To test this hypothesis we measured the N3 pK_a values of fC and caC by ^{13}C NMR and IR spectroscopy. Careful analysis of the data yields an opposite site assignment of the two caC pK_a values (N3 and COOH) with respect to the previous suggestions[23,24] and we assign the more acidic pK_a to N3. Subsequent IR and UV measurement of the stability of modified-cytosine-containing dsDNA oligomers confirmed that fC and caC oligomers are destabilized with respect to the unmodified oligomer with caC-oligomer stability being pH-dependent. These findings provide a chemical basis for distinguishing fC and caC from C, mC, and hmC in the DNA duplex that could be used for selective recognition and excision by TDG.

8.3 Results and Discussion

8.3.1 Both fC and caC Favor an Amino-Keto Tautomeric State

To test whether fC or caC could exist in the rare imino-keto tautomeric form, we used vibrational spectroscopy since different tautomers are expected to give distinct vibrational fingerprints.[25] We focused on the frequency window for in-plane base vibrations (1450-1800 cm^{-1}) which includes carbonyl stretches and ring breathing modes that mix C=C, C=N stretching, and ND_2 bending. As a first step we acquired temperature-dependent Fourier transform infrared (FTIR) spectra since the coexistence of multiple tautomers can result in spectral shifts and isosbestic points depending on their equilibrium thermodynamic properties.[26] Both fC and caC (Fig. 8.2a,b) exhibit minimal changes under physiological conditions, suggesting that only one tautomeric form is predominant.

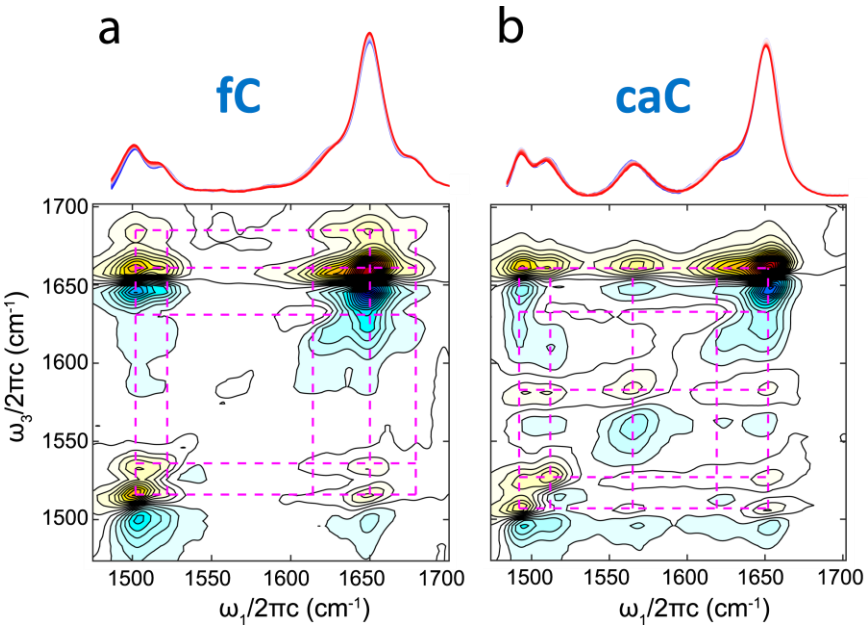

Figure 8.2: Temperature dependent FTIR spectra of (a) fC and (b) caC, pD 7.3, ranging from 10-95 °C (blue-red). 2D IR spectra with ZZYY polarization of fC and caC are aligned beneath the temperature ramp spectra.

Comparing the experimental FTIR spectra of fC and caC against the known amino-keto spectrum of canonical 2'-deoxycytidine (Fig. 8.5a), similarities such as the intense carbonyl mode at ~1650 cm^{-1} and the peak pattern and intensity of the two ring-modes at ~1500 cm^{-1} suggest that the cytidine analogs are likewise amino-keto tautomers. To be more definitive in this assignment, we used DFT to calculate an IR spectrum for each of the possible tautomers of fC and caC. The fC calculations were performed in the gas phase while the caC calculations included three explicit D_2O molecules. The tautomers included were the amino-keto (KA), imino-keto (KI), and imino-enol (EI) tautomers. Both the cis and trans isomers of the imino forms were considered. Calculated spectra were determined using Gaussian 09[27] and are shown in Fig. 8.3. For both fC and caC, only the KA spectrum reproduces the two low frequency ring-modes at ~1500 cm^{-1}. In addition, the KA spectrum best reproduces the peak pattern of the remaining modes. Overall the amino-keto spectrum best matches the experimentally measured spectrum for both nucleosides, and we assign the predominant tautomer to be amino-keto.

2D IR spectra of the fC and caC free nucleosides provide direct evidence the amino-keto tautomer is the only appreciable form. Ultrafast 2D IR spectroscopy reports on the coupling between molecular vibrations. By correlating excitation (ω_1) and detection (ω_3) frequencies, mixtures of tautomers can be separately resolved before they exchange through the distinct cross peak patterns unique to each tautomer. Previous studies have shown that for a single tautomer of a nucleobase or nucleobase analog, cross peaks exist between all of the in-plane base vibrations due to the delocalization of these modes.[28,29] This is also the case for both fC and caC, as seen in their 2D IR spectra plotted in Fig. 8.2a, b, respectively. The diagonal peaks in the 2D spectrum mirror the peaks in the linear FTIR spectrum, each consisting of an oppositely signed doublet (red above blue).

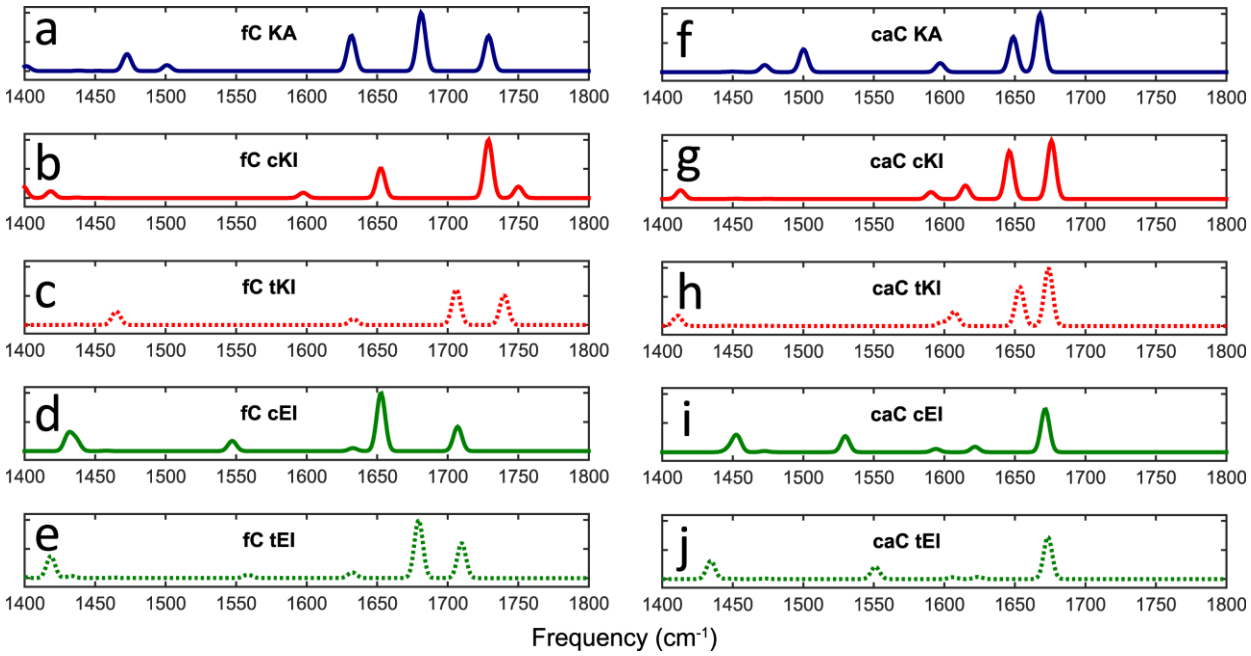

Figure 8.3: DFT calculated spectra for the tautomers of fC and caC, including the amino-keto (KA), cis imino-keto (cKI), trans imino-keto (tKI), cis imino-enol (cEI), and trans imino-enol (tEI), (a-e) for fC and (f-j) for caC. Calculations were run in Gaussian 09 using the B3LYP hybrid functional with the 6-31G(d,p) basis set.

The gridlines help to illustrate that cross-peaks are observed between all the diagonal peaks, indicating the presence of a single dominant tautomer species for both nucleosides. In the event that multiple tautomers were present, one would expect to see multiple overlapping grid patterns lacking cross-peaks to one another.[30] We have also considered the possibility of tautomerism in singly protonated caC, but we find no evidence for tautomers other than the dominant amino-keto species. Together the temperature-dependent FTIR and 2D IR spectra provide direct experimental evidence arguing against the presence of multiple fC and caC tautomers under physiological conditions. This result is consistent with computational predictions that the amino-keto tautomer of fC and caC is the most stable species.[23]

8.3.2 Measurement of N3 pK_as by ^{13}C NMR

The extent of hydrogen bond weakening due to the -CHO and -COOH substituents can be correlated with changes in the pK_a at the N3 site. If our hypothesis is correct, both fC and caC should demonstrate increased N3 acidity. In the past, these pK_as have been determined by pH dependent UV spectra,[23,24] but site-specific assignment is difficult since the carboxyl group of caC complicates the investigation by introducing a second pK_a not present in the other cytosine derivatives. We reassessed the pK_a values of hmC, fC, and caC by recording the ^{13}C NMR spectra of the corresponding ^{13}C-labeled free nucleosides as a function of pH. The label was inserted at the exocyclic carbon atom connected to C5 of cytosine.

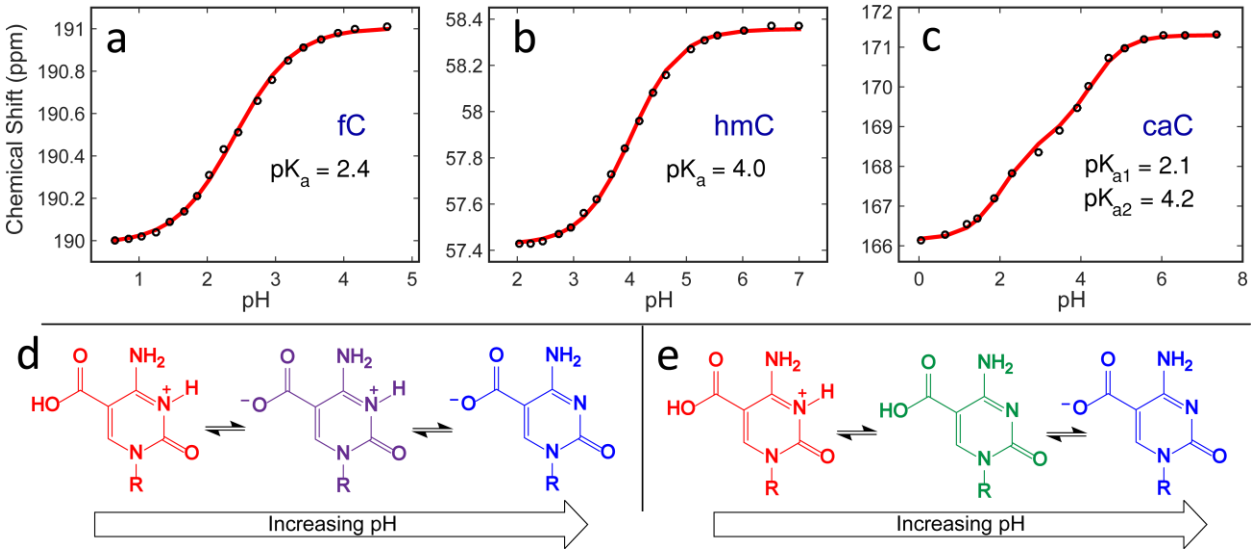

Figure 8.4: Chemical shift vs pH titration profiles obtained from ^{13}C NMR measurements of ^{13}C-labeled fC, hmC, and caC (a, b, and c, respectively). Possible neutral species for caC include the Zwitterionic species protonated only at N3 (purple, d) or the species protonated at the carboxyl group (green, e).

We recorded ^{13}C NMR spectra in the pH range 0.5 to 8 and tracked the chemical shift of the labeled carbon for the nucleosides fC, hmC, and caC. For both fC and hmC, plotting chemical

shift vs pH results in a single-transition titration curve that is readily fit to the Henderson-Hasselbalch equation, yielding a pK_a value of 2.4 for fC and 4.0 for hmC (Fig. 8.4a,b). These pK_a values are comparable to those obtained by UV measurements[23] and indicate that the more electron-withdrawing formyl substituent in fC lowers the N3 pK_a significantly in contrast to C and hmC, consistent with our reasoning.

For caC the chemical shift vs pH curve results in two transitions with pK_a values 4.2 and 2.1. Although these values are similar to the pK_as measured by UV,[23] it is difficult to conclusively assign which pK_a corresponds to N3 because there are two possible neutral species of caC depending on which site protonates first (Fig. 8.4d,e). Since the carboxylic proton is much closer to the ^{13}C-labeled carbon than the N3 proton, we expect the greater change in chemical shift to be associated with the carboxylic proton. We found that the change in chemical shift around pH 4.2 (~3 ppm) is greater than the shift around pH 2.1 (~2 ppm), suggesting that the pK_a of 4.2 should be assigned to the carboxylic group while the pK_a of 2.1 should be assigned to N3. These assignments, however, are not definitive and are opposite of previous assignments in the literature.[23,24]

8.3.3 Determination and Site-Assignment of the pK_as of 5-Formylcytidine and 5-Carboxylcytidine by FTIR Spectroscopy

To independently examine these conclusions, we measured the pK_a values of fC and caC through pH-dependent FTIR spectroscopy. Because a mixture of protonated and deprotonated species exists at each pH point, we employ singular value decomposition (SVD) analysis and the maximum entropy method described in Chapter 4 to disentangle the pH-dependent spectra and reconstruct pure component spectra that individually represent each of the contributing species. These reconstructed spectra can then be compared directly against DFT calculations to assign the

structure of each protonation state and the resulting titration curves can be fit to the Henderson-Hasselbalch equation to determine pK_as. As a control on this method we assigned the pK_a of 2'-deoxycytidine to be 4.5 and the pK_a of fC to be 2.4 (Fig. 8.5), consistent with previous reports.[31,32]

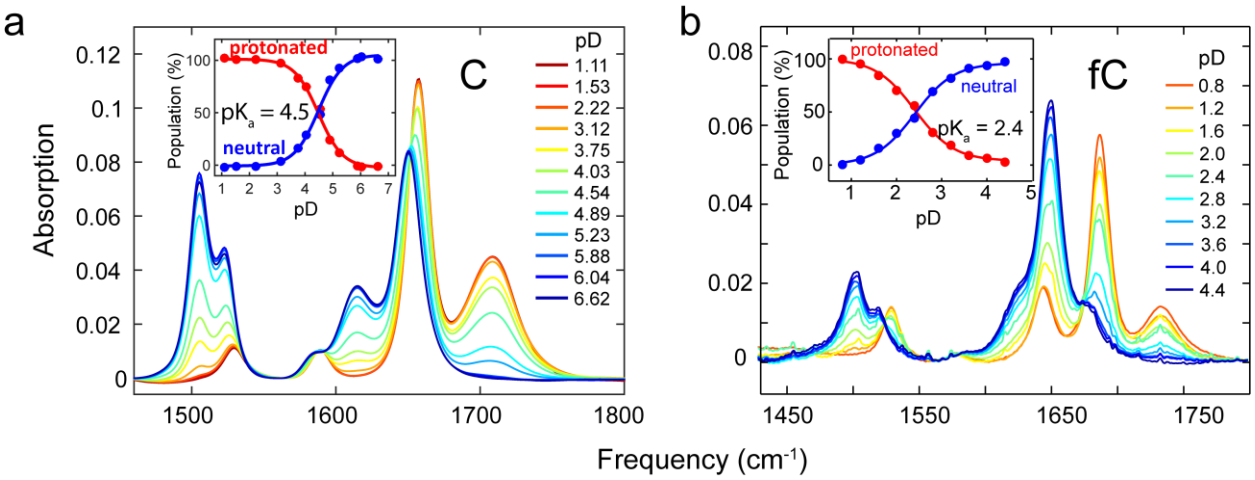

Figure 8.5: (a) pH-dependent FTIR spectra of 2′-dC at 3mg/mL and at 24 °C. The inset shows the titration curves for the protonated and neutral 2′-dC species obtained from the SVD analysis. (b) The corresponding pH-dependent FTIR spectra of fC.

At pD 4.4 the two highest frequency modes of fC at 1680 cm^{-1} and 1651 cm^{-1} (Fig. 8.5b) are assigned based on DFT calculations to the coupled C=O stretching modes of the carbonyl group on the cytosine ring and the formyl group. We report the pD since deuterated water is used as the solvent to improve sample transmission. The 1680 cm^{-1} mode carries more formyl C=O character while the 1651 cm^{-1} mode displays a greater contribution from the cytosine C2=O. As the pD decreases, two carbonyl stretches grow in at higher frequency at 1735 cm^{-1} and 1688 cm^{-1}. A similar blue shift is observed for the C=O stretch of cytidine when protonated at N3, as seen in Fig. 8.5a. Furthermore the two shoulder peaks around the 1651 cm^{-1} mode are separated and better resolved with decreasing pD, consistent with the DFT calculated spectrum for N3 protonated fC

(Fig. 8.6). Therefore we assign the pH-dependent spectral changes to be the result of protonation at N3. Close correlation of the changes in experimental spectra with the changes in DFT calculated spectra between fC and fC^+ provide a definitive assignment of the pK_a of 2.4 to N3.

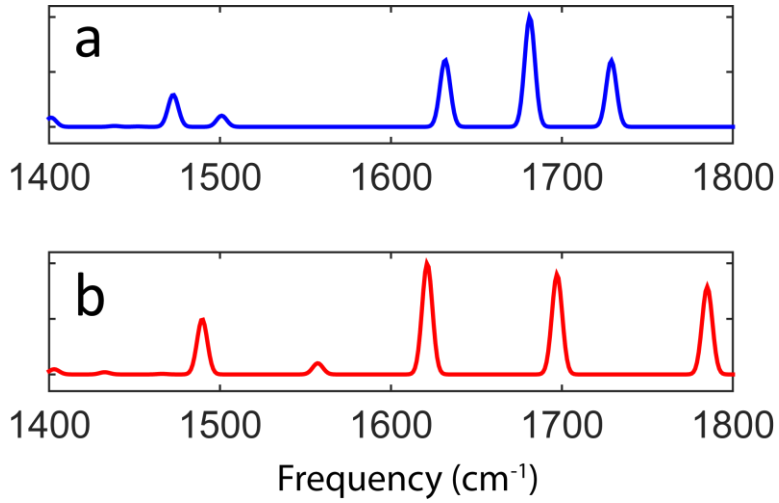

Figure 8.6: DFT calculated spectra for (a) N3 deprotonated and (b) N3 protonated fC. Calculations were run in Gaussian 09 using the B3LYP hybrid functional with the 6-31G(d,p) basis set.

Turning to caC, one is faced with the more complicated problem of site-specific assignment of multiple pK_as. As a result we adopted an isotope labeling strategy similar to the ^{13}C NMR experiments in which a ^{13}C isotope label was inserted at the exocyclic carbon atom connected to C5 of cytosine. The pD-dependent FTIR spectra for unlabeled (UL) caC and ^{13}C-labeled caC are presented in Fig. 8.7a and b, respectively. At pD 7.4 the two coupled C=O stretches of UL caC give rise to the main carbonyl mode at 1655 cm^{-1} and a weaker band at 1567 cm^{-1}, as assigned by DFT. In general the spectra of ^{13}C labeled caC are similar to UL caC, except that the 1567 cm^{-1} carbonyl peak red shifts to 1540 cm^{-1}, indicating that this mode has significant contribution from the labeled carboxyl group.

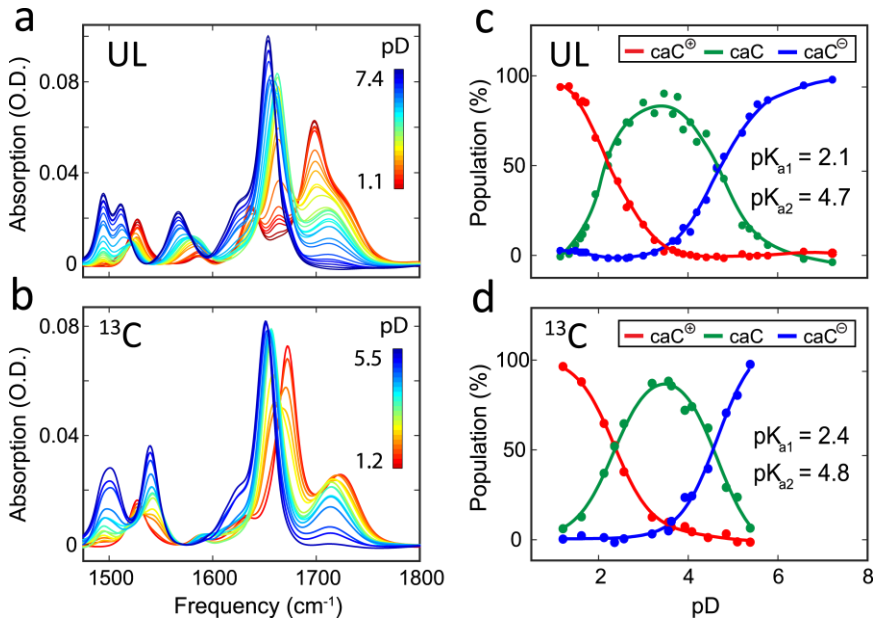

Figure 8.7: pD-dependent FTIR spectra of (a) unlabeled and (b) ^{13}C labeled (exocyclic carbonyl carbon) caC. (c,d) Titration curves of caC cation (red), neutral (green) and anion (blue) species derived from the SVD analysis.

We performed SVD analysis and reconstruction of pure component spectra corresponding to the cationic, neutral, and anionic caC species for both unlabeled and ^{13}C labeled caC. The reconstructed spectra are plotted in Fig. 8.8a-c, with the UL caC and the ^{13}C labeled caC represented by solid and dashed lines, respectively. The corresponding population fractions for the three species as a function of pD are plotted in Fig. 8.7c and d. Through this analysis the two pK_a values of caC were determined to be 2.1 and 4.7 from the UL caC spectra and, in reasonable agreement, 2.4 and 4.8 from the ^{13}C-labeled caC spectra.

To assign the molecular origin of the two pK_a values we compared the experimental spectra (Fig. 8.8a-c) with DFT calculated spectra (Fig. 8.8d-g) for both UL and ^{13}C labeled caC. The pink arrows in Fig. 8.8 highlight frequency shifts upon isotopic labeling while the orange bars indicate peaks that are unaffected by the label. In the calculations, caC molecules with –1, 0, and +1 charges were solvated by three explicit water molecules near the hydrogen bond donor/acceptor sites. Two

different isomers of neutral caC were considered: one protonated at the carboxyl group (Fig. 8.8e, green) and another protonated at the N3 atom (Fig. 8.8g, purple).

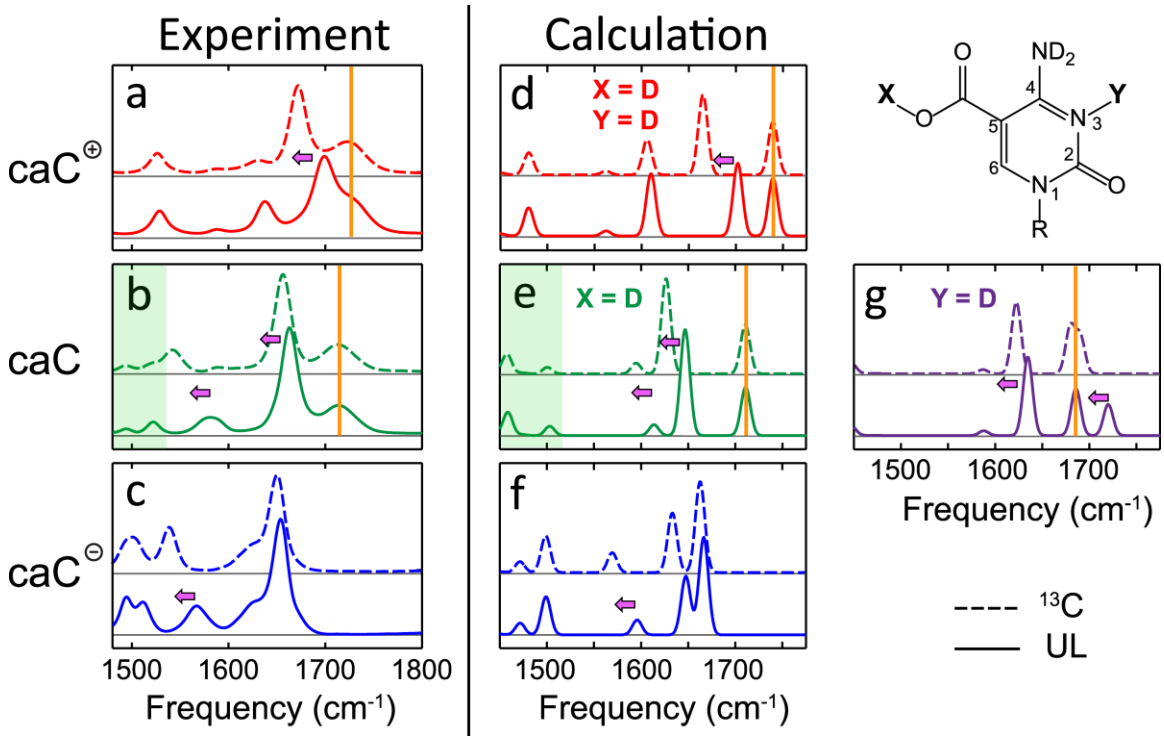

Figure 8.8: Comparison between the experimental (left) and DFT calculated (right) spectra for caC cation (red), neutral (green/purple) and anion (blue) species. Both unlabeled (solid lines) and ^{13}C labeled (dashed lines) caC spectra are shown. Pink arrows highlight frequency shifts upon isotopic labeling while orange bars highlight frequencies that are unaffected by the label.

As a check on the validity of our DFT calculated spectra, we first compared the cationic and anionic experimental spectra (Fig. 8.8a,c) against their calculated spectra (Fig. 8.8d,f). Since these species correspond to either complete protonation or deprotonation of the nucleobase, there is no ambiguity in molecular structure. Overall we find a close match in the peak pattern, peak intensities, and ^{13}C isotope shift between the experimental and calculated spectra for both the caC

cation and anion. This provides strong support for the use of DFT calculations to assign these vibrational spectra, and therefore we turn to assigning the neutral caC species with pK_a 4.7.

The neutral species of caC can be protonated at one of two sites: either the carboxyl group or the N3 of cytosine. As seen in Fig. 8.8e and g, DFT calculations predict distinct spectra for these two possible structures. However the spectrum calculated for the isomer with a protonated exocyclic carboxyl group (Fig. 8.8e) best reproduces the experimental spectra, displaying a similar C=O peak pattern and the presence of low frequency ring modes between 1450-1550 cm^{-1} (highlighted by the green shading in Fig. 8.8). The ^{13}C labeled spectrum for neutral caC (Fig. 8.8b, dashed line) demonstrates that upon isotope labeling the 1713 cm^{-1} peak does not shift but the 1657 cm^{-1} peak red shifts, indicating that these peaks involve mostly C2=O and carboxyl C=O character, respectively. This isotope-induced frequency shift is in excellent agreement with the calculated spectra for the neutral caC molecule protonated at the carboxyl group (Fig. 8.8e). In contrast, the calculated spectra for the neutral caC molecule protonated at N3 (Fig. 8.8g) predicts that the highest frequency mode is mostly carboxyl C=O stretch (seen to red shift upon ^{13}C labeling), but this pattern does not match the experimental observation. Moreover, the lower frequency delocalized ring vibrations around 1500 cm^{-1} are not reproduced for the N3-protonated structure. In light of these results we assign the pK_a of 4.7 to the carboxyl group and the pK_a of 2.1 to the N3 position.

Our assignment of the 2.1 pK_a of caC to N3 is opposite of previous assignments that were based on the similar isosbestic points between the UV spectra of 2′-deoxycytidine and caC and chemical analogies to other aromatic compounds possessing a carboxyl group with a vicinal amine.[23,24] Our assignment supports the hypothesis that the electron-withdrawing substituent -COOH lowers the pK_a of N3 and destabilize G-caC base pairs.

8.3.4 Stability of DNA Duplexes Containing 5-Formlycytidine and 5-Carboxylcytidine

In order to further test that both G-fC and G-caC base pairs form less stable hydrogen bonds than the canonical G-C base pair, we studied the thermal stability of dsDNA oligonucleotides containing different cytosine modifications using IR and UV spectroscopy. To accentuate the difference in melting temperature (T_m) we used a self-complementary dsDNA oligomer containing six G-X base pairs with sequence 5'-TAXGXGXGTA-3', where X denotes C, mC, hmC, fC, or caC. Temperature-dependent FTIR spectra measured at pD 7.3 (Fig. 8.9) were analyzed using SVD and the resulting melting curves were fit to the two-state model described in Chapter 1. The analogous UV measurements were also collected, but a single frequency intensity at 260 nm was tracked as a function of temperature and this trace was fit to the same two-state model. Melting temperatures for the set of dsDNA are listed in Table 8.1. The ~10 °C difference in T_m's measured by the two techniques is explained by the oligomer concentration difference between the two methods (1000 μM for IR vs 4 μM for UV). Fig. 8.10 shows the melting curves fit to each data set as well as a comparison of the melting temperature trend measured by each technique. The oligomer where X = hmC is omitted for clarity, as the T_m of this oligomer is equal to the T_m for X = C.

The spectral region 1610-1720 cm^{-1} contains mostly carbonyl stretches and the ring modes of A and T at 1625 cm^{-1} and 1630 cm^{-1}, respectively. Since our DNA sequence is 60% GC content the main spectral variation with temperature stems from the carbonyl stretches of G and C. At low temperature, the guanine C=O stretch blue shifts significantly from 1662 cm^{-1} to 1684 cm^{-1} while the cytosine C=O stretch stays at 1651 cm^{-1}. These spectral features are consistent with literature reports on IR absorption frequencies of DNA double helices.[33] Furthermore the splitting of these two carbonyl peaks is due to the vibrational coupling of these modes upon duplex formation.[34,35]

As the temperature increases and the duplex melts the guanine C=O red shifts back to 1662 cm^{-1} and overlaps with the cytosine C=O, resulting in a broad absorption band.

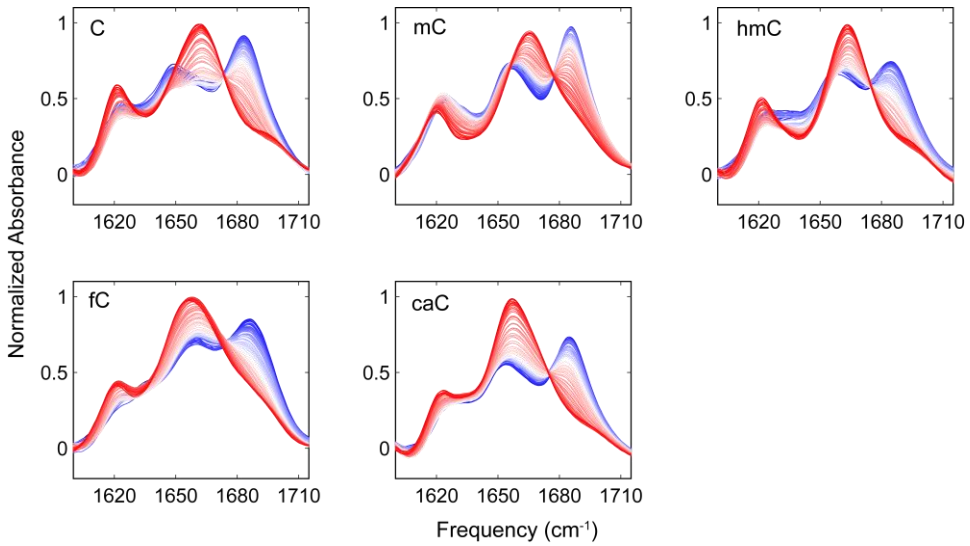

Figure 8.9: Temperature dependent FTIR for the five oligonucleotides (sequence 5'-TAXGXGXGTA-3', where X denotes a cytosine analog). The temperature was ramped between 10 °C in dark blue to 95 °C in dark red.

Since the cytosine analogs have similar absorption spectra in the frequency range of interest the oligonucleotide spectra containing modified cytosines appear quite similar (Fig. 8.9). At low temperatures a blue-shifted guanine C=O stretch is observed for all five oligomers, indicating the formation of duplex DNA. As the temperature increases all spectra exhibit features consistent with duplex melting: loss of the ~1680 cm^{-1} mode and gain of the ~1660 cm^{-1} mode. A closer examination reveals subtle differences between the data sets. For example at low temperature the carbonyl peak splitting for the fC oligomer is smaller than the splitting observed for the 2′-deoxycytidine (C) unmodified oligomer, with the lower frequency C=O mode measured at ~1660 cm^{-1} for fC instead of at ~1650 cm^{-1}. Moreover the higher frequency C=O mode of the fC oligo at 1686 cm^{-1} is broader than that of the other oligos. Both observations suggest that the 5fC

oligomer duplex is not as stable, as weaker base-pairing would lead to smaller vibrational splitting and the broader C=O peaks imply greater solvent fluctuation around the carbonyl groups, consistent with a destabilized duplex.

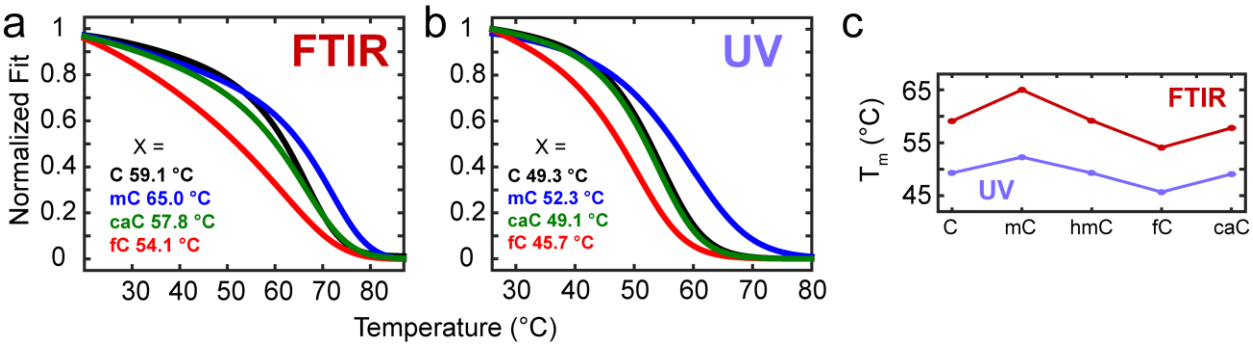

Figure 8.10: Melting curves obtained at physiological pH and fit to a two-state model to determine T_m for each of the oligonucleotides (a) obtained from the 2nd SVD component of the FTIR spectra and (b) tracking the UV intensity at 260 nm. (c) Relative T_m trends between the methods.

Consistent with our hypothesis of weakened N3 hydrogen bonding, we find the fC oligomer to be less stable than the unmodified oligomer, having a significant 5 °C and 3 °C decrease in T_m from IR and UV measurements, respectively. Once again the story surrounding caC proves more complicated. Our experiments show that the caC oligomer has an equal (UV) or slightly lowered (IR) T_m compared to the unmodified oligomer at neutral pH, but in light of our caC pK_a assignments one would expect that the protonation state of the carboxyl group could influence the properties of the base pair. This possibility is explored in detail below with pH-dependent melting studies.

Currently no clear consensus exists in the literature regarding the influence of naturally occurring cytosine derivatives on the stability of dsDNA. Our finding of nearly equal stability for unmodified and caC-containing oligonucleotides at neutral pH is consistent with several past

reports,[36,37] but it is possible that discrepancies with reports of increased stability[24] could be due to the different sequences and experimental conditions employed by each study. This raises an interesting question regarding the sequence context of the modification. In this study as well as Raiber et al.'s, we include the modification in symmetrically modified CpG domains which have been proposed to be the most biologically relevant context.[38] Münzel et al.[37] considered a sequence with a CpG domain, but with only one modified base, while Sumino et al.[24] did not include the modifications in a CpG domain.

Table 8.1: Summary of past reports of the influence of modified cytosines on the T_m of DNA oligonucleotides compared to the present report.

	C	**mC**	**hmC**	**fC**	**caC**
Our IR	59	65	59	54	58
Our UV	49	52	49	46	49
Sumino et al.[24]	49	-	-	-	51
Raiber et al.[36]	53.5	60	58	54.5	53
Münzel et al.[37]	51.4		51.0	52.7	51.3
Thalhammer et al.[39]	59	65	61	-	-

Our analysis suggests that the mC oligomer is more stable than the unmodified oligomer, with a T_m that is 6 °C and 3 °C higher in the IR and UV measurements, respectively. These results agree with the UV melting measurement by Thalhammer et al. where a 6 °C stabilizing effect for mC was reported for the same DNA sequence[39] and with reports of the stabilizing influence of mC from Raiber et al.[36] We find the hmC oligomer has an equal T_m to the unmodified sequence, while previous findings suggest nearly equal or slightly increased stability.[36,37,39] With regard to the influence of caC on duplex stability, both Raiber et al.[36] and Münzel et al.[37] reported equal T_m for

oligomers containing caC or C, while Sumino et al.[24] found that caC has a ~2 °C stabilizing effect. Our experiments suggest that the caC oligomer has an equal (UV) or slightly lowered (IR) T_m compared to the unmodified oligomer, which is more consistent with the results obtained by Raiber et al.[36] and Münzel et al.[37] In contrast, the fC oligomer is less stable than the unmodified oligomer, having a significant 5 °C and 3 °C decrease in T_m from IR and UV measurements, respectively. This result is inconsistent with previous reports that fC-containing oligomers show ~1 °C increased stability.[36,37]

Many factors, such as differing DNA sequences or buffer and sample concentrations may influence the T_m of DNA oligonucleotides. In addition it is important to note how T_m is defined and determined in each experiment. Defining T_m as the inflection point of the melting curve by considering the maximum of the first derivative or zero of the second derivative can lead to discrepancies in T_m, especially when the melting curve deviates from a sigmoid symmetric about the inflection point. An analytical report concluded that the most consistent and meaningful approach, assuming two state behavior, defines T_m as the temperature at which half of the base pairs are dissociated,[40] which is the approach we have adopted here. In Chapter 9, this question is revisited in detail beyond the two state assumption, but it is helpful to outline the ongoing discrepancies here.

8.3.5 Evaluating the pH Dependence of T_m for the X = caC Sequence

It has been reported that the excision of caC by TDG is acid catalyzed while the excision of fC is pH independent.[23] To determine whether the pH dependence in excision rate is correlated with pH dependent stability of the fC and caC oligomers, we repeated the infrared T_m determination for these duplexes in a pD 3.7 solution prepared at identical salt and buffer concentration as the

previous measurements. For the fC oligomer, we observed no pD dependence for the T_m, while we observed a 7 °C drop in T_m for the caC oligomer (Fig. 8.11). The destabilization of the caC oligomer relative to the fC oligomer is likely due to the influence of protonation at the carboxyl group of caC. Therefore fC, with only the N3 site to protonate, displays no pD dependence. In general, for caC, one would expect that the increased positive charge due to protonation at the carboxyl group would lower the pK_a at N3 and destabilize the base pair, consistent with the observed reduction in T_m at decreased pD.

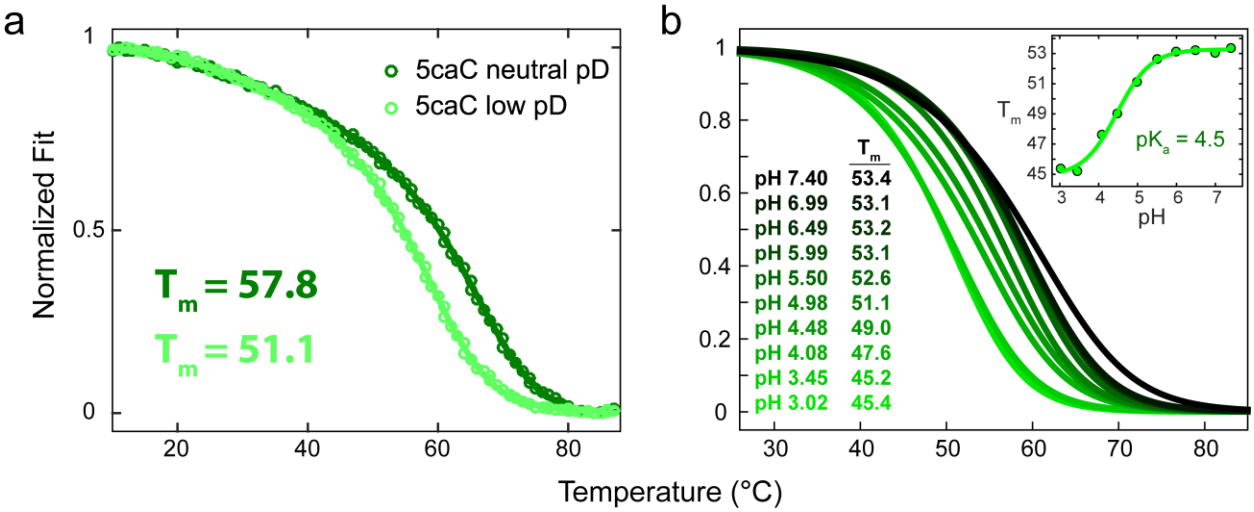

Figure 8.11: (a) FTIR melting curves for the X = caC oligonucleotide at pD 7.2 and pD 3.7 (b) UV melting curves of the 5caC oligomer as a function of pH and inset, the T_m vs pH trace fit to the Henderson-Hasselbalch equation revealing a consistent pK_a of 4.5.

To further evaluate the pH dependence of the X = caC oligomer's T_m, we carried out a series of UV melting experiments as a function of pH. The oligomer concentration and salt concentration were the same as the previous UV experiments (Fig. 10b), however, to increase the buffer capacity the buffer concentration was increased from 10 to 100 mM. The change in sodium

cation concentration accounts for the ~4 °C difference in T_m for the caC oligomer measured at similar pH above. Fig. 8.11 shows the pH dependence of the X = caC oligomer's melting curve and melting temperature. As seen in the inset, the T_m plateaus around 53 °C above pH = 6, decreases with decreasing pH, and then plateaus around 45 °C below pH = 3.5. Fitting this profile to the Henderson-Hasselbalch equation results in a pK_a of 4.5, consistent with the pK_a measured for the carboxyl group of the caC free nucleoside. These observations support a picture in which increasing protonation of the carboxyl group of caC (pK_a = 4.7) within the duplex weakens the caC-G base pairs, accounting for the behavior of the T_m with decreasing pH. These findings suggest that previous reports of acid catalyzed excision of caC could be explained in part by the influence of increasing protonation of the caC nucleobase at the exocyclic carboxyl group leading to a weakening of the caC-G base pair.

8.4 Conclusion

Our studies have revealed two observations that have direct consequences for the mechanism of base recognition by TDG. First, we assign the lower pK_a of caC to N3 instead of the carboxyl group based on direct site-specific assignment of the pK_a values through IR spectroscopy measurement and DFT calculations. Second, using two different techniques we provide a complete data set reporting the influence of the naturally occurring cytosine modifications on dsDNA stability in order to provide a robust survey of the stability trend. Specifically, we find that at neutral pH the T_m of a caC-containing oligomer is not significantly different from the analogous C-containing oligomer while the T_m of a fC-containing oligomer is significantly lower. Furthermore, we measured the T_m of the fC-containing oligomer to be pH-independent while we observed the T_m of the caC-containing oligomer to drop below that of

the fC-containing oligomer with decreasing pH as the carboxyl groups are increasingly protonated. This influence can explain the pH dependence of both the T_m for caC-containing oligomers as well as the limited TDG activity towards caC at physiological pH, since some small percentage of the carboxyl groups will be transiently protonated and thus capable of flipping into the active site of TDG. These results demonstrate that an electron-withdrawing substituent at C5 decreases the electron density at N3 (and thus the pK_a) such that the hydrogen bonding capacity of the base is weakened. Furthermore, we believe that weakened base-pairing facilitates extrahelical flipping of the modified base for recognition and excision by TDG. Our proposal can explain the previous finding that the excision of caC is acid catalyzed, with caC serving as an even more effective TDG substrate than fC at low pH.[23] It should be noted that the electron-withdrawing properties of the 5-formyl and 5-carboxyl groups may also affect base stacking and hydrophobicity due to a shift in the electronic distribution on the base and the observed destabilization is likely not due solely to weakened N3 hydrogen bonding, but also to these more global changes.

In addition to new insights regarding the effect of N3 acidity, we emphasize the importance of past findings by Maiti et al. regarding the influence of the formyl and carboxyl groups on glycosidic bond stability as well as critical interactions between fC and caC with the enzyme. We believe our new insight regarding the nucleobases and DNA duplex are complementary with these past results.[23] Maiti et al. reported an apparent pK_a of 5.75 for caC when bound in the enzyme-substrate complex, but they assign this pK_a to protonation at N3. In light of our IR/DFT analysis, we believe this apparent pK_a corresponds to protonation at the carboxyl group of caC. In the presence of the enzyme this elevated apparent pK_a would allow for more protonation of the carboxyl group and can further help explain the limited TDG activity towards caC under neutral conditions.

8.5 Acknowledgements

I thank Qing Dai for helpful discussions, synthesizing the modified cytosine samples, and taking the ^{13}C NMR and UV measurements. I thank Sam Peng for helpful discussions and for his mentorship at the early stages of this project.

8.6 References

1. Lindahl, T., Instability and decay of the primary structure of DNA. *nature* **1993,** *362* (6422), 709-715.

2. Stivers, J. T.; Jiang, Y. L., A mechanistic perspective on the chemistry of DNA repair glycosylases. *Chemical reviews* **2003,** *103* (7), 2729-2760.

3. Morgan, M. T.; Bennett, M. T.; Drohat, A. C., Excision of 5-Halogenated Uracils by Human Thymine DNA Glycosylase: Robust Activity for DNA Contexts other than CpG. *Journal of Biological Chemistry* **2007,** *282* (38), 27578-27586.

4. Hitomi, K.; Iwai, S.; Tainer, J. A., The intricate structural chemistry of base excision repair machinery: implications for DNA damage recognition, removal, and repair. *DNA repair* **2007,** *6* (4), 410-428.

5. Wiebauer, K.; Jiricny, J., Mismatch-specific thymine DNA glycosylase and DNA polymerase beta mediate the correction of GT mispairs in nuclear extracts from human cells. *Proceedings of the National Academy of Sciences* **1990,** *87* (15), 5842-5845.

6. Bhutani, N.; Burns, D. M.; Blau, H. M., DNA demethylation dynamics. *Cell* **2011,** *146* (6), 866-872.

7. He, Y.-F.; Li, B.-Z.; Li, Z.; Liu, P.; Wang, Y.; Tang, Q.; Ding, J.; Jia, Y.; Chen, Z.; Li, L., Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **2011,** *333* (6047), 1303-1307.

8. Klose, R. J.; Bird, A. P., Genomic DNA methylation: the mark and its mediators. *Trends in biochemical sciences* **2006,** *31* (2), 89-97.

9. Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y., Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **2011,** *333* (6047), 1300-1303.

10. Tahiliani, M.; Koh, K. P.; Shen, Y.; Pastor, W. A.; Bandukwala, H.; Brudno, Y.; Agarwal, S.; Iyer, L. M.; Liu, D. R.; Aravind, L., Conversion of 5-methylcytosine to 5-

hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **2009,** *324* (5929), 930-935.

11. Ito, S.; D'Alessio, A. C.; Taranova, O. V.; Hong, K.; Sowers, L. C.; Zhang, Y., Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **2010,** *466* (7310), 1129-1133.

12. Pfaffeneder, T.; Hackner, B.; Truß, M.; Münzel, M.; Müller, M.; Deiml, C. A.; Hagemeier, C.; Carell, T., The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angewandte Chemie International Edition* **2011,** *50* (31), 7008-7012.

13. Wu, S. C.; Zhang, Y., Active DNA demethylation: many roads lead to Rome. *Nature reviews Molecular cell biology* **2010,** *11* (9), 607-620.

14. Maiti, A.; Drohat, A. C., Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of CpG sites. *Journal of Biological Chemistry* **2011,** *286* (41), 35334-35338.

15. Song, C.-X.; Szulwach, K. E.; Dai, Q.; Fu, Y.; Mao, S.-Q.; Lin, L.; Street, C.; Li, Y.; Poidevin, M.; Wu, H., Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **2013,** *153* (3), 678-691.

16. Shen, L.; Wu, H.; Diep, D.; Yamaguchi, S.; D'Alessio, A. C.; Fung, H.-L.; Zhang, K.; Zhang, Y., Genome-wide analysis reveals TET-and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **2013,** *153* (3), 692-706.

17. Nabel, C. S.; Jia, H.; Ye, Y.; Shen, L.; Goldschmidt, H. L.; Stivers, J. T.; Zhang, Y.; Kohli, R. M., AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nature chemical biology* **2012,** *8* (9), 751-758.

18. Raiber, E.-A.; Beraldi, D.; Ficz, G.; Burgess, H. E.; Branco, M. R.; Murat, P.; Oxley, D.; Booth, M. J.; Reik, W.; Balasubramanian, S., Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome biology* **2012,** *13* (8), R69.

19. Zhang, L.; Lu, X.; Lu, J.; Liang, H.; Dai, Q.; Xu, G.-L.; Luo, C.; Jiang, H.; He, C., Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nature chemical biology* **2012,** *8* (4), 328-330.

20. Hashimoto, H.; Hong, S.; Bhagwat, A. S.; Zhang, X.; Cheng, X., Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic acids research* **2012,** *40* (20), 10203-10214.

21. Bennett, M. T.; Rodgers, M.; Hebert, A. S.; Ruslander, L. E.; Eisele, L.; Drohat, A. C., Specificity of human thymine DNA glycosylase depends on N-glycosidic bond stability. *Journal of the American Chemical Society* **2006,** *128* (38), 12510-12519.

22. Hansch, C.; Leo, A.; Unger, S. H.; Kim, K. H.; Nikaitani, D.; Lien, E. J., Aromatic substituent constants for structure-activity correlations. *Journal of medicinal chemistry* **1973,** *16* (11), 1207-1216.

23. Maiti, A.; Michelson, A. Z.; Armwood, C. J.; Lee, J. K.; Drohat, A. C., Divergent mechanisms for enzymatic excision of 5-formylcytosine and 5-carboxylcytosine from DNA. *Journal of the American Chemical Society* **2013,** *135* (42), 15813-15822.

24. Sumino, M.; Ohkubo, A.; Taguchi, H.; Seio, K.; Sekine, M., Synthesis and properties of oligodeoxynucleotides containing 5-carboxy-2′-deoxycytidines. *Bioorganic & medicinal chemistry letters* **2008,** *18* (1), 274-277.

25. Miles, H. T., Tautomeric forms in a polynucleotide helix and their bearing on the structure of DNA. *Proceedings of the National Academy of Sciences* **1961,** *47* (6), 791-802.

26. Peng, C. S.; Baiz, C. R.; Tokmakoff, A., Direct observation of ground-state lactam–lactim tautomerization using temperature-jump transient 2D IR spectroscopy. *Proceedings of the National Academy of Sciences* **2013,** *110* (23), 9243-9248.

27. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 09 Rev. A.02,* Wallingford, CT, 2009.

28. Peng, C. S.; Tokmakoff, A., Identification of lactam–lactim tautomers of aromatic heterocycles in aqueous solution using 2D IR spectroscopy. *The journal of physical chemistry letters* **2012,** *3* (22), 3302-3306.

29. Peng, C. S.; Jones, K. C.; Tokmakoff, A., Anharmonic vibrational modes of nucleic acid bases revealed by 2D IR spectroscopy. *Journal of the American Chemical Society* **2011,** *133* (39), 15650-15660.

30. Peng, C. S.; Fedeles, B. I.; Singh, V.; Li, D.; Amariuta, T.; Essigmann, J. M.; Tokmakoff, A., Two-dimensional IR spectroscopy of the anti-HIV agent KP1212 reveals protonated and neutral tautomers that influence pH-dependent mutagenicity. *Proceedings of the National Academy of Sciences* **2015,** *112* (11), 3229-3234.

31. La Francois, C. J.; Jang, Y. H.; Cagin, T.; Goddard, W. A.; Sowers, L. C., Conformation and proton configuration of pyrimidine deoxynucleoside oxidation damage products in water. *Chemical research in toxicology* **2000,** *13* (6), 462-470.

32. Karino, N.; Ueno, Y.; Matsuda, A., Synthesis and properties of oligonucleotides containing 5-formyl-2′-deoxycytidine: in vitro DNA polymerase reactions on DNA templates containing 5-formyl-2′-deoxycytidine. *Nucleic acids research* **2001,** *29* (12), 2456-2463.

33. Banyay, M.; Sarkar, M.; Gräslund, A., A library of IR bands of nucleic acids in solution. *Biophysical chemistry* **2003,** *104* (2), 477-488.

34. Krummel, A. T.; Mukherjee, P.; Zanni, M. T., Inter and intrastrand vibrational coupling in DNA studied with heterodyned 2D-IR spectroscopy. *The Journal of Physical Chemistry B* **2003,** *107* (35), 9165-9169.

35. Lee, C.; Cho, M., Vibrational dynamics of DNA. II. Deuterium exchange effects and simulated IR absorption spectra. *The Journal of chemical physics* **2006,** *125* (11), 114509.

36. Raiber, E.-A.; Murat, P.; Chirgadze, D. Y.; Beraldi, D.; Luisi, B. F.; Balasubramanian, S., 5-Formylcytosine alters the structure of the DNA double helix. *Nature Structural and Molecular Biology* **2015,** *22* (1), 44-49.

37. Münzel, M.; Lischke, U.; Stathis, D.; Pfaffeneder, T.; Gnerlich, F. A.; Deiml, C. A.; Koch, S. C.; Karaghiosoff, K.; Carell, T., Improved Synthesis and Mutagenicity of Oligonucleotides Containing 5-Hydroxymethylcytosine, 5-Formylcytosine and 5-Carboxylcytosine. *Chemistry-A European Journal* **2011,** *17* (49), 13782-13788.

38. Xu, L.; Chen, Y. C.; Chong, J.; Fin, A.; McCoy, L. S.; Xu, J.; Zhang, C.; Wang, D., Pyrene-Based Quantitative Detection of the 5-Formylcytosine Loci Symmetry in the CpG Duplex Content during TET-Dependent Demethylation. *Angewandte Chemie International Edition* **2014,** *53* (42), 11223-11227.

39. Thalhammer, A.; Hansen, A. S.; El-Sagheer, A. H.; Brown, T.; Schofield, C. J., Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chemical Communications* **2011,** *47* (18), 5325-5327.

40. Owczarzy, R., Melting temperatures of nucleic acids: discrepancies in analysis. *Biophysical chemistry* **2005,** *117* (3), 207-215.

Chapter 9

The Effect of Epigenetic Cytosine Modifications on CpG Domain Opening in Duplex DNA

9.1 Introduction

The previous chapter discussed the influence of epigenetic cytosine modifications on some of the fundamental properties of DNA at both the level of the nucleobase and oligonucleotide duplex, with a primary focus at the nucleobase level. This chapter seeks to revisit and resolve the influence of modified cytosines on base pairing and to investigate the effect of each modification on the barrier to opening CpG domains through temperature jump (T-jump) kinetic studies. In this context CpG is a shorthand for a 5′-deoxycytidine-phosphate-deoxyguanosine-3′ motif along the linear sequence of DNA. The self-complementary DNA oligonucleotide first introduced in the previous chapter serves as a model system with sequence 5′-TAXGXGXGTA-3′ where X is either cytosine (C), 5-methylcytosine (mC), 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), or 5-carboxylcytosine (caC). These bases represent all of the known participants in the active cytosine demethylation pathway introduced in the previous chapter, where mC is sequentially oxidized to hmC, fC, and finally caC by the ten-eleven translocation (TET) family of enzymes. The bases fC and caC are selectively excised by thymine DNA glycosylase (TDG) and the resulting abasic site is repaired with canonical C by base excision repair (BER) enzymes, thus closing the demethylation cycle. Fig. 9.1 provides a summary of the proposed active demethylation pathway.

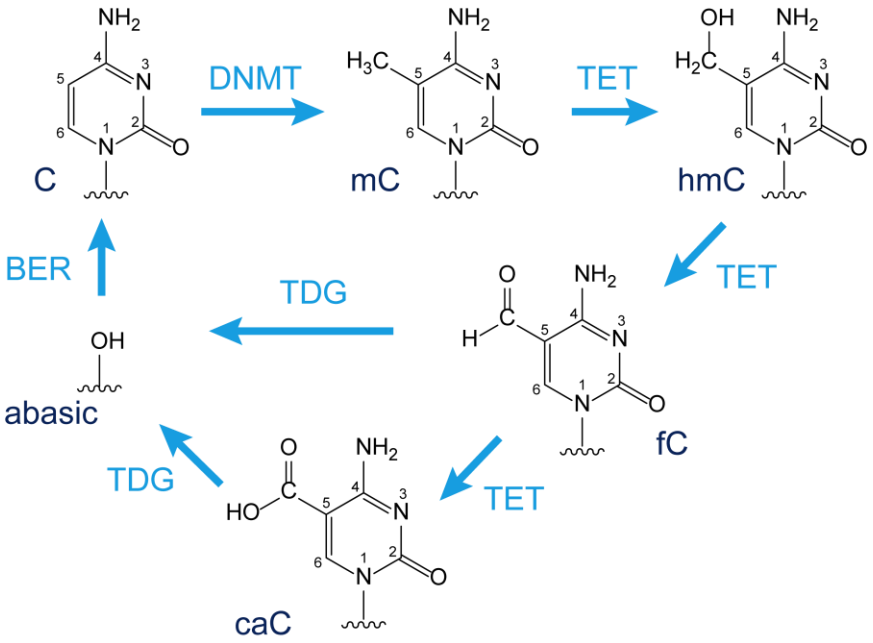

Figure 9.1: Summary of the active cytosine demethylation pathway. C is methylated at the 5 position by DNA methyltransferases (DNMT) to become mC. Ten eleven translocation (TET) enzymes sequentially oxidize mC to hmC, fC, and caC. The fC and caC bases can be excised by thymine DNA glycosylase (TDG) resulting in an abasic site. Base excision repair (BER) enzymes replace the abasic site with C, thereby closing the cycle.

As of yet, the influence of these epigenetic cytosine modifications on the fundamental biophysics, thermodynamics, and kinetics of DNA hybridization has not been fully resolved and several potential biological implications remain unexplored.[1] Despite previous suggestions to the contrary,[2] it has been shown that cytosine derivatives do not appear to deviate the solution structure of double stranded DNA significantly away from the canonical B form.[3] However, evidence suggests that fC and hmC increase while mC decreases the flexibility of duplex DNA and molecular dynamics simulations suggest this is accompanied by a corresponding increase/decreases in structural fluctuations about average B form DNA.[4] There is no evidence that cytosine modifications shift the tautomeric equilibrium relative to C, as previously proposed to account for the selective activity of TDG,[5] and the keto-amino form predominates in all cases.[6,7]

Both fC and caC appear to alter the nucleobase electronics as measured through N1 acidity and leaving group ability of the nucleobase.[6] It has been proposed that this could explain the selective excision of these bases by TDG as well as the strange features relative to canonical B form DNA observed in the ultraviolet circular dichroism (UV CD) spectrum of oligonucleotides containing these modifications.[3]

As discussed in the previous chapter, modified cytosines appear to have only a modest effect on DNA duplex stability and no consensus has been reached on the thermodynamic trend. Different researchers alternately report the same modifications to be stabilizing or destabilizing, in some cases even for the same sequence under similar conditions. Attempting to resolve this ongoing debate is the focus of this chapter. It is likely that understanding the influence of cytosine modification on the dynamics of base pair opening and the kinetics of DNA dissociation is the more relevant question when evaluating the possibility of a pre-excision recognition step in the active cytosine demethylation pathway. However, as discussed in Chapter 8, to date we and others have primarily focused on the influence of cytosine modification on DNA thermodynamics, as reported through the melting temperature (T_m). In this chapter it is demonstrated how the common measure of duplex stability via T_m breaks down to varying degrees for DNA oligonucleotides containing cytosine derivatives. A more detailed analysis based on a statistical description of base pairing is proposed. Finally, T-jump IR experiments are employed to characterize the influence of each of the modifications on the dissociation barrier to opening CpG domains and the potential biological implications of the observed trend is discussed.

9.2 Temperature Dependent FTIR Reveals Modified Cytosines Exhibit Distinct Influences on DNA Hybridization

9.2.1 Comparison to Standard Melting Temperature Analysis

Revisiting the temperature dependent Fourier transform infrared (FTIR) melting experiments first discussed in Chapter 8 for the sequence 5′-TAXGXGXGTA-3′ where X is either C, mC, hmC, fC, or caC, the FTIR spectrum was tracked across 3-98 °C. The 1530-1725 cm^{-1} frequency range contains in-plane ring vibrations and carbonyl stretches that are sensitive to the base stacking and hydrogen bonding interactions that mediate DNA hybridization. Melting curves that reflect the global changes to the IR spectrum in this frequency range were obtained through singular value decomposition (SVD), where the second SVD vector is assumed to contain the melting curve, as discussed in Chapter 6. Sample conditions were 1 mM oligonucleotide in 20 mM sodium phosphate buffer (pD 7.2) plus 16 mM NaCl. Fig. 9.2a-e shows the FTIR temperature series for each of the oligonucleotides and Fig. 9.2f contains the melting curve extracted from each of the spectra. The FTIR spectrum in this frequency range is clearly sensitive to the presence of modified cytosine bases, most noticeably for X = fC and caC since these bases introduce additional carbonyl modes that absorb in this frequency range.

It would be standard practice at this point to assume a model that describes the DNA dimer to monomer transition and fit the melting curves to determine T_m as a proxy for DNA duplex stability. Indeed we and others have taken this exact approach in the past, as discussed at length in the previous chapter. However, all of the standard models assume a two state all-or-nothing description of base pairing where all possible base pairs are either fully intact (dimers) or fully broken (monomers). This assumption is typically well justified for short canonical sequences, which generally display melting curves that are sharply transitioning sigmoids symmetric about their inflection points.

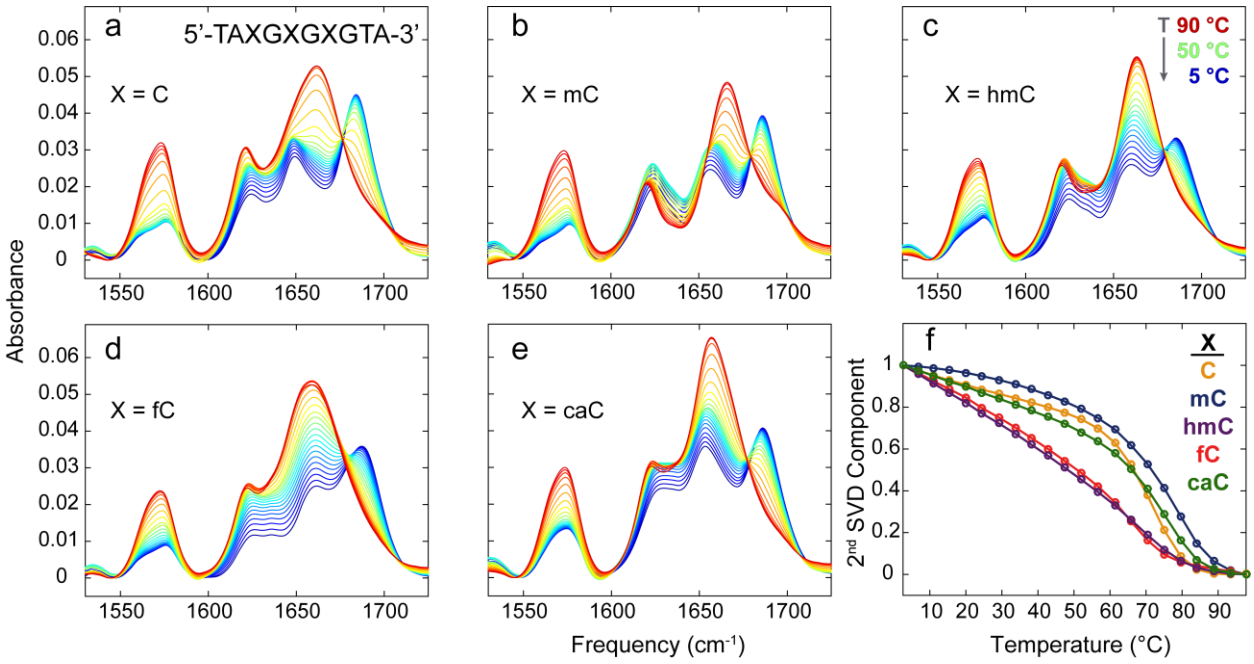

Figure 9.2: (a-e) FTIR temperature series from 3-98 °C for 5′-TAXGXGXGTA-3′ where X = C, mC, hmC, fC, or caC as indicated in each panel. Temperature runs from cold in blue to hot in red. (f) The normalized second SVD component corresponding to each of the set of spectra in panels a-e.

It is clear in Fig. 9.2f that the standard all-or-nothing description breaks down for many of the oligonucleotides containing cytosine modifications, most notably for X = fC and hmC. The upper baseline of the melting curve is severely sloped in both cases. Sloping baselines in melting curve analysis are a common feature and are attributed to such factors as changes in solvent transmission, sample evaporation, and drifts in spectrometer lamp intensity. It is therefore common practice to subtract linear fits to the baselines to correct for these influences, as discussed in Chapter 6. However, the degree to which the upper baselines are sloped for these oligonucleotides is so great that it is likely indicative of some change to the DNA dehybridization process itself. In addition, the severe slope further complicates the already subjective task of baseline fitting, since it is difficult to resolve where the upper baseline ends. Such a consideration can influence the melting temperature, since T_m is commonly defined as the temperature at which the melting curve

is equal to 0.5 and such a large baseline correction can shift this point by several degrees depending on how the baselines are defined. For example, fitting and subtracting sloping baselines to the X = C and the X = fC melting curves in Fig. 9.2f results in T_m's of 72 °C and 71°C, respectively. These are essentially identical values of T_m even though inspection of the melting curves prior to baseline correction suggests the X = fC transition is shifted to lower temperature relative to the canonical sequence.

9.2.2 A Statistical Description of Base Pairing Accounts for the Shape of the Melting Curve

For this set of sequences the standard melting curve analysis in which T_m is used as a proxy for duplex stability appears to break down. It is therefore unsurprising that there is currently no consensus for trends in T_m as a function of cytosine modification. A model which can account for the anomalous features observed for these sequences is clearly required in order to properly evaluate the influence of cytosine modification on DNA dehybridization. We therefore turn to a statistical description in which the melting curve is assumed to report on the overall fraction of intact base pairs in the DNA ensemble, $\theta(T)$. As defined in Chapter 5, this quantity can be expressed as a product of an internal $\theta_{int}(T)$ and external $\theta_{ext}(T)$ base pair fraction, where $\theta_{int}(T)$ is the average fraction of intact base pairs among DNA dimers and $\theta_{ext}(T)$ is the fraction of duplexed oligonucleotides out of the total number of all DNA strands. Fig. 9.3a shows a simulated melting curve for the X = C sequence obtained using the DNA lattice model presented in Chapter 5. In brief, the model is a simple statistical extension of the nearest neighbor (NN) model[8] and considers all possible combinations of broken and intact base pairs in the sequence explicitly. In Fig. 9.3a, $\theta_{ext}(T)$ is plotted as the gray dashed curve while $\theta_{int}(T)$ is plotted as the orange dashed curve. The simulated melting curve, plotted as the solid orange line, is the product of $\theta_{int}(T)$ and $\theta_{ext}(T)$. The

model reasonably reproduces the measured curve for the canonical sequence, as can be seen through comparison with the experimental melting curve plotted in orange in Fig. 9.3b.

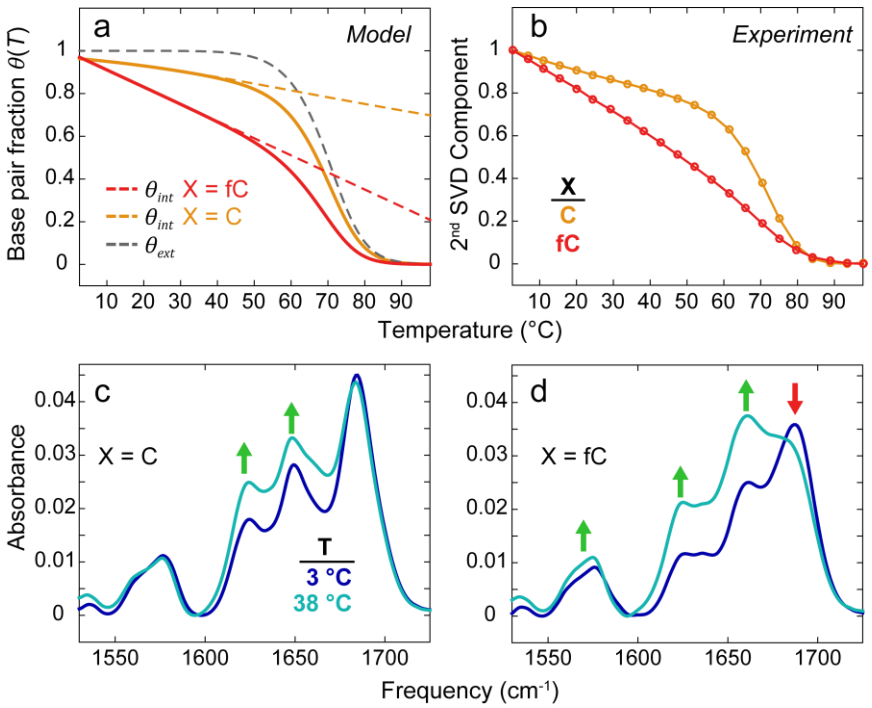

Figure 9.3: (a) Simulated melting curves for the X = C and X = fC sequences. The dashed gray curve shows the external fraction of intact base pairs $\theta_{ext}(T)$ while the color coded dashed curves show the internal fraction of intact pairs $\theta_{int}(T)$ for each of the sequences. The solid lines plot the simulated melting curves and represent the product $\theta_{int}(T)\theta_{ext}(T)$. (b) Experimental melting curves for the X = C and X = fC sequences from Fig. 9.2f. FTIR spectra for the (c) X = C and (d) X = fC sequences illustrating that the X = fC sequence shows more signs of lost base pairing compared to the X = C sequence between 3-38 °C.

This statistical description of base pairing adequately describes the canonical sequence, but the model is not designed to account for the influence of cytosine modifications on DNA hybridization directly. However, by modeling the canonical sequence one can gain insight into the origin of the shape of experimentally measured melting curves with respect to the underlying ensemble of oligonucleotides. The asymmetric distortion of the melting curve at low temperature

is caused by $\theta_{int}(T)$. Furthermore, $\theta_{int}(T)$ is approximately a linear function of temperature with negative slope. It is $\theta_{ext}(T)$ that dictates the overall sigmoidal shape of the melting curve and the position of the inflection point of $\theta_{ext}(T)$ largely sets where the melting transition is centered. This insight can rationalize the anomalous shape of the melting curve observed for some of the sequences in Fig. 9.2f. For example, for the X = fC sequence it would appear that $\theta_{ext}(T)$ is not perturbed significantly with respect to the X = C sequence since following baseline subtraction the transitions are centered in the same temperature range, as discussed above. The primary difference between the melting curves of the two sequences is the shape of the curve at low temperature, with the X = fC sequence exhibiting a severely sloping baseline relative to the X = C sequence. This observation suggests that $\theta_{int}(T)$ for the X = fC sequence drops off more rapidly with increasing temperature than for the X = C sequence. Fitting a line to $\theta_{int}(T)$ for the canonical sequence and then adjusting the slope such that it is around 2.7 times steeper results in reasonable agreement between the simulated and experimental melting curves for the X = fC sequence, as seen by comparing the solid red curves in Fig. 9.3a and b. The $\theta_{int}(T)$ for the X = fC sequence is plotted as the red dashed line in Fig. 9.3a.

If the above explanation for the anomalous shape of the melting curves is correct and the average fraction of intact base pairs among DNA dimers $\theta_{int}(T)$ is dropping off more rapidly with temperature for sequences containing some of the cytosine modifications, then one would expect this to be evident in the mid-IR spectrum since the vibrations in this frequency range are sensitive reporters of base pairing. Returning to the FTIR temperature series in Fig. 9.2, there is clear evidence for our proposed picture in which cytosine modification can shift $\theta_{int}(T)$. For example, Fig. 9.3c and d show the FTIR spectrum at 3 °C and 38 °C for the X = C and the X = fC sequences, respectively. These temperature points lie on the upper baseline of the melting curve for both

sequences, so one would expect that changes to the spectrum in this temperature range are dominated by shifts in the average base pair contact number in dimers rather than dissociation into monomers. Upon hybridization, the G carbonyl mode at 1665 cm^{-1} blue shifts to 1685 cm^{-1} and the linewidth sharpens. Loss of this feature and intensity growth near 1665 cm^{-1} is therefore indicative of a loss of GC base pairing. The G ring modes around 1572 cm^{-1} are suppressed upon the formation of a stacked, hydrogen bonded base pair. Intensity growth in these peaks is therefore also an indication of the loss of GC base pairing. In Fig. 9.3c, there is virtually no change in either of these G features, suggesting that for the canonical sequences there is little loss of GC base pairing between 3 °C and 38 °C. In contrast, for the X = fC sequence there is noticeable intensity loss at 1685 cm^{-1} accompanied by gain at 1665 cm^{-1} and 1572 cm^{-1}, suggesting a reduction in GC base pairing that is consistent with the more rapid drop off in $\theta_{int}(T)$ proposed for this sequence above. Losses and gains in intensity are indicated by the red and green arrows in Fig. 9.3d. Both sequences show intensity gains at 1621 cm^{-1} and around 1650 cm^{-1}. These frequencies are congested with A, T, and C absorptions, but in this case the intensity gain is likely a reflection of a loss in AT base pairing because the 1621 cm^{-1} A ring mode exhibits a large intensity increase upon AT base pair dissociation and the lack of response for the previously discussed G features suggests that, at least in the case of the canonical sequence, any changes observed in this range do not originate from a loss of GC pairing.

9.2.3 Summary of Thermodynamic Results from a Statistical Description of Base Pairing

One must conclude that the standard melting curve analysis that relies on trends in T_m to evaluate DNA duplex stability is an inadequate approach when evaluating the influence of cytosine modifications on the stability of CpG domains. Certain modifications, most notably fC and hmC,

lead to a decrease in the average number of intact base pairs within DNA dimers even if the overall fraction of dimers relative to the total concentration of DNA strands at a given temperature is not perturbed significantly. This subtle point offers an explanation for the ongoing discrepancies in T_m analysis. For the canonical sequence in which X = C, there does not appear to be a significant loss of GC contacts prior to dissociation into monomer strands and $\theta_{int}(T)$ appears to primarily reflect a loss of AT base pairing.

Table 9.1: Summary of results analyzing the melting curves in Fig. 9.2f in terms of a statistical model of base pairing.

X =	**Overall dimer stability relative to X = C sequence**	**Average fraction of intact base pairs within the dimer relative to X = C sequence**
mC	Shifted to higher T	Similar
hmC	Similar	Sharply reduced
fC	Similar	Sharply reduced
caC	Shifted to slightly higher T	Reduced

For the X = mC sequence, the transition region of the melting curve is unambiguously shifted to higher temperature, consistent with the literature consensus that cytosine methylation results in a slight increase in thermodynamic stability, as discussed in Chapter 8. This conclusion is also consistent with the statistical description of the melting curves presented here in which $\theta_{ext}(T)$ is shifted to higher temperature. The melting curve for the X = caC sequence appears to be slightly shifted toward higher temperature, but the low temperature baseline also appears more highly sloped than in the canonical case. These observations suggest that the inflection point of $\theta_{ext}(T)$ for this sequence is shifted to a slightly higher temperature range relative to the canonical sequence, but that $\theta_{int}(T)$ is more steeply sloped. Taken together, the X = caC modification would

appear to have a slight stabilizing influence on the DNA dimer as a whole, but at the same time this modification leads to an overall reduction in the average base pair contact number within the dimer ensemble at a given temperature. Table 9.1 includes a summary of results after analyzing the melting curves for each of the sequences in the context of a statistical description of oligonucleotide melting.

9.3 Temperature Jump Experiments Reveal Modified Cytosines Tune the Barrier to Dissociating CpG Domains

9.3.1 Motivation and Approach for T-jump Experiments

Up to this point the discussion has focused on how epigenetic cytosine modifications perturb the equilibrium of local base pairing as well as the overall dehybridization of DNA. However, the more relevant question with regard to a possible pre-excision recognition of certain cytosine modifications over others centers on the influence of the nucleobases on the barrier to opening CpG domains, motivating a kinetic study of these sequences. One would expect that a modification that lowers the barrier to CpG domain opening could be preferentially recognized by an enzyme in several ways. The modified CpG domain could spend a larger proportion of the time dehybridized and oriented outward toward the surrounding environment, serving as a recognition point for enzyme binding. The reduced barrier could also facilitate the flipping of the base into the active site as the enzyme searches along the DNA strand. To quantify the influence of epigenetic cytosine modification on the rate of CpG domain opening we studied the same set of sequences using T-jump IR spectroscopy in analogy to the studies of canonical DNA sequences presented in Chapters 6 and 7. Here we apply these methods to quantify how the kinetics of CpG domain dissociation vary as a function of cytosine modification.

For each sequence, six T-jumps of ΔT = ~15 °C were sampled across the melting curve to map the kinetics across the dimer to monomer transition. For example, Fig. 9.4a plots the T-jumps sampled in relation to the melting curve for the X = C sequence as color-coded lines linking the initial (T_i) and final (T_f) temperature for each jump. Following the T-jump, the response of the DNA oligonucleotide ensemble was tracked using a sequence of three mid-IR pulses to probe the 1500-1720 cm^{-1} in-plane nucleobase vibrations through the transient heterodyne-detected dispersed vibrational echo (t-HDVE), as described in detail in Chapter 3. In brief the t-HDVE spectra reported here are equivalent to pump probe thermal difference spectra sampled along each point of the temperature profile imparted by the T-jump pulse. Fig. 9.4b shows an example set of t-HDVE spectra for the T_i = 65 °C and the T_f = 80 °C T-jump for the canonical sequence. The HDVE spectrum contains both positively signed ground state bleach (GSB) and negatively signed excited state absorption (ESA) features corresponding to the 0→1 and 1→2 vibrational transitions, respectively. The t-HDVE spectrum can therefore contain both positive and negative growth in amplitude in response to the T-jump.

To illustrate the amplitude changes in time following the T-jump pulse, three kinetic traces tracked at single frequency slices through the t-HDVE data set in Fig. 9.4b are plotted in Fig. 9.4c. The features at 1545 cm^{-1} and 1572 cm^{-1} correspond to the ESA and GSB of the G ring modes centered around 1572 cm^{-1} discussed in the FTIR spectrum above. The feature at 1669 cm^{-1} is dominated by the growth of the GSB of the G carbonyl mode that is blue shifted to 1685 cm^{-1} in a Watson-Crick pair but is centered near 1665 cm^{-1} in the unpaired base. The slight frequency shift in the maximum with respect to the peak amplitude in the FTIR spectrum is due to spectral overlap between the ESA and GSB of multiple vibrational modes in this frequency range. These peaks all report on the opening of the CpG domain in the sequence in response to the T-jump. Fig. 9.4c

shows a consistent pattern in the kinetic traces tracked at these frequencies. First there is an initial rise away from the zero baseline dominated by dehybridization with a time constant of around 100 μs followed by a return to the baseline that tracks the thermal relaxation and rehybridization of the DNA strands with a time constant of around 4 ms.

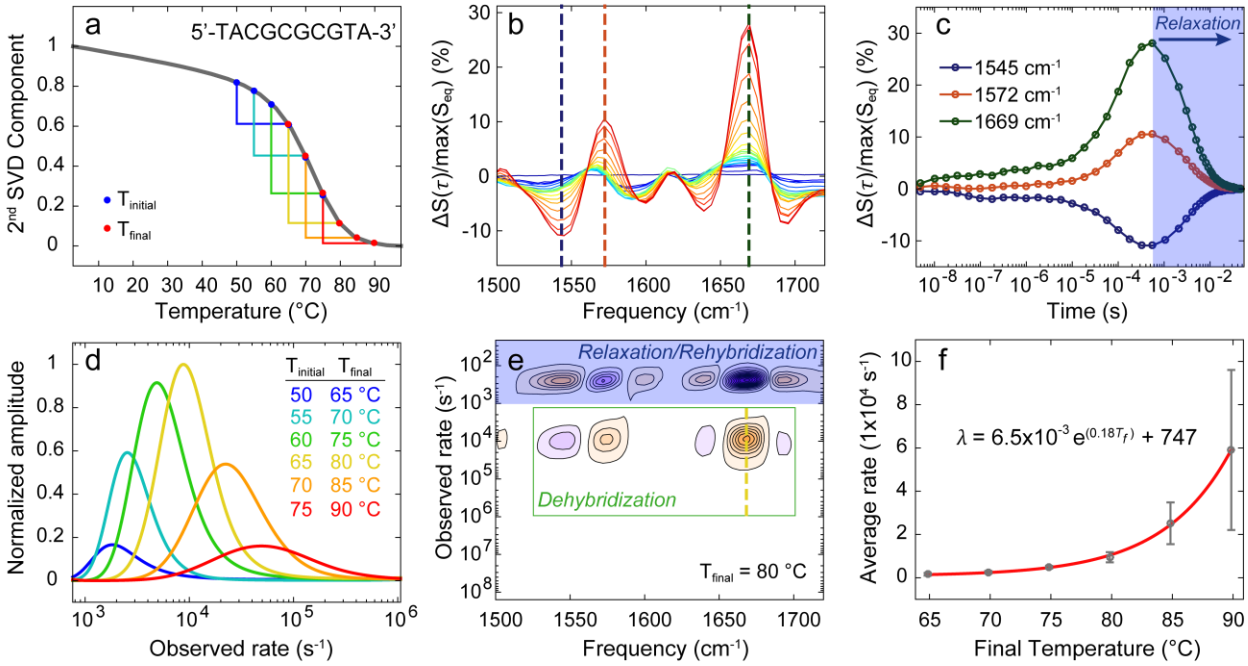

Figure 9.4: (a) Temperature range for the T-jump experiments sampled across the melting transition of the X = C sequence. Each T-jump is color coded with the initial and final temperatures denoted by a blue and red point on the curve. (b) Illustrative t-HDVE spectra for the X = C sequence prepared at the same sample conditions as for the FTIR experiments and with T_i = 65 °C and T_f = 80 °C. (c) Select time slices through the t-HDVE spectrum in panel b demonstrating kinetic traces tracked at 1545, 1572, and 1669 cm^{-1}. (d) The evolution of the rate spectrum at 1669 cm^{-1} between 10^3-10^6 s^{-1} tracked across the set of T-jump measurements indicated in panel a. These distributions of rates correspond to the DNA dimer to monomer transition and are normalized to the rate distribution for the T_i = 65 °C jump. (e) The full rate domain representation of the t-HDVE set in panel b. The ranges corresponding to dehybridization and thermal relaxation/rehybridization are indicated. The dashed yellow line shows the location of the frequency slice plotted in panel d. (f) The amplitude weighted average observed rate for the dimer to monomer transition as a function of T_f taken across the entire frequency range indicated by the green box in panel e. Error bars represent the standard deviation across the frequency range weighted by the rate amplitude.

It is common practice to assume a functional form in the time domain and then fit kinetic traces in order to extract time constants. Alternatively, the inverse Laplace transform method described in Chapter 4 can be applied to the time domain t-HDVE data to produce a rate map off of which observed time constants can be read directly. An example of the rate domain representation of t-HDVE data is illustrated in Fig. 9.4e for the t-HDVE data plotted in Fig. 9.4b. The sign in the rate domain directly reflects the sign of the features in the time domain t-HDVE spectra. Orange features in the rate spectrum correspond to positive rate amplitude while purple features correspond to negative rate amplitude. Fig. 9.4d shows illustrative slices through the rate spectrum at 1669 cm^{-1} between 10^3-10^6 s^{-1} for the full set of T-jumps on the X = C sequence. This range contains the distribution of observed rates for the dehybridization response of the duplex following the T-jump. The peak of the rate spectrum shifts towards faster rate as temperature increases, as one would expect. Furthermore, the integrated rate amplitude tracks the trend suggested for the changes in signal amplitude between each T_i and T_f on the melting curve.

It is interesting to note changes in the width and symmetry of the rate distribution. As the temperature increases, the distribution of rates broadens and becomes more symmetric about the maximum, while at low temperature there is a distinct skew towards faster rates. To visualize trends in the observed rate with increasing temperature, it is convenient to track the average across the 1530-1700 cm^{-1} frequency range and the 10^3-10^6 s^{-1} rate range corresponding to the dimer to monomer transition. The average rate across this range, indicated by the green box in Fig. 9.4e, is determined by weighting the rate by the magnitude of the rate amplitude at each frequency. The resulting amplitude weighted average observed rate as a function of T_f is plotted in Fig. 9.4f and demonstrates an exponential temperature dependence.

Both the trends in rate amplitude and the shape of the rate distribution are consistent with the interpretation of the melting curves above. One would expect that a heterogeneous dimer ensemble that gives rise to the sloping upper baseline in the melting curve would also be reflected in the dehybridization kinetics as a broad distribution of rates. At low temperature, dimers with a higher fraction of intact base pairs are favored (larger $\theta_{int}(T)$). Therefore much of the dissociation observed following the T-jump originates from highly hybridized dimer configurations with a smaller number originating from partially dehybridized dimers, consistent with the skewed rate distributions centered toward slower rates observed at low temperature in Fig. 9.4d. However, as the temperature increases dimer heterogeneity increases ($\theta_{int}(T)$ decreases) thus resulting in a broadened rate distribution that is more symmetric since the dimer configurations are increasingly more evenly distributed about the mean fraction of intact base pairs. These arguments assume that dimers with more intact base pairs result in slower dissociation timescales, which seems reasonable given the additional barrier that must be overcome with each additional base pair that must be broken.

9.3.2 Stretched Exponential Kinetics and Heterogeneous Duplex Structures

If the proposed correlation between a heterogeneous distribution of dehybridization rates and increased configurational diversity in the dimer ensemble is a reasonable interpretation of both the thermodynamic and kinetic results, then one would expect a clear correspondence between the shape of the melting curve and the heterogeneity of timescales observed in the T-jump experiments. This comparison across the oligonucleotide set would ideally be conducted at a fixed temperature and at a fixed point relative to the dimer/monomer equilibrium. However, both of these conditions cannot be met simultaneously for all of the sequences due to the shifts observed

in the melting curves in Fig. 9.2f. Comparing the T-jump with T_i = 65 °C for all of the sequences offers a reasonable compromise since this measurement jumps across the inflection point of all of the melting curves and samples the temperature range corresponding to the sharpest signal variation. Fig. 9.5a plots the kinetic trace tracked at 1669 cm^{-1} for each of the sequences. To quantify heterogeneity in the kinetics of the dimer to monomer transition, we fit stretched exponentials to these time traces since this is the most direct assessment of the experimental data. However, rather than fitting time constants for the dehybridization and relaxation response of the system, we obtain these characteristic rates from the peak of the rate spectrum as discussed above to reduce the number of unknowns. Therefore the only fit parameters in the time domain are the amplitude and stretching parameter, β for the rise and decay of each kinetic trace.

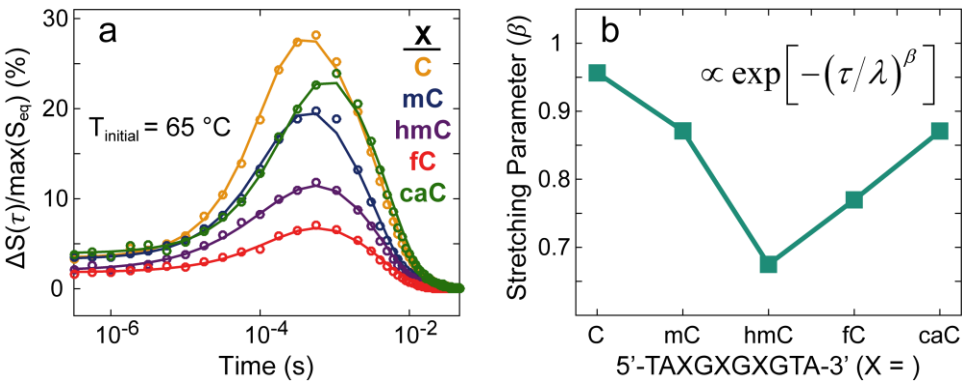

Figure 9.5: (a) Kinetic traces tracked at 1669 cm^{-1} for each of the sequences with T_i = 65 °C and ΔT = ~15 °C. Color coded points represent the experimental time traces. Solid lines represent stretched exponential fits to the data. (b) The stretching parameter β for the rise in the kinetic traces in panel a corresponding to the dehybridization response of the system as a function of cytosine modification.

The value of the stretching parameter can range from β = 0 to 1. The single exponential kinetics expected for a homogeneous process corresponds to the case where β = 1. Decreasing

values of β indicate increasingly stretched exponential kinetics. Assuming that a stretched exponential in this case is a reflection of an essentially continuous sum of overlapping single exponential processes, the value of β can be used as a simple proxy for the heterogeneity of the dehybridization kinetics as a function of cytosine modification. Fig. 9.5b shows the stretching parameter obtained from fitting the signal rise in Fig. 9.5a, corresponding to the dehybridization response of the ensemble following the T-jump. For the X = C sequence, $\beta = 0.96$ suggesting that the canonical sequence results in essentially single exponential kinetics. For the remaining sequences, the trend in β follows the trends discussed in Table 9.1 for the influence of each modification on the average fraction of intact base pairs among dimers. The X = hmC and fC sequences exhibit comparatively stretched exponential kinetics with $\beta = 0.67$ and 0.77, respectively, while the X = mC and caC sequences both exhibit only mildly stretched exponential kinetics with $\beta = 0.87$ in both cases. Therefore those sequences that contain cytosine modifications which give rise to the most anomalous melting curves also appear to exhibit the most heterogeneous dehybridization kinetics, consistent with the increase in variety within the dimer ensemble proposed for these sequences above.

9.3.3 Evaluating the Dissociation Barrier as a Function of Cytosine Modification

To characterize the influence of cytosine modification on the barrier to opening CpG domains, we apply an Arrhenius description following the approach described in Chapter 7 to describe the dehybridization of canonical DNA sequences. The observed rate constant is related to rate constants for dimer association k_a and dissociation k_d. Plotting the logarithm of these rate constants versus the reciprocal of T_f produces Arrhenius plots where the slope is proportional to the activation energy for association E_a and dissociation E_d. Fig. 9.6a-e shows the Arrhenius plots

for each of the sequences. The anti-Arrhenius behavior for the association process is well documented and this result is discussed in detail in Chapter 7. Here the focus is on the dissociation process since the opening of the CpG domain is required for enzymatic action to occur, whether that is oxidation by TET or base excision by TDG. Fig. 9.6f plots the trend in the dissociation barrier as a function of cytosine modification.

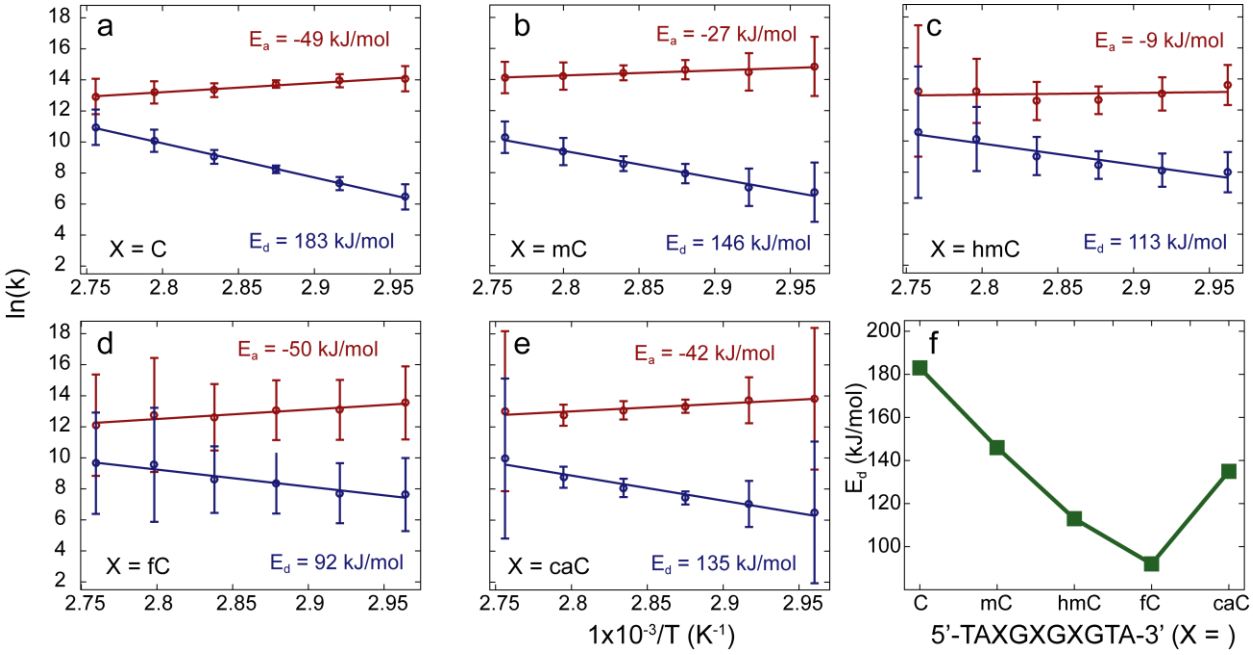

Figure 9.6: (a-e) Arrhenius plot for the sequence indicated in each panel. The trend in the association rate with temperature is plotted in red while the trend in the dissociation rate is plotted in blue. Error bars represent the propagation of the error bars in Fig. 3f through the Arrhenius analysis. (f) The dissociation barrier determined from the Arrhenius analysis as a function of cytosine modification.

The canonical sequence exhibits the largest dissociation barrier at 183 kJ/mol. Despite the increased thermodynamic stability of the X = mC sequence relative to the canonical sequence, E_d is lowered by about 40 kJ/mol by this cytosine modification. The X = caC sequence also has a lowered dissociation barrier compared to the X = C sequence but by about 50 kJ/mol. The

X = hmC and fC sequences display the most dramatic reduction in E_d, with barriers of 113 kJ/mol and 92 kJ/mol, respectively.

It is interesting to note the similarity between the trend in the stretching parameter in Fig. 9.5b and the trend in the dissociation barrier in Fig. 9.6f. This correspondence is consistent with the assignment of the origins of stretched exponential kinetics and anomalous melting curves discussed above. Since dimer heterogeneity corresponds to duplex configurations that have a reduced average number of intact base pairs, one would expect those sequences exhibiting the most severely sloped upper baseline in their melting curves and the most stretched exponential kinetics would also correspond to the lowest average dissociation barrier for the dimer to monomer transition. All of these conditions are most exemplified by the X = hmC and fC sequences. This result is consistent with a recent study that reported increased flexibility in DNA containing fC and hmC pairs and attributed this property to increased structural fluctuations around these pairs.[4]

9.4 Conclusions and Potential Biological Implications

Taking all of the results in this chapter into consideration, we propose the energy diagram in Fig. 9.7 to summarize the influence of cytosine modification on the thermodynamics and kinetics of DNA dehybridization. The monomer state is selected as the reference state for each of the sequences. The degree to which the dimer state is stabilized relative to the monomer state does not vary significantly as a function of cytosine modification. However, X = mC does modestly stabilize the dimer with respect to X = C, as evidenced by the shift in the melting curve to higher temperature. Likewise the X = caC modification demonstrates a mild stabilizing influence on the dimer, but this effect is less pronounced than for the X = mC sequence consistent with the comparatively minor shift of the X = caC melting curve towards higher temperature. The X = C,

hmC, and fC sequences all appear to have dimer states similar in energy, in line with the discussion of the melting curves for these sequences above in which the fraction of dimers out of the total number of DNA strands as a function of temperature is not significantly perturbed for these sequences.

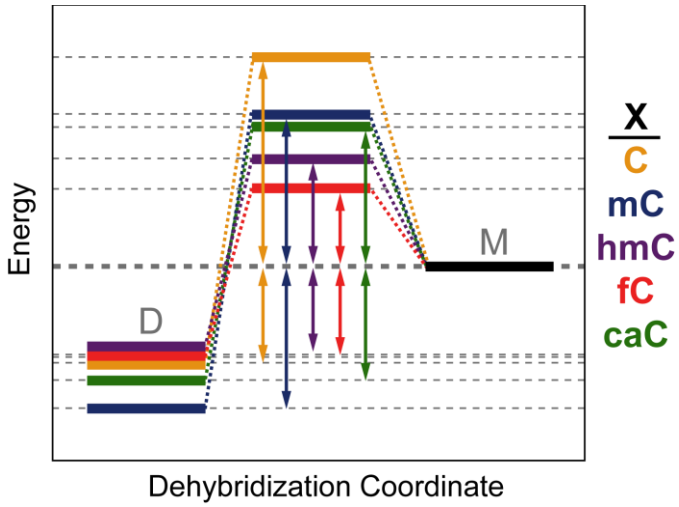

Figure 9.7: Proposed energy diagram based on the thermodynamic and kinetic results in this chapter.

Considering dehybridization thermodynamics alone is insufficient to explain the influence of these modifications on duplex DNA. In fact one would likely conclude that the dimer is more or less insensitive to cytosine modification if T_m is the only point of comparison across the oligonucleotide set. Therefore it is unsurprising that no clear consensus on T_m trends, other than the mild stabilization due to methylation, presently exists in the literature. In light of the results presented here, it would appear that the most significant influence of cytosine modification on the fundamental biophysical properties of DNA are kinetic rather than thermodynamic. All of the cytosine modifications appear to lower the barrier to dehybridization compared to canonical C,

with hmC and fC resulting in a considerable reduction in the barrier height. A reduced barrier height to opening the XpG domain likely translates into the nucleobases in this domain spending more time on average dehybridized and oriented outwards towards the surrounding environment. This picture is consistent with the reduction in the average number of intact base pairs among dimers observed in the melting curves for the X = hmC, fC, and, to a lesser extent, caC sequences. It is also consistent with the degree to which the sequences exhibit stretched exponential kinetics, since stretched exponential kinetics are interpreted as a reflection of the increased heterogeneity of dimers in the T_i equilibrium ensemble giving rise to a spectrum of closely spaced exponential dissociation processes.

In light of these results, it seems reasonable to postulate that many of the modified cytosine bases could exhibit varying degrees of preferential recognition by the enzymes which act upon them due to the fact that the barrier to dissociating the XpG domain is lowered. The modified nucleobases with lower dissociation barriers likely spontaneously flip out of the duplex more frequently than canonical C resulting in more time spent on average open to the solvent, ready to flag down a passing enzyme. Further still, once an enzyme encounters a modified cytosine, the lowered barrier to dehybridization would help facilitate the flipping of the base fully out of the DNA duplex and into the active site of the enzyme. This effect likely contributes in part to the overall enzymatic activity towards a particular cytosine modification, although there are undoubtedly other critical factors at play. For instance, when considering the ability of TDG to excise a nucleobase, the N1 acidity and leaving group ability of the base have been shown to be the essential properties that determine whether or not a base can be removed from the duplex.[6] Therefore even though hmC and fC appear to produce a similar reduction in the barrier to opening a CpG domain, only fC is excised. In addition, TDG exhibits excision activity towards caC despite

a comparatively modest reduction in the dehybridization barrier. Activity towards caC is lower with respect to fC at physiological pH, but activity towards caC increases and even surpasses fC excision as the pH is reduced to below ~6.5.[6] In the future, it would be an interesting extension of the present study to characterize the caCpG dissociation barrier at proton concentrations greater than physiological pH in order to evaluate if the pH dependence of the dissociation barrier correlates with the pH dependence of TDG activity.

9.5 Acknowledgements

I thank Qing Dai for helpful discussions and for synthesizing the modified DNA oligonucleotides.

I thank Ryan Menssen and Brennan Ashwood for careful reading of this chapter.

9.6 References

1. Hardwick, J. S.; Lane, A. N.; Brown, T., Epigenetic Modifications of Cytosine: Biophysical Properties, Regulation, and Function in Mammalian DNA. *BioEssays* **2018,** *40* (3), 1700199.

2. Raiber, E.-A.; Murat, P.; Chirgadze, D. Y.; Beraldi, D.; Luisi, B. F.; Balasubramanian, S., 5-Formylcytosine alters the structure of the DNA double helix. *Nature Structural and Molecular Biology* **2015,** *22* (1), 44-49.

3. Hardwick, J. S.; Ptchelkine, D.; El-Sagheer, A. H.; Tear, I.; Singleton, D.; Phillips, S. E.; Lane, A. N.; Brown, T., 5-Formylcytosine does not change the global structure of DNA. *Nature Structural and Molecular Biology* **2017,** *24* (6), 544-552.

4. Ngo, T. T.; Yoo, J.; Dai, Q.; Zhang, Q.; He, C.; Aksimentiev, A.; Ha, T., Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nature communications* **2016,** *7*, 10813.

5. Karino, N.; Ueno, Y.; Matsuda, A., Synthesis and properties of oligonucleotides containing 5-formyl-2′-deoxycytidine: in vitro DNA polymerase reactions on DNA templates containing 5-formyl-2′-deoxycytidine. *Nucleic acids research* **2001,** *29* (12), 2456-2463.

6. Maiti, A.; Michelson, A. Z.; Armwood, C. J.; Lee, J. K.; Drohat, A. C., Divergent mechanisms for enzymatic excision of 5-formylcytosine and 5-carboxylcytosine from DNA. *Journal of the American Chemical Society* **2013,** *135* (42), 15813-15822.

7. 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 chemical biology* **2015,** *11* (2), 470-477.

8. SantaLucia, J., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences* **1998,** *95* (4), 1460-1465.