Exploring the folding and assembly of nucleic acids by thermal denaturation

Robert W. Harkness, V

Department of Chemistry, McGill University

Montréal, Québec, Canada

Submitted March 2018

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

This thesis is dedicated to my parents Bob and Chris Harkness, who raised me to always move forward in the face of adversity.

Acknowledgements

First, I am tremendously grateful to my supervisor, Dr. Anthony Mittermaier, for his excellent mentorship on my journey through doctoral studies. I have learned an incredible amount from him over the past six years, both in biophysics and the management of the trepidations of academia. Being exposed to Tony's consistently high level of scientific rigor and way of thinking has been a great honour and pleasure. I will move forward in my scientific career striving to always apply the same high level of rigor that Tony applies in his work. Tony has been a fantastic inspiration to improve myself as a scientist and pushed me to always do the best we possibly could in everything we produced. I am grateful for Tony's patience and instruction as I developed the mathematical and biophysical tools I required for addressing the scientific barriers we faced. Six years ago, we started a new research area on nucleic acid dynamics together, which has blossomed into an active and exciting program with a bright future in the Mittermaier laboratory. Under Tony's supervision, I have had the opportunity to explore what has really been a highly stimulating scientific sandbox. My doctoral work has turned out to be exactly what I wanted it to be; starting something entirely new and seeing how far we could push ourselves in developing useful tools for the scientific communities we have entered and contributing to the understanding of nucleic acid biophysics.

Thank you to the members of my Ph. D. committee, Dr. Hanadi Sleiman and Dr. Karine Auclair, for their support and questions during my yearly meetings which served to strengthen the work I carried out. I am deeply indebted to both Dr. Sleiman and Dr. Masad Damha for permitting me to use their nucleic acid synthesizers and thermal denaturation instruments which formed the basis for a major portion of my doctoral work. I am further grateful for our fruitful collaborations on nucleic acid biophysics. I give special thanks to Dr. Graham Hamblin for teaching me how to use the MerMade nucleic acid synthesizer that produced many of the samples I performed experiments with. Thank you to my collaborators at York University, Dr. Philip Johnson and Sladjana Slavkovic for the chance to work on interesting aptamer systems and develop the thermolabile ligand technique. I am grateful to Dr. Julian Adams for the David Noble Harpp Fellowship I received in my first year at McGill. It really helped me to get started in graduate school. I give thanks to Dr. Annick Guyot and Dr. Anne Noronha for their help and opportunities provided by the McGill CREATE Training Program in Bionanomachines, and to Dr. Naoki

Sugimoto and the staff scientists at the Frontier Institute for Biomolecular Engineering Research (FIBER) in Kobe, Japan, who helped me tremendously during my research exchange. I am grateful to Chantal Marotte for her help in enrolling me in McGill Chemistry, and for her help in administrative matters over the years. Thank you to any other McGill faculty, students, or staff who have helped me during my time at McGill.

I thank my mother and father, Chris and Bob Harkness, for raising me to have what I needed to complete the work described in this thesis. My parents told me that I should go to university from a young age, and while I did not always know what I should do with myself, particularly as a university student, their urgings nucleated the path that brought me to where I am today. My parents have been a continual source of inspiration to keep moving forward in the face of adversity in life and during my doctoral studies. In science it is often difficult to find the strength to try again when an experiment does not work, or a computer program bug remains elusive. One must maintain internal motivation, and for this I am grateful to my parents.

I give thanks to my two wonderful brothers, Matthew and Charles, for being there to talk when stress mounted, and for being the providers of laughter and comfort. My brothers helped me to keep going and to put everything in perspective when traversing certain scientific barriers became treacherous. You boys helped me to remember that life should not be just about working as hard as you can. I always look forward to our time together.

I am profoundly grateful to my girlfriend Mei-Ni Belzile, for listening to me every day, being so kind during the making of this thesis, and importantly when I couldn't seem to put the lines through the points. You have been an everlasting source of love and support, and a caring listener throughout these final years of my Ph. D. My being of pure light. You inspired me to never give up and to remember the reasons which prompted me to begin my doctoral work. You always inspire me to improve myself. Without you, I would never have reached this point. I am further grateful to the Belzile-Chen family for their support, delicious and relaxing meals, and stimulating conversations.

I thank the members of the Mittermaier laboratory who have helped me along the way. Thank you to Justin "Spanky" Di Trani, my *de facto* Ph. D. brother, for his mathematical help, great laughs and companionship, and showing me that $\widehat{I_x}$ is diagonal, as well as 0 = 0 *Q. E. D.* I

have also had the privilege of working with a number of motivated and insightful undergraduate students during my Ph. D., whom I thank for their hard work and help. Thank you to the members of the Cosa and Damha laboratories for their help and camaraderie during the challenges of doctoral studies. Thank you to Hala Abou Assi for her continued help and discussions over the years. Thank you to Dr. R. Stan Brown of Queen's University, for saying that he would make a scientist out of me when I told him I was studying for the MCAT while employed as a research associate during the summer of my second undergraduate year. Thank you to my high school physics teacher, Mr. Jim Strachan, for his advice when I was unsure of going to graduate school.

Abstract

Nucleic acids are highly dynamic biomolecules that regulate biological function and are widely used to construct nanoscale materials. Nucleic acid folding and assembly dynamics modulate how they interact with the environment to achieve specific molecular outcomes. These motions are highly complex and often occur via transient and low-populated intermediate states which can evade experimental detection. Consequently, a full understanding of nucleic acid dynamics and their role in function has not been achieved, largely due to the limitations of contemporary techniques. The available methodologies for studying nucleic acid dynamics in detail are costly in terms of time and labour, and require extensive user expertise. Methods that provide quantitative information on nucleic acid dynamics in a rapid and straightforward manner are therefore highly desirable for understanding and ultimately controlling their function in biology and biomaterials.

To address the aforementioned challenges, this thesis explores three global data fitting analyses that we developed to harness the under-utilized potential in thermal denaturation datasets for nucleic acids. In Chapter 2, we detail a global fitting analysis for equilibrium thermal denaturation experiments in application to the complex folding dynamics of guanine quadruplexes (GQs). We demonstrated that these motions influence GQ stability and functions such as gene expression. Chapter 3 describes a combined experimental and global fitting method using nucleic acids with heat-sensitive ligands to extract the suite of folding and binding parameters to the initial and thermally-converted ligand product. In Chapter 4, we developed an approach that utilizes non-equilibrium thermal melting and annealing experiments in characterizing the assembly of supramolecular nucleic acids via transient and low-populated intermediates. Importantly, all of the methods described in this thesis are rapid (requiring as little as one day of experimentation and analysis time), low-cost, and provide quantitative information on nucleic acid dynamics with a level of detail that is not easily accessible to current techniques. Furthermore, these approaches can be applied to systems of virtually any complexity. Taken together, these methods substantially expand the toolkit available to nucleic acid researchers and pave the way for robust, facile characterizations of nucleic acid folding and assembly dynamics.

Résumé

Les acides nucléiques sont des biomolécules hautement dynamiques qui régularisent la fonction biologique et qui sont largement utilisées dans la construction de nano biomatériaux. Les dynamiques de repliement et d'assemblage d'acides nucléiques déterminent comment ils interagissent avec leur environnement pour en arriver à des résultats moléculaires spécifiques. Ces mouvements sont très complexes et se produisent souvent en passant par des états intermédiaires transitoires et peu peuplés qui peuvent échapper à la détection expérimentale. Par conséquent, une compréhension complète de la dynamique des acides nucléiques et de leur rôle fonctionnel n'a pas pu être achevée, en grande partie à cause des limitations des techniques actuelles. Les méthodologies disponibles pour l'étude détaillée de la dynamique des acides nucléiques sont coûteuses en temps et en travail et elles requièrent une expertise poussée de l'utilisateur. Les méthodes qui fournissent de l'information quantitative sur la dynamique des acides nucléiques d'une manière rapide et directe deviennent ainsi fortement souhaitables afin de comprendre et ultimement contrôler leur fonction en biologie et dans les biomatériaux.

Afin de s'attaquer à ces défis, cette thèse explore trois analyses globales d'ajustement de données que nous avons développées pour saisir le potentiel sous-utilisé des bases de données de dénaturation thermique des acides nucléiques. Dans le Chapitre 2, nous décrivons en détails une analyse d'ajustement global pour l'équilibre thermique d'expériences de dénaturation en application à la dynamique complexe de repliement des quadruplexes de la guanine (GQs). Nous avons ainsi démontré que ces mouvements influencent la stabilité de GQ et des fonctions comme l'expression du gène. Le Chapitre 3 décrit la combinaison d'une méthode expérimentale et d'ajustement global qui utilise des acides nucléiques avec des ligands thermosensibles pour en extraire une suite de paramètres de repliement et de liaison au produit initial et à celui converti par ligand thermosensible. Dans le Chapitre 4, nous avons développé une approche qui utilise des expériences de fusion et refonte thermiques en non-équilibre pour caractériser l'assemblage d'acides nucléiques supramoléculaires en passant par des intermédiaires transitoires et peu peuplés. Il est important de mentionner que toutes les méthodes décrites dans cette thèse sont rapides (ne requérant aussi peu qu'un seul jour d'expérimentation et d'analyse), peu coûteuses et qu'elles fournissent de l'information quantitative sur la dynamique des acides nucléiques avec un niveau de détails qui n'est pas facilement accessible avec les techniques actuelles. De plus, ces approches peuvent être appliquées à des systèmes de virtuellement toutes les complexités. Prises dans leur ensemble, ces méthodes augmentent substantiellement les outils disponibles aux chercheurs sur les acides nucléiques et pavent la voie à des caractérisations robustes et faciles du repliement des acides nucléiques et de la dynamique des assemblages.

Table of Contents

List of Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables

Author contributions

Chapter 1: Introduction

A large portion of this chapter was adapted with permission from Harkness, R. W., V, and Mittermaier, A. K. G-quadruplex dynamics. *Biochimica et Biophysica Acta – Proteins and Proteomics* **1865(11B)**, 1544-1554 (2017). I performed all background research for this review article, created or adapted the figures with permissions, and co-wrote the manuscript with Dr. Mittermaier.

Chapter 2: G-register exchange dynamics in guanine quadruplexes

Chapter 2 was reproduced with permission from Harkness, R. W., V, and Mittermaier, A. K. G-register exchange dynamics in guanine quadruplexes. *Nucleic Acids Research* **44(8)**, 3481-3494 (2016). **Cover article***. Dr. Anthony Mittermaier conceived the study and I performed all experiments, developed the analysis with Dr. Mittermaier, wrote all MATLAB software for global fitting and other data analyses, and created all figures. The article was co-written by myself and Dr. Mittermaier.

Chapter 3: Rapid characterization of biomolecular folding and binding interactions with thermolabile ligands by DSC

Chapter 3 is adapted with permissions from two articles: Harkness, R. W., V, *et al*. Rapid characterization of folding and binding interactions with thermolabile ligands by DSC. *Chemical Communications* **52**, 13471-13474 (2016), and Harkness, R. W., V, *et al.* Measuring biomolecular DSC profiles with thermolabile ligands to rapidly characterize folding and binding interactions. *The Journal of Visualized Experiments* **129**, e55959 (2017). In these works, samples were prepared and shared by Sladjana Slavkovic and Dr. Philip Johnson of York University. I performed all experiments and developed the global fitting analyses and simulations with Dr. Mittermaier. These two articles were co-written by myself and Dr. Mittermaier. Dr. Johnson provided guidance in data analysis and the manuscript preparations.

Chapter 4: Mapping the energy landscapes of supramolecular assembly by thermal hysteresis

Chapter 4 is currently in review at Nature Communications for consideration as an article. The study was conceived by Dr. Mittermaier and Dr. Hanadi Sleiman. The experimental portion

of the study was carried out entirely by Nicole Avakyan. I developed the analysis and interpreted the results with Dr. Mittermaier, and wrote all MATLAB software for analyzing, simulating, and global fitting thermal hysteresis data. I created all figures except for the 3D molecular structures that were designed by Dr. Andrea Greschner. The manuscript was co-written by myself and Dr. Mittermaier, with guidance from Nicole Avakyan and Dr. Sleiman.

List of abbreviations

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
A	Adenine
U	Uracil
C	Cytosine
G	Guanine
T	Thymine
I	Hypoxanthine/inosine
d	Deoxy
WC	Watson-Crick
HG	Hoogsteen
GQ	G-quadruplex
GR	G-register
WT	Wild-type
BER	Base excision repair
NMR	Nuclear magnetic resonance
DPFGSE	Double pulsed field gradient spin echo
AFM	Atomic force microscopy
DSC	Differential scanning calorimetry
ITC	Isothermal titration calorimetry
UV-Vis	UV-Visible
TH	Thermal hysteresis
RSS	Residual sum of squares
CD	Circular dichroism
FRET	Förster resonance energy transfer
IR	Infrared

EM	Electron microscopy
LC	Liquid chromatography
QTOF	Quadrupole time-of-flight
ESI	Electrospray ionization
ALS	Amyotrophic lateral sclerosis
DMS	Dimethyl sulfate
CPG	Controlled pore glass
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
D_2O	Deuterium oxide
K_2HPO_4	Dipotassium phosphate
KH_2PO_4	Monopotassium phosphate
CA	Cyanuric acid
Tris	Tris(hydroxymethyl)aminomethane
$MgCl_2$	Magnesium chloride
NaCaco	Sodium cacodylate
NaCl	Sodium chloride
TEMED	Tetramethylethylenediamine
PAGE	Polyacrylamide gel electrophoresis
TBE	Tris-boric acid-ethylenediaminetetraacetic acid

Chapter 1: Introduction

1.1. The importance of nucleic acid dynamics

Nucleic acids are highly dynamic biomolecules that carry the genetic information to generate and sustain life. Nucleic acids are also used as a key building material in biotechnology applications. Over the last roughly 65 years, it has become increasingly clear that nucleic acid dynamics are intimately related to their function. A robust understanding of nucleic acid dynamics is therefore of paramount importance to our understanding of biology, the development of medical treatments, and the rational design of biomaterials. Yet, the folding and assembly of nucleic acids are currently challenging to characterize, since these dynamics are highly complex and often occur through transient and low-populated intermediates that can escape experimental detection. In the first half of this introduction, I will emphasize the importance of studying nucleic acids by reviewing their structure, dynamics, and the interplay between dynamics and function. The second half describes the basics of using thermal denaturation methodologies to acquire quantitative information on nucleic acid dynamics. These fundamentals are then built upon in later chapters, where three global fitting methods are described that we have developed in application to simple thermal denaturation datasets. These analyses have allowed us to gain insight into a wide array of nucleic acid dynamics with a level of detail that is not readily accessible to contemporary techniques.

1.2. Nucleic acid structure

1.2.1. Nucleotides

Nucleic acid structures are biopolymers composed of sequences of nucleotide monomers[1]. The nucleotides in nucleic acid structures contain a nitrogenous base, a pentose sugar, and a phosphate group (Figure 1.1a). In ribonucleic acid (RNA), the nitrogenous bases are adenine (A),

uracil (U), cytosine (C), and guanine (G). Deoxyribonucleic acid (DNA) contains thymine (T) instead of U. The bases are further subdivided into the bicyclic purines (A and G) and the monocyclic pyrimidines (C, T, and U). The bases are attached to the C1' of the sugars through a glycosidic linkage. The rotation of the base about the glycosidic bond gives rise to two main base orientations – *anti* and *syn*. These orientations are most easily visualized for the bulky purine bases. For example, an A nucleotide in the *anti* conformation has the base flipped away from the sugar. In contrast, the *syn* conformation corresponds to the A being flipped toward the sugar (Figure 1.1b). These orientations are less obvious for the monocyclic pyrimidines, however the base orientation can be visually identified by the proximity of the base C2 oxygen to the sugar. *Syn* pyrimidines have the C2 oxygen flipped close to the sugar, whereas *anti* pyrimidines have the C2 oxygen flipped away. The ability to interconvert between *syn* and *anti* conformations is important for adopting different nucleic acid topologies (see Section 1.2.3).

The sugars in DNA and RNA differ by one hydroxyl group. In DNA, the deoxy (d) prefix refers to the absence of the 2'-hydroxyl group on the sugar (deoxyribose) relative to RNA (ribose). Nucleotide sugars adopt puckered conformations to relieve strain and steric clash between the adjacent substituent groups[1]. Sugar puckering plays an important role in dictating the types of nucleic acid structures that a given sequence can form (see Section 1.2.2). Sugar puckers are defined by the position of the C2' and C3' with respect to the plane formed by C1'-O4'-C4'. The major sugar puckers are commonly referred to as North and South. In North puckering, the C3' is above the plane, on the same side as the base (termed *endo*), with the C2' situated on the opposite side of the sugar, away from the base (termed *exo*). In South puckering, these orientations are reversed (Figure 1.1c). Nucleic acid sequences contain multiple nucleotide monomers connected via phosphodiester linkages between the 3' and 5' hydroxyl groups of adjacent nucleotides. This

sugar-phosphate connectivity is known as the backbone of a nucleic acid sequence. By convention, nucleic acid sequences are written in the 5' to 3' direction, meaning the sequence begins with the nucleotide that has a free 5' end.

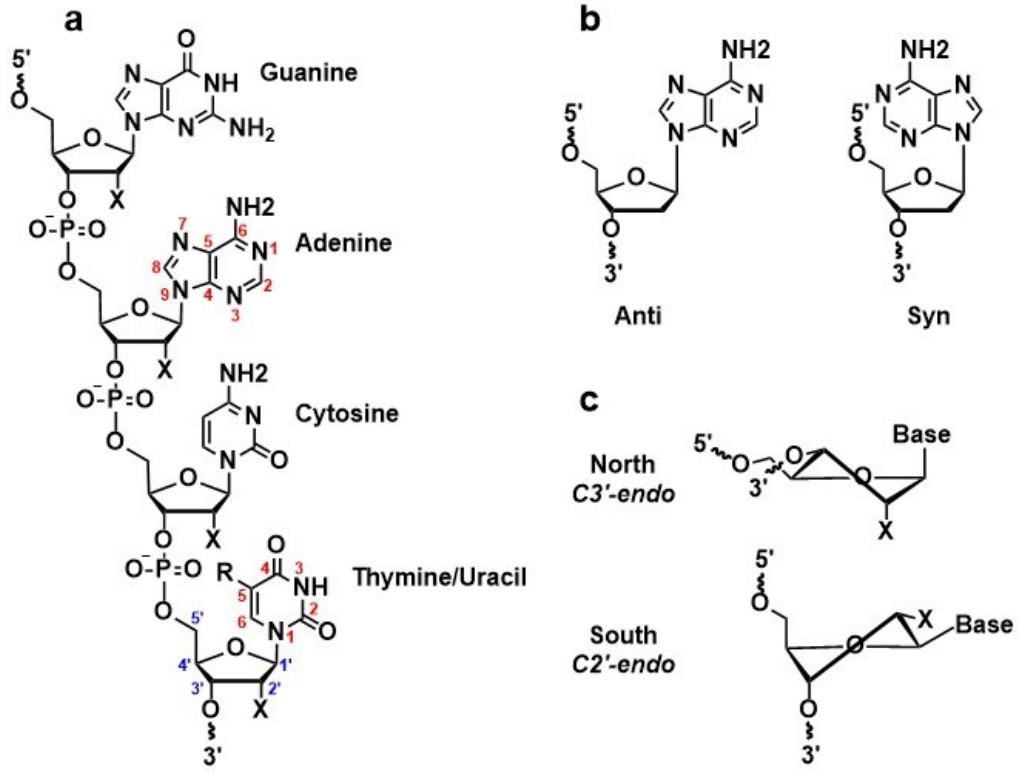


Figure 1.1. Nucleotide structure. (a) A nucleotide polymer. The base and sugar position numberings are shown in red and blue respectively. On the sugar, X = OH and H in RNA and DNA respectively. R = H and CH_3 in U and T respectively. (b) *Anti* and *syn* glycosidic bond angles shown for an adenosine moiety. (c) North and South sugar puckers.

1.2.2. The double helix

The double helix is the fundamental structure of the genome of an organism[2]. Double helices are also key structural elements acquired in the folding of single-stranded DNA[1] and RNA molecules[3] and form the basis for many nucleic acid-based nanotechnology applications[4]. In 1953, James Watson and Francis Crick published the structure of the canonical DNA double helix[2]

(Figure 1.2a, B-form) based on Edwin Chargaff's base proportions[5] and the X-ray crystallography experiments of Maurice Wilkins and Rosalind Franklin[6]. The canonical DNA double helix contains two strands arranged in an antiparallel fashion, i.e. one strand runs in the opposite direction to the other. The sugar-phosphate backbone of each strand winds around the exterior of the structure. The winding of the two strands creates the major and minor grooves[1] which are important for protein and small molecule ligand binding[7-9]. The two strands are held together by intermolecular hydrogen bonds between the bases of the two complementary sequences. In DNA, A hydrogen bonds with T, and G with C. AT base pairs contain two hydrogen bonds, while GC base pairs contain three and thus are thermodynamically more favorable (Figure 1.2b). The nitrogenous bases also form favorable stacking interactions within the duplex interior[10]. The burial of the bases inside the structure protects them from damage by endogenous mutagens[11], preserving the genetic code. As well, DNA duplex structures typically have South sugar puckers and *anti* base orientations[12]. In RNA double helices, A pairs with U instead of T, and the sugar puckers and base orientations are usually North and *anti* respectively[12]. The GC and AT arrangements found in DNA duplexes are now known as the standard Watson-Crick base pairs. The bases can also adopt the Hoogsteen (HG) arrangement[13] by using the alternate face of the purines to form hydrogen bonds with the pyrimidines. However, in the context of duplex structures, this requires an energetically unfavorable flipping of the purine into the *syn* orientation and therefore the WC mode is favored (see Sections 1.2.3 and 1.3.2 for more discussion of HG base pairing). The favorable base stacking and hydrogen bonding interactions in duplex structures are countered by repulsive electrostatic interactions between the closely situated phosphate groups[14]. In biological solutions, this effect is mitigated by cations such as Mg^{2+} that coat the sugar-phosphate backbone[15]. Additionally, the

entropic cost of duplex formation is reduced by the favorable release of water molecules to the bulk solution upon association of the two complementary strands[16].

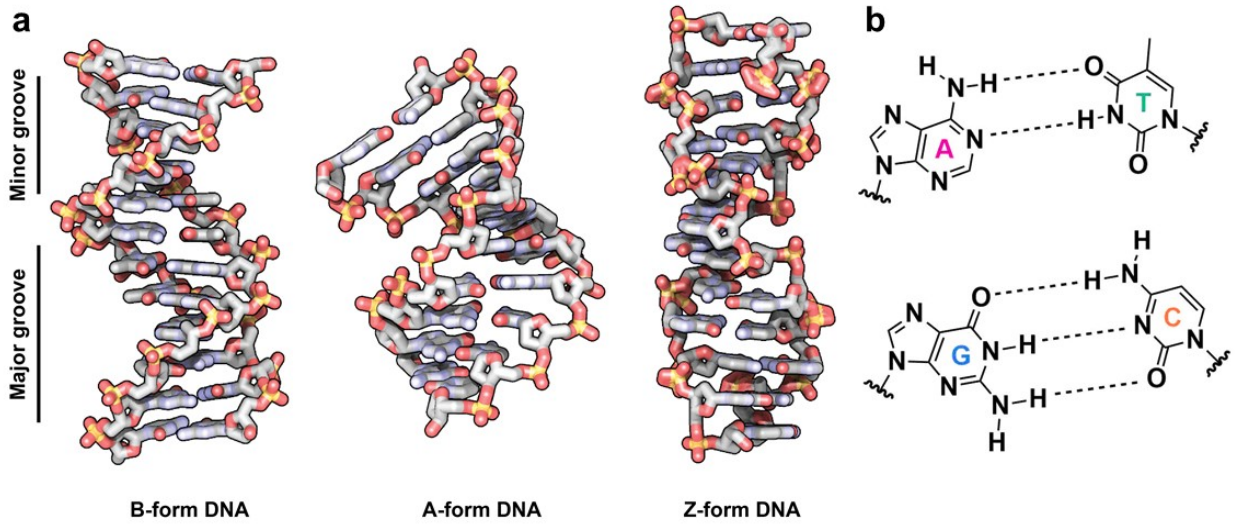


Figure 1.2. Double helix structure. (a) Three forms of double-helical DNA. The B- and A-form are right-handed helices (PDB IDs 1FQ2 and 4IZQ respectively). The Z-form is a left-handed helix (PDB ID 4OCB). (b) WC base pairing. AT and GC base pairs contain two and three hydrogen bonds respectively.

Interestingly, double helices can adopt multiple forms that are implicated in biological function. Three of these double-helical conformations in DNA are the B- A- and Z-forms[17-19] (Figure 1.2a). The B- and A-forms are right-handed double helices, with South and North sugar puckers respectively[20]. The Z-form is a left-handed double helix typically adopted by duplexes of alternating South pyrimidines and North purines that drive the formation of the zig-zag backbone[21]. RNA double helices usually adopt the A-form, since the presence of the 2'-hydroxyl group makes it energetically unfavorable to adopt the South pucker in the B-conformation[12]. The B-form of DNA solved by Watson and Crick is thought to be predominant in biology as it is energetically most favorable. The A-form of DNA is favored under dehydrating conditions and

protects DNA in response to desiccation[22]. Z-DNA occurs near transcription start sites and is thought to act as a cis-element for gene expression[21]. This is supported by a number of transcription factors that bind to and stabilize the Z-DNA conformation[23]. Double helical DNA and RNA can also accommodate bulges[24] and hairpins[25] where local portions of the duplex are single stranded, revealing binding sites for protein partners[11, 26]. The exposure of single strands may also lead to the formation of G-quadruplex and i-motif structures, as discussed in the following two sections.

1.2.3. G-quadruplexes

G-quadruplexes (GQ) are four-stranded, helical nucleic acid structures formed by G-rich DNA and RNA sequences[27-29] (Figure 1.3a, left). Biologically-occurring GQs typically form from single strands containing four G-repeats (termed G-tracts) that are connected by variable sequences (termed loops). These usually follow 5'-$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$-3'[30] where N is A, C, G, T, or U, although divergent sequences can also adopt a GQ fold[31, 32]. For example, sequences with longer (G_{4-7}), shorter (G_2), or uneven G-tracts also form GQ structures[33-35]. GQs can also form from multiple strands via intermolecular interactions (Chapter 4). GQ sequences are found throughout the genome and in mRNA, notably in gene promoters, telomeres, and telomeric or virus RNA[29, 36-38]. GQs are thought to form when single strands are exposed during DNA replication and transcription events, and have been implicated in regulation of gene expression, protein translation, and proteolysis[39-43], pointing to wide-ranging roles in biological function. GQs are also important in biotechnology, forming a key structural component of aptamers[44, 45] and catalytic DNA[46, 47], among other applications.

The canonical four G_3-tract GQ is formed by assembling the G-tracts into three stacked G-tetrads while adopting energetically favorable loop interactions[48, 49] (Figure 1.3a, left). G-tetrads

are planar arrangements of four Gs engaged in HG hydrogen bonding interactions, where each G in the tetrad is rotated roughly 90° with respect to the adjacent ones (Figure 1.3a, right). The second and third G tetrads are rotated relative to the first. This rotation gives rise to the helical nature of GQs, which are almost always right handed, with interesting exceptions[50, 51]. Loop sequences can form intra- and inter-loop interactions[52, 53], in addition to loop-core arrangements where longer loops form capping structures and stack onto outer G-tetrads[54, 55]. GQs may also feature flanking sequences to the 5'- and 3'-ends that form stabilizing capping interactions with the GQ core[36, 56] (Figure 1.3a, left).

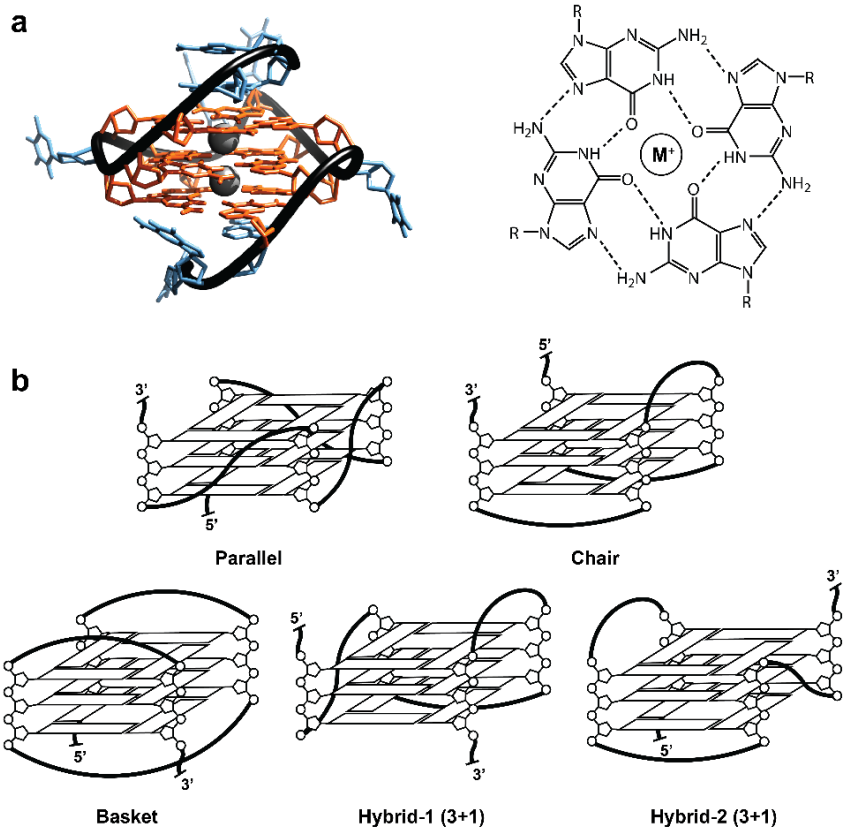


Figure 1.3. GQ structure. (a) Solution NMR structure of the c-myc Pu22 sequence 5'-TGAGGGTGGGTAGGGTGGGTAA-3' (PDB ID: 1XAV) with G-tetrad and coordinated metal depicted immediately to the right. In the 3D structure, core G residues are colored orange, loop and flanking residues are colored blue. Bound K^+ ions are colored grey and the backbone is colored black. (b) GQ topologies. Reproduced from Harkness and Mittermaier[57] with permission.

GQ assembly requires suitable cations that are bound between each pair of tetrads, coordinated by the carbonyl groups of the tetrad Gs[58] (Figure 1.3a left). A variety of cations can play this structural role including Ca^{2+}, Pb^{2+}, Sr^{2+}, Na^{+}, and K^{+}[59-63], although the last two are considered most biologically relevant as their intracellular concentrations (~10 mM and 140 mM in mammalian cells respectively[64]) are much larger than those of the others. Typically K^{+} is bound with highest affinity[65, 66]. Interestingly, NH_4^{+} is readily coordinated in the GQ core and $^{15}NH_4^{+}$ has been used to probe cation dynamics within GQs by NMR spectroscopy[67, 68]. Cations may also be coordinated by loop sequences and can play a role in loop dynamics[69, 70].

GQs fold into different topologies that are characterized by the relative orientations of the four G-tracts and the types of loop motifs that connect them[27, 71] (Figure 1.3b). G-tracts can be aligned in the GQ core in the same or opposing directions, known as parallel and antiparallel respectively. When all four G-tracts are aligned in the same direction, the GQ is parallel or 4 + 0 topology (4 parallel + 0 antiparallel G-tracts, Figure 1.3b parallel). GQs with G-tracts aligned in alternating opposing directions are antiparallel or 2 + 2 topology (2 parallel + 2 antiparallel G-tracts, Figure 1.3b chair, basket). Topologies where 3 of 4 G-tracts are parallel while the 4th is antiparallel are called hybrid or 3 + 1 (3 parallel + 1 antiparallel G-tract, Figure 1.3b hybrid-1 and hybrid-2). Loops that connect parallel G-tracts are termed double-chain-reversal or propeller type (because they appear similar to a propeller blade, Figure 1.3b parallel) and are typically formed by short loop sequences of 1-2 nucleotides in length[49, 72-74]. Edge-wise (Figure 1.3b chair) and diagonal (Figure 1.3b basket) loops connect antiparallel G-tracts. These types of loops run along the edge or bridge the diagonal of the top or bottom face of the GQ and are typically formed by longer loop sequences of 3+ nucleotides in length[75, 76]. GQ topologies can also be differentiated by the glycosidic bond angles for Gs within the GQ core. A parallel GQ has all

anti glycosidic angles, while an antiparallel or hybrid GQ has mixed *syn* and *anti* glycosidic bond angles[27]. The determinants of adopting one topology over another are not currently clear, although G-tract length, loop length and composition, and formation of favorable loop interactions have been implicated in influencing topological preferences[33, 49, 77].

1.2.4. i-Motifs

i-Motifs are four-stranded structures formed by C-rich DNA sequences[37, 78] (Figure 1.4a). Their structure consists of two parallel-stranded duplexes intercalated in an antiparallel orientation[79]. The formation of an i-motif is coupled to the protonation of the Cs at the N3 position, which allows the core of the structure to be stabilized by unusual hemi-protonated CC^+ base pairings (Figure 1.4b). In theory, an i-motif can form in any region of double-stranded genomic DNA where a GQ-forming sequence is located, since the complementary strands at these loci have tracts of C residues[80]. However, unlike GQs, i-motifs have garnered less attention for their potential roles in biological function owing to their sensitivity to the solution pH[78]. For example, the C N3 pK_a is roughly 4.5[78] which corresponds to the protonation equilibrium lying approximately 200-fold to the deprotonated side at neutral biological pH. At face value, this suggests that i-motif structures are not easily formed *in-vivo*. Yet, i-motif structures adopted by sequences from the promoter regions of oncogenes have been shown to be stable at neutral pH[78, 81], suggesting i-motifs may take part in regulating transcription. The discovery of several proteins that bind to i-motif sequences has provided further evidence for their role in biology[82-84]. Despite the apparent barrier to their formation in-vivo, the pH sensitivity of i-motifs has made them a key element in nanotechnology applications as pH sensors[85, 86].

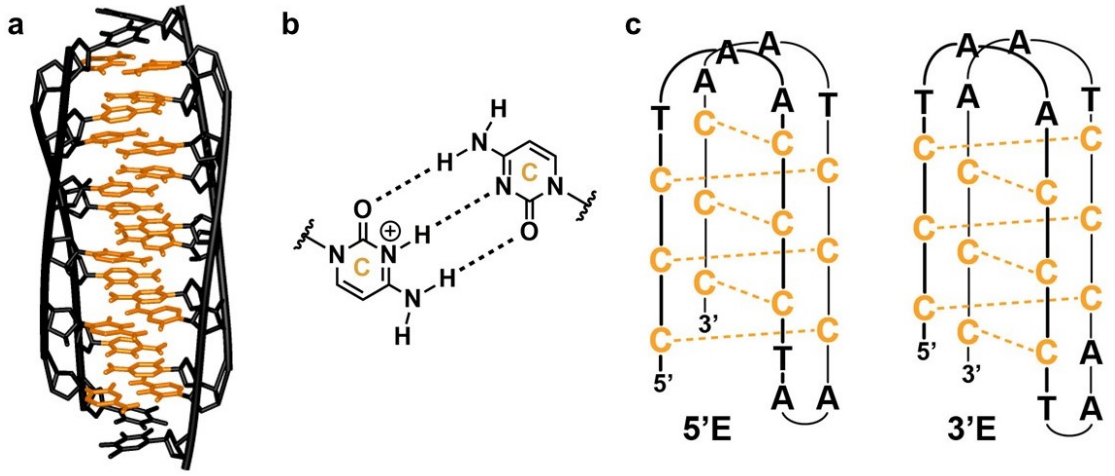


Figure 1.4. i-Motif structure. (a) A tetramolecular i-motif (PDB ID 2N89). (b) CC^+ base pairing. (c) Intercalation topologies for the telomeric i-motif structure.

The stability of i-motifs is also highly dependent on the length of the C-tracts and loops, and loop composition[87, 88]. The Waller group has shown that a minimum of 5 Cs per tract are necessary to have stable i-motif formation under physiological conditions[88]. Presumably, this is due to the intercalation pattern of i-motifs that precludes extensive stacking interactions between the aromatic heterocycles of consecutive bases in the core structure[78]. Interestingly, this detrimental effect is offset by the strength of the CC^+ base pairs, which have stronger hydrogen bonding interaction energies compared to their GC counterparts[89]. Additional stabilization comes from a hydrogen bonding network between the sugars of the backbone[90]. Like GQs, i-motifs can adopt different topologies. These are known as the 5'E and 3'E forms[91]. i-Motifs in the 5'E topology have the outermost CC^+ pair at the 5' end of the sequence. In contrast, the 3'E topology has the outermost CC^+ pair at the 3' end (Figure 1.4c).

Abou Assi *et al.* have demonstrated that i-motifs can be formed at neutral pH and physiological temperatures by introducing 2'-arabinofluoro modifications[92, 93], allowing their study in model systems that mimic their biological context where a GQ, i-motif, and duplex are

putatively in equilibrium. In a recent landmark study toward elucidating the formation propensity of unmodified i-motifs *in-vivo*, efforts by Trantirek and coworkers have shown that pre-formed i-motif structures are stable in the nuclei of living mammalian cells by in-cell nuclear magnetic resonance (NMR) spectroscopy[94]. They suggested that i-motif formation is stabilized by an excluded volume effect derived from the crowded biological milieu. In this environment, the C N3 pK_a could also be highly perturbed from its *in-vitro* value, making i-motif formation feasible at biological pH. The identification of stable i-motifs *in-vivo*, combined with the ability to study these structures in simulated biological environments points to a bright future for i-motif research.

1.2.5. Aptamers

Nucleic acid aptamers are single-stranded DNA or RNA molecules that are developed to bind with high specificity and affinity to a target ligand[95]. The term aptamer is derived from the Latin word *aptus* (to fit) and the Greek word *meros* (part), coined to reflect the binding specificity of these structures. To date, aptamers that can bind to ligands as small as metal ions, or as large as proteins or complete viruses have been identified[96-98]. Aptamers can adopt a wide range of structural motifs such as double-helical stem loops and pseudoknots, bulges, and GQs[95, 99]. Local secondary structures within an aptamer strand acquire tertiary contacts by folding upon themselves or with elements formed by distal parts of the sequence. The formation of tertiary contacts is important for generating the shape complementary that imbues ligand specificity[95]. The ability to engineer ligand binding has generated significant interest in using aptamers for sensing and biomedical applications. Synthetic aptamers made from DNA and RNA are used to detect drugs in colorimetric assays and have found success as medical diagnostic tools and treatments[100-102]. For instance, an aptamer based on a GQ scaffold is in clinical trials as an anti clotting agent in

heart surgery[95] owing to its picomolar affinity for thrombin[103], an enzyme involved in the blood clotting process. GQ-based aptamers have also been applied in metal sensing[104-106], since the folding of GQs is coupled to coordination of multiple metal ions in the central channel.

Typically, aptamers are developed to bind a single ligand, ensuring a specific response in sensing applications. However, aptamers can develop off-target interactions that complicate their usage. In an effort to understand the structural bases of specificity and high affinity DNA aptamer:ligand interactions for drug sensing applications, the Johnson group has employed a dual NMR and isothermal titration calorimetry (ITC) approach in studies of the cocaine-binding aptamer and libraries of its sequence variants[107-111] (Figure 1.5a). Interestingly, the cocaine-binding aptamer interacts with quinine roughly 30- to 40-fold more strongly than cocaine, the ligand for which it was initially selected to bind. The capacity to bind two structurally different ligands makes the cocaine-binding aptamer a useful model system for understanding the determinants of ligand binding promiscuity. The cocaine-binding aptamer features three double-helical stems with WC base pairing that are connected by a three-way junction. The aptamer junction core contains an internal bulge situated next to tandem GA mispairs where cocaine and quinine bind[108]. The Johnson group has suggested that quinine binding is driven by extensive base stacking interactions between the bicyclic aromatic ring and the core of the aptamer. The monocyclic ring of cocaine likely does not stack as efficiently into the aptamer core, leading to reduced affinity. Engineering aptamers with core structures that tightly accommodate the smaller cocaine ring could therefore lead to increased cocaine affinity and total specificity for more rigorous application in drug sensing.

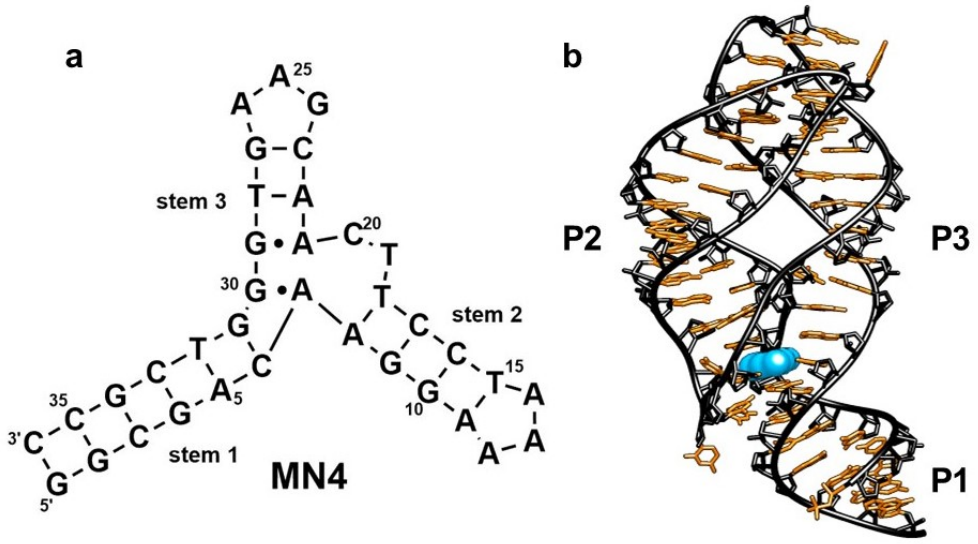


Figure 1.5. Synthetic and biological aptamer structure. (a) The cocaine-binding aptamer MN4. Cocaine and quinine bind at the three-way junction. Reproduced from Harkness *et al.*[112] with permission. (b) 3D structure of a G riboswitch from the soil bacterium *Bacillus subtilis* (PDB ID 1Y27). The bound G is shown in blue.

Aptamers initially gained attention in 1990 for their ease of development by *in-vitro* procedures[113-115]. More recently, it was discovered that aptamers were present in biology long before methods to generate synthetic aptamers were brought to bear. In the early 2000s, several groups discovered the existence of naturally occurring aptamer structures called riboswitches that are adopted by mRNA sequences[116-119]. Riboswitches have evolved to interact with small molecule metabolites in the cell in order to regulate gene expression. There are riboswitches that bind to nucleotides, vitamins, amino acids, and even fluoride anions[120]. Typically, riboswitch structures occur in the 5'-UTR of bacterial mRNA where they act as cis-regulatory elements for mRNA translation by sensing metabolite levels[3]. For instance, in the absence of a given metabolite, certain riboswitches adopt a conformation that conceals the mRNA translation start site. Upon metabolite binding, the riboswitch structure reorganizes to expose this motif. Like other large RNA structures,

riboswitches are organized according to two general structure principles[3]. These are (i) the formation of coaxial helical stacks, and (ii) the parallel packing arrangement of distal helices and helical stacks. As an example, the G riboswitch (Figure 1.5b) contains P1, P2, and P3 helices. The P1 and P3 helices form a nearly linear coaxial stack that is packed against the P2 helix in the 3D structure[121]. Riboswitches are further organized by parallel packing of GA_3 tetraloops and their receptor motifs[3, 122]. These structural arrangements facilitate the formation of metabolite binding cavities on riboswitch exteriors. In the G riboswitch, G binds at the three-way junction formed by the P1-P3 helices. By comparison with the cocaine-binding aptamer, this demonstrates how aptamers of completely different origins (synthetic vs. biological) can remarkably adopt similar ligand docking elements.

1.2.6. Higher-order nucleic acid assemblies

As described in the previous sections, nucleic acid sequences can adopt a variety of complex structures that are ultimately encoded by their primary sequence. Over the last roughly 40 years, scientists have exploited this property in order to generate novel, higher-order (multimeric) nucleic acid structures with pre-defined properties[4, 123]. The field of nucleic acid nanotechnology has developed from the ability to control rigidity, structure, and binding interactions by programming the primary nucleic acid sequence. These efforts have been greatly aided by the development of modern automated nucleic acid synthesizers that permit the facile and large-scale production of user-defined nucleic acid sequences. Furthermore, employing chemical modifications and developing methods to systematically guide the formation of novel folds has significantly expanded the structural diversity of nucleic acids[124-127]. Nucleic acids have become

the basis for many widely-used nanotechnology applications, where a diverse array of higher-order structures and functionalities have been achieved beyond those strictly found in nature[128].

DNA origami is one of the best-known examples of user-control over higher-order nucleic acid structure formation. The seminal paper by Paul Rothemund[129] describes how a single, several kilobase long strand of DNA from a bacteriophage can be folded into shapes such as squares, stars, and triangles using hundreds of short "staple strands" that form local double-helices with the longer strand at specified positions via WC base pairing interactions. The staple strands cause the longer strand to bend at defined locations in its sequence, resulting in a macroscopic pattern that can be visualized by atomic force microscopy (AFM). Importantly, this can be performed in a single one-pot step where all sequences are simultaneously mixed. The structures produced in this manner can be used as scaffolds for molecular electronics and circuits[129], among many other applications[130]. Complex DNA architectures with unique functionalities can additionally be generated by iterative addition of their components. In these cases, single strands or pre-formed nucleic acid motifs are added one at a time[131, 132] to yield the desired structural element without competition from other potential binding partners. The structure that forms can then be purified or the excess strands washed away for a new round of assembly. In this manner, structural components such as triangular double-helical "rungs" with free sticky ends can be installed into larger architectures (Figure 1.6a). Notable examples of this sequential approach include the assembly of high aspect ratio DNA nanotubes[124, 133, 134] and cages[131] for controlled delivery and release of drugs and gold nanoparticles.

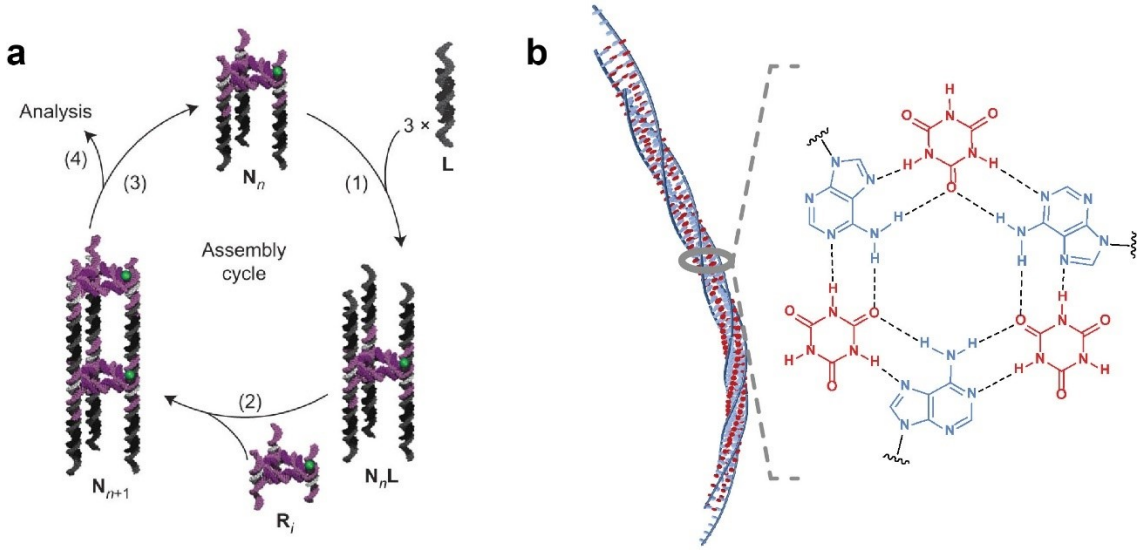


Figure 1.6. Supramolecular nucleic acid assemblies. (a) Step-wise assembly of DNA nanotubes. DNA rungs and linking duplexes are added sequentially to generate nanotubes of arbitrary length. Adapted from Hariri *et al.*[124] with permissions. (b) A DNA fiber (left) formed by the co-assembly of CA and poly(dA) strands into stacked, hexameric rosette hydrogen bonding motifs (right).

Unmodified nucleic acid sequences may also be coaxed into forming higher-order assemblies without the addition of extra strands in one-pot or sequential protocols. Instead, small molecules can be used to initiate the assembly of macromolecular structures from single-stranded nucleic acids. In a recent report by Avakyan *et al.*, strands of poly(A) were discovered to form fibers of roughly one micron long in the presence of cyanuric acid (CA)[135]. Both DNA and RNA sequences were found to adopt these fiber structures. The core structure of the fibers consists of hexameric "rosette" hydrogen bonding arrangements, facilitated by CA. Each rosette contains three A bases (each from a separate poly(A) strand), interspersed by three CA molecules. The rosette faces are hydrophobic and participate in extensive stacking interactions within the core of the fibers which drives their assembly (Figure 1.6b). The discovery of self-assembling nucleic acid fibers has important implications for novel biological structures, since they suggest a hidden layer

of complexity in the nucleic acid code. For example, fiber structures could be derived from the poly(A) tails in mRNA in the presence of intracellular small molecules[135]. Intramolecular assemblies guided by small molecules could also form in tracts of chromosomal poly(A) DNA, much like how GQs fold from single strands in the presence of cations.

1.3. Nucleic acid dynamics

1.3.1. Hierarchical motions in nucleic acids

The static molecular pictures presented in the previous sections are highly informative about the general aspects of nucleic acid structure. However, these views belie one fundamentally important feature of nucleic acids – their structures evolve over time, leading to highly dynamic molecular ensembles. These conformational dynamics occur over a range of length and timescales that are important for nucleic acid function[11, 136-139]. Nucleic acid dynamics are also strongly sequence-dependent, meaning that one portion of a given structure may experience drastically different motions than another[140-142]. An upstream sequence can have unique propensities to bend, twist, and open relative to a downstream sequence, for example[140, 143, 144]. Nucleic acid dynamics can also be cooperative, where the dynamics at one site influence dynamics at distal sites[35]. The collective motions available to even short duplex elements are potentially daunting yet, much like protein dynamics[145], nucleic acid motions can be more easily understood by organizing them into a length and timescale hierarchy containing macro and microscale dynamics clustered according to slow (μs – ms+) and fast (ps – ns) timescales[136]. In this view, many iterations of a fast timescale process occur within one iteration of a slower timescale conformational excursion. It is important to note that these timescales of motion are strongly system- and structure-dependent, and therefore

the following groupings are general descriptions that serve to illustrate the hierarchical dynamics in nucleic acids.

On a macro scale, cellular double-helical DNA is coiled into nucleosomes by complexing with histone proteins[16]. The nucleosomes are packaged together to yield massive DNA:protein fibers. These fibers are further assembled into chromosomes of roughly 1400 nm in width[146] (Figure 1.7). The successive wrapping and compaction allows the roughly two meters of chromosomal DNA per cell to fit inside the cell nucleus that is only approximately 6 microns in diameter. By analogy, this corresponds to packing 40 kilometres of fine thread into a tennis ball[147]. Furthermore, chromosome compaction varies as a function of the cell cycle; chromosomes are most compacted during metaphase when they can be seen with a light microscope and are more loosely wound during interphase[148]. The organization/reorganization of DNA at this level requires an enormous amount of conformational change and therefore occurs on the slower timescale over a matter of ms to minutes[137, 148, 149]. Although the sheer magnitude of the compaction ratio that cellular DNA experiences is highly impressive, this packaging poses a problem, since the bending of duplex DNA into tightly-packed chromosomes can preclude access to sections of the genetic code that need to be expressed at a particular time. To circumvent this, DNA also locally unwinds from histone proteins while in the context of nucleosomes and globally-compacted chromosomes in a sequence-dependent manner[150], typically on the ms timescale[149]. The ability of DNA to safely and reversibly undergo these tremendous deformations is critical to the integrity of the genome and how it is read by molecular machinery.

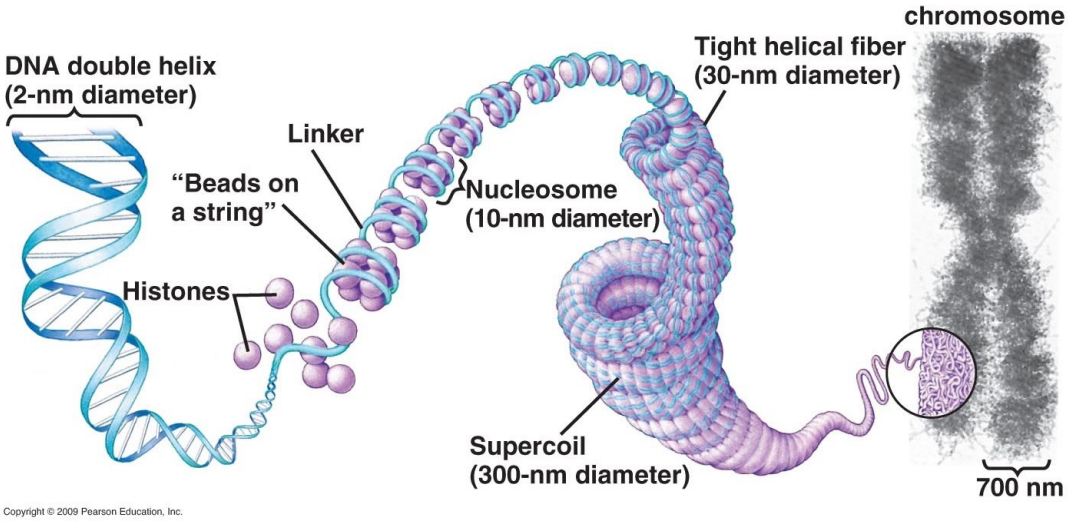


Figure 1.7. Macroscale organization of DNA into nucleosomes and chromosomes. Reproduced from Campbell "Essential Biology with Physiology" Chapter 8: Cellular reproduction.

On the microscopic scale, DNA and RNA molecules undergo dynamics at the level of their tertiary, secondary, and primary structure[57, 136, 139]. These dynamics are important for recognition by DNA- and RNA-interacting proteins[11], among other functions[120, 151]. Motions at this level can be grouped into both slow (μs to ms+) and fast (ps to ns) timescales. The slow timescale motions consist of tertiary and secondary structure dynamics, the formation of non-standard base pairing arrangements, and duplex "breathing" events, where base pairs transiently open via disruption of stacking and hydrogen bonding interactions[11]. Duplex breathing creates transient local single-stranded regions which contain one or more bases. The lower stability of AT/U base pairs (two hydrogen bonds) predisposes them to breathing with a greater propensity than the more stable GC pairings (three hydrogen bonds)[152-155]. This can lead to locally "melted" portions of duplex structures surrounded by GC-rich regions[11]. The strength of base stacking interactions also influences breathing events. For example, pyrimidine-pyrimidine steps have weaker stacking interactions than purine-purine steps[156] and are thus more flexible. Furthermore, the transient opening of double-helical elements is critical to the adoption of other secondary structures. The

release of longer sequences in DNA can lead to the formation of GQs and i-motifs in the case of G- and C-rich strands[80, 157], for instance. Moving to more rapid dynamics, the fast timescale motions include torsional sampling by the sugar-phosphate backbone[158], conformational exchange between sugar puckers (North-South equilibria)[141], and interhelical motions[136]. One of many highly interesting cases of faster timescale motions in nucleic acids is given by the collective helix dynamics of the human immunodeficiency virus (HIV)-1 transactivation response (TAR) element RNA (Figure 1.8a). TAR is required for HIV replication and for this reason is a major drug target. TAR exists as an ensemble of two helical elements with variable intervening bend angles. The bending and twisting motions of the two TAR helices are correlated and these dynamics give TAR "directional flexibility" which ligands use to bind in a "tertiary capture" mechanism to specific pre-existing helix orientations[159, 160] (Figure 1.8b).

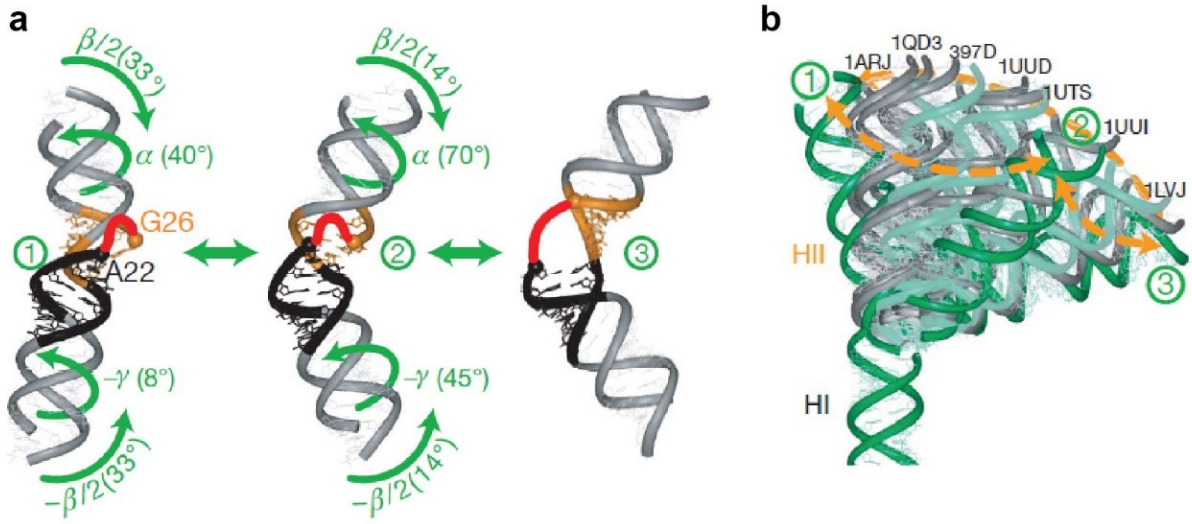


Figure 1.8. Correlated helix motions in HIV-1 TAR RNA. (a) Helix I (HI, black) and Helix II (HII, orange) exist as an ensemble of three conformations with correlated bending and twisting angles. (b) The ligand-bound forms of TAR (grey) are pre-existing conformers within the free TAR ensemble (green). Adapted from Zhang *et al.*[160] with permissions.

The Al-Hashimi group has been at the forefront of elucidating the motions in nucleic acid structures at atomic resolution[159-165], which are particularly amenable to visualization by NMR spectroscopy. In a 2011 NMR dynamics report, they revealed sequence-dependent motions over both the fast and slow timescales that modulate the propensity for B-DNA to adopt the Z-conformation[166]. They studied the solution B-form of a 15-mer duplex, Z-JXN, that, in the presence of a Z-DNA stabilizing protein, crystallizes into a B-Z junction; a duplex containing both the B- and Z-forms of DNA. Peculiarly, an AT base pair is extruded at the interface of the B- and Z-portions of Z-JXN in its crystal form, which allows the B- and Z-segments to stack into a continuous double-helical element. Moreover, the Z-portion of the sequence extends beyond its CG repeats to include a CC step which is predicted to be highly energetically unfavorable[167]. The B-Z interface thus represented an interesting target for investigating local DNA dynamics and their role in the B-Z interconversion. Al-Hashimi and coworkers found that the base pairs in and around the B-Z interface in the B-form of Z-JXN exhibited conformational dynamics consistent with the disruption of base pairing and stacking interactions. In particular, the base and the sugar of the extruded A and T respectively were found to be highly dynamic, pointing to an underlying propensity for Z-DNA formation. Furthermore, the A bases at or near the junction exhibiting conformational exchange were part of CA and TA steps, which are known to have weaker stacking interactions relative to AA or GG steps[156] and are therefore more flexible. They further examined the role of the CG repeats that adopt the Z-DNA conformation in the crystal structure by comparing the B-form Z-JXN duplex with a sequence variant where the CG repeats were disrupted. NMR measurements of the disrupted variant revealed decreases in dynamics at residues several base pairs away in the B-Z interface of the sequence. This suggested that the CG repeats modulate the dynamics of their neighboring sequences. Taken together, their results highlight how nucleic acid

dynamics are encoded by the primary sequence and reveal the influences that distal sequence elements can have on the dynamic characteristics of other portions of nucleic acid structure.

1.3.2. Transient high energy base pairings in duplex DNA and RNA

The structure of the DNA duplex proposed by Watson and Crick was remarkable owing to its complementary GC and AT base pairings, which immediately suggested a mechanism for genetic replication[2] – one strand dictates the sequence of the other. Watson and Crick's structure used the four nucleobases in their most probable tautomeric forms, i.e. keto rather than enol. These arrangements gave similar shapes to GC and AT base pairs, allowing any sequence to be easily accommodated within the double-helix. While captivating, the WC structure was met with substantial skepticism as evidence for alternative base pairing conformations in double-helical nucleic acids began to accumulate[13, 168]. For example, the observation that A and T derivatives frequently crystallized into HG base pairs instead of WC cast considerable doubt over the nature of base pairing in the double helix. Over time, it has become clear that nucleic acid structures do not rely solely on WC base pairing. In fact, WC and non-WC base pairs are now known to coexist within DNA and RNA, each with unique structural and functional roles[136, 161-163]. These include HG base pairing in DNA duplexes, base pair reshuffling and isomerization in RNA, and nucleobase tautomerizations (keto-enol equilibria) or ionizations that yield WC-like mispairs. These non-canonical base pairs are often transient and high energy, populated by undergoing an energetically unfavorable structural change from the lower energy WC state. The transient adoption of non-canonical high energy base pairs has important implications for propagating the information contained within nucleic acids because they change the local properties of the double helix[169] (among other outcomes, see below), thereby influencing interactions with DNA-binding

proteins. Transient, high-energy base pairs thus add a deeper layer of complexity to the genetic code, which we are just beginning to understand. This section will discuss the recent discovery and experimental characterization of transient HG base pairs and rare tautomers in duplex DNA and RNA as an exciting new example of nucleic acid dynamics.

Karst Hoogsteen initially identified HG base pairs in crystals of 1-methylthymine and 9-methyladenine in 1959[13], beginning a search for the role of this hydrogen bonding pattern in duplex structure. Several subsequent crystallographic reports identified DNA duplexes that formed HG base pairs[169-171], yet investigations of the same duplexes by solution NMR found them to contain only the WC mode[169, 172]. This suggested that HG base pairs were merely an artifact of crystal packing. Yet, HG base pairs were also found in complexes of duplex DNA with antibiotics[173] and proteins[169, 174, 175], pointing to a potential role in regulating DNA transactions. It was not until 2011 that the WC base pairs within naked, canonical duplex DNA were revealed to be in dynamic equilibrium with their HG counterparts by solution NMR spectroscopy[161]. In these duplex excursions, the purine rotates roughly 180° around the glycosidic bond to adopt the syn conformation (Figure 1.9a). In doing so, the purine uses its so-called HG face to form hydrogen bonds with the pyrimidine. This constricts the diameter of the double-helix around the HG site and increases the local negative charge density of the backbone. The increase in negative charge density is perhaps balanced in part by the formation of GC HG base pairs, which are coupled to protonation of the C N3. The protonation event makes GC HG base pairs sensitive to the local pH environment in an analogous manner to i-motif folding. In contrast to the GC transition which has a net loss of one hydrogen bond, AT HG base pairs maintain the same number of hydrogen bonds as in their WC form (two). Nikolova *et al.* discovered that HG base pairs occur with populations of ~0.1-2.7% and lifetimes of ~0.1-2.6 ms depending on the nearest-neighbor sequence[161, 176].

Remarkably, they found that the transition from a WC to a HG base pair is accompanied by an activation free energy of similar magnitude to a duplex breathing event, suggesting that HG base pair formation could be coupled to transient duplex opening. Moreover, the relatively small unfavorable free energy difference between the HG and WC states (~3 kcal mol^{-1}) explained how HG base pairs could be observed in certain experimental conditions, but not others (as described above). The outcomes of their studies have far reaching implications, since the sequence-dependent adoption of a single HG base pair in a sea of WC DNA provides a unique marker for recognition by transcription factors. Furthermore, oxidative lesions are known to trap HG base pairs, giving repair enzymes an easily identifiable motif against the background of WC base pairs.

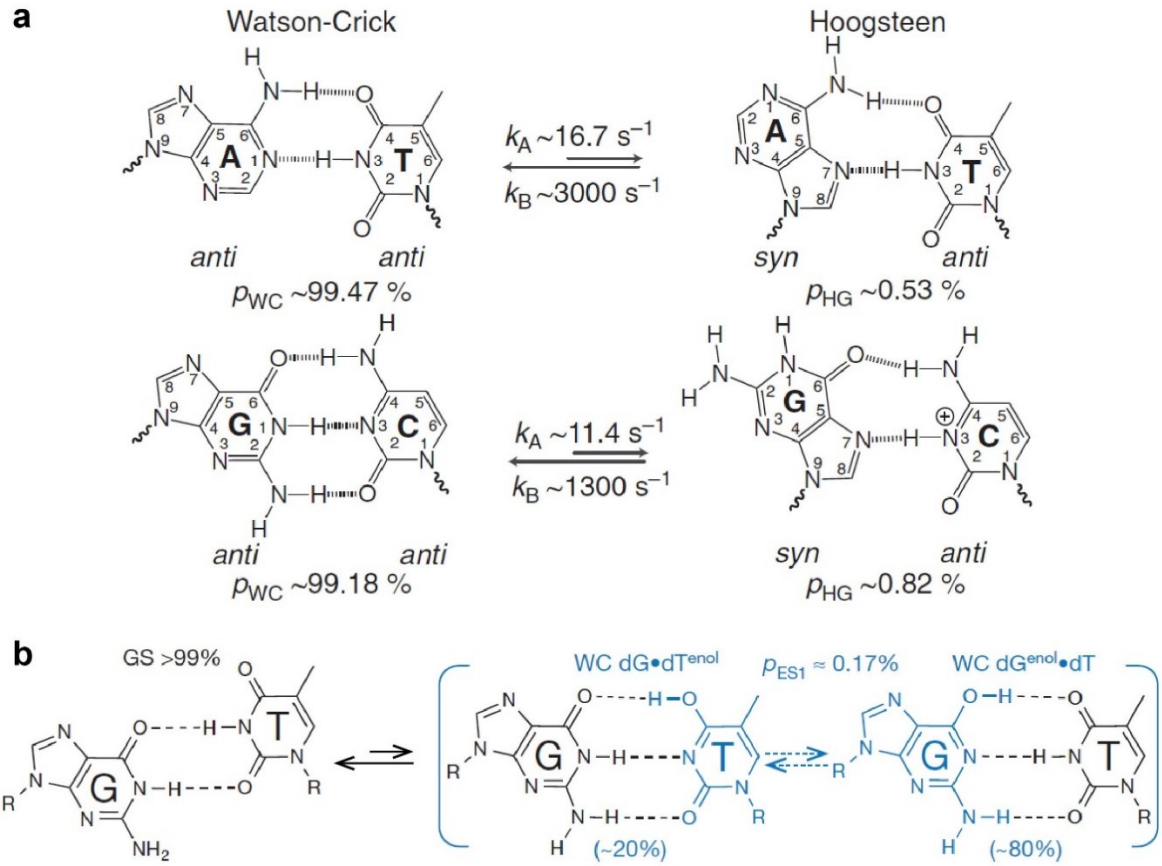


Figure 1.9. Transient high energy base pairs in duplex DNA. (a) Transient HG base pairs. WC base pairs in duplex DNA are in dynamic equilibrium with the HG arrangement which requires rotation of the purine into the *syn* conformation. The adoption of HG base pairs is energetically unfavorable and therefore the equilibrium lies nearly 100-fold to the WC side. Adapted from Alvey *et al.*[176] with permission. (b) Transient WC-like mispairs. GT wobble pairs tautomerize to their ~0.2% populated enol form that is a rapidly equilibrating 80:20 mixture of enol G paired with T and G paired with enol T. Adapted from Kimsey *et al.*[163] with permission.

Beyond HG base pairing, the nucleobases in duplex nucleic acids are further capable of undergoing an energetically unfavorable tautomerization from their more stable keto states to their enol versions. These rare tautomers are thought to be responsible for spontaneous mutations and translation errors[177], since they permit WC-like geometries between non-canonical GT/GU base mismatches, thereby fooling the checkpoint mechanisms of polymerases[163]. The role of base

tautomers in biology has remained elusive however, since they are experimentally challenging to detect and have only been observed in a handful of crystal structures[178, 179]. Recently (2015/2018), the Al-Hashimi group discovered that GT/GU pairings which normally adopt a non-WC "wobble" conformation with two hydrogen bonds in duplex DNA and RNA can transiently tautomerize, yielding WC-like mispairs stabilized by three hydrogen bonds[162, 163] (Figure 1.9b). The tautomerization process includes deprotonation of either a G N1 or T/U N3 with concomitant protonation of the G O6 or T/U O4 and sliding of the bases into the WC-like geometry. Kimsey *et al.* found that these WC-like mispairs in DNA and RNA occur with populations on the order of 0.1-0.2% and lifetimes of 0.2-0.4 ms at pH 6.9. A second chemical exchange process from the ground state GT/U wobble pairs was also identified, where the T/U N3 becomes deprotonated to form anionic GT/U mispair arrangements. These were characterized by populations of ~0.04 and 0.2% at pH ~8 in DNA and RNA respectively. Interestingly, computational analysis of the GT/U tautomer equilibrium further revealed that the ~0.1-0.2% WC-like mispair is a rapidly equilibrating weighted average of G paired with enol T/U and enol G paired with T/U. Toward understanding the presumed role of GT/U mispairs in polymerase errors, Kimsey *et al.* compared the results of their analysis with the literature values for polymerase nucleotide and amino acid misincorporation probabilities, finding that these frequencies track with their NMR-derived populations. Taken at face value, this suggested that competition from WC-like mispairs is a key determinant of misincorporation at the levels of replication and translation. Combined with the recent identification of transient HG base pairing in naked duplex DNA, WC-like mispairs add a rich layer of complexity to the genetic code and emphasize how we are still gaining new insights into the dynamic characteristics of even exceptionally well-studied nucleic acid structures like the double helix. Since the functional roles of duplex dynamics (in the context of polymerase active

sites, for example) remain largely unexplored, there is still much to understand regarding the structural diversity afforded by the relatively small set of four nucleobases that comprise canonical nucleic acid sequences.

1.3.3. GQ dynamics

1.3.3.1. An overview of GQ conformational exchange

In the case of G-rich primary sequences, the dynamical repertoire available to nucleic acids can be greatly extended by folding into GQ structures. Intriguingly, many GQ sequences can exchange among multiple folded conformations of similar energies, producing a highly dynamic folded structural ensemble. The potentially large number of energetically similar conformational states for an individual GQ sequence creates an interesting landscape for biological and biophysical study. Despite the existence of a rich literature on GQ structure and stability, their dynamics are not well understood. GQ dynamics are challenging to study by standard biophysical techniques. For example, solution NMR spectra of dynamic GQs can be broadened beyond identification of any clear resonances[35, 74, 180], preventing structural characterization and application of standard NMR dynamics experiments. In addition, studying GQ folding with thermal melting experiments often yields multiple or broad transitions[181-183] which require application of sophisticated statistical thermodynamic approaches to appropriately extract thermodynamic information[184-187]. For these reasons, it is experimentally attractive to suppress GQ dynamics with nucleotide substitutions[49, 56, 74, 142], or to study GQs that naturally populate a single ground conformation. In many cases, the properties of lesser-populated states in GQ ensembles are entirely ignored, meaning that their roles in folding, stability, and biological function are unknown. Conformational exchange processes in proteins and RNA can feature transient

excursions to high energy, low-populated structural states that can play important roles in a variety of cellular processes[164, 188-191], and GQ dynamics are likely similarly implicated. The importance of GQ dynamics is starting to be recognized and evidence is accumulating for its contribution to GQ function. The following sections will discuss important examples of conformationally heterogeneous GQs and their characterization by several biophysical methodologies.

1.3.3.2. GQ folding

The folding mechanisms for GQs have been a topic of considerable debate in recent years. Some GQs have been found to fold in an all-or-none (two state) manner[35, 183, 192], while others are thought to proceed through well-populated folding intermediates[181, 182, 193, 194]. Double-stranded hairpins from folding of two G-tracts, or triple-stranded triplex structures from assembly of three G-tracts have been proposed as intermediates in GQ folding. Non-native topologies have additionally been implicated as GQ folding intermediates[195]. Furthermore, GQ structures featuring long, non-native loops have been identified in off pathway folding[194]. Strand-shifted G-tract conformations, where G-tracts are out of native alignment with respect to each other[196], and partially cation bound states[197] are also thought to be populated in GQ folding. Thus GQs appear to exhibit tremendous diversity in their folding mechanisms and obvious sequence determinants have yet to emerge (see below)[77]. Despite these challenges, the modular nature of GQ sequences presents an avenue for kinetic and thermodynamic evaluation of folding intermediates because they may be easily synthesized by truncating a G-tract (to generate a triplex-forming sequence, for example[181, 198]), or trapped by making G>X (X = dT, dA, or deoxyinosine (dI)) substitution mutations in the desired GQ sequence[35]. Complete kinetic and thermodynamic descriptions of GQ

folding can be obtained by characterizing analogues of the GQ folding intermediates individually and analyzing the data together with those of the wild-type molecule[181].

1.3.3.3. Topology exchange

Topology exchange in GQs is characterized by the switching of G-tract orientations (parallel/antiparallel) in the GQ core, with concomitant structural reorganization of the adjoining loops. Topology exchange manifests in ^{1}H NMR spectra as sets of resonances for multiple GQ structures in slow exchange[193, 199] (Figure 1.10a) and as broadening for GQ isomers in intermediate exchange respectively[35]. Topology exchange can occur during GQ folding[193, 195], in response to different cations[200], as part of native ensemble dynamics[35, 49], or with added ligands[201] or molecular crowding agents[49, 202]. The populations of adopted topologies are thought to depend on loop and stacking interactions from 5' and 3' flanking sequences[49, 193, 203].

Bessi *et al.* investigated topology exchange in folding of the Tel24 sequence 5'-TTGGG(TTAGGG)$_3$A-3' by time-resolved NMR spectroscopy[193]. Collecting 1D ^{1}H spectra of Tel24 upon K^+-induced folding revealed imino proton resonances for hybrid-1 and -2 topologies in slow exchange on the chemical shift timescale (Figure 1.10a). Following peaks characteristic of each topology as a function of time permitted extraction of folding and unfolding rate constants from an analysis of peak intensities. It was found that Tel24 folds rapidly into the hybrid-2 topology with slow rearrangement to the hybrid-1 topology via a loosely-ordered hairpin ensemble (U) that is not stabilized by hydrogen bonds (Figure 1.10b-d). Minor signals for low-populated (<5%) early folding intermediates were evident in the measured ^{1}H spectra, however these were attributed to different or partially folded capping structures interacting with the hybrid core. Though the hybrid-1 and-2 topologies had similar populations 1-2 hours after injection of

K^+ into the Tel24 sequence, the hybrid-1 topology became dominant after ~20 hours. Interestingly, signal for residual unfolded molecules was measured months after induction of Tel24 folding. In theory, the latent population of the unfolded state could be captured by *in-vivo* GQ-binding proteins through a conformational selection mechanism.

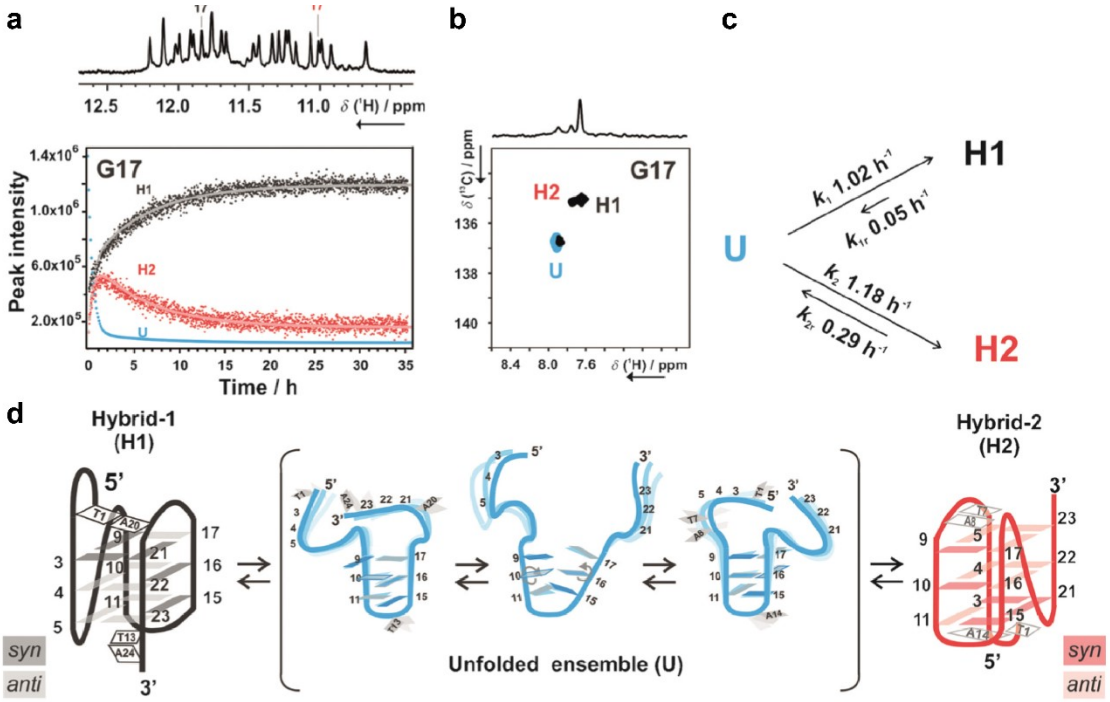


Figure 1.10. Topology exchange in Tel24 GQ folding. a) Kinetic traces describing the intensity of the imino peaks of residue G17 as a function of time at 298 K. Imino signals of G17 in the hybrid-1 (H1, black) and hybrid-2 conformation (H2, red) are marked on the 1D ^{1}H spectrum recorded 2.5 h after inducing folding at 298 K. b) Aromatic region of the ^{1}H, ^{13}C HSQC spectrum of the sample selectively labeled at G17 before the induction of folding (cyan) and 3 days after folding induction (black) at 298 K. Label color code: black for the H1 conformation, red for the H2 conformation, and cyan for the unfolded state (U). c) Kinetic model proposed to fit the data. Average rate constants obtained from the global fitting of well-resolved imino peaks are shown. Fitting curves are displayed in (a) as solid lines and result from the global fitting of the kinetic traces with the proposed mechanism. d) Folding topologies involved in the proposed folding mechanism for Tel24. Proposed capping structures for H2 are in gray. Adapted from Bessi *et al.*[193] with permission.

1.3.3.4. G-register exchange

G-register (GR) exchange dynamics occur when GQs can populate multiple structures by shifting the alignments of their G-tracts. We refer to each strand-shifted state as a GR isomer. These dynamics usually appear when a GQ sequence contains unequal numbers of Gs, giving rise to different subsets of Gs that can form GQ cores[35]. A recent bioinformatic analysis of G-rich promoter sequences found that ensembles of >10 different GQ structures are possible from many promoter GQ sequences, however most promoter GQ sequences that can undergo GR exchange dynamics have between 2-8 GR isomers[35]. For example, the c-myc Pu18 sequence 5'-AGGGTGGGGAGGGTGGGG-3' exchanges between four GR isomers because there are 1x2x1x2 = 4 ways for the G-tracts to align themselves in formation of a three-tetrad GQ core (the G_3-tracts have one register, whereas the G_4-tracts have two). The study of GR exchange dynamics in GQs will be discussed in depth in Chapter 2, where new thermal denaturation methodologies were developed to rapidly capture information on large GR exchange ensembles. Below is a brief review of other examples from the literature where GR exchange dynamics were investigated.

In the case of the human CEB1 minisatellite GQ 5'-AGGGGGG$(AGGG)_2$TGG-3', which forms an intermolecular stacked dimer in solution (Figure 1.11a), Adrian *et al.* employed ^{1}H-^{1}H NOESY/EXSY and ROESY NMR spectroscopy to study molecular dynamics[204]. Because the first and last G-tracts in each subunit contain greater and fewer numbers of Gs than required to complete the GQ core respectively, Gs from the first G-tract effectively fill in the vacant position in the final G-tract. However, as there are multiple additional Gs within the first G-tract, this creates structural ambiguity. Conformational exchange in this structure was identified by peaks for a low-populated state (7%) in 1D ^{1}H NMR experiments. NOESY in conjunction with ROESY confirmed several crosspeaks corresponding to exchange between two conformational states, with the largest

chemical shift changes involving the G4 H8 and G3 H1' base and sugar protons respectively. The dynamics localized to the 5' end of the dimer subunits, mostly involving A1, G2, G3, and G4. The large upfield chemical shift change for the G4 H8 proton indicated a change from a propeller loop position (major form) to a *syn* G-tetrad bound conformation (minor form). The chemical shift of the G4 H8 minor conformation also had a relatively strong intra-residue crosspeak with the sugar H1' proton, indicating that the base adopts the *syn* conformation in the minor state. Furthermore, the G3 H1' chemical shifts implied shifting from a G-tetrad bound position in the major form to being above the GQ core in the minor form, in agreement with the G4 data. Further support for the 5' dynamics came from elimination of the NOESY exchange crosspeaks with G4>T and G(2,4)>T sequence mutations that quenched the structural ambiguity in the first G-tract (Figure 1.11b). Taken together, these results strongly suggested the dimer experiences a base-swapping motion at the 5' end of each subunit such that either the 3rd or 4th G in the first G-tract participate in GQ core formation.

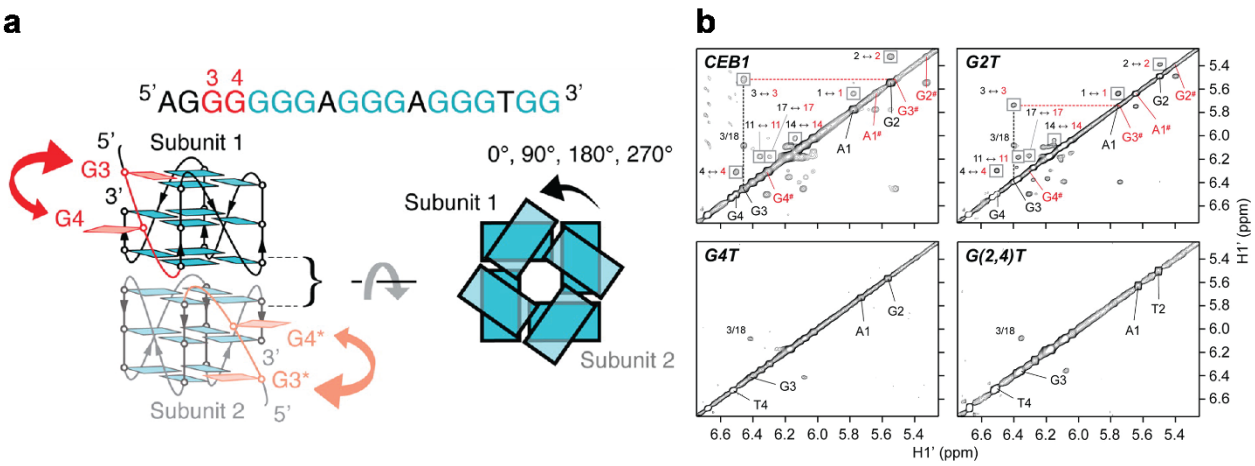


Figure 1.11. Base swapping dynamics in the human CEB1 minisatellite GQ. (a) The GQ sequence (top) forms a stacked dimer (bottom) where the subunits can be rotated with respect to each other. G3 and G4 undergo a swapping motion in completing the gap created by the last G-tract. (b) NOESY spectra show chemical exchange crosspeaks in wild-type CEB1 and CEB1 G2T (top left and right respectively). The absence of exchange crosspeaks in dynamically-quenched mutant CEB1 structures G4T and G(2,4)T (bottom left and right respectively) reveals G3 and G4 undergo the swapping motion. Adapted from Adrian *et al.*[204] with permission.

Recently, "spare tire" GQ sequences containing five G-tracts[151] have been demonstrated to undergo GR exchange dynamics. "Spare tire" GQs can form different GQ structures depending on which four of the five available G-tracts form the four-stranded core structure. In this instance the GR exchange dynamics feature complete replacement of one G-tract with another. "Spare-tire" GR exchange dynamics were found in the VEGF GQ 5'-$G_4CG_3C_2G_5CG_4TC_3G_2CG_4$-3' by dimethyl sulfate (DMS) footprinting. DMS methylates solution-exposed Gs at the N7 position, while Gs participating in hydrogen bonding in the GQ core are protected[205]. After exposure to DMS, the GQ sequence is fragmented at the methylated guanine residues by treatment with piperidine. The GQ fragments are then resolved by gel electrophoresis and the fragmentation pattern is used to determine which Gs are involved in the GQ core structure. DMS footprinting is sensitive to GQ dynamics because structural fragments may show partial methylation, indicative

of exchange processes where Gs are transiently exposed to the solution. The DMS footprint of the VEGF GQ showed that in the absence of oxidative damage, no significant protection of Gs from alkylation was observed, consistent with a heterogeneous mixture of folds (presumably from GR exchange dynamics). When an oxidative stress lesion was introduced to the core structure, it was found that the fifth G-tract substitutes (as a "spare tire") for an oxidatively-damaged G-tract which is then flipped out to be restored in the base-excision repair (BER) process[151]. As new classes of GQs are investigated, it is becoming clear that shifting G-tract alignments can be as subtle as movements of the G-tract by one position, or they can feature extensive structural rearrangements such as replacement of an entire G-tract, with the potential for diverse biological roles.

1.3.3.5. Oligomer exchange

The planar surfaces of the external 5' and 3' G-tetrads present exceptional contacts for stacking and oligomerization interactions. GQ monomers can associate with each other at these faces to form intermolecular homodimers that are in conformational exchange with the dissociated GQ monomers. Higher order oligomeric states are indeed possible because the dimers may continue to oligomerize at their external G-tetrads, however we consider dimeric GQs here for simplicity. Oligomer exchange can be intra or intermolecular depending on the number of GQ subunits contained within a sequence. For example, the sequence 5'-$(TGGGNGGG)_2$-3' is likely to form an intermolecular dimer with other subunits in solution, while 5'-$(TGGGNGGGT)_4$-3' contains two GQ subunits and in theory can form an intramolecular dimer because the additional subunit is covalently attached (the T residues at the ends of each sequence disrupt further oligomerization). NMR is highly useful for studying oligomeric interactions in GQs because

signals can be observed for both the monomeric and oligomeric states, permitting calculation of affinity constants and probing of oligomer dynamics[206].

GGA repeats exist in promoter regions of genes such as c-myb, which codes for a transcription factor regulating proliferation, differentiation, and survival of haematopoietic progenitor cells[207]. Solution NMR spectroscopy and DMS footprinting have shown that GGA repeats adopt unusual dimeric GQ structures[207-209] that can repress c-myb promoter activity[207]. For example, 5'-$(GGA)_8$-3' folds into an intramolecularly-stacked dimer consisting of two $(GGA)_4$ subunits[209]. As there are only two Gs per G-tract, each GQ subunit contains two stacked G-tetrads instead of the usual three. The structure additionally differs from canonical GQs because loop As from each subunit hydrogen bond with their respective G-tetrads at the dimerization interface, forming mixed G/A-heptads. The GQ is formally known by the T:H:H:T (T = tetrad, H = heptad) motif, originally solved by Matsugami *et al*[208, 209]. The conformational dynamics of this peculiar structure were not extensively explored, though NOEs for A6, A9, A18, and A21 participating in the heptad interface were observed that suggested small fractions of *syn* A conformations in equilibrium with the predominant *anti* forms. This observation points to some form of conformational rearrangement at the heptad interface, possibly the loss of heptad structure or dimer destacking.

The T:H:H:T GQ-forming sequence in the c-myb promoter region was found to undergo intramolecular dimer exchange by DMS footprinting[207]. In this case, the GQ sequence 5'-$(GGA)_3$GGTCAC$(GGA)_4$GAA$(GGA)_4$-3' has three subunit repeats and therefore three possibilities for forming the T:H:H:T dimer structure (subunits 1 + 2, 1 + 3, or 2 + 3). The pattern of DMS protection for this sequence suggested all three possible dimers are populated in dynamic equilibrium. DNA and RNA polymerase stop assays were further performed to test the effects of

dimer formation on DNA and RNA polymerization. Polymerase stop assays are used to measure where a DNA or RNA polymerase stops polymerization of a nascent nucleic acid strand, in this case in response to formation of a GQ. Two distinct DNA and RNA polymerase arrest sites were observed for the c-myb sequence. The position of the two sites could be explained by formation of the 1 + 3 or 2 + 3 dimers at the first site and the 1 + 2 dimer at the second site. Furthermore, polymerase arrest at site 1 was only partial; the most arrest occurred at site 2 consistent with subunits 1 and 2 forming the strongest dimeric interaction. The requirement for a dimeric structure in disruption of polymerase read-through was tested by deleting two of the three possible dimer-forming subunits. The resulting sequence could only adopt a T:H GQ. The stop assays using the T:H structure showed no DNA polymerase arrest and only partial RNA polymerase arrest, suggesting that intramolecular dimerization is necessary for GQ stabilization and interruption of polymerization.

1.3.4. Nucleic acid dynamics and biological function

Nucleic acid dynamics are intimately tied to their biological function. As briefly discussed in the previous sections, nucleic acid dynamics operate across a wide range of length and timescales with diverse functional outcomes at each level. Consider for a moment the breadth of consequences that arise from undergoing relatively simple base pair breathing events. Transient opening events are key conformational excursions that expose the bases to solution, thereby enabling transcription factors to bind and initiate transcription[210, 211]. Base pair opening events are even now thought of as a form of transcriptional self-regulation by the DNA duplex because the sequences that comprise transcription start sites have higher breathing propensities[211-213]. Without these motions, the information contained within DNA would be permanently buried within the

interior of the double helix, precluding access to the bases by molecular machinery[11]. However, the opening of the double helix is also problematic for the integrity of the genetic code. Mutagens such as the highly reactive formaldehyde that is endogenously produced during nucleosome remodeling[214] can modify exposed bases. Therefore, the timescale of breathing dynamics, i.e. the lifetimes of open base pairs, need to be long enough to permit recognition by transcription factors but not so long that the bases are overly vulnerable to mutation by damaging agents like formaldehyde. In the event that DNA does become damaged, base pair opening dynamics are important for checking the integrity of the DNA code by repair enzymes which are thought to scan DNA and interrogate flipped out bases[215-217]. Interestingly, the putative adoption of HG base pairs in breathing events seems to have dual advantages for preserving the integrity of the genome. One benefit is to act as a backup for conservation of the double helix structure in response to base methylating agents. If the WC face of A or G is modified, the bases can flip into their HG orientations to hydrogen bond with their partner bases in the duplex core. The other possible advantage is to facilitate recognition by DNA repair proteins since HG base pairs provide a unique recognition motif against the WC background[161]. Furthermore, HG base pairs are known to be less stable than their WC counterparts and presumably, open more readily for recognition and repair. Interestingly, HG base pairs have also been shown to promote the modification of neighboring Cs involved in WC GC pairs by formaldehyde[214], the biological reason for which remains unexplored. Taken together, base pair breathing events appear to be important for both the expression and the integrity of the information contained within nucleic acids.

As alluded to in the last paragraph, the motions that nucleic acids experience have evolved to coincide with the timescales of binding and catalysis by the molecular machinery that acts on nucleic acids to uphold and process the information contained within them[138, 218]. The

synchronization of nucleic acid and other biomacromolecular dynamics enables biological processes to be carried out within suitable timeframes. An interesting example of the interconnectedness of nucleic acid dynamics at multiple levels in the central dogma is found in the comparison of the rate of RNA polymerization by RNA polymerase II and the rates of spontaneous unwrapping/rewrapping of DNA in nucleosomes[219]. Li *et al.* have shown that DNA unwrapping from nucleosomes occurs with a rate constant of approximately 4 s^{-1}, whereas RNA polymerase II generates RNA at 23 nucleotides s^{-1}. The similarity between these rates means that transcription is not massively delayed by the exposure of DNA from nucleosomes. This is because the DNA will unwrap many times within the duration it takes for the polymerase to transcribe the accessible linker DNA between nucleosomes. Intriguingly, Li *et al.* also showed that nucleosomes rewrap more quickly with a rate constant of ~20-90 s^{-1}, which could hinder the entry of RNA polymerase II into nucleosomal DNA since the unwrapped lifetime is on the order of 10-50 ms, meaning that the polymerase barely advances by one nucleotide within this span. This in turn suggests that the RNA polymerase can wait several unwrapping and rewrapping cycles before advancing into the nucleosome DNA. While apparently detrimental to the speed of RNA polymerization, the more rapid rewrapping of nucleosomal DNA appears to have beneficial outcomes: the pausing of RNA polymerase in response to rewrapping can facilitate the co-transcriptional folding of nascent RNA molecules[218, 220], which, considering riboswitches, in turn regulates mRNA translation. The slowing of RNA elongation by nucleosomal DNA dynamics may additionally permit the alternative splicing of mRNA to generate protein isoforms[221]. Alternative splicing is also regulated in some cases by riboswitch folding[222]. In this view, the dynamics of nucleosome DNA propagate to influence multiple biological functionalities.

There is an emergent view that nature uses the pre-existing conformational excursions of RNA to achieve functional endpoints[136, 165], rather than driving the adoption of new conformational states in response to molecular cues. In this way, pre-existing equilibria between multiple structural states need only be tuned by protein binding, for example, to bring about sensitive control over the conformations adopted by RNA. Recent studies further suggest that effector binding to RNA does not stabilize a single bound state[165]. Instead, a binding partner simply guides RNA dynamics by influencing the energy barriers between sets of dynamic ensembles that may correspond to alternate secondary structures with unique interhelical motions. This allows RNA to maintain dynamics for regulatory purposes. Another important aspect of the ensemble view of nucleic acids is that the coupling of conformational dynamics across multiple motional levels imbues functional complexity[136]. For example, changes to secondary and tertiary structure can be linked; the packing of a helix into an alternate tertiary conformation can provide the energy to stabilize unfavorable secondary structures. In RNA structures such as riboswitches, the ability to combine dynamics in this manner permits the propagation of a binding signal from the aptamer domain to the expression platform, thereby regulating mRNA translation by the ribosome[120]. We have additionally shown how GR exchange in GQs can be coupled to topological interconversion (Chapter 2), expanding the diversity of loop ensembles displayed as potential protein recognition motifs[35]. The principle of shifting balances between pre-existing conformational ensembles is likely similarly applied in the less-well studied dynamics of DNA structures like GQs, where the balance between distinct sets of interconverting topologies could be shifted by GQ-interacting proteins such as nucleolin[223].

There is growing evidence that GQs play a variety of regulatory roles in biology[28]. Furthermore, GQs have been implicated in the development of neurodegenerative disorders such as amyotrophic lateral sclerosis (ALS)[224]. In many studies of biological GQ-forming sequences,

the investigated GQ is highly dynamic, with the potential for modulating cellular outcomes. Yet, the effect of GQ dynamics on function is often neglected, with the result that relatively little is currently known about how GQ dynamics influence biological processes. We have recently shown that GQ dynamics can strongly influence their stability (Chapter 2)[35], potentially modulating gene expression and mRNA translation. Furthermore, sliding motions of the four strands occur cooperatively and can be coupled to topology exchange. Interestingly, parallel intermolecular telomeric GQs are good substrates for human telomerase, while intramolecular antiparallel topologies are not[225]. Since telomeric GQs can populate parallel[226] and antiparallel topologies[203], topology exchange dynamics could directly influence telomere extension by telomerase. Similarly, oligomer exchange may also play a role in regulatory processes that is currently not well understood. Tandem GQ sequences appear throughout the genome[207, 226], and nucleic acid fragments containing tandem GQs often show tight intramolecular GQ/GQ interactions, potentially influencing their biological functions. The exchange between different dimeric GQ states in the c-myb promoter differentially arrests DNA and RNA polymerases, providing proof-of-principle for this idea[207]. As well, GQ dynamics may be involved in DNA damage repair. In "spare tire" promoter GQs which contain five G-tracts, oxidative damage to one G-tract results in its being replaced with the fifth or "spare tire" G-tract, preserving the four-stranded structure[151]. The damaged tract can thus be corrected via base excision repair. The idea of structural redundancy in GQ folding applies to other types of dynamic GQs, such as those with multiple GR isomers or that form oligomer ensembles. Because these GQ sequences have the ability to adopt multiple stable conformations, they may be less sensitive to DNA damage at individual residues. With the link between GQ dynamics and biological function becoming apparent, an understanding of the determinants of GQ dynamics (and nucleic acid dynamics in general) is critical to our

understanding of biology and the development of therapeutics targeting transient and low-populated GQ conformers. GQ dynamics likely nuance their functional capacities in a similar manner to the better understood conformational excursions in RNA.

1.3.5. Nucleic acid dynamics and biotechnology

The conformational excursions that nucleic acids undergo in biology are also relevant to their applications in biotechnology. In analogy to biological RNA dynamics, the ability to undergo reversible conformational transitions between flexible and more ordered states in response to ligand binding makes synthetic aptamers an attractive component for biosensors[227, 228]. A single-stranded aptamer that is highly disordered in its free state yet undergoes rapid coupled folding and binding in response to doxorubicin, a potent chemotherapeutic, has shown promise in the real-time detection of drug molecules in human blood with sub-minute resolution[227]. The disordered free state is likely able to undergo sweeping motions similar to those observed in HIV-1 TAR RNA that encourage the capture of doxorubicin molecules from solution and accelerate the kinetics of detection. The end goal of this technology is to bring about the personalized monitoring of drug levels in patients by physicians immediately after drug administration. At its most fundamental level, the widespread application of this type of personalized medicine ultimately rests on the inherent dynamic capabilities of nucleic acids.

Beyond single-stranded applications, many interesting implementations of nucleic acids rely on the ability to assemble higher-order structures from multiple strands[4]. This process requires the strands to anneal in the correct registers to yield the desired structure. Physically, structural motifs like double helices anneal by finding initial contacts and subsequently "zipping up" from that nucleation site[229]. Often, the initial contacts are not in the correct register and the strand must

take extra steps to fold into the proper orientation. Taking breathing dynamics as an important illustration, these motions permit the sliding of a DNA strand into the correct register with its complementary sequence since the duplex can locally melt and reanneal into the correct orientation. Backbone and looped sequence motions additionally facilitate the search for the desired duplex register. In controlled release applications such as the delivery of drugs to target sites by DNA nanotubes, breathing dynamics are of critical importance. To release cargo, so-called "eraser" strands are added that compete for interactions with one of the strands in the double-helical elements that make up the nanotube structure[133]. This "strand-displacement" reaction[230, 231] results in the loss of the structural sequences from the nanotube (they leave by forming duplexes with the eraser strand), causing the nanotube to lose rigidity and release the encapsulated cargo. The intrinsic breathing dynamics of the DNA assembly promote the invasion of the eraser strand by transiently exposing interaction sites[230]. This leads to an important conclusion – the final loaded assembly must be thermodynamically favorable enough so as to outcompete the formation of other structures, yet not so stable that breathing fluctuations are far too rare and transient for strand invasion and cargo release to occur.

The assembly of nucleic acid structures from many dynamic, competing strands raises an important question; how do these sequences manage to adopt the desired structure with so many possible folding pathways? The answer is that in many cases the strands do not reach the target conformation and instead form defective structures. This is one of the major obstacles to generating complex higher-order nucleic acid architectures[232]. The dynamic nature of nucleic acid strands, combined with the presence of potentially large numbers of binding partners means that the assembly process can be easily misdirected by the undesired assembly of sequences within off-target folding pathways of similar energy to the intended one. Simply designing a target structure

as a global thermodynamic minimum may fail entirely when unpredictable and unknown defect structures that are lower in energy than the desired assembly occur. Defect formation is particularly problematic in generating DNA origami patterns, since the seminal protocol entails a one-pot mixing of bacteriophage DNA and hundreds of staple strands[129]. Because defects are such a common occurrence, the development of methods to obviate defect formation in the assembly of nucleic acid nanostructures is an active area of research[134, 233]. In the absence of knowledge of nucleic acid dynamics, it becomes difficult to rationally design primary sequences that yield the desired structure with minimal side products. Moreover, without quantitatively understanding the dynamics that control the assembly of unprecedented nucleic acid architectures such as CA-mediated poly(A) fibers, the synthesis of new biomaterials with unique capabilities remains one of trial-and-error. In both of these endeavours, a strong grasp of the underlying nucleic acid dynamics is of paramount importance.

1.4. Folding experiments

In general, the folding dynamics of nucleic acids are complicated. As illustrated in Section 1.3, even relatively small nucleic acid structures like short duplexes and GQs (~5-10 kDa) have remarkably intricate folding behavior which can be difficult to address experimentally. The thermal unfolding and refolding of nucleic acids represents one extremely powerful approach toward elucidating their complex folding dynamics and relating these to function, since the variation of the experimental temperature causes nucleic acids to pass through their folding trajectories[234]. Thermal denaturation profiles can be deconvoluted with statistical thermodynamics in order to extract the equilibrium populations and thermodynamic parameters for transitioning between conformations[184]. These datasets additionally offer the ability to extract folding rates and

kinetic populations[235]. Thermal denaturation experiments can be performed on nucleic acids in their naturally occurring forms, without the need for labeling of the biomolecule with reporter moieties such as a fluorescent tag that can modify their natural folding. Furthermore, a typical thermal denaturation experiment requires little sample (~nmol amounts), is straightforward to implement, and can be performed in one day. Taken together, thermal denaturation experiments offer a rapid and cost-effective approach for complete kinetic and thermodynamic exploration of nucleic acid folding and assembly processes compared to other much more laborious methodologies such as NMR spectroscopy.

A standard nucleic acid thermal denaturation experiment entails measuring a real-time physical observable that reports on conformational transitions in response to variation of the experimental temperature. Suitable observables include absorbance, circular dichroism (CD), chemical shift, or a calorimetric heat flow. A thermal denaturation dataset has three main components: low (folded) and high (unfolded) temperature baselines, and a transition region where folding/unfolding occurs (Figure 1.12a). One major advantage to performing thermal denaturation experiments is that the complexity of the folding process is immediately evident in the shapes of thermal profiles (Figure 1.12b). The population of folding intermediates can appear as multiple "wiggles" in the thermal transition, for example. In this view, thermal denaturation experiments are highly robust since they represent both an initial screening method for dynamics and can be extended to obtain quantitative information on complex folding equilibria. In this thesis, the two primary thermal denaturation methodologies utilized are absorbance spectroscopy and differential

scanning calorimetry (DSC), which have distinct and complementary benefits. These are introduced in the following sections.

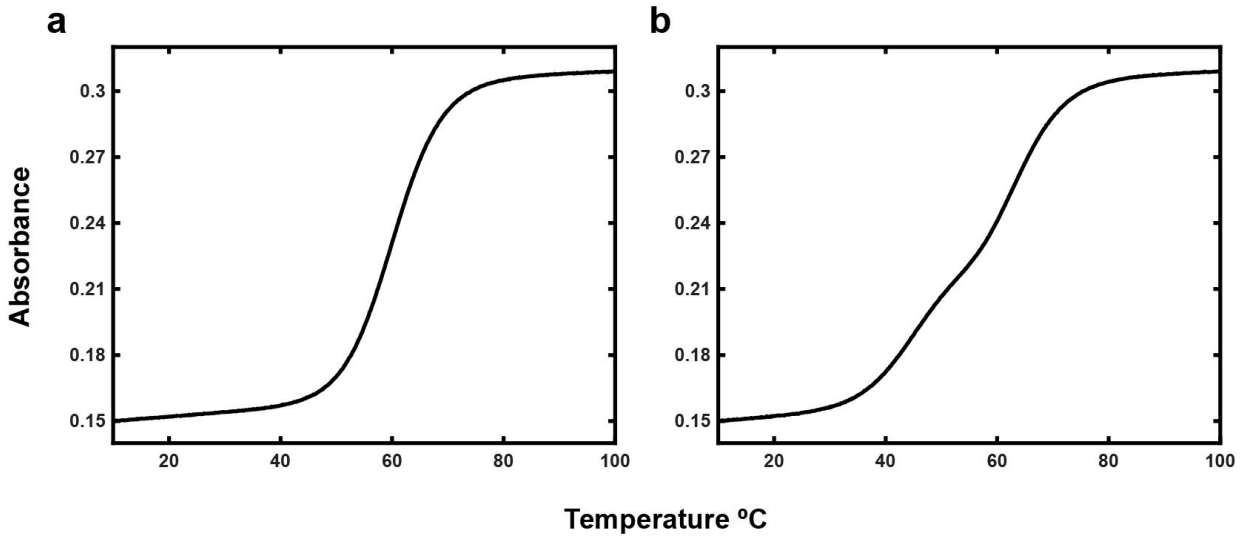


Figure 1.12. Simulated examples of thermal denaturation data. (a) A two-state absorbance unfolding profile. (b) A multi-state absorbance unfolding profile.

1.4.1. Absorbance spectroscopy

Absorbance spectroscopy applied to nucleic acids relies on the absorption of photons by the nucleobases. In particular, the bases within nucleic acid structures strongly absorb light with wavelengths in the 250-290 nm region. The absorbance of nucleic acids changes substantially as a function of their structural state. It follows that the change in absorbance in response to variation of the experimental temperature can be used as a sensitive reporter of fluctuations in nucleic acid structure. Additionally, different wavelengths are responsive to different aspects of 3D structure[236], giving access to conformational complexity by comparing absorbance changes at multiple wavelengths (see Section 1.5.1). Absorbance-based thermal denaturation experiments have added advantages in their sample requirements and throughput. Because the nucleobases have strong inherent absorbances, absorbance experiments can be performed with nmol amounts of material.

As well, modern absorbance spectrophotometers typically have the capability of measuring multiple samples (up to 12) at multiple wavelengths in tandem, drastically accelerating data collection times relative to other thermal denaturation methodologies (DSC) that are restricted to analysis of one sample per experiment.

In an absorbance-based thermal denaturation experiment, an absorbance spectrophotometer shines a beam of light that passes through a monochromator to select for the desired wavelength. The monochromatic light is then passed through the nucleic acid sample contained within a quartz cuvette in the sample block. The absorption of photons by the nucleobases can be used to report on nucleic acid structure, as well as on dynamics when measured while the temperature is continuously varied at a set scan rate. The absorbance is given by

$$A = \varepsilon c l \tag{1.1}$$

Where ε is the molar extinction coefficient, c is the concentration of the nucleic acid, and l is the cuvette pathlength that the input light must traverse. Importantly, the extinction coefficient in Equation 1.1 changes as a function of the conformational state of the nucleic acid[237]. At low temperature, the nucleic acid is folded and the absorbance signal is dominated by that of the native state. By varying the experimental temperature, the nucleic acid can be induced to pass through distinct states which can be observed via their corresponding extinction coefficients. The resulting temperature-dependent absorbance profiles can be analyzed with models to quantitatively describe nucleic acid folding.

1.4.2. Differential scanning calorimetry

DSC is an extremely powerful methodology for evaluating nucleic acid dynamics because it directly measures the heats associated with folding transitions[234]. DSC is particularly advantageous for studying spectroscopically opaque or invisible nucleic acid systems (e.g. short duplexes with poor absorbance changes accompanying folding), since the folding and binding transitions of nucleic acids are typically accompanied by the absorption or release of heat, regardless of their spectroscopic properties. The most powerful aspect of DSC is that the measured heats are directly linked to thermodynamic theory (Section 1.5.4), obviating the need for application of quantitative models that assume the folding mechanism *a-priori*. This property is not shared by any other physical observable, other than volume changes[184]. Furthermore, the heats measured by DSC can be used to estimate sequential folding enthalpies in a straightforward manner, providing immediate clues as to the nature of the structural rearrangements in nucleic acid dynamics.

A DSC instrument (Figure 1.13, left) measures the heat required to unfold a biomolecule at constant pressure as the temperature is continuously increased at a constant scan rate[234]. This is detected as isobaric heat capacity of the sample (C_p), which is the temperature derivative of the enthalpy (the heat measured at constant pressure)

$$C_p = \frac{d}{dT} H \,. \tag{1.2}$$

Thus, the enthalpy at any point throughout the thermal denaturation transition can be obtained by integrating the sample's heat capacity as a function of temperature. The instrument contains a reference and sample cell that are filled with buffer and nucleic acid in buffer respectively. The cells are machined to be thin, wound capillaries of small volumes that serve two purposes (Figure

1.13, right). The first is to maximize the sample's surface area and permit rapid thermal equilibration as the temperature is changed. The second is to enable fast and sensitive detection of small heats that are absorbed or released by the nucleic acid.

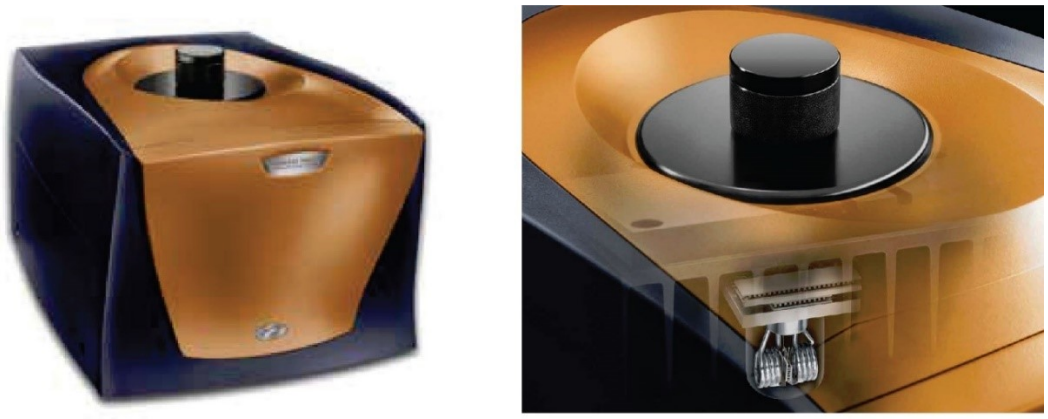


Figure 1.13. A modern nano-DSC instrument . A magnified view of the reference and sample capillaries is shown on the right. Images were obtained from the TA instruments nanocalorimeter brochure at http://www.tainstruments.com/wp-content/uploads/Nano_DSC.pdf.

The instrument measures the difference in power supplied to the sample and reference cells throughout the temperature scan[238]. In practice, it is impossible to have reference and sample cells that are identically machined. The small differences between the physical states of the cells give rise to a baseline artefact that must be subtracted out prior to analysis of DSC data. This is done by measuring a buffer "baseline" experiment with buffer in both cells before running the experiment for the desired sample. The baseline experiment is then subtracted from the subsequent sample experiment in order to eliminate any signal due to the physical mismatch of the cells.

Modern nanocalorimeters are known as "compensation" calorimeters, since they apply a compensating power to maintain the sample cell at the same temperature as the reference cell[234]. In an isothermal titration calorimeter (ITC), this is done at a fixed temperature, while in a DSC

this is done throughout a temperature scan. As the nucleic acid begins to unfold in a DSC scan, heat is absorbed and therefore an increase in power must be supplied to the sample cell to maintain a constant rate of temperature change relative to the reference cell. The power difference between the sample and reference cells is proportional to the apparent excess heat capacity relative to the solvent[238]

$$\Delta C_p^{app.} = \Delta P \frac{dt}{dT} \tag{1.3}$$

Where ΔP is the power difference and dt/dT is the inverse temperature scan rate. The apparent excess heat capacity relative to the solvent is given by the difference in heat capacities between the sample and solvent (water) of the same volume

$$\Delta C_p^{app.} = C_p^{sample} m^{sample} - C_p^{solvent} \Delta m^{solvent} \tag{1.4}$$

Where C_p^{sample} and $C_p^{solvent}$ are the partial specific and specific heat capacities of the sample and solvent respectively, and m^{sample} and $\Delta m^{solvent}$ are the masses of the sample and solvent displaced by the sample respectively. The heat capacity of the sample is given by

$$C_p^{sample} = \frac{\Delta P \frac{dt}{dT}}{C_T V_{cell}} + C_p^{solvent} \phi \rho \tag{1.5}$$

Where ϕ is the partial specific volume of the sample, ρ is the density of the solvent, C_T is the total concentration of the nucleic acid sample, and V_{cell} is the volume of the sample cell. The right-hand term in Equation 1.5 accounts for the heat capacity associated with the solvent displaced by the nucleic acid.

1.5. Folding analysis

1.5.1. Two-state folding tests

The investigation of nucleic acid dynamics by thermal denaturation begins with the question of how many states are required to describe the experimental data. One approach for discriminating between two-state and multi-state folding in nucleic acids involves characterizing the thermal unfolding transition by measuring spectroscopic absorbances at two different wavelengths (260 and 295 nm for example) and performing a dual-wavelength parametric test[236]. When a correlation plot of the two datasets produces a straight line, the folding transition is said to be two-state. Significant curvature in the correlation plot may indicate the presence of well-populated folding intermediates. Care must be taken in this analysis, since curvature can also result from the linear baselines obtained at each wavelength having different slopes. This may be due to effects other than the formation of folding intermediates, for example from different responses to solvent thermal expansion. In this case, it is appropriate to scale and overlay the two datasets to see if the unfolding transition coincides, indicating a two-state process[35].

The two-state folding assumption may also be tested by calculating the ratio of the van ‘t Hoff to calorimetric enthalpies using heat capacity data obtained from DSC[239, 240]. The van ‘t Hoff enthalpy assumes a two-state unfolding transition in analysis of the DSC peak shape. It is calculated by converting the heat capacity data to a fraction folded progress curve and finding the slope of the curve at the melting temperature (T_m), corresponding to the 50% unfolded point for a two-state transition. The van ‘t Hoff enthalpy for an intramolecular two-state folding process can be calculated from thermal denaturation data as

$$\Delta H_{VH} = 4RT_m^2 \frac{d\theta}{dT} \tag{1.6}$$

where R is the ideal gas constant and $d\theta/d$T is the slope of the fraction folded curve at the T_m. The van 't Hoff enthalpy can be calculated in a similar manner for folding processes of other molecularities. In contrast, the calorimetric enthalpy is a model-independent measure of the total enthalpy required to unfold a biomolecule, accounting for unfolding all structural states that might exist along the folded to unfolded trajectory. The calorimetric enthalpy is obtained by integrating the DSC heat capacity profile using an appropriate baseline. When the ratio of the van 't Hoff and calorimetric enthalpies is unity, the unfolding process is considered to be two state. If the ratio is <1, additional states are required to describe the data, because the measured enthalpy is larger than that found by assuming two-state unfolding. If the ratio is >1, there could be sample aggregation or inactivation[241], because less enthalpy is measured than found in the van 't Hoff analysis. When more than two states are detected (i.e. the ratio is > 1) statistical thermodynamic analyses may be applied to determine the minimum number of states required to account for the DSC data[184-187, 239, 240].

1.5.2. Multi-state models

Thermal denaturation profiles for nucleic acids are analyzed by assuming they populate discrete states along their folding trajectories[234] (Figure 1.14). Within this framework, the folding process is described by free energy wells separated by barriers along a reaction coordinate. The wells correspond to the free energy for a given conformational state. The relative heights of the energy wells govern the equilibrium populations of the nucleic acid ensemble, i.e. lower energy wells have greater equilibrium populations than elevated wells. The barrier heights dictate the rates at which each nucleic acid conformation traverses the barriers[165], meaning how quickly the populations equilibrate. A large barrier is energetically difficult to overcome and has a

concomitantly slow rate of barrier traversal. Conversely, small barriers are less unfavorable to cross and have higher proportions of molecules traversing the barrier per unit time. This section will address models where the nucleic acid populations rapidly equilibrate with respect to the temperature scan rate throughout the thermal denaturation experiment. This is a typical assumption in many thermal denaturation analyses. In this case, the folding process can be treated with equilibrium constants.

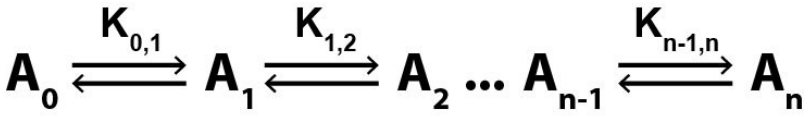


Figure 1.14. A multi-state model for intramolecular folding. The folded state A_0 is assumed to be in equilibrium with the folding intermediates A_1 to A_{n-1} and the unfolded state A_n. Equilibria between each state are given by the equilibrium constant *K*, where the indices correspond to an equilibrium between successive states in the unfolding trajectory.

As the temperature is increased in a thermal denaturation experiment, the energies of each conformation change. The relative positions of energy wells and barriers shift, inducing a redistribution of the conformational populations. The rearrangement of the free energies as a result of temperature variation can be simply demonstrated for a two-state process (Figure 1.15). At low temperatures, the folding free energy is favorable ($\Delta G < 0$), leading to an excess of the nucleic acid in the folded state. Through the transition, the free energies become similar ($\Delta G \approx 0$) and the two states are nearly equally populated. At high temperatures, the folding free energy becomes unfavorable ($\Delta G > 0$) and the unfolded state is predominant.

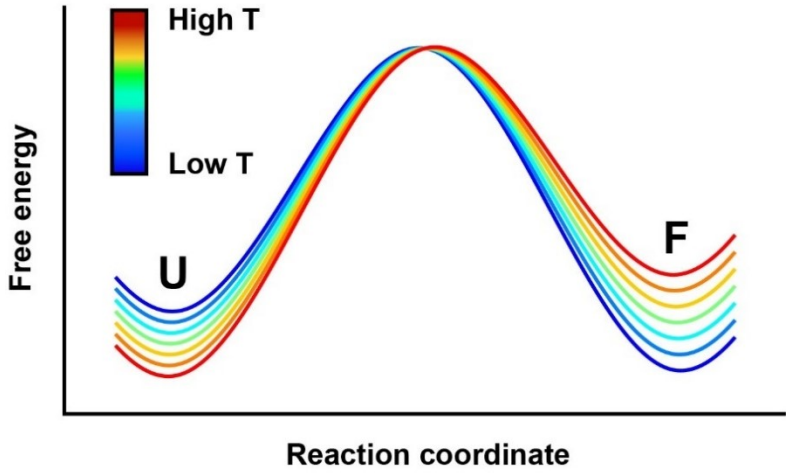


Figure 1.15. A simulated two-state free energy diagram for nucleic acid folding as a function of temperature. At low temperature (dark blue), the free energy difference between F and U is negative (favorable), therefore the F state is predominant. At high temperature (red), the free energy difference between F and U is positive (unfavorable) and the U state is highly populated.

The equations that describe nucleic acid folding at thermal equilibrium with respect to the temperature scan rate are

$$K_{0,i} = \prod_{i=1}^{n} K_{i-1,i} \tag{1.7}$$

$$\Delta H_{0,i} = H_i - H_0 = \sum_{i=1}^{n} \Delta H_{i-1,i} \tag{1.8}$$

$$\Delta S_{0,i} = S_i - S_0 = \sum_{i=1}^{n} \Delta S_{i-1,i} \tag{1.9}$$

$$\Delta C_{p,0,i} = C_{p,i} - C_{p,0} = \sum_{i=1}^{n} \Delta C_{p,i-1,i} \tag{1.10}$$

Where $K_{0,i}$, $\Delta H_{0,i}$, $\Delta S_{0,i}$, and $\Delta C_{p,0,i}$ are the equilibrium constant, change in enthalpy, change in entropy, and change in heat capacity between the i^{th} state and the reference state respectively. The

reference state is usually taken to be the fully folded form of the nucleic acid. In what follows, the indices and explicit temperature dependences of the variables will be omitted for clarity, unless they are required. Equilibrium constants are calculated according to

$$K = \exp\left(\frac{-\Delta G}{RT}\right) = \exp\left(\frac{-\Delta H}{RT} + \frac{\Delta S}{R}\right) \tag{1.11}$$

Where ΔG is the Gibbs free energy difference between the i^{th} state and the reference state. Changes in enthalpy, entropy, and heat capacity are related by

$$\Delta H(T) = \Delta H_0 + \Delta C_p (T - T_0) \tag{1.12}$$

$$\Delta S(T) = \Delta S_0 + \Delta C_p \ln\left(\frac{T}{T_0}\right) \tag{1.13}$$

Where ΔH_0 and ΔS_0 are the changes in enthalpy and entropy at the reference temperature T_0. When the ΔC_p for a particular transition is zero, the changes in enthalpy and entropy become temperature-independent, for example $\Delta H = \Delta H_0$ at all temperatures. The partition function which accounts for the populations of each state as a function of temperature is

$$Q = 1 + \sum_{i=1}^{n} K_{0,i} = \frac{C_T}{C_0} \tag{1.14}$$

Which has been defined relative to the concentration of the reference state, C_0. The summation runs over the number of conformations beyond the reference state. The populations of the reference and subsequent states respectively are given by

$$P_0 = \frac{1}{Q} \tag{1.15}$$

$$P_i = \frac{K_{0,i}}{Q} \tag{1.16}$$

which at all temperatures satisfy the identity

$$P_0 + \sum_{i=1}^{n} P_i = 1. \tag{1.17}$$

At the reference temperature, $Q \approx 1$ because the equilibrium constants are small (i.e. the free energies for populating states beyond the reference are highly unfavorable) and thus the reference state is the dominant population. As the temperature is raised, the free energies for populating additional states become more favorable and the partition function becomes dominated by the equilibrium constants corresponding to transitions to other states. Assuming an intramolecular two-state process, it follows from this analysis that the equilibrium constant becomes unity when the nucleic acid is 50% folded (at the T_m), i.e. $Q = 1 + 1 = 2$ and therefore the population of the folded state $P_F = 0.5$.

In thermal denaturation experiments, the measured data is an average over the physical observables for each conformation (their respective signals such as absorbance), weighted by their relative populations at each temperature[184]. The signal in a thermal denaturation experiment is thus given by

$$< \alpha >= \sum_{i=0}^{n} \alpha_i P_i \tag{1.18}$$

Where $<\alpha>$ is the average observable, α_i is the observable corresponding to the i^{th} state, and P_i is the fraction of nucleic acid in the i^{th} state. The summation runs over the total number of states populated in the folding trajectory, and can be arbitrarily expanded to account for highly complex

folding ensembles. The low and high temperature baselines in a thermal denaturation experiment correspond to the temperature-dependences of the initial (folded) and final (unfolded) physical observables in Equation 1.18. The physical observables for intermediate states may not correspond to a clear baseline in the transition region, however these can be estimated as fractions of the total signal change[234], or set to be identical to the folded state as in Chapter 4. Baselines in biomolecular thermal denaturation profiles often appear linear due to gradual increases in conformational fluctuations of the folded and unfolded states and their interactions with solvent molecules[242]. These types of baselines can be approximated by a first order polynomial (a straight line)[35]. In all cases, it is critically important to remember that the baselines in thermal denaturation experiments represent the temperature-dependence of the physical observable for a given biomolecular state (or set of states), and therefore the application of mathematical functions to account for the baselines is not merely a cosmetic one. One approach to extracting the thermodynamic parameters governing nucleic acid folding is to apply Equation 1.18 in a non-linear least-squares fitting routine. This yields an optimized simulated thermal denaturation dataset that best describes how the populations of each state and their respective experimental observables vary with temperature.

1.5.3. Analysis of DSC data

The analysis of DSC data is unique among thermal denaturation methodologies because the measured heat capacity data is effectively the derivative of the population-weighted average sample enthalpy[234]

$$C_p^{sample} = \frac{d}{dT} H^{sample} = \frac{d}{dT} \sum_{i=0}^{n} H_i P_i \quad . \tag{1.19}$$

Thus, the average heat capacity of the sample (Figure 1.16) is given by

$$C_p^{sample} = C_{p,0} + \sum_{i=1}^{n} \Delta H_i \frac{d}{dT} P_i + \sum_{i=1}^{n} \Delta C_{p,i} P_i \ . \qquad (1.20)$$

The term for the heat capacity of the reference state, $C_{p,0}$, represents the initial baseline of the DSC experiment. DSC baselines transition sigmoidally from $C_{p,0}$ to the heat capacity of the final state, which is usually the unfolded baseline, according to the sum of the first and last terms in Equation 1.20. The difference between the final and initial baselines is the total ΔC_p for the unfolding transition, assuming temperature-independent baselines like those simulated in Figure 1.16.

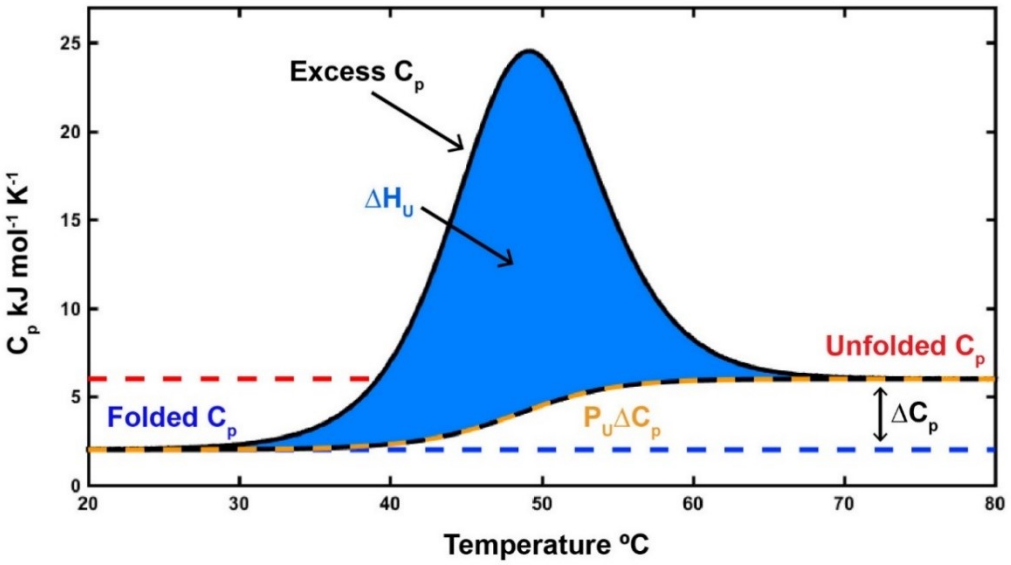


Figure 1.16. A two-state DSC thermogram simulated using Equation 1.20 with a positive ΔC_p of unfolding and the folded state as the reference. The folded and unfolded C_p baselines are dashed purple and red lines respectively. The ΔC_p for the transition (black double-headed arrow) is the difference between the two C_p baselines. For a biomolecule with a positive ΔC_p of unfolding, the observed C_p baseline for the transition is given by $P_U \Delta C_p$ (orange dashed line), which transitions sigmoidally from the folded to the unfolded C_p baselines. The excess C_p is the profile given by the black line above $P_U \Delta C_p$. The ΔH_U is the area under the excess C_p profile (blue), which can be obtained by integrating the excess C_p.

The average excess heat capacity of the biomolecule is important in several DSC analyses (Section 1.5.1 and 1.5.4). The average sample excess heat capacity is given by

$$C_{p,excess}^{sample} = \sum_{i=1}^{n} \Delta H_i \frac{d}{dT} P_i \ . \tag{1.21}$$

The integration of the average excess C_p gives the calorimetric enthalpy $\Delta H_{cal} = \Delta H_U$, which is the change in enthalpy for fully unfolding the biomolecule. This can be used to test for two-state folding by comparison with the van 't Hoff enthalpy change[234].

1.5.4. Model-free deconvolution of complex folding processes by DSC

The analysis of multi-state folding processes, as evidenced by multiple and potentially overlapping thermal transitions, can be difficult as the number of states required to describe the folding process is not initially clear. Furthermore, the number of parameters needed to capture the folding behavior in a fitting routine is potentially large, which can preclude the accurate estimation of their true values. Fortunately, DSC is an extremely powerful methodology for circumventing these problems in analysis of nucleic acid folding. A thermodynamic recursion analysis can be applied to DSC data directly which permits the extraction of the number of states, their folding thermodynamic parameters, and their populations, entirely in a model-free manner[184]. The analysis relies on the statistical thermodynamic relationship between the average excess enthalpy and the partition function

$$\langle \Delta H \rangle = RT^2 \frac{d \ln(Q)}{dT} = \int_{T_0}^{T} C_{p,excess}^{sample} dT \ . \tag{1.22}$$

From this expression, the partition function is given by

$$Q = \exp\left(\int_{T_0}^{T} \frac{\langle \Delta H \rangle}{RT^2} dT \right). \tag{1.23}$$

Thus, the partition function for unfolding the biomolecule can be obtained from experimental excess heat capacity measured by DSC. Starting at low temperature, the nucleic acid is folded and thus $Q \approx 1$. Therefore, the population of the folded state can be calculated as

$$P_F = \frac{1}{Q} = \exp\left(-\int_{T_0}^{T} \frac{\langle \Delta H \rangle}{RT^2} dT \right). \tag{1.24}$$

Conversely, the nucleic acid is fully unfolded at high temperature, and the population of the unfolded state can be obtained by performing the integration from high to low temperature (relative to the unfolded baseline) according to

$$P_U = \exp\left(-\int_{T_U}^{T} \frac{\Delta H_{F,U} - \langle \Delta H \rangle}{RT^2} dT \right) \tag{1.25}$$

where $\Delta H_{F,U}$ is the total enthalpy change through the transition and T_U is the temperature corresponding to the unfolded baseline, i.e. where $\Delta H_{F,U} = <\Delta H>$. For a two-state transition, $1 - P_F - P_U = 0$ at all temperatures. If this calculation reveals a leftover portion of the total fraction of biomolecules (roughly >5-10%), folding intermediates exist and the model-free recursion analysis can be applied to calculate their thermodynamic parameters and populations, as illustrated in more detail below. Performing this check can be particularly useful for datasets that have somewhat broadened thermal transitions and may not exhibit visually obvious non-two-state behavior. The DSC deconvolution routine for a simulated multi-state DSC transition will be given here as an example. The DSC were simulated according to Figure 1.17.

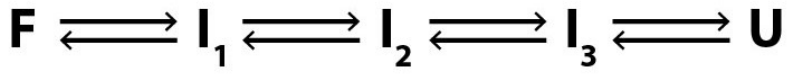


Figure 1.17. A complex folding process where the biomolecule populates three folding intermediates along the unfolding trajectory.

The simulated DSC profile for this model clearly exhibits multiple folding transitions and therefore $1 - P_F - P_U$ is non-zero (Figure 1.18a). After calculating the populations of the folded and unfolded states according to Equations 1.24 and 1.25, the starting point for the deconvolution analysis of these types of DSC thermograms is to calculate a new average excess enthalpy normalized by the population of molecules that do not reside in the reference folded state

$$\varphi_1 = \frac{\langle \Delta H \rangle}{1 - P_F} \quad . \tag{1.26}$$

For a two-state process, φ_1 is equal to $\Delta H_{F,U}$ at all temperatures

$$\varphi_1 = \frac{\langle \Delta H \rangle}{1 - P_F} = \frac{\Delta H_{F,U} P_U}{1 - P_F} = \frac{\Delta H_{F,U} P_U}{P_U} = \Delta H_{F,U} \, . \tag{1.27}$$

In more complex folding scenarios, φ_1 is a function with a lower limit equal to the enthalpy change $\Delta H_{F,1}$ associated with populating the first intermediate state I_1 from the folded state (Figure 1.18b light orange line). Assuming the situation in Figure 1.17

$$\varphi_1 = \frac{\sum_{i=1}^{U} \Delta H_{F,i} P_i}{\sum_{i=1}^{U} P_i} = \frac{\Delta H_{F,1} P_1 + \sum_{i=2}^{U} \left(\Delta H_{F,1} + \Delta H_{1,i} \right) P_i}{P_1 + \sum_{i=2}^{U} P_i} = \Delta H_{F,1} + \frac{\sum_{i=2}^{U} \Delta H_{1,i} \frac{P_i}{P_1}}{1 + \sum_{i=2}^{U} \frac{P_i}{P_1}} \tag{1.28}$$

where the ΔHs in the right-hand side of Equation 1.28 have now been defined relative to the first intermediate state. The right-hand term of Equation 1.28 is the average excess enthalpy where the

reference state has been set as the first folding intermediate, i.e. it is mathematically equivalent to taking I_1 as if it were the folded reference for all later populations (I_2, I_3, and U). Therefore

$$\varphi_1 = \Delta H_{F,1} + \langle \Delta H_1 \rangle . \tag{1.29}$$

Subtracting the minimum of φ_1 (corresponding to $\Delta H_{F,1}$) eliminates the thermodynamic contribution of the first transition, leaving a subsystem containing n-1 energy states with I_1 as the reference. A new partition function can be calculated according to

$$Q_1 = \exp\left(\int_{T_0}^{T} \frac{\langle \Delta H_1 \rangle}{RT^2} dT \right) \tag{1.30}$$

And the relative population of the first intermediate state = $[I_1]/C_T$ (Figure 1.18c) is calculated with

$$P_1 = P_F \frac{Q-1}{Q_1} . \tag{1.31}$$

A new normalized excess enthalpy, φ_2, can be calculated as

$$\varphi_2 = \frac{\langle \Delta H_1 \rangle}{1 - P_{1*}} = \Delta H_{1,2} + \langle \Delta H_2 \rangle \tag{1.32}$$

where P_{1*} is not the true population of P_1 as calculated by Equation 1.31, instead it is the population as the reference for all later states = $1/Q_1$. The minimum of φ_2 ($\Delta H_{1,2}$) can be subtracted to obtain $\langle \Delta H_2 \rangle$, and the process repeated to calculate the enthalpy changes for transitioning between each successive state (Figure 1.18b orange and red lines).

Assuming that enthalpy and entropy are temperature-independent, the temperatures at which the populations of states i and i-1 (calculated according to Equations 1.24, 1.25, and 1.31) are equal can be taken as the T_ms for each transition (Figure 1.18c colored circles). The associated changes in entropy between each state can be calculated as $\Delta H/T_m$. The C_p profile using the parameters extracted from the deconvolution procedure can then be calculated (Equation 1.20) for

comparison with the original (Figure 1.18d). This type of model-free recursion analysis can also be applied to association reactions such as duplex formation[187]. With accurate heat capacity data and proper baseline correction, complete thermodynamic characterizations of multi-state folding landscapes can be obtained in a model-free manner, providing a wealth of information for further biophysical investigation. The model-free parameters provided by the DSC recursion analysis can also be applied as initial estimates in direct fitting of the dataset to provide a cross-validation of the extracted parameters and populations.

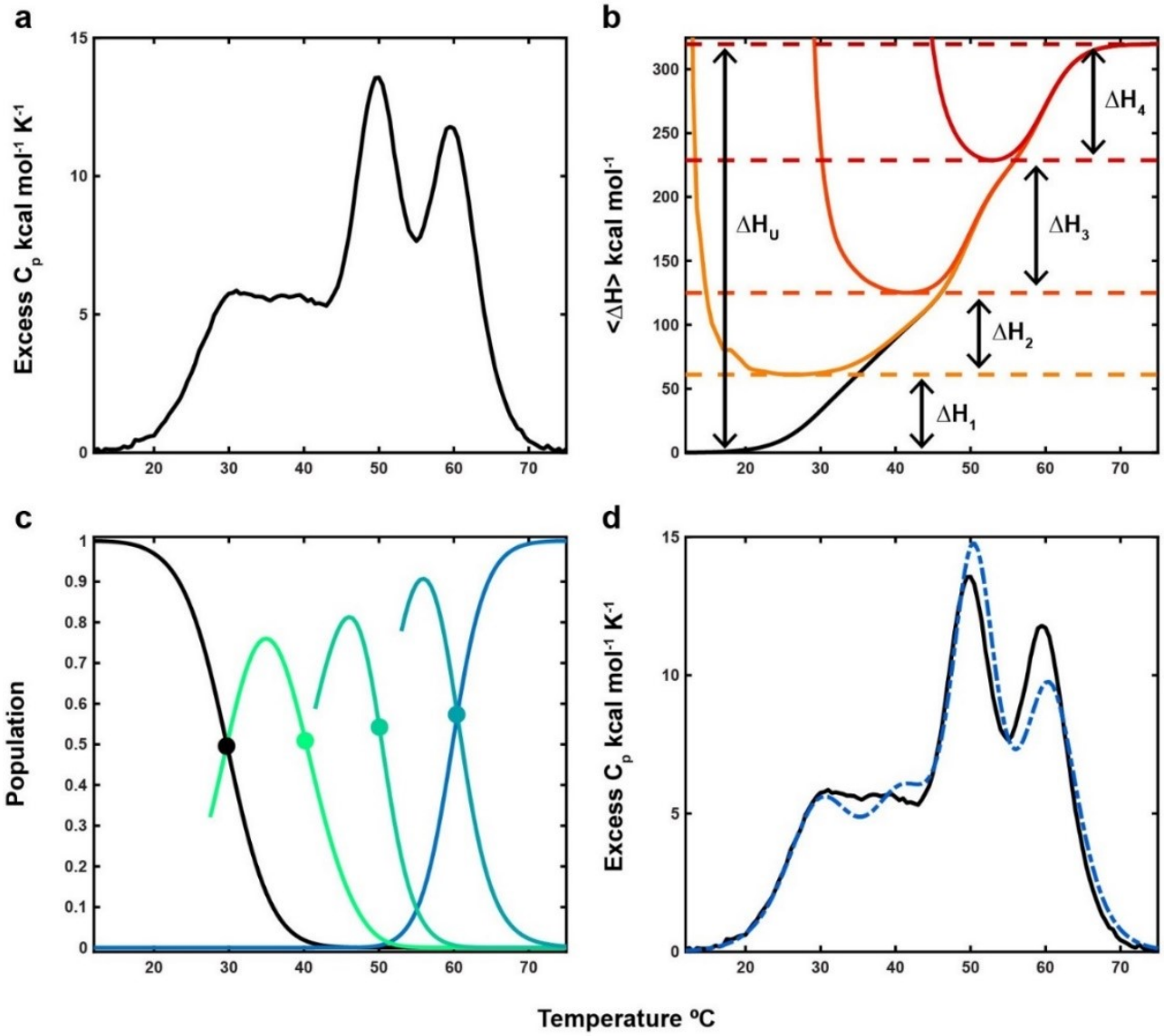


Figure 1.18. Model-free deconvolution of a complex DSC profile. (a) A simulated multi-state excess C_p DSC thermogram. Simulation parameters were obtained from Freire and Biltonen[184] in order to reproduce the deconvolution process for this thesis. (b) $\langle \Delta H \rangle$ functions calculated using the recursive deconvolution process. The total and subsystem $\langle \Delta H \rangle$s are shown as black and colored lines respectively. Light orange to red indicates the first to last subsystem $\langle \Delta H \rangle$ in the deconvolution process. The ΔHs for transitioning between each state are calculated from the differences in the minima of the $\langle \Delta H \rangle$ profiles, shown as the short black double-headed arrows. The total ΔH for unfolding, ΔH_U, is indicated by the long black double-headed arrow. The large fluctuation in the magnitudes of the colored $\langle \Delta H \rangle$ profiles at low temperature are from artefacts in the numerical integration procedure. (c) Populations of each state calculated from the recursion analysis. The population of the folded state is shown in black, and light green to blue indicates the intermediate to unfolded state populations. The colored circles indicate the estimated T_ms of each population that are used in conjunction with the sequential ΔHs from (b) to calculate the ΔS for each transition. The intermediate populations are truncated at lower temperatures as a result of the artefact in numerical integration of the $\langle \Delta H \rangle$ functions. (d) Simulated excess C_p profiles using the original thermodynamic parameters (black line) and the results of the deconvolution procedure (dashed-dotted blue line).

1.5.5. Binding polynomials

Nucleic acid folding often contains association events where strands assemble into higher order complexes, for example in duplex annealing or in metabolite binding to riboswitches. The analysis of these types of events are more complicated since the folding depends on the concentration of the binding partner, in contrast to an intramolecular folding process. As a result, the partition functions for binding events are polynomials in terms of the concentration of the binding partner[243, 244], hence the term "binding polynomial". Binding polynomials are extremely powerful for analyzing folding and binding dynamics because they may be written for systems of virtually any complexity. They have been used extensively to describe allostery in dimeric proteins

that undergo folding and binding transitions[245, 246], for example. The application of binding polynomials to nucleic acid folding is most simply demonstrated by considering the formation of a duplex AB from strands A and B (Figure 1.19).

$$\mathbf{A + B \rightleftharpoons AB}$$

Figure 1.19. A two-state equilibrium model for the formation of a heteroduplex AB from strands A and B**.**

The mass conservation equations for strands A and B are

$$A_T = [A] + [AB] \tag{1.33}$$

$$B_T = [B] + [AB] \tag{1.34}$$

where A_T and B_T are the total concentrations of strands A and B. The equilibrium constant is

$$K = \frac{[AB]}{[A][B]} \tag{1.35}$$

which can be solved for $[AB]$ and substituted into Equation 1.33, giving the partition function upon dividing by the $[A]$

$$Q = 1 + K[B] = \frac{A_T}{[A]}\,. \tag{1.36}$$

This is a first order binding polynomial in $[B]$, where the reference state = 1 as been set to the free A strand. In order to calculate the duplex and free strand populations for application to thermal denaturation profiles, the concentrations of the free and duplex states must be determined. Because there are three equations and the three unknown concentration variables $[A]$, $[B]$, and $[AB]$, this

system of equations can be solved for one of the concentrations, from which all other concentrations can be determined. Rewriting the equilibrium constant and substituting Equations 1.33 and 1.34 yields

$$[AB]^2 + [AB]\left(-A_T - B_T - K^{-1}\right) + A_T B_T = 0 \tag{1.37}$$

Since this is a quadratic equation, the [*AB*] can be solved according to

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{1.38}$$

where substitution of $x = [AB]$, $a = 1$, $b = (-A_T - B_T - K^{-1})$, and $c = A_T B_T$ gives

$$[AB] = \frac{1}{2}\left(A_T + B_T + K^{-1} - \sqrt{\left(A_T + B_T + K^{-1}\right)^2 - 4A_T B_T} \right) \tag{1.39}$$

as the real, positive root. Using the [*AB*], the populations of the free and duplex states are calculated with Equation 1.36. This solution is identical to that for analysis of single site ITC binding isotherms[247], where a small molecule inhibitor binds to a protein target. In an ITC analysis, the concentration of the bound state is calculated as a function of ligand concentration at a single temperature. In contrast, the analysis of thermal denaturation profiles with binding polynomials requires calculating the concentration of the folded bound state as a function of temperature at a single binding partner concentration. The binding polynomials for more complicated folding and binding scenarios can have higher order terms for the binding partner (detailed in Chapter 3), which may not be possible to solve for analytically. In these instances, it is possible to resort to numerical approaches such as Newton's method to calculate the required concentrations.

1.5.6. Folding kinetics

1.5.6.1. Thermal hysteresis

The theory of multi-state folding models that has been presented thus far assumes that the folding processes rapidly equilibrate with respect to the temperature scan rate. This means that the net rates of interconversion for each state are close to zero throughout a thermal denaturation experiment and their corresponding equilibrium populations can be calculated with equilibrium constants dictated by standard state free energy changes. The assumption of equilibrium with respect to the scan rate can be checked by examining a forward and reverse scan in a thermal denaturation experiment. If the two scans are superimposable, i.e. the shapes are closely similar and the difference between the transition temperatures in each direction is not more than ~1 °C, the biomolecule can be said to be at equilibrium throughout the experiment (Figure 1.20a). However, many nucleic acids have slow folding and unfolding kinetics (such as i-motifs and intermolecular GQs) and their populations are not at equilibrium throughout a thermal scan. The forward and reverse thermal scans for a nucleic acid with slow folding kinetics will have large differences between the apparent transition temperatures; the apparent T_ms of the forward and reverse scans are shifted to higher and lower values respectively relative to their equilibrium positions (Figure 1.20b). This phenomenon is known as thermal hysteresis (TH)[235].

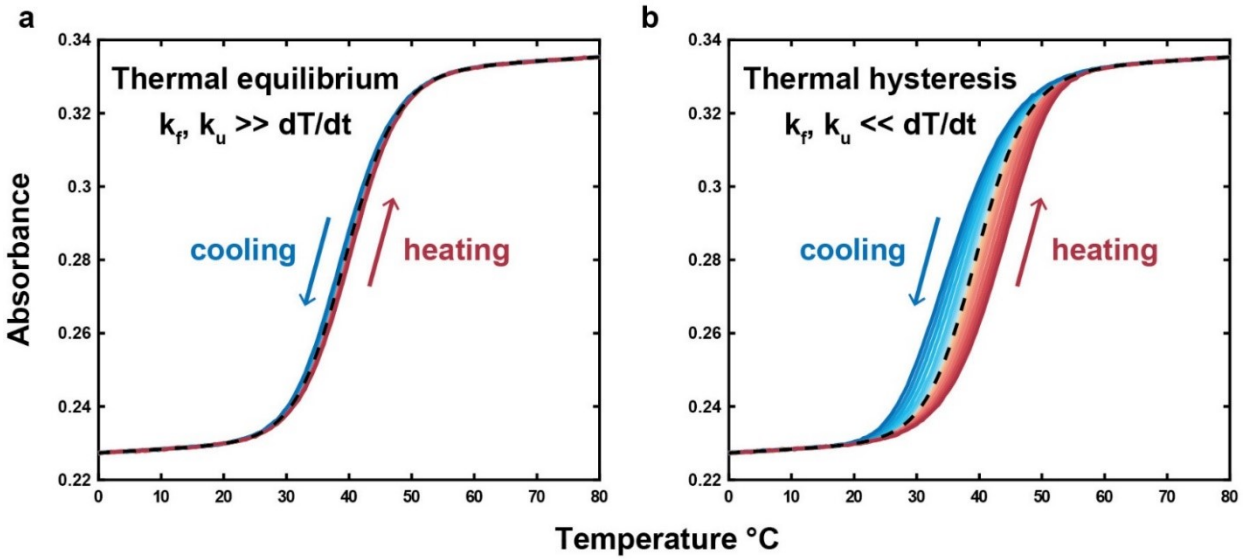


Figure 1.20. Simulated two-state equilibrium and TH folding profiles as a function of temperature scan rate. (a) A thermal equilibrium dataset where heating and cooling scans are essentially superimposable at all temperature scan rates. The folding kinetics are rapid relative to the scan rate. (b) A TH dataset where the heating and cooling scans are not superimposable because the folding kinetics are slow with respect to the scan rate. In (a,b), dark to light blue and dark red to orange indicates fastest to slowest cooling and heating scan rates respectively. Dashed black lines indicate the equilibrium profile calculated using an equilibrium constant. In (a), the profiles closely overlay with their equilibrium position. In (b), the scans only become close to their equilibrium positions at ultra-slow scan rates.

TH is a direct result of slow equilibration kinetics with respect to the scan rate; the populations lag behind their equilibrium values, leading to suppression of the thermal transition in each direction. The extent of TH can be modified by changing the temperature scan rate. Slower scan rates attenuate TH by allowing additional time for the system to equilibrate. Conversely, faster scan rates increase the amount of TH by pushing the system further out of equilibrium as a result of less equilibration time per unit temperature. In some cases, it is possible to bring the biomolecule close to equilibrium by decreasing the magnitude of the scan rate, although this

depends on the kinetics of the system at hand. Nucleic acids that exhibit TH in their thermal denaturation experiments can still be analyzed with discrete state models, however they cannot be analyzed with equilibrium thermodynamics approaches. Instead, they must be analyzed with kinetics equations in order to extract the rate constants and barrier energies governing the transitions between conformations.

1.5.6.2. Analysis of TH experiments

TH has largely been used to study two-state folding in nucleic acids[235, 248, 249]. It has also been useful in analyzing the kinetics of unfolding in monomolecular protein systems that undergo irreversible multi-step denaturation by DSC[250]. In Chapter 4, a new methodology for mapping the folding landscapes of supramolecular assemblies by TH will be presented that builds on these fundamentals. The basics of TH analysis will be given here as a primer. All analyses of TH data rely on a simple, yet highly important relationship between the time and temperature derivatives of the biomolecular concentrations[235]. Using the inverse scan rate dt/dT

$$\frac{d}{dT}C = \frac{d}{dt}C\frac{dt}{dT} \tag{1.40}$$

where C is the concentration of a given biomolecular state. Using this conversion, standard differential equations accounting for the rates of interconversion of biomolecular states with respect to time can be applied to TH datasets. In the first step of a traditional TH analysis, the data are converted to fraction folded with appropriate baseline selection and the temperature derivatives are calculated as the slope of the fraction folded with respect to temperature. Next, rate equations for the folding trajectory are written and converted to their temperature derivatives using Equation

1.40. Taking a two-state folding process as a simple example, the rate equations governing the interconversion of the folded (θ_F) and unfolded (θ_U) fractions with respect to temperature are

$$\frac{d}{dT}\theta_F = \frac{d}{dt}\theta_F \frac{dt}{dT} = \left(k_F\theta_U - k_U\theta_F\right)\frac{dt}{dT} \tag{1.41}$$

$$\frac{d}{dT}\theta_U = \frac{d}{dt}\theta_U \frac{dt}{dT} = \left(-k_F\theta_U + k_U\theta_F\right)\frac{dt}{dT} \tag{1.42}$$

where k_F and k_U are the folding and unfolding rate constants respectively. This is a system of coupled differential equations that can be solved to extract the values of the rate constants through the transition regions of the TH experiment, since the θ_F, θ_U, $d\theta_F/dT$, and $d\theta_U/dT$ can be calculated from the experimental data directly. The activation energies governing the transition between the folded and unfolded states can be obtained from the linear form of the Arrhenius equation where a plot of ln(k) versus 1/T is a straight line with a slope of -E_a/R and an intercept of ln(A)[235]. When TH datasets appear highly complex with multiple overlapping or broad transitions, the data can be fit directly with kinetic models where the number of states and fit parameters are optimized in a computer program. This procedure is described in detail in Chapters 3 and 4.

1.6. Global fitting analysis

The work presented in this thesis makes extensive use of a fitting technique known as global analysis. In global fitting analysis, multiple datasets are fit to a single, unifying physical model that is thought to describe the underlying behavior of the system in question. For example, thermal denaturation profiles for a supramolecular assembly process collected at multiple scan rates and concentrations may be globally fit by applying one model to the entirety of the data. The adjustable parameters in the fit are shared in the calculation of each data type during the

minimization procedure (hence the global nature of the fit). This yields a single set of optimized parameters that best captures the shapes of all experimental profiles. In contrast, a traditional fitting approach entails carrying out separate fits to individual datasets and consequently multiple parameter sets are obtained, with no direct link between them. There are two major advantages to performing global fitting analyses relative to individual fits. These are: (i) that excellent agreement between the globally-fitted and all experimental data ensures the proposed model is a plausible description of the underlying physical phenomena[251]. To illustrate, separate fits of two different folding models to a single thermal denaturation profile may result in closely similar fit qualities, precluding the identification of the correct model based on the goodness of fit. Expanding the breadth of the data by varying an experimental parameter that influences the folding (other than temperature, e.g. concentration) and performing a global fit provides additional information to exclude models that only give good fits under one or even a subset of the explored conditions[252]. (ii) Performing global fits increases the accuracy of the extracted fit parameters[251], since the greater abundance of experimental data collected over multiple conditions acts as a fitting constraint and therefore the number of parameter combinations that can describe the global dataset is reduced.

1.7. Thesis objectives

Nucleic acid dynamics are highly complex. Conventional techniques to study nucleic acid folding and assembly are costly in terms of time and sample, often requiring special instrumentation and extensive user expertise. The study of conformational rearrangements in nucleic acid ensembles is therefore out of easy reach in many cases. Thermal denaturation experiments represent a facile approach for swift and cost-effective generation of large datasets on nucleic acid dynamics because they are relatively high-throughput and need small amounts of

sample. Furthermore, thermal denaturation instruments such as the absorbance spectrophotometer are near-ubiquitous, making these types of experiments widely accessible. Thermal denaturation datasets for nucleic acids are particularly amenable to global analysis since nucleic acid folding and assembly dynamics typically need to be described by multiple physical parameters governing the transitions between each state. Global fitting analyses therefore provide a unique opportunity to robustly define the underlying dynamic behavior manifested across multiple thermal denaturation datasets for a nucleic acid ensemble. This thesis will explore several novel global fitting analyses that we have developed to extract detailed information on nucleic acid dynamics from thermal denaturation experiments. The global fitting methodologies presented in the following chapters are particularly rapid, low-cost, and provide information on nucleic acid dynamics over a wide temperature range. These methods deliver a level of thermodynamic and kinetic detail that is difficult to access with current techniques, and in a fraction of the time and input required.

In Chapter 2, we developed a new global fitting analysis applied to GQs within the promoter regions of human genes that we demonstrated to exist as large ensembles of GR isomers (up to 12). This allowed us to quantitate the influence that GR exchange dynamics have on GQ stability. We revealed that populating multiple GR isomers substantially improves the GQs melting temperature, which can directly influence gene expression. We additionally performed a bioinformatic analysis that suggests GR exchange dynamics are widespread in the human genome as a mechanism for modulating promoter GQ stability and protein interactions. Chapter 3 details an experimental approach using thermolabile ligands to generate a series of ligand-bound aptamer DSC profiles from a single experiment. These are analyzed simultaneously with the curve for the unbound transition in a global fitting analysis to extract the thermodynamics governing folding

and binding to two distinct ligands in a competition-type scenario. This enabled us to drastically reduce the experimental time and sample required for folding and binding analyses by DSC. Computer simulations of more complicated scenarios involving thermolabile ligand binding by DSC were also presented as a visual guide for users of our method. In Chapter 4, we developed a combined model-free and global fitting method for multi-scan rate TH profiles and applied it to the assembly of a tetramolecular GQ and the polymerization of poly(A) fibers. Our global fits revealed that tetramolecular GQ formation occurs through a sequential folding pathway with dominant free energy barriers that shift in response to temperature. In application to poly(A) fiber assembly, we demonstrated that fiber formation is cooperative and elongation is driven by the favorable addition of a fourth strand to an unstable trimeric nucleus containing three poly(A) monomers. Viewed as a whole, we have used global fitting analyses to successfully describe a diverse cross section of nucleic acids that fold and assemble via widely divergent mechanisms.

1.8. References

1. Blackburn, G.M. & Royal Society of Chemistry (Great Britain) Nucleic acids in chemistry and biology, Edn. 3rd. (RSC Pub., Cambridge; 2006).
2. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
3. Garst, A.D., Edwards, A.L. & Batey, R.T. Riboswitches: Structures and mechanisms. *Cold Spring Harbor perspectives in biology* **3**, a003533 (2011).
4. Seeman, N.C. & Sleiman, H.F. DNA nanotechnology. *Nature Reviews Materials* **3**, 17068 (2017).
5. Chargaff, E., Lipshitz, R. & Green, C. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J. Biol. Chem.* **195**, 155-160 (1952).
6. Olby, R.C. The path to the double helix. (University of Washington Press, Seattle; 1974).
7. Bewley, C.A., Gronenborn, A.M. & Clore, G.M. Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 105-131 (1998).
8. Wei, D., Wilson, W.D. & Neidle, S. Small-molecule binding to the DNA minor groove is mediated by a conserved water cluster. *J. Am. Chem. Soc.* **135**, 1369-1377 (2013).
9. Hamilton, P.L. & Arya, D.P. Natural product DNA major groove binders. *Nat Prod Rep* **29**, 134-143 (2012).
10. Guckian, K.M. et al. Experimental Measurement of Aromatic Stacking Affinities in the Context of Duplex DNA. *J. Am. Chem. Soc.* **118**, 8182-8183 (1996).
11. von Hippel, P.H., Johnson, N.P. & Marcus, A.H. Fifty years of DNA "breathing": Reflections on old and new approaches. *Biopolymers* **99**, 923-954 (2013).
12. Rich, A. The double helix: a tale of two puckers. *Nat. Struct. Biol.* **10**, 247-249 (2003).
13. Hoogsteen, K. The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallographica* **12**, 822-823 (1959).
14. Bloomfield, V.A. DNA condensation by multivalent cations. *Biopolymers* **44**, 269-282 (1997).
15. Gueroult, M., Boittin, O., Mauffret, O., Etchebest, C. & Hartmann, B. Mg2+ in the major groove modulates B-DNA structure and dynamics. *PLoS One* **7**, e41704 (2012).
16. Nelson, D.L., Cox, M.M. & Lehninger, A.L. Lehninger principles of biochemistry, Edn. 7th edition. (W.H. Freeman, New York; 2017).
17. DiMaio, F. et al. Virology. A virus that infects a hyperthermophile encapsidates A-form DNA. *Science* **348**, 914-917 (2015).
18. Wang, A.H. et al. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**, 680-686 (1979).
19. Wing, R. et al. Crystal structure analysis of a complete turn of B-DNA. *Nature* **287**, 755-758 (1980).

20. Potaman, V.N. & Sinden, R.R. in Madame Curie Bioscience Database (Landes Bioscience, Austin, TX; 2000-2013).
21. Wang, G. & Vasquez, K.M. Z-DNA, an active element in the genome. *Front Biosci* **12**, 4424-4438 (2007).
22. Whelan, D.R. et al. Detection of an en masse and reversible B- to A-DNA conformational transition in prokaryotes in response to desiccation. *J R Soc Interface* **11**, 20140454 (2014).
23. Oh, D.B., Kim, Y.G. & Rich, A. Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. *Proc Natl Acad Sci U S A* **99**, 16666-16671 (2002).
24. Jiao, Y., Stringfellow, S. & Yu, H. Distinguishing "looped-out" and "stacked-in" DNA bulge conformation using fluorescent 2-aminopurine replacing a purine base. *J. Biomol. Struct. Dyn.* **19**, 929-934 (2002).
25. Cheung, A.K. A stem-loop structure, sequence non-specific, at the origin of DNA replication of porcine circovirus is essential for termination but not for initiation of rolling-circle DNA replication. *Virology* **363**, 229-235 (2007).
26. Bikard, D., Loot, C., Baharoglu, Z. & Mazel, D. Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol. Mol. Biol. Rev.* **74**, 570-588 (2010).
27. Phan, A.T. Human telomeric G-quadruplex: structures of DNA and RNA sequences. *FEBS J.* **277**, 1107-1117 (2010).
28. Rhodes, D. & Lipps, H.J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* **43**, 8627-8637 (2015).
29. Huppert, J.L. & Balasubramanian, S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35**, 406-413 (2007).
30. Huppert, J.L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908-2916 (2005).
31. Li, X.M. et al. Guanine-vacancy-bearing G-quadruplexes responsive to guanine derivatives. *Proc Natl Acad Sci U S A* **112**, 14581-14586 (2015).
32. Mukundan, V.T. & Phan, A.T. Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. *J. Am. Chem. Soc.* **135**, 5017-5028 (2013).
33. Rachwal, P.A., Brown, T. & Fox, K.R. Effect of G-tract length on the topology and stability of intramolecular DNA quadruplexes. *Biochemistry* **46**, 3036-3044 (2007).
34. Smirnov, I. & Shafer, R.H. Effect of loop sequence and size on DNA aptamer stability. *Biochemistry* **39**, 1462-1468 (2000).
35. Harkness, R.W.,V & Mittermaier, A.K. G-register exchange dynamics in guanine quadruplexes. *Nucleic Acids Res.* **44**, 3481-3494 (2016).
36. Dai, J. et al. Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation. *Nucleic Acids Res.* **35**, 2440-2450 (2007).
37. Phan, A.T. & Mergny, J.L. Human telomeric DNA: G-quadruplex, i-motif and Watson-Crick double helix. *Nucleic Acids Res.* **30**, 4618-4625 (2002).

38. Murat, P. et al. G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. *Nat. Chem. Biol.* **10**, 358-364 (2014).
39. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. & Hurley, L.H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* **99**, 11593-11598 (2002).
40. Ou, T.M. et al. G-quadruplexes: targets in anticancer drug design. *ChemMedChem* **3**, 690-713 (2008).
41. Arora, A. et al. Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. *RNA* **14**, 1290-1296 (2008).
42. Arora, A. & Suess, B. An RNA G-quadruplex in the 3' UTR of the proto-oncogene PIM1 represses translation. *RNA Biol* **8**, 802-805 (2011).
43. Endoh, T., Kawasaki, Y. & Sugimoto, N. Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor alpha. *Nucleic Acids Res.* **41**, 6222-6231 (2013).
44. Gatto, B., Palumbo, M. & Sissi, C. Nucleic acid aptamers based on the G-quadruplex structure: therapeutic and diagnostic potential. *Curr. Med. Chem.* **16**, 1248-1265 (2009).
45. Shastri, A. et al. An aptamer-functionalized chemomechanically modulated biomolecule catch-and-release system. *Nat Chem* **7**, 447-454 (2015).
46. Wang, C. et al. Enantioselective Diels-Alder reactions with G-quadruplex DNA-based catalysts. *Angew Chem Int Ed Engl* **51**, 9352-9355 (2012).
47. Jiang, H.X., Kong, D.M. & Shen, H.X. Amplified detection of DNA ligase and polynucleotide kinase/phosphatase on the basis of enrichment of catalytic G-quadruplex DNAzyme by rolling circle amplification. *Biosens Bioelectron* **55**, 133-138 (2014).
48. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **34**, 5402-5415 (2006).
49. Lim, K.W. et al. Coexistence of two distinct G-quadruplex conformations in the hTERT promoter. *J. Am. Chem. Soc.* **132**, 12331-12342 (2010).
50. Chung, W.J. et al. Structure of a left-handed DNA G-quadruplex. *Proceedings of the National Academy of Sciences* **112**, 2729-2733 (2015).
51. Fu, B. et al. Right-handed and left-handed G-quadruplexes have the same DNA sequence: distinct conformations induced by an organic small molecule and potassium. *Chemical Communications* **52**, 10052-10055 (2016).
52. Phan, A.T., Kuryavyi, V., Burge, S., Neidle, S. & Patel, D.J. Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Am. Chem. Soc.* **129**, 4386-4392 (2007).
53. Ghimire, C. et al. Direct Quantification of Loop Interaction and π–π Stacking for G-Quadruplex Stability at the Submolecular Level. *J. Am. Chem. Soc.* **136**, 15537-15544 (2014).

54. Zhang, Z., Dai, J., Veliath, E., Jones, R.A. & Yang, D. Structure of a two-G-tetrad intramolecular G-quadruplex formed by a variant human telomeric sequence in K(+) solution: insights into the interconversion of human telomeric G-quadruplex structures. *Nucleic Acids Res.* **38**, 1009-1021 (2010).
55. Agrawal, P., Hatzakis, E., Guo, K., Carver, M. & Yang, D. Solution structure of the major G-quadruplex formed in the human VEGF promoter in K(+): insights into loop interactions of the parallel G-quadruplexes. *Nucleic Acids Res.* **41**, 10584-10592 (2013).
56. Ambrus, A., Chen, D., Dai, J., Jones, R.A. & Yang, D. Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry* **44**, 2048-2058 (2005).
57. Harkness, R.W.,V & Mittermaier, A.K. G-quadruplex dynamics. *Biochim. Biophys. Acta* **1865**, 1544-1554 (2017).
58. Šket, P., Črnugelj, M. & Plavec, J. Identification of mixed di-cation forms of G-quadruplex in solution. *Nucleic Acids Res.* **33**, 3691-3697 (2005).
59. Lee, M.P.H., Parkinson, G.N., Hazel, P. & Neidle, S. Observation of the Coexistence of Sodium and Calcium Ions in a DNA G-Quadruplex Ion Channel. *J. Am. Chem. Soc.* **129**, 10106-10107 (2007).
60. Liu, W. et al. Kinetics and mechanism of G-quadruplex formation and conformational switch in a G-quadruplex of PS2.M induced by Pb2+. *Nucleic Acids Res.* **40**, 4229-4236 (2012).
61. Chen, F.M. Strontium(2+) facilitates intermolecular G-quadruplex formation of telomeric sequences. *Biochemistry* **31**, 3769-3776 (1992).
62. Deng, H. & Braunlin, W.H. Kinetics of Sodium Ion Binding to DNA Quadruplexes. *J. Mol. Biol.* **255**, 476-483 (1996).
63. Haider, S., Parkinson, G.N. & Neidle, S. Crystal structure of the potassium form of an Oxytricha nova G-quadruplex. *J. Mol. Biol.* **320**, 189-200 (2002).
64. Lodish, H., Berk, A. Zipursky, S. L., Matsudaira, P., Baltimore, D., Darnell, J. in Molecular Cell Biology, Edn. 4th edition (W. H. Freeman, New York; 2000).
65. Campbell, N.H. & Neidle, S. G-quadruplexes and metal ions. *Met Ions Life Sci* **10**, 119-134 (2012).
66. Hardin, C.C., Perry, A.G. & White, K. Thermodynamic and kinetic characterization of the dissociation and assembly of quadruplex nucleic acids. *Biopolymers* **56**, 147-194 (2000).
67. Podbevsek, P., Hud, N.V. & Plavec, J. NMR evaluation of ammonium ion movement within a unimolecular G-quadruplex in solution. *Nucleic Acids Res.* **35**, 2554-2563 (2007).
68. Šket, P., Virgilio, A., Esposito, V., Galeone, A. & Plavec, J. Strand directionality affects cation binding and movement within tetramolecular G-quadruplexes. *Nucleic Acids Res.* **40**, 11047-11057 (2012).

69. Wei, D., Parkinson, G.N., Reszka, A.P. & Neidle, S. Crystal structure of a c-kit promoter quadruplex reveals the structural role of metal ions and water molecules in maintaining loop conformation. *Nucleic Acids Res.* **40**, 4691-4700 (2012).
70. Islam, B. et al. Extended molecular dynamics of a c-kit promoter quadruplex. *Nucleic Acids Res.* **43**, 8673-8693 (2015).
71. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res* **34**, 5402-5415 (2006).
72. Hazel, P., Huppert, J., Balasubramanian, S. & Neidle, S. Loop-Length-Dependent Folding of G-Quadruplexes. *J. Am. Chem. Soc.* **126**, 16405-16415 (2004).
73. Dai, J., Chen, D., Jones, R.A., Hurley, L.H. & Yang, D. NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region. *Nucleic Acids Res.* **34**, 5133-5144 (2006).
74. Phan, A.T., Modi, Y.S. & Patel, D.J. Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter. *J. Am. Chem. Soc.* **126**, 8710-8716 (2004).
75. Amrane, S. et al. A novel chair-type G-quadruplex formed by a Bombyx mori telomeric sequence. *Nucleic Acids Res.* **37**, 931-938 (2009).
76. Lim, K.W. et al. Structure of the human telomere in K(+) solution: a stable basket-type G-quadruplex with only two G-tetrad layers. *J. Am. Chem. Soc.* **131**, 4301-4309 (2009).
77. Lane, A.N., Chaires, J.B., Gray, R.D. & Trent, J.O. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.* **36**, 5482-5515 (2008).
78. Benabou, S., Avino, A., Eritja, R., Gonzalez, C. & Gargallo, R. Fundamental aspects of the nucleic acid i-motif structures. *RSC Advances* **4**, 26956-26980 (2014).
79. Gehring, K., Leroy, J.L. & Gueron, M. A tetrameric DNA structure with protonated cytosine.cytosine base pairs. *Nature* **363**, 561-565 (1993).
80. Dai, J., Hatzakis, E., Hurley, L.H. & Yang, D. I-Motif Structures Formed in the Human c-MYC Promoter Are Highly Dynamic–Insights into Sequence Redundancy and I-Motif Stability. *PLOS ONE* **5**, e11647 (2010).
81. Simonsson, T., Pribylova, M. & Vorlickova, M. A Nuclease Hypersensitive Element in the Human c-myc Promoter Adopts Several Distinct i-Tetraplex Structures. *Biochem. Biophys. Res. Commun.* **278**, 158-166 (2000).
82. Kang, H.J., Kendrick, S., Hecht, S.M. & Hurley, L.H. The transcriptional complex between the BCL2 i-motif and hnRNP LL is a molecular switch for control of gene expression that can be modulated by small molecules. *J. Am. Chem. Soc.* **136**, 4172-4185 (2014).
83. Du, Z. et al. Crystal structure of the first KH domain of human poly(C)-binding protein-2 in complex with a C-rich strand of human telomeric DNA at 1.7 A. *J. Biol. Chem.* **280**, 38823-38830 (2005).
84. Eid, J.E. & Sollner-Webb, B. ST-1, a 39-kilodalton protein in Trypanosoma brucei, exhibits a dual affinity for the duplex form of the 29-base-pair subtelomeric repeat and its C-rich strand. *Mol. Cell. Biol.* **15**, 389-397 (1995).

85. Dembska, A., Bielecka, P. & Juskowiak, B. pH-Sensing fluorescence oligonucleotide probes based on an i-motif scaffold: a review. *Analytical Methods* **9**, 6092-6106 (2017).
86. Nesterova, I.V. & Nesterov, E.E. Rational design of highly responsive pH sensors based on DNA i-motif. *J. Am. Chem. Soc.* **136**, 8843-8846 (2014).
87. Fleming, A.M. et al. 4n–1 Is a "Sweet Spot" in DNA i-Motif Folding of 2′-Deoxycytidine Homopolymers. *J. Am. Chem. Soc.* **139**, 4682-4689 (2017).
88. Wright, E.P., Huppert, J.L. & Waller, Zoë A E. Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. *Nucleic Acids Res.* **45**, 2951-2959 (2017).
89. Yang, B. & Rodgers, M.T. Base-pairing energies of proton-bound heterodimers of cytosine and modified cytosines: implications for the stability of DNA i-motif conformations. *J. Am. Chem. Soc.* **136**, 282-290 (2014).
90. Berger, I., Egli, M. & Rich, A. Inter-strand C-H...O hydrogen bonds stabilizing four-stranded intercalated molecules: stereoelectronic effects of O4' in cytosine-rich DNA. *Proc Natl Acad Sci U S A* **93**, 12116-12121 (1996).
91. Lieblein, A.L., Buck, J., Schlepckow, K., Furtig, B. & Schwalbe, H. Time-resolved NMR spectroscopic studies of DNA i-motif folding reveal kinetic partitioning. *Angew Chem Int Ed Engl* **51**, 250-253 (2012).
92. Assi, H.A. et al. Stabilization of i-motif structures by 2'-beta-fluorination of DNA. *Nucleic Acids Res.* **44**, 4998-5009 (2016).
93. Abou Assi, H., El-Khoury, R., Gonzalez, C. & Damha, M.J. 2'-Fluoroarabinonucleic acid modification traps G-quadruplex and i-motif structures in human telomeric DNA. *Nucleic Acids Res.* **45**, 11535-11546 (2017).
94. Dzatko, S., Krafcikova, M., Hansel-Hertsch, R., Fessl, T., Fiala, R., Loja, T., Krafcik, D., Mergny, J-L., Foldynova-Trantirkova, S., Trantirek, L. Evaluation of stability of DNA i-motifs in the nuclei of living mammalian cells. *Angewandte Chemie International Edition* (2017).
95. Zhou, J. & Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.* **16**, 440 (2017).
96. Qu, H. et al. Rapid and Label-Free Strategy to Isolate Aptamers for Metal Ions. *ACS Nano* **10**, 7558-7565 (2016).
97. Shiratori, I. et al. Selection of DNA aptamers that bind to influenza A viruses with high affinity and broad subtype specificity. *Biochem. Biophys. Res. Commun.* **443**, 37-41 (2014).
98. Rahimi, F. & Bitan, G. Selection of aptamers for amyloid beta-protein, the causative agent of Alzheimer's disease. *J Vis Exp* (2010).
99. Mayer, G. The chemical biology of aptamers. *Angew Chem Int Ed Engl* **48**, 2672-2689 (2009).
100. Strehlitz, B., Reinemann, C., Linkorn, S. & Stoltenburg, R. Aptamers for pharmaceuticals and their application in environmental analytics. *Bioanalytical Reviews* **4**, 1-30 (2012).

101. Stojanovic, M.N. & Landry, D.W. Aptamer-based colorimetric probe for cocaine. *J. Am. Chem. Soc.* **124**, 9678-9679 (2002).
102. Ni, X., Castanares, M., Mukherjee, A. & Lupold, S.E. Nucleic acid aptamers: clinical applications and promising new horizons. *Curr. Med. Chem.* **18**, 4206-4214 (2011).
103. Bock, L.C., Griffin, L.C., Latham, J.A., Vermaas, E.H. & Toole, J.J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* **355**, 564-566 (1992).
104. Catherine, A.T., Shishido, S.N., Robbins-Welty, G.A. & Diegelman-Parente, A. Rational design of a structure-switching DNA aptamer for potassium ions. *FEBS Open Bio* **4**, 788-795 (2014).
105. Hoang, M., Huang, P.-J.J. & Liu, J. G-Quadruplex DNA for Fluorescent and Colorimetric Detection of Thallium(I). *ACS Sensors* **1**, 137-143 (2016).
106. Ma, D.-L. et al. Utilization of G-Quadruplex-Forming Aptamers for the Construction of Luminescence Sensing Platforms. *ChemPlusChem* **82**, 8-17 (2017).
107. Slavkovic, S., Altunisik, M., Reinstein, O. & Johnson, P.E. Structure-affinity relationship of the cocaine-binding aptamer with quinine derivatives. *Bioorg. Med. Chem.* **23**, 2593-2597 (2015).
108. Reinstein, O. et al. Quinine binding by the cocaine-binding aptamer. Thermodynamic and hydrodynamic analysis of high-affinity binding of an off-target ligand. *Biochemistry* **52**, 8652-8662 (2013).
109. Neves, M.A.D., Slavkovic, S., Churcher, Z.R. & Johnson, P.E. Salt-mediated two-site ligand binding by the cocaine-binding aptamer. *Nucleic Acids Res.* **45**, 1041-1048 (2017).
110. Neves, M.A., Reinstein, O., Saad, M. & Johnson, P.E. Defining the secondary structural requirements of a cocaine-binding aptamer by a thermodynamic and mutation study. *Biophys. Chem.* **153**, 9-16 (2010).
111. Neves, M.A., Reinstein, O. & Johnson, P.E. Defining a stem length-dependent binding mechanism for the cocaine-binding aptamer. A combined NMR and calorimetry study. *Biochemistry* **49**, 8478-8487 (2010).
112. Harkness, R.W.,V, Slavkovic, S., Johnson, P.E. & Mittermaier, A.K. Rapid characterization of folding and binding interactions with thermolabile ligands by DSC. *Chem Commun (Camb)* **52**, 13471-13474 (2016).
113. Ellington, A.D. & Szostak, J.W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818 (1990).
114. Robertson, D.L. & Joyce, G.F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344**, 467 (1990).
115. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510 (1990).

116. Winkler, W.C., Cohen-Chalamish, S. & Breaker, R.R. An mRNA structure that controls gene expression by binding FMN. *Proceedings of the National Academy of Sciences* **99**, 15908-15913 (2002).
117. Winkler, W., Nahvi, A. & Breaker, R.R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952-956 (2002).
118. Mironov, A.S. et al. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747-756 (2002).
119. Nahvi, A. et al. Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9**, 1043 (2002).
120. Furtig, B., Nozinovic, S., Reining, A. & Schwalbe, H. Multiple conformational states of riboswitches fine-tune gene regulation. *Curr. Opin. Struct. Biol.* **30**, 112-124 (2015).
121. Serganov, A. et al. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.* **11**, 1729-1741 (2004).
122. Sudarsan, N. et al. Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science* **321**, 411-413 (2008).
123. Aldaye, F.A., Palmer, A.L. & Sleiman, H.F. Assembling Materials with DNA as the Guide. *Science* **321**, 1795-1799 (2008).
124. Hariri, A.A., Hamblin, G.D., Gidi, Y., Sleiman, H.F. & Cosa, G. Stepwise growth of surface-grafted DNA nanotubes visualized at the single-molecule level. *Nature Chemistry* **7**, 295 (2015).
125. Carneiro, K.M.M., Aldaye, F.A. & Sleiman, H.F. Long-Range Assembly of DNA into Nanofibers and Highly Ordered Networks Using a Block Copolymer Approach. *J. Am. Chem. Soc.* **132**, 679-685 (2010).
126. Avakyan, N., Conway, J.W. & Sleiman, H.F. Long-Range Ordering of Blunt-Ended DNA Tiles on Supported Lipid Bilayers. *J. Am. Chem. Soc.* **139**, 12027-12034 (2017).
127. Chidchob, P., Edwardson, T.G., Serpell, C.J. & Sleiman, H.F. Synergy of Two Assembly Languages in DNA Nanostructures: Self-Assembly of Sequence-Defined Polymers on DNA Cages. *J. Am. Chem. Soc.* **138**, 4416-4425 (2016).
128. Seeman, N.C. Nanomaterials Based on DNA. *Annu. Rev. Biochem.* **79**, 65-87 (2010).
129. Rothemund, P.W. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297-302 (2006).
130. Endo, M., Yang, Y. & Sugiyama, H. DNA origami technology for biomaterials applications. *Biomaterials Science* **1**, 347-360 (2013).
131. Yang, H. et al. Metal–nucleic acid cages. *Nature Chemistry* **1**, 390 (2009).
132. Aldaye, F.A. & Sleiman, H.F. Sequential Self-Assembly of a DNA Hexagon as a Template for the Organization of Gold Nanoparticles. *Angewandte Chemie International Edition* **45**, 2204-2209 (2006).
133. Lo, P.K. et al. Loading and selective release of cargo in DNA nanotubes with longitudinal variation. *Nature Chemistry* **2**, 319 (2010).

134. Hamblin, G.D. et al. Simple Design for DNA Nanotubes from a Minimal Set of Unmodified Strands: Rapid, Room-Temperature Assembly and Readily Tunable Structure. *ACS Nano* **7**, 3022-3028 (2013).
135. Avakyan, N. et al. Reprogramming the assembly of unmodified DNA with a small molecule. *Nat Chem* **8**, 368-376 (2016).
136. Mustoe, A.M., Brooks, C.L. & Al-Hashimi, H.M. Hierarchy of RNA functional dynamics. *Annu. Rev. Biochem.* **83**, 441-466 (2014).
137. Hubner, M.R. & Spector, D.L. Chromatin dynamics. *Annu Rev Biophys* **39**, 471-489 (2010).
138. Cruz, J.A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136**, 604-609 (2009).
139. Perez, A., Luque, F.J. & Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* **129**, 14739-14745 (2007).
140. Gueron, M. & Leroy, J.L. Studies of base pair kinetics by NMR measurement of proton exchange. *Methods Enzymol.* **261**, 383-413 (1995).
141. Nikolova, E.N., Bascom, G.D., Andricioaei, I. & Al-Hashimi, H.M. Probing sequence-specific DNA flexibility in a-tracts and pyrimidine-purine steps by nuclear magnetic resonance (13)C relaxation and molecular dynamics simulations. *Biochemistry* **51**, 8654-8664 (2012).
142. Hsu, S.T. et al. A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics. *J. Am. Chem. Soc.* **131**, 13399-13409 (2009).
143. Geggier, S. & Vologodskii, A. Sequence dependence of DNA bending rigidity. *Proc Natl Acad Sci U S A* **107**, 15421-15426 (2010).
144. Matsumoto, A. & Olson, W.K. Sequence-Dependent Motions of DNA: A Normal Mode Analysis at the Base-Pair Level. *Biophys. J.* **83**, 22-41 (2002).
145. Frauenfelder, H., Sligar, S.G. & Wolynes, P.G. The energy landscapes and motions of proteins. *Science* **254**, 1598-1603 (1991).
146. Annunziato, A. DNA packaging: nucleosomes and chromatin. *Nature Education* **1**, 26 (2008).
147. Alberts, B. Molecular biology of the cell, Edn. 4th. (Garland Science, New York; 2002).
148. Cooper, G.M. The cell: a molecular approach, Edn. 2nd. (ASM Press; Sinauer Associates, Washington, D.C.; Sunderland, Mass.; 2000).
149. Buning, R. & van Noort, J. Single-pair FRET experiments on nucleosome conformational dynamics. *Biochimie* **92**, 1729-1740 (2010).
150. Eslami-Mossallam, B., Schiessel, H. & van Noort, J. Nucleosome dynamics: Sequence matters. *Adv Colloid Interface Sci* **232**, 101-113 (2016).
151. Fleming, A.M., Zhou, J., Wallace, S.S. & Burrows, C.J. A Role for the Fifth G-Track in G-Quadruplex Forming Oncogene Promoter Sequences during Oxidative Stress: Do These "Spare Tires" Have an Evolved Function? *ACS Central Science* **1**, 226-233 (2015).

152. Guéron, M. & Leroy, J.-L. in Nucleic Acids and Molecular Biology. (eds. F. Eckstein & D.M.J. Lilley) 1-22 (Springer Berlin Heidelberg, Berlin, Heidelberg; 1992).
153. Snoussi, K. & Leroy, J.L. Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry* **40**, 8898-8904 (2001).
154. Bhattacharya, P.K., Cha, J. & Barton, J.K. 1H NMR determination of base-pair lifetimes in oligonucleotides containing single base mismatches. *Nucleic Acids Res.* **30**, 4740-4750 (2002).
155. Leijon, M. & Graslund, A. Effects of sequence and length on imino proton exchange and base pair opening kinetics in DNA oligonucleotide duplexes. *Nucleic Acids Res.* **20**, 5339-5343 (1992).
156. Friedman, R.A. & Honig, B. A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophys. J.* **69**, 1528-1535 (1995).
157. Seenisamy, J. et al. The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. *J. Am. Chem. Soc.* **126**, 8702-8709 (2004).
158. Trieb, M. et al. Dynamics of DNA: BI and BII Phosphate Backbone Transitions. *The Journal of Physical Chemistry B* **108**, 2470-2476 (2004).
159. Al-Hashimi, H.M. NMR studies of nucleic acid dynamics. *J. Magn. Reson.* **237**, 191-204 (2013).
160. Zhang, Q., Stelzer, A.C., Fisher, C.K. & Al-Hashimi, H.M. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature* **450**, 1263 (2007).
161. Nikolova, E.N. et al. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470**, 498-502 (2011).
162. Kimsey, I.J. et al. Dynamic basis for dG*dT misincorporation via tautomerization and ionization. *Nature* **554**, 195-201 (2018).
163. Kimsey, I.J., Petzold, K., Sathyamoorthy, B., Stein, Z.W. & Al-Hashimi, H.M. Visualizing transient Watson-Crick-like mispairs in DNA and RNA duplexes. *Nature* **519**, 315-320 (2015).
164. Dethoff, E.A., Petzold, K., Chugh, J., Casiano-Negroni, A. & Al-Hashimi, H.M. Visualizing transient low-populated structures of RNA. *Nature* **491**, 724-728 (2012).
165. Dethoff, E.A., Chugh, J., Mustoe, A.M. & Al-Hashimi, H.M. Functional complexity and regulation through RNA dynamics. *Nature* **482**, 322-330 (2012).
166. Bothe, J.R., Lowenhaupt, K. & Al-Hashimi, H.M. Sequence-specific B-DNA flexibility modulates Z-DNA formation. *J. Am. Chem. Soc.* **133**, 2016-2018 (2011).
167. Ho, P.S., Ellison, M.J., Quigley, G.J. & Rich, A. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.* **5**, 2737-2744 (1986).
168. Mathews, F.S. & Rich, A. The molecular structure of a hydrogen bonded complex of N-ethyl adenine and N-methyl uracil. *J. Mol. Biol.* **8**, 89-95 (1964).

169. Nikolova, E.N. et al. A historical account of Hoogsteen base-pairs in duplex DNA. *Biopolymers* **99**, 955-968 (2013).
170. Pous, J. et al. Stabilization by extra-helical thymines of a DNA duplex with Hoogsteen base pairs. *J. Am. Chem. Soc.* **130**, 6755-6760 (2008).
171. Abrescia, N.G., Thompson, A., Huynh-Dinh, T. & Subirana, J.A. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc Natl Acad Sci U S A* **99**, 2806-2811 (2002).
172. Abrescia, N.G., Gonzalez, C., Gouyette, C. & Subirana, J.A. X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA. *Biochemistry* **43**, 4092-4100 (2004).
173. Wang, A.H. et al. The molecular structure of a DNA-triostin A complex. *Science* **225**, 1115-1121 (1984).
174. Rice, P.A., Yang, S., Mizuuchi, K. & Nash, H.A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87**, 1295-1306 (1996).
175. Patikoglou, G.A. et al. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.* **13**, 3217-3230 (1999).
176. Alvey, H.S., Gottardo, F.L., Nikolova, E.N. & Al-Hashimi, H.M. Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. *Nat Commun* **5**, 4786 (2014).
177. Watson, J.D. & Crick, F.H. The structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**, 123-131 (1953).
178. Bebenek, K., Pedersen, L.C. & Kunkel, T.A. Replication infidelity via a mismatch with Watson-Crick geometry. *Proc Natl Acad Sci U S A* **108**, 1862-1867 (2011).
179. Wang, W., Hellinga, H.W. & Beese, L.S. Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proc Natl Acad Sci U S A* **108**, 17644-17648 (2011).
180. Ambrus, A. et al. Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res.* **34**, 2723-2735 (2006).
181. Gray, R.D., Buscaglia, R. & Chaires, J.B. Populated intermediates in the thermal unfolding of the human telomeric quadruplex. *J. Am. Chem. Soc.* **134**, 16834-16844 (2012).
182. Boncina, M., Lah, J., Prislan, I. & Vesnaver, G. Energetic basis of human telomeric DNA folding into G-quadruplex structures. *J. Am. Chem. Soc.* **134**, 9657-9663 (2012).
183. Dettler, J.M., Buscaglia, R., Le, V.H. & Lewis, E.A. DSC deconvolution of the structural complexity of c-MYC P1 promoter G-quadruplexes. *Biophys. J.* **100**, 1517-1525 (2011).
184. Freire, E. & Biltonen, R.L. Statistical mechanical deconvolution of thermal transitions in macromolecules. I. Theory and application to homogeneous systems. *Biopolymers* **17**, 463 (1978).
185. Spink, C.H. The deconvolution of differential scanning calorimetry unfolding transitions. *Methods* **76**, 78-86 (2015).

186. Freire, E. & Biltonen, R.L. Statistical mechanical deconvolution of thermal transitions in macromolecules. II. General treatment of cooperative phenomena. *Biopolymers* **17**, 481-496 (1978).
187. Freire, E. & Biltonen, R.L. Statistical mechanical deconvolution of thermal transitions in macromolecules. III. Application to double-stranded to single-stranded transitions of nucleic acids. *Biopolymers* **17**, 497-510 (1978).
188. Baldwin, A.J. & Kay, L.E. NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* **5**, 808-814 (2009).
189. Fawzi, N.L., Ying, J., Torchia, D.A. & Clore, G.M. Kinetics of Amyloid β Monomer-to-Oligomer Exchange by NMR Relaxation. *J. Am. Chem. Soc.* **132**, 9948-9951 (2010).
190. Kimsey, I.J., Petzold, K., Sathyamoorthy, B., Stein, Z.W. & Al-Hashimi, H.M. Visualizing transient Watson-Crick-like mispairs in DNA and RNA duplexes. *Nature* **519**, 315-320 (2015).
191. Latham, M.P., Sekhar, A. & Kay, L.E. Understanding the mechanism of proteasome 20S core particle gating. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 5532-5537 (2014).
192. Mergny, J.L., De Cian, A., Ghelab, A., Sacca, B. & Lacroix, L. Kinetics of tetramolecular quadruplexes. *Nucleic Acids Res.* **33**, 81-94 (2005).
193. Bessi, I., Jonker, H.R., Richter, C. & Schwalbe, H. Involvement of Long-Lived Intermediate States in the Complex Folding Pathway of the Human Telomeric G-Quadruplex. *Angew Chem Int Ed Engl* **54**, 8444-8448 (2015).
194. Koirala, D. et al. Long-loop G-quadruplexes are misfolded population minorities with fast transition kinetics in human telomeric sequences. *J. Am. Chem. Soc.* **135**, 2235-2241 (2013).
195. Gray, R.D., Trent, J.O. & Chaires, J.B. Folding and unfolding pathways of the human telomeric G-quadruplex. *J. Mol. Biol.* **426**, 1629-1650 (2014).
196. Stadlbauer, P., Krepl, M., Cheatham, T.E., 3rd, Koca, J. & Sponer, J. Structural dynamics of possible late-stage intermediates in folding of quadruplex DNA studied by molecular simulations. *Nucleic Acids Res.* **41**, 7128-7143 (2013).
197. Mashimo, T., Yagi, H., Sannohe, Y., Rajendran, A. & Sugiyama, H. Folding pathways of human telomeric type-1 and type-2 G-quadruplex structures. *J. Am. Chem. Soc.* **132**, 14910-14918 (2010).
198. Limongelli, V. et al. The G-Triplex DNA. *Angewandte Chemie International Edition* **52**, 2269-2273 (2013).
199. Phan, A.T. & Patel, D.J. Two-Repeat Human Telomeric d(TAGGGTTAGGGT) Sequence Forms Interconverting Parallel and Antiparallel G-Quadruplexes in Solution: Distinct Topologies, Thermodynamic Properties, and Folding/Unfolding Kinetics. *J. Am. Chem. Soc.* **125**, 15021-15027 (2003).

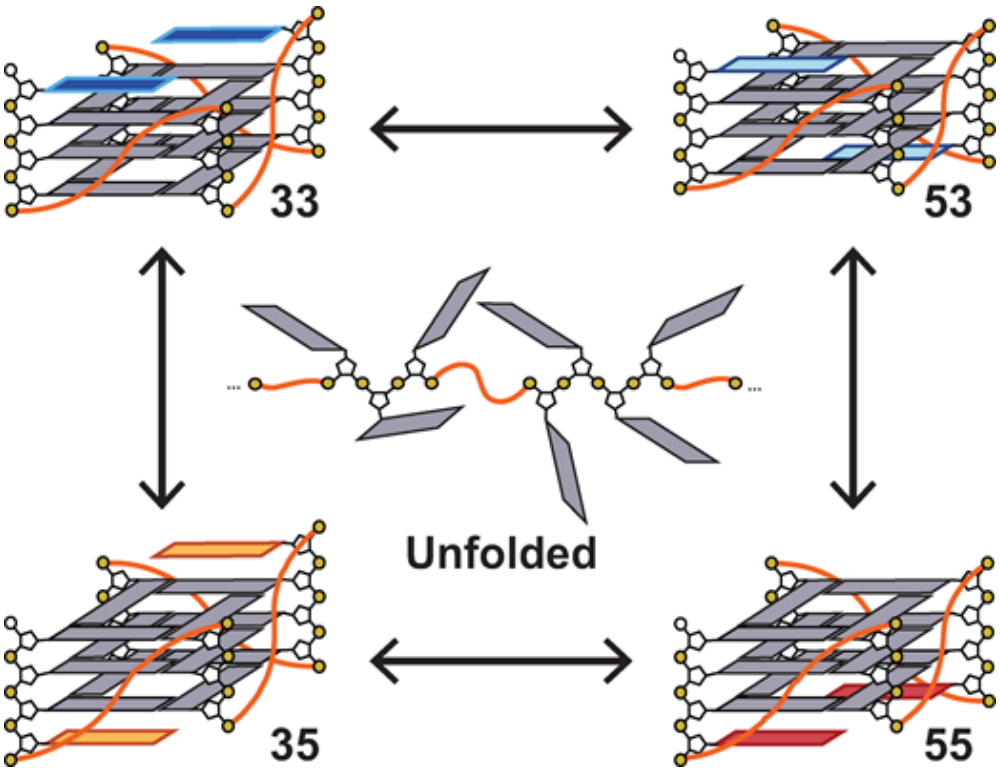
200. Ambrus, A. et al. Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic acids research* **34**, 2723-2735 (2006).
201. Wang, Z.-F., Li, M.-H., Chen, W.-W., Hsu, S.-T.D. & Chang, T.-C. A novel transition pathway of ligand-induced topological conversion from hybrid forms to parallel forms of human telomeric G-quadruplexes. *Nucleic Acids Res.* **44**, 3958-3968 (2016).
202. Buscaglia, R. et al. Polyethylene glycol binding alters human telomere G-quadruplex structure by conformational selection. *Nucleic Acids Res.* **41**, 7934-7946 (2013).
203. Dai, J., Carver, M. & Yang, D. Polymorphism of human telomeric quadruplex structures. *Biochimie* **90**, 1172-1183 (2008).
204. Adrian, M. et al. Structure and Conformational Dynamics of a Stacked Dimeric G-Quadruplex Formed by the Human CEB1 Minisatellite. *J. Am. Chem. Soc.* **136**, 6297-6305 (2014).
205. Sun, D. & Hurley, L.H. Biochemical Techniques for the Characterization of G-Quadruplex Structures: EMSA, DMS Footprinting, and DNA Polymerase Stop Assay. *Methods in molecular biology (Clifton, N.J.)* **608**, 65-79 (2010).
206. Do, N.Q. & Phan, A.T. Monomer–Dimer Equilibrium for the 5′–5′ Stacking of Propeller-Type Parallel-Stranded G-Quadruplexes: NMR Structural Study. *Chemistry – A European Journal* **18**, 14752-14759 (2012).
207. Palumbo, S.L. et al. A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity. *Nucleic Acids Res.* **36**, 1755-1769 (2008).
208. Matsugami, A. et al. An intramolecular quadruplex of (GGA)4 triplet repeat DNA with a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad, and its dimeric interaction 1. *J. Mol. Biol.* **313**, 255-269 (2001).
209. Matsugami, A., Okuizumi, T., Uesugi, S. & Katahira, M. Intramolecular Higher Order Packing of Parallel Quadruplexes Comprising a G:G:G:G Tetrad and a G(:A):G(:A):G(:A):G Heptad of GGA Triplet Repeat DNA. *J. Biol. Chem.* **278**, 28147-28153 (2003).
210. Alexandrov, B.S. et al. DNA breathing dynamics distinguish binding from nonbinding consensus sites for transcription factor YY1 in cells. *Nucleic Acids Res.* **40**, 10116-10123 (2012).
211. Choi, C.H. et al. DNA dynamically directs its own transcription initiation. *Nucleic Acids Res.* **32**, 1584-1590 (2004).
212. Alexandrov, B.S. et al. Toward a Detailed Description of the Thermally Induced Dynamics of the Core Promoter. *PLoS Comp. Biol.* **5**, e1000313 (2009).
213. Alexandrov, B.S. et al. DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation. *Nucleic Acids Res.* **38**, 1790-1795 (2010).

214. Bohnuud, T. et al. Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde. *Nucleic Acids Res.* **40**, 7644-7652 (2012).
215. Friedman, J.I. & Stivers, J.T. Detection of Damaged DNA Bases by DNA Glycosylase Enzymes. *Biochemistry* **49**, 4957-4967 (2010).
216. Yang, C.-G. et al. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature* **452**, 961-965 (2008).
217. Slupphaug, G. et al. A nucleotide-flipping mechanism from the structure of human uracil–DNA glycosylase bound to DNA. *Nature* **384**, 87 (1996).
218. Steinert, H. et al. Pausing guides RNA folding to populate transiently stable RNA structures for riboswitch-based transcription regulation. *Elife* **6** (2017).
219. Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nat. Struct. Mol. Biol.* **12**, 46-53 (2005).
220. Perdrizet, G.A., 2nd, Artsimovitch, I., Furman, R., Sosnick, T.R. & Pan, T. Transcriptional pausing coordinates folding of the aptamer domain and the expression platform of a riboswitch. *Proc Natl Acad Sci U S A* **109**, 3323-3328 (2012).
221. Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M. & Bustamante, C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* **325**, 626-628 (2009).
222. Cheah, M.T., Wachter, A., Sudarsan, N. & Breaker, R.R. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature* **447**, 497-500 (2007).
223. González, V., Guo, K., Hurley, L. & Sun, D. Identification and Characterization of Nucleolin as a c-myc G-quadruplex-binding Protein. *J. Biol. Chem.* **284**, 23622-23635 (2009).
224. Simone, R., Fratta, P., Neidle, S., Parkinson, G.N. & Isaacs, A.M. G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome. *FEBS Lett.* **589**, 1653-1668 (2015).
225. Moye, A.L. et al. Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat Commun* **6**, 7643 (2015).
226. Parkinson, G.N., Lee, M.P.H. & Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **417**, 876-880 (2002).
227. Ferguson, B.S. et al. Real-time, aptamer-based tracking of circulating therapeutic agents in living animals. *Sci Transl Med* **5**, 213ra165 (2013).
228. Xiao, Y., Lubin, A.A., Heeger, A.J. & Plaxco, K.W. Label-Free Electronic Detection of Thrombin in Blood Serum by Using an Aptamer-Based Sensor. *Angewandte Chemie International Edition* **44**, 5456-5459 (2005).
229. Ouldridge, T.E., Sulc, P., Romano, F., Doye, J.P. & Louis, A.A. DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic Acids Res.* **41**, 8886-8895 (2013).

230. Zhang, D.Y. & Seelig, G. Dynamic DNA nanotechnology using strand-displacement reactions. *Nature Chemistry* **3**, 103 (2011).
231. Zhang, D.Y. & Winfree, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.* **131**, 17303-17314 (2009).
232. Pinheiro, A.V., Han, D., Shih, W.M. & Yan, H. Challenges and opportunities for structural DNA nanotechnology. *Nat Nanotechnol* **6**, 763-772 (2011).
233. Wagenbauer, K.F., Wachauf, C.H. & Dietz, H. Quantifying quality in DNA self-assembly. *Nature Communications* **5**, 3691 (2014).
234. Privalov, P.L. & Potekhin, S.A. Scanning microcalorimetry in studying temperature-induced changes in proteins. *Methods Enzymol.* **131**, 4-51 (1986).
235. Mergny, J.L. & Lacroix, L. Analysis of thermal melting curves. *Oligonucleotides* **13**, 515-537 (2003).
236. Wallimann, P., Kennedy, R.J., Miller, J.S., Shalongo, W. & Kemp, D.S. Dual Wavelength Parametric Test of Two-State Models for Circular Dichroism Spectra of Helical Polypeptides: Anomalous Dichroic Properties of Alanine-Rich Peptides. *J. Am. Chem. Soc.* **125**, 1203-1220 (2003).
237. Puglisi, J.D. & Tinoco, I. in Methods Enzymol., Vol. 180 304-325 (Academic Press, 1989).
238. Biltonen, R.L., Freire, E. & Brandts, J.F. Thermodynamic characterization of conformational states of biological macromolecules using differential scanning calorimetry. *CRC critical reviews in biochemistry* **5**, 85-124 (1978).
239. Privalov, P.L. in Pure and Applied Chemistry, Vol. 52 479 (1980).
240. Privalov, P.L. Microcalorimetry of Macromolecules: The Physical Basis of Biological Structures. *Journal of Solution Chemistry* **44**, 1141-1161 (2015).
241. Pagano, B. et al. Differential scanning calorimetry to investigate G-quadruplexes structural stability. *Methods* **64**, 43-51 (2013).
242. Privalov, P.L. & Dragan, A.I. Microcalorimetry of biological macromolecules. *Biophys. Chem.* **126**, 16-24 (2007).
243. Wyman, J. in Advances in Protein Chemistry, Vol. 4. (eds. M.L. Anson & J.T. Edsall) 407-531 (Academic Press, 1948).
244. Freire, E., Schön, A. & Velazquez-Campoy, A. in Methods Enzymol., Vol. 455 127-155 (Academic Press, 2009).
245. Freiburger, L. et al. Substrate-dependent switching of the allosteric binding mechanism of a dimeric enzyme. *Nat. Chem. Biol.* **10**, 937-942 (2014).
246. Freiburger, L.A. et al. Competing allosteric mechanisms modulate substrate binding in a dimeric enzyme. *Nat. Struct. Mol. Biol.* **18**, 288-294 (2011).
247. Wiseman, T., Williston, S., Brandts, J.F. & Lin, L.-N. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Anal. Biochem.* **179**, 131-137 (1989).

248. Mergny, J.L. & Lacroix, L. Kinetics and thermodynamics of i-DNA formation: phosphodiester versus modified oligodeoxynucleotides. *Nucleic Acids Res.* **26**, 4797-4803 (1998).
249. Hatzakis, E., Okamoto, K. & Yang, D. Thermodynamic stability and folding kinetics of the major G-quadruplex and its loop isomers formed in the nuclease hypersensitive element in the human c-Myc promoter: effect of loops and flanking segments on the stability of parallel-stranded intramolecular G-quadruplexes. *Biochemistry* **49**, 9152-9160 (2010).
250. Sanchez-Ruiz, J.M. Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophys. J.* **61**, 921-935 (1992).
251. Drobnak, I., Vesnaver, G. & Lah, J. Model-based thermodynamic analysis of reversible unfolding processes. *J Phys Chem B* **114**, 8713-8722 (2010).
252. Freiburger, L.A., Auclair, K. & Mittermaier, A.K. Elucidating protein binding mechanisms by variable-c ITC. *ChemBioChem* **10**, 2871-2873 (2009).

Chapter 2: G-register exchange dynamics in guanine quadruplexes



2.1. Preface

Many GQ sequences contain different number of Gs in their G-tracts, leading to an ensemble of GQ isomers undergoing GR exchange. GR exchange dynamics can in theory influence the thermal stability of the GQ ensemble by reducing the entropy penalty for folding and may modulate binding interactions with partner proteins by changing the structural motifs displayed to the cellular environment. The complexity of these dynamics makes them difficult to address by conventional techniques for investigating conformational exchange, precluding an in-depth understanding of their effects on biological function. In this chapter, a method for analyzing GR exchange by global fitting of thermal denaturation profiles for GQ ensembles is presented and discussed. Using the approach, GR exchange dynamics are revealed to entropically stabilize GQs and occur cooperatively, potentially influencing gene expression. Evidence for the widespread occurrence of GR exchange dynamics as a regulatory mechanism for human genes is additionally presented.

2.2. Abstract

GQs are 4-stranded DNA structures formed by tracts of stacked, HG-hydrogen bonded Gs. GQs are found in gene promoters and telomeres where they regulate gene transcription and telomere elongation. Though GQ structures are well-characterized, many aspects of their conformational dynamics are poorly understood. For example, when there are surplus Gs in some of the tracts, they can slide with respect to one another, a process we term GR exchange. These motions could in principle entropically stabilize the folded state, crucially benefitting GQs as their stabilities are closely tied to biological function. We have developed a method for characterizing GR exchange where each isomer in the wild-type conformational ensemble is trapped by mutation

and thermal denaturation data for the set of trapped mutants and wild-type are analyzed simultaneously. This yields GR isomer populations as a function of temperature, quantifies conformational entropy, and sheds light on correlated sliding motions of the G-tracts. We measured entropic stabilizations from GR exchange up to 14.3±1.6 J mol^{-1} K^{-1}, with melting temperature increases up to 7.3±1.6 °C. Furthermore, bioinformatic analysis suggests a majority of putative human GQ sequences are capable of GR exchange, pointing to the generality of this phenomenon.

2.3. Introduction

DNA GQs are structures adopted by dG-rich nucleic acids. They are composed of four G-tracts of three or more consecutive dGs connected by loop sequences. The tracts come together to form G-tetrads (Figure 2.1a), planes of four HG-hydrogen bonded dGs that are stacked to form the core GQ structure[1]. GQs are structurally diverse and dynamic molecules; they may adopt and exchange amongst different topologies (Figure 2.1b,c)[2, 3] and oligomeric states[4]. DNA GQs are found in telomeres and gene promoters, where they regulate telomere elongation and gene transcription[5, 6]. In addition, several functional RNA GQs have been discovered, hinting at a general regulatory role for GQs in biology[7, 8]. Importantly, the thermodynamic stability of GQs has been found to correlate with their regulatory functions[9-11]. Thus, a thorough understanding of GQ function is predicated on characterizing the determinants of GQ stability and dynamics. Moreover, GQs are known to interact with proteins and these dynamics may play a role in molecular recognition.

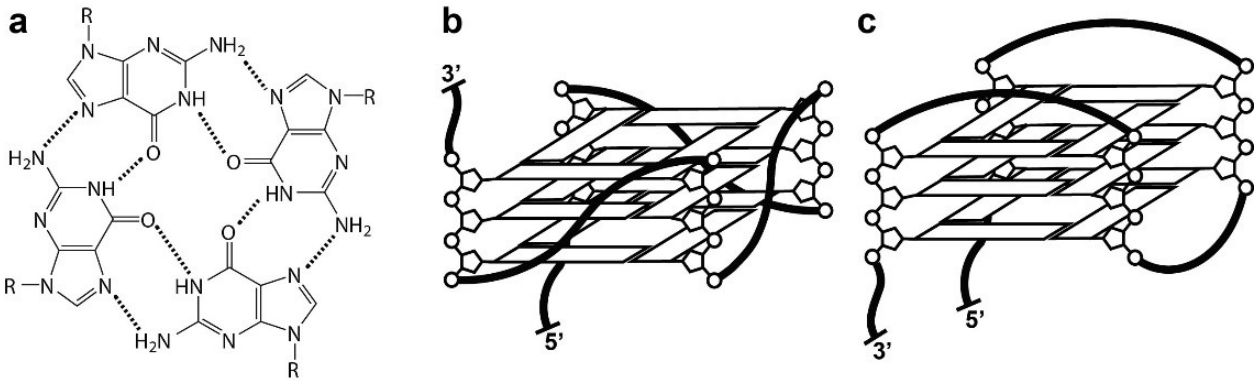


Figure 2.1. GQ structure. (a) The G-tetrad, a planar arrangement of four HG-hydrogen bonded G residues that forms the layers of a GQ. (b) Propeller-type GQ topology with parallel G-tracts. (c) Basket-type GQ with antiparallel G-tracts.

Previous studies have investigated how topology, G-tetrad number, and loop length affect GQ stability[12-14]. However, one distinctive feature of GQs has attracted little attention to date. When G-tracts contain different numbers of dGs, as is the case for several known promoter GQs[15-18], several folded isomers exist for the same sequence, where each isomer contains a different subset of dGs participating in the stacked G-tetrad core. This type of GQ can potentially undergo dynamics characterized by G-tracts sliding with respect to one another. These dynamics contribute to conformational entropy of the folded state and expose different recognition motifs to potential binding partners, thereby influencing GQ stability and possibly function. We refer to these strand-shifted GQs as GR isomers and the exchange amongst these isomers as GR exchange. For example, the c-myc Pu18 promoter GQ contains two dG_3 and two dG_4 tracts[19]. Thus, it can form a total of four different GQ structures, with each dG_4 tract contributing either the 5' or 3' dG to the GQ core. It has been shown using NMR and DMS foot printing experiments that all four possible GR isomers are formed at equilibrium and interconvert rapidly[6, 20-22]. If each GR isomer were equally populated, this would lead to a four-fold increase in the stability of the folded state compared to a

GQ with a single GR isomer, corresponding to an entropic stabilization of $\Delta\Delta S = R\ln(4)$ where R is the ideal gas constant. However, the actual populations of the different c-myc GR isomers are not known. Individual GR isomers are challenging to characterize thermodynamically because they are potentially large in number, they interconvert on a range of timescales, and they are difficult to distinguish from one another experimentally. Therefore, the entropic contribution of GR exchange to GQ stability has not yet been quantitatively determined.

Mutations are commonly used to quench GQ conformational exchange, facilitating structural characterization and functional studies[23, 24]. In this approach, surplus dG residues in G-tracts are substituted with bases that are not stably incorporated into G-tetrads, such as dT or dI, or are truncated from the sequence entirely if at the 5' or 3' end. This produces a unique core of 12 dG residues (in the case of four G_3 G-tracts) that cannot slide with respect to one another, although exchange between different topologies is still possible[25]. In principle, the thermal stabilities of the trapped mutants can be related directly to the thermal stabilities of the corresponding wild-type GR isomers. However, this assumes implicitly that the mutations cause minimal thermodynamic perturbation, beyond trapping the GQ in a single register. The validity of this assumption has not been quantitatively tested to date, which limits our ability to use this approach to unravel GR exchange dynamics.

We have developed an experimental method that yields the populations of all wild-type exchanging GR isomers as a function of temperature, while at the same time testing whether trapped GQ mutants are good thermodynamic mimics of the corresponding GR isomers. The method is based on a global analysis of thermal denaturation data, and the principle that a macromolecule populates all possible conformations in proportion to each conformation's thermodynamic stability[26]. We fit data for the wild-type and trapped mutants simultaneously, in

such a way that good agreement is obtained if and only if the trapped mutant GQs are thermodynamically equivalent to the corresponding wild-type GR isomers. We applied the method to study GQs found in the promoter regions of genes encoding the vascular endothelial growth factor A (VEGFA), the c-myc transcription factor (c-myc Pu18), and the Pim1 kinase (PIM1), which have 2, 4, and 12 GR isomers, respectively (Table 2.1 and Supplementary Figure 2.1). Using a combination of DSC, UV-Visible (UV-Vis), CD, NMR spectroscopy, and our new global fitting method we showed that: (i) The trapped mutant GQs are structurally similar to the corresponding wild-type GR isomers; (ii) The trapped mutant GQs fold in an effectively two-state manner; (iii) Wild-type and trapped mutant GQ thermal melt data agree with our global fitting method; (iv) Trapping mutations do not perturb stability beyond locking the GQ in a single state, therefore the fitted GR isomer populations are meaningful; (v) GR exchange can provide substantial stabilization to the GQ folded state; and (vi) A large majority of naturally-occurring GQs can potentially undergo GR exchange.

2.4. Results

2.4.1. Systematically trapping GR isomers with mutations

We studied three promoter GQ sequences, referred to here as VEGFA, c-myc Pu18, and PIM1, which contain 1, 2, and 4 surplus dG residues, respectively. We note that it could be possible for surplus dG residues to be accommodated in the form of a G-pentad or G-hexad, rather than leading to GR exchange. However, the pentad and hexad arrangements observed to date contain four dG residues and either one or two dA residues. Furthermore, these unusual structures exist only in the context of dimers[27, 28]. We selected monomeric GQs[29] in which it is likely that surplus dG residues lead to GR exchange rather than G-pentad or G-hexad formation. Therefore, the

VEGFA, c-myc Pu18, and PIM1, GQs populate 2, 4, and 12 GR isomers, respectively. The sequences were systematically trapped in structures mimicking each of their GR isomers by substituting all surplus dG residues with dT or dI (Table 2.1 and Supplementary Figure 2.1). In the following we refer to these sequence variants as trapped mutants, where we use the term mutant to indicate alterations of naturally occurring DNA sequences. We note that the difference between dI and dG is the presence of a hydrogen atom versus an amino group at the 2-position, respectively. Replacing dG with dI in a G-tetrad therefore comes at the cost of one hydrogen bond, estimated at several kcal mol^{-1}[30, 31]. This translates into a greater than 100:1 preference for dG relative to dI in the core of the GQ. In the case of dT, the preference for dG is even stronger. Therefore we consider both dI- and dT-containing mutants to be effectively trapped in a single GR isomer with dGs located in the core and dI or dT residues in the loops or flanking regions. For instance in VEGFA (5'-**$G_1G_2G_3$-**A_4-**$G_5G_6G_7$-**T_8T_9-**$G_{10}G_{11}G_{12}G_{13}$-**T_{14}-**$G_{15}G_{16}G_{17}$-3'), the mutation 10dG>dX locks the third G-tract in a 5' shifted position, while 13dG>dX locks it in a 3' shifted position, relative to the core of 3 G-tetrads (dX=dT or dI). For the c-myc Pu18 GQ (5'-A_1-**$G_2G_3G_4$-**T_5-**$G_6G_7G_8G_9$-**A_{10}-**$G_{11}G_{12}G_{13}$-**T_{14}-**$G_{15}G_{16}G_{17}G_{18}$-3'), the double mutations 6dG,15dG>dX, 6dG,18dG>dX, 9dG,15dG>dX, and 9dG,18dG>dX lock the second and fourth G-tracts in the 5'5', 5'3', 3'5', and 3'3' registers relative to the core, respectively. In what follows, we will refer to the wild-type GR isomers corresponding to these trapped mutants as 55, 53, 35, and 33. The PIM1 GQ 5'-**$G_1G_2G_3$-**C_4-**$G_5G_6G_7G_8$-**C_9-**$G_{10}G_{11}G_{12}G_{13}G_{14}$-**$C_{15}$-**$G_{16}G_{17}G_{18}G_{19}$-**3' has 12 possible GR isomers. To mutationally trap this sequence in a single register isomer, four substitutions per mutant sequence are required: one in the second G-tract, two in the third G-tract, and one in the fourth G-tract. For example, in the third G-tract, we employed the following pairs of mutations to force adoption of a single register with respect to a GQ core of three G-tetrads: 5'-…$X_{10}X_{11}$**$G_{12}G_{13}G_{14}$**…-3', 5'-

…X_{10}**$G_{11}G_{12}G_{13}$**X_{14}…-3’, and 5’-…**$G_{10}G_{11}G_{12}$**$X_{13}X_{14}$…-3’, where (…) denotes the rest of the PIM1 sequence in the 5’ and 3’ directions.

Table 2.1 Wild-type and trapped mutant GQ sequences.

Name	Sequence	T_m °C [a]
VEGFA-WT	**5′-GGGAGGGTTGGGGTGGG-3′**	**61.2±0.1**
VEGFA-1	5′-GGGAGGGTTIGGGTGGG-3′	58.1±0.1
VEGFA-2	5′-GGGAGGGTTGGGITGGG-3′	57.7±0.1
c-mycPu18 WT	**5′-AGGGTGGGGAGGGTGGGG-3′**	**79.9±0.2[b]**
c-myc Pu18 55	5′-AGGGTTGGGAGGGTTGGG-3′	66.2±0.1
	5′-AGGGTIGGGAGGGTIGGG-3′	62.00±0.04
c-myc Pu18 53	5′-AGGGTTGGGAGGGTGGGT-3′	69.4±0.1
	5′-AGGGTIGGGAGGGTGGGI-3′	65.60±0.04
c-myc Pu18 35	5′-AGGGTGGGTAGGGTTGGG-3′	73.6±0.1
	5′-AGGGTGGGIAGGGTIGGG-3′	74.40±0.04
c-myc Pu18 33	5′-AGGGTGGGTAGGGTGGGT-3′	77.3±0.1
	5′-AGGGTGGGIAGGGTGGGI-3′	77.20±0.03
PIM1-WT	**5′-GGGCGGGGCGGGGGCGGGG-3′**	**68.8±1.6[b]**
PIM1-1	5′-GGGCTGGGCTTGGGCTGGG-3′	54.5±0.1
	5′-GGGCIGGGCIIGGGCIGGG-3′	50.0±0.1
PIM1-2	5′-GGGCTGGGCTTGGGCGGGT-3′	56.7±0.2
	5′-GGGCIGGGCIIGGGCGGGI-3′	53.4±0.2
PIM1-3	5′-GGGCTGGGCTGGGTCTGGG-3′	52.5±0.1
	5′-GGGCIGGGCIGGGICIGGG-3′	49.1±0.1
PIM1-4	5′-GGGCTGGGCTGGGTCGGGT-3′	55.7±0.2
	5′-GGGCIGGGCIGGGICGGGI-3′	50.2±0.2
PIM1-5	5′-GGGCTGGGCGGGTTCTGGG-3′	55.9±0.2
	5′-GGGCIGGGCGGGIICIGGG-3′	49.4±0.4
PIM1-6	5′-GGGCTGGGCGGGTTCGGGT-3′	56.4±0.2
	5′-GGGCIGGGCGGGIICGGGI-3′	52.9±0.2
PIM1-7	5′-GGGCGGGTCTTGGGCTGGG-3′	52.4±0.1
	5′-GGGCGGGICIIGGGCIGGG-3′	53.6±0.1
PIM1-8	5′-GGGCGGGTCTTGGGCGGGT-3′	61.2±0.1
	5′-GGGCGGGICIIGGGCGGGI-3′	61.8±0.1
PIM1-9	5′-GGGCGGGTCTGGGTCTGGG-3′	51.4±0.2
	5′-GGGCGGGICIGGGICIGGG-3′	51.1±0.1
PIM1-10	5′-GGGCGGGTCTGGGTCGGGT-3′	54.7±0.2
	5′-GGGCGGGICIGGGICGGGI-3′	55.3±0.1
PIM1-11	5′-GGGCGGGTCGGGTTCTGGG-3′	51.8±0.2
	5′-GGGCGGGICGGGIICIGGG-3′	48.4±0.3
PIM1-12	5′-GGGCGGGTCGGGTTCGGGT-3′	55.8±0.5
	5′-GGGCGGGICGGGIICGGGI-3′	53.3±0.2

[a] derived from UV-Vis analysis.

[b] Average T_m from global fitting of dT and dI trapped mutant datasets.

2.4.2. Structural analyses

We used a combination of CD and ^{1}H NMR spectroscopy to structurally characterize the wild-type GQs and their trapped mutants. In the case of both c-myc Pu18 and VEGFA, the CD spectra of all trapped mutants closely overlay those of the wild-type GQs (Figure 2.2a,b and Supplementary Figure 2.2a-c). The sharp maxima at 265 and minima at 240 nm imply that the wild type GQs and all mutants adopt parallel topologies[32], as expected. Furthermore, the ^{1}H NMR spectra of the wild-type GQs closely correspond to a superposition of the trapped mutant spectra (Supplementary Figure 2.3a,b, Supplementary Figure 2.4a). This strongly supports the idea that the wild-type GQs populate a mixture of conformational isomers, each of which resembles one of the trapped mutants, with exchange occurring relatively slowly on the NMR chemical shift timescale (≥seconds). In other words, the trapped mutants are good structural mimics of the wild-type GR isomers, at this level of resolution.

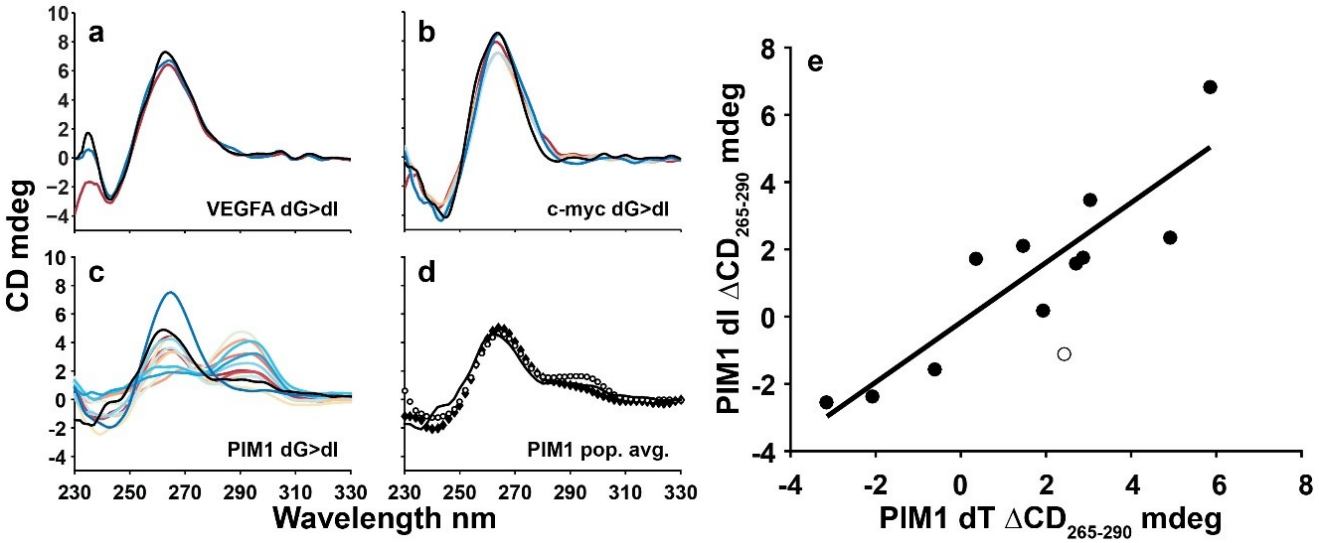


Figure 2.2. GQ CD spectra. CD spectra for wild-type (black) and trapped mutant (colored) VEGFA dG>dI (a), c-myc Pu18 dG>dI (b), and PIM1 dG>dI (c) GQs. In (c) and (d), the solid line corresponds to the corrected wild-type PIM1 CD spectrum. In (d) the population-weighted average spectrum of the dG>dT and dG>dI trapped mutants are shown with filled and empty symbols, respectively. (e) Comparison of $\Delta CD_{265–290}$ (CD(265 nm)–CD(290 nm)) signals for dT and dI trapped mutants of the PIM1 GQ. The empty circle indicates data for the outlying PIM1-10 trapped mutant (Supplementary Figure 2.6). The line indicates the best linear fit to the 11 filled data points (R=0.90).

Although the CD spectrum of the PIM1 wild-type GQ (Supplementary Figure 2.5) is consistent with a predominantly parallel topology, it exhibits a small shoulder at 290 nm (Figure 2.2c,d and Supplementary Figure 2.2d-f), indicating that some members of the conformational ensemble contain antiparallel strands[3]. Interestingly, while the CD spectra of many mutants closely match that of the wild-type GQ, several others are quite different, with maxima near 290 nm and shoulders at 265, indicating that these trapped mutants adopt predominantly antiparallel topologies. The corresponding dG>dT and dG>dI spectra closely superimpose, with only one exception (Supplementary Figure 2.6). In order to quantify this agreement, we have compared the

difference in ellipticities at 265 and 290 nm ($\Delta CD_{265–290}$ = CD(265 nm) – CD(290 nm)) for dG>dT and dG>dI mutants. Positive, zero, and negative $\Delta CD_{265–290}$ values are consistent with predominantly parallel, mixed, and predominantly antiparallel topologies respectively. The $\Delta CD_{265–290}$ values obtained for the dG>dT and dG>dI trapped mutants agree closely, yielding a correlation coefficient of R = 0.90 (Figure 2.2e). This implies that the same topology is obtained regardless of whether dT or dI residues are present at the mutated loop positions. It seems likely that the wild-type GR isomer, which contains dG at these positions, would prefer the same topology favoured by the corresponding dT and dI trapped mutants as well. This idea is supported by the excellent agreement between the CD spectrum of the wild-type GQ and the averaged spectra of the dG>dT and dG>dI trapped mutants (Figure 2.2d). In other words, GR exchange is linked to topological dynamics with some GR isomers favouring parallel topologies and other GR isomers favouring antiparallel topologies.

^{1}H NMR spectra of the wild-type PIM1 GQ and many of the trapped mutants are highly broadened, indicative of conformational dynamics on the microsecond to millisecond timescales (Supplementary Figure 2.3c, Supplementary Figure 2.4b)[33]. Similarly broadened ^{1}H NMR spectra have been previously observed for GQs that interconvert between different GR isomers[17] and/or topologies[34]. In the case of the PIM1 trapped mutants, the broadening is likely due to interconversion between parallel and various antiparallel topologies, as discussed above. For the wild-type PIM1 GQ, the extensive spectral broadening is likely due to dynamics among different topologies within a given GR isomer (as observed for the trapped mutants), as well as interconversion between different GR isomers. The population-weighted average NMR spectrum of the trapped mutants is in good agreement with that of the wild-type, consistent with the notion

that the trapped mutants are good structural mimics of the corresponding wild-type GR isomers (using globally-fit populations).

2.4.3. Trapped mutant folding

The thermodynamic stabilities of the wild-type GQs and trapped mutants were characterized by DSC and UV-Vis thermal melts, shown in Figure 2.3 and Figure 2.4. In what follows, the data for the trapped mutants were analyzed assuming a two-state model comprising only fully folded and fully unfolded states. While two-state models have previously been applied to GQ folding[29, 35, 36], some GQs are known to populate folding intermediates[36-39], therefore the two-state assumption warrants closer examination. The DSC and UV-Vis data show simple monophasic melting transitions, which provides some indication that folding is two-state, since GQs that populate folding intermediates such as G-triplexes often yield DSC thermograms and UV unfolding traces with distinctive shoulders or multiple distinct transitions[37, 40-42]. Nevertheless, more complicated folding processes can be present despite the absence of multiple or broad transitions[38, 43-45], and further tests are required. A classic approach for establishing two-state folding is to separately analyze the shape of a DSC thermogram, which yields the van 't Hoff enthalpy (ΔH^{VH}) and entropy (ΔS^{VH}), and the overall magnitude of the thermogram i.e. the areas under the C_p and C_p/T traces, which yield the calorimetric enthalpy (ΔH^{cal}) and entropy (ΔS^{cal})[46-48], respectively. When folding is two-state, $\Delta H^{VH} = \Delta H^{cal}$ and $\Delta S^{VH} = \Delta S^{cal}$. We performed this test on c-myc Pu18 GQ trapped mutant DSC data and obtained excellent agreement (Supplementary Table 2.2). We further applied the DSC deconvolution method of Freire and Biltonen[36, 49] and found again that these thermograms are consistent with two-state unfolding (Supplementary Figure 2.7). Additionally, we compared c-myc Pu18 UV-Vis thermal melt data collected at 260 and 295

nm. When folding is two-state these sets of data are expected to coincide, while they are likely to diverge when intermediates are present[38, 44]. The 260 and 295 nm melting profiles are nearly superimposable (Supplementary Figure 2.8) and yield identical folding parameters (Supplementary Figure 2.9a,b). Together this gives us confidence that the c-myc Pu18 GQ trapped mutants fold in a two-state manner.

The situation is more complicated for the PIM1 trapped mutants, as the CD and NMR data show that some of them populate multiple folded topologies, and thus there are formally more than two thermodynamic states accessible to these molecules. Nevertheless, if the stabilities of the different topologies are similar, then folding can still be considered a pseudo-two-state process in which the population of the unfolded state increases and the populations of all folded sub-states decrease in concert as the temperature is raised. In order to test whether this approximation holds, we compared UV-Vis data of PIM1 trapped mutants obtained at 260 and 295 nm (Supplementary Figure 2.10). Data at the two wavelengths exhibit quite different baseline slopes, perhaps due in part to topology exchange[25, 50]. Nevertheless, the melting transition regions coincide closely, and stability analyses performed on either the 260 or 295 nm data yield virtually the same results (Supplementary Figure 2.9c,d). We have therefore treated folding of the PIM1 trapped mutants as pseudo-two-state. The melt is well-defined at all temperatures by effective two-state folding parameters, in which the "fraction folded", plotted in Figure 2.4e refers to the fraction of chains adopting any folded topology.

2.4.4. Globally fitting GQ thermal denaturation data

We developed a global fitting method for characterizing GR exchange that can be applied to any set of thermal denaturation data (in this study DSC and UV-Vis). In a standard

thermodynamic analysis, unfolding data for a single sample are fitted independently of all other samples to yield a single set of thermodynamic parameters. In our method, data for a complete set of trapped mutants and the wild-type GQ are analyzed simultaneously to yield a combined set of thermodynamic parameters describing the stabilities of all mutants, as well as the populations of all wild-type GR isomers (see Sections 2.7.6 and 2.7.7). Our method is based on the assumption that the free energy difference between the folded and unfolded states of a trapped mutant is equal to the free energy difference between the corresponding folded GR isomer and the unfolded state of the wild-type GQ. To illustrate, data for a wild-type GQ with two GR isomers (such as VEGFA) would be analyzed together with those of the two corresponding trapped mutants. If at a given temperature, the folded:unfolded ratio for mutant A is 1:1, and that of mutant B is 2:1, then the assumption is that for the wild-type GQ, the (isomer A):(isomer B):(unfolded) ratio is 1:2:1. This relationship between the stabilities of trapped mutants and those of the corresponding wild-type GR isomers is applied across the full temperature range. Violations of this assumption lead to poor agreement across the dataset in the global analysis (see below).

Notably, we observe excellent agreement between the experimental data and global fits suggesting that the trapped mutants are good thermodynamic mimics of the corresponding wild-type GR isomers. This implies that our description of wild-type GQ conformational sampling is accurate. For example, Figure 2.3a,c shows complete sets of DSC data for the c-myc Pu18 GQ, comprising thermograms for the wild-type and four trapped mutants employing dG>dT (Figure 2.3a) and dG>dI (Figure 2.3c) substitutions. Interestingly, the wild-type GQ is more thermally stable than any of the trapped mutants, with thermal upshifts of the melting point of as much as 18 °C relative to the least stable of the trapped mutants. This increase in stability is caused by entropic stabilization of the wild-type GQ folded state due to the presence of multiple GR isomers[19, 20]. The

entropically driven thermal up-shift indicates that the wild-type populates multiple GR isomers at equilibrium and that the interconversion rate is relatively rapid (≤minutes). If the wild-type populated just one GR isomer, then we would expect its DSC thermogram to overlay that of the corresponding trapped mutant (i.e. no shift in melting temperature). If the wild-type exchanged very slowly (≥minutes) among the four GR isomers, the wild-type DSC trace would be the population-weighted average of the four traces of the trapped mutants, which would appear as a slight thermal down-shift from the most stable mutant (Supplementary Figure 2.11). Figure 2.3b,d shows the extracted populations of the four wild-type GR isomers as a function of temperature. Notably, very similar values are obtained with either dG>dT or dG>dI mutations, suggesting that this method is relatively insensitive to the precise chemistry of the substitute residues. The results show that the wild-type conformational ensemble is highly skewed with dominant populations of the 33 and 53 GR isomers. Interestingly, the proportions of 53, 35, and 55 GR isomers increase with increasing temperature, however they never represent more than ~30-40, 10, and 5% respectively.

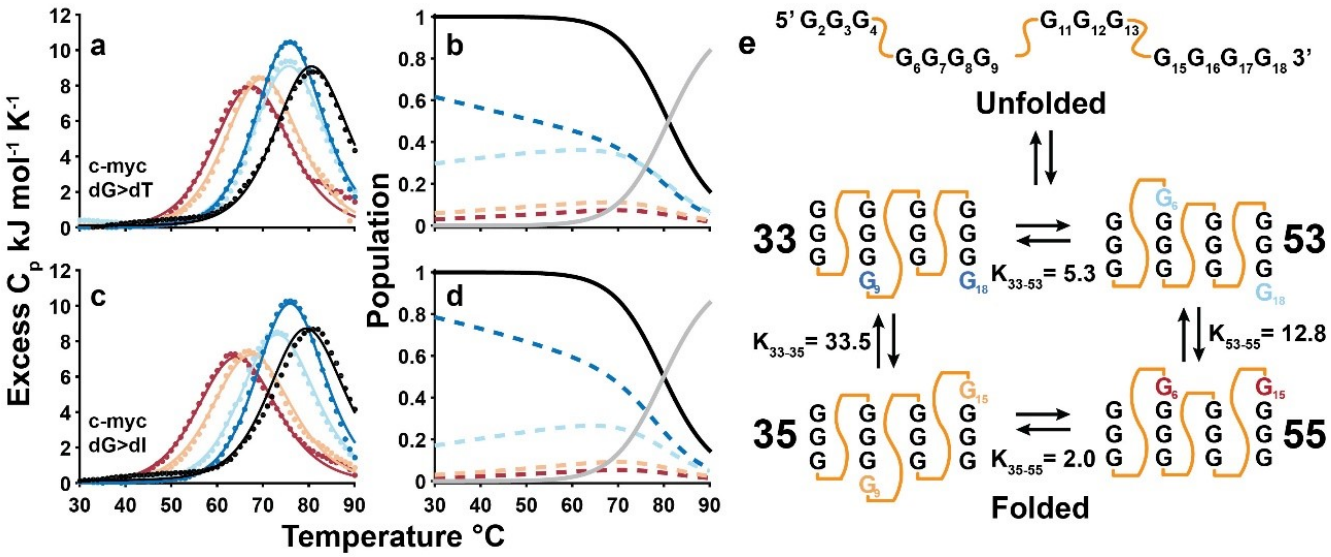


Figure 2.3. Global fits of GQ DSC thermal denaturation data. Thermograms (points) and global best-fit lines (lines) obtained for the complete set of c-myc Pu18 wild-type, dT-containing (a), dI-containing (c) trapped mutant GQs. Black and coloured lines correspond to data for the wild-type and trapped mutant GQs, respectively. The GR isomer populations (i.e. the fraction of wild-type DNA chains adopting each conformation) were extracted from the global fits (Equation 2.11) and are plotted in the panels immediately to the right, for dT-containing (b) and dI-containing (d) trapped mutants. Coloured dashed lines indicate the populations of each folded GR isomer, while black and grey lines correspond to the total population of the folded state and the population of the unfolded state, respectively. (e) Cartoon of the c-myc Pu18 GQ undergoing exchange between 4 folded GR isomers and the unfolded state. $K_{X\text{-}Y}=P_X/P_Y$ is the equilibrium constant for isomers X and Y, while the values indicated are for the dG>dI trapped mutants. Equilibrium constants have been expressed as >1 for ease of comparison.

UV-Vis absorbance spectroscopy is a more high-throughput method for thermal analysis than DSC, requires approximately 100-fold lower sample concentrations, and is equally amenable to global analyses of GR exchange. Due to these advantages, it was possible to apply UV-Vis spectroscopy to a range of different GQs and trapped mutants (see Section 2.7.7, Supplementary Figure 2.12 and Supplementary Figure 2.13 for sets of raw and globally analyzed melting profiles

of dG>dT and dG>dI trapped mutants respectively). Figure 2.4a shows absorbance melting profiles of the wild-type VEGFA GQ and a complete set of two dG>dI trapped mutants. Figure 2.4c shows melting profiles for the wild-type c-myc Pu18 GQ and complete sets of four trapped mutants employing dG>dI substitutions. Figure 2.4e shows melting profiles for the wild-type PIM1 GQ and complete set of twelve trapped mutants employing dG>dI substitutions. The extracted populations of the GR isomers comprising the wild-type ensembles are plotted in the panels to the right (Figure 2.4b,d,f). Importantly the UV-Vis results for the c-myc Pu18 GQ (Figure 2.4d) closely match those obtained by DSC (Figure 2.3b,d), indicating that the global fitting approach can be robustly applied to different experimental modalities. In addition, this confirms that the c-myc Pu18 GQ is monomeric, as the same stabilities were obtained at very different concentrations, 10 versus 150 μM for UV-Vis and DSC, respectively. In all cases, the conformational ensemble is skewed. In the case of VEGFA, the two GR isomers are populated with a 3:2 ratio. For c-myc Pu18, one of the four GR isomers represents more than 60% of the ensemble, while for PIM1, just one of the twelve GR isomers is populated to 50%. This has implications for the entropy contribution of GR exchange, as discussed below.

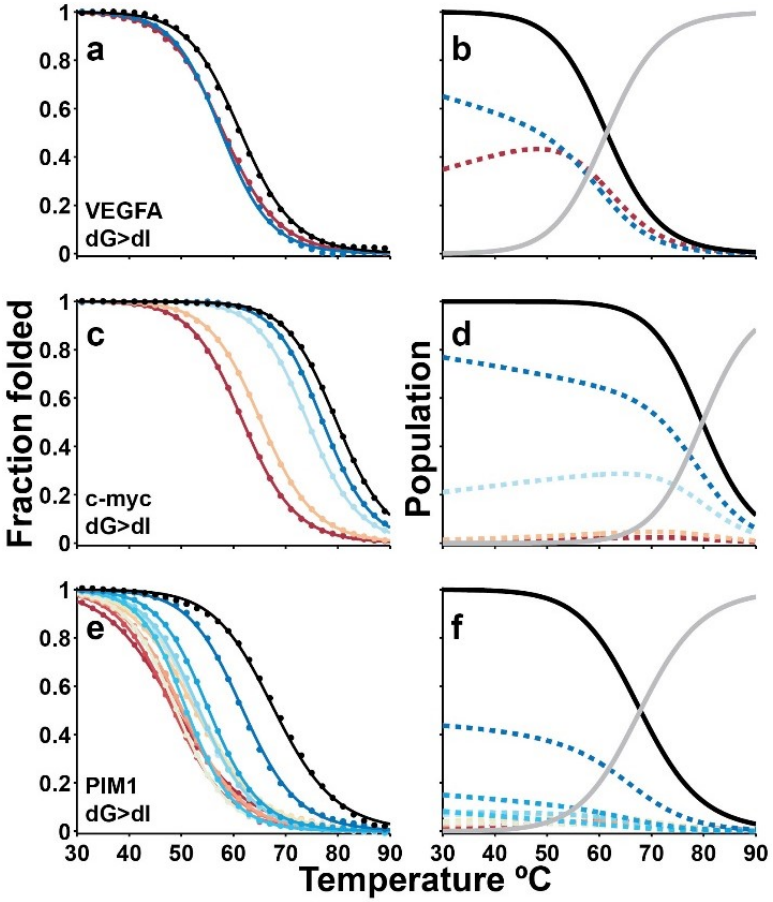


Figure 2.4. Global fits of GQ UV-Visible thermal denaturation data. Thermograms (points) and global best-fit (lines) for complete sets of wild-type and dI-containing trapped mutants of VEGFA (a), c-myc Pu18 (c), and PIM1 (e) GQs. Black and coloured lines correspond to data for the wild-type and trapped mutant GQs, respectively. The GR isomer populations (i.e. the fraction of wild-type DNA chains adopting each conformation) were extracted from the global fits (Equation 2.11) and are plotted in the panels immediately to the right for VEGFA (b), c-myc Pu18 (d), and PIM1 (f) GQs. Coloured dashed lines indicate the populations of each folded GR isomer, while black and grey lines correspond to the total population of the folded state and the population of the unfolded state, respectively.

2.4.5. Trapped mutants as thermodynamic mimics of GR isomers

The key assumption of the global analysis is that thermal stability of each trapped mutant relative to its unfolded state is identical to that of the corresponding wild-type GR isomer relative

to the wild-type unfolded state. However, the trapped mutants and the GR isomers differ; the wild-type contains one or more dG residues in loop positions whereas trapped mutants contain dT or dI substitutions. Therefore, it is of paramount importance to determine if the trapped mutants are good thermodynamic mimics of the wild-type GR isomers. Fortunately, the global fitting approach is well-suited to evaluating whether or not this is the case. If the mutations lead to thermodynamic perturbations, then the complete set of thermal denaturation data for the wild-type and trapped mutants will be mutually inconsistent. This will be reflected in poor agreement between experimental and calculated heat capacities or spectroscopic absorbances. We used Monte Carlo computer simulations to evaluate the ability of the global fitting approach to detect such inconsistencies (Section 2.7.8). Using the thermodynamic parameters of the c-myc Pu18 GQ as a starting point, we generated synthetic sets of thermal denaturation data in which the thermodynamic folding parameters of the trapped mutants differed from those of the corresponding GR isomers, i.e. the synthetic datasets were in violation of the key assumption stated above. We then subjected the synthetic data sets to global analysis. The simulations indicate the differences between the folding parameters of the trapped mutants and the corresponding wild-type GR isomers are no more than a maximum of ±1.5 kJ mol^{-1} for c-myc Pu18 (Figure 2.5a,b and Supplementary Figure 2.14). Using the same analysis, we find the maximum difference to be approximately ±2.5 kJ mol^{-1} for PIM1 (Figure 2.5c and Supplementary Figure 2.15).

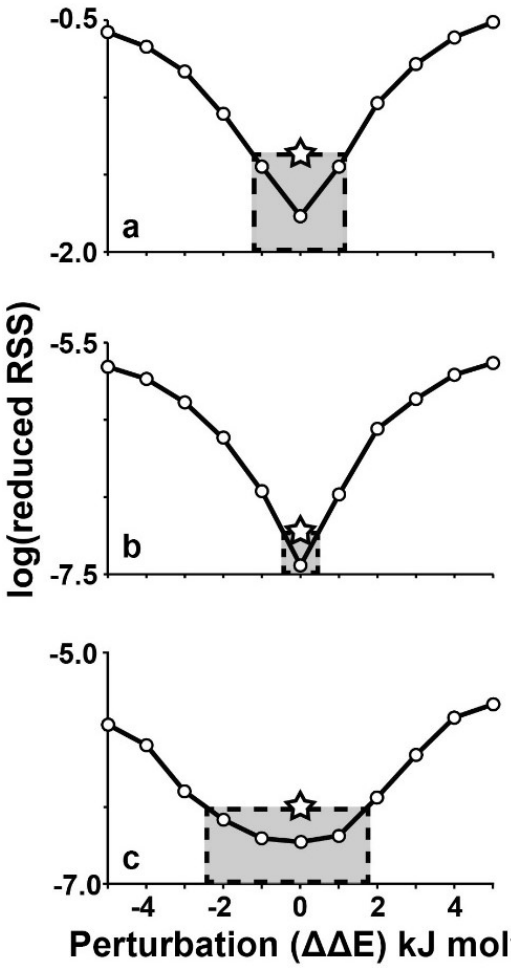


Figure 2.5. Sensitivity of the global fit to thermodynamic perturbations. Simulations are shown for c-myc Pu18 GQ (a) DSC and (b) UV-Vis data, and (c) PIM1 UV-Vis data using dG>dI mutations (see Supplementary Figure 2.14 and Supplementary Figure 2.15 for dG>dT simulations). Increasing thermodynamic mismatch between trapped mutants and their corresponding GR isomers ($\Delta\Delta E$) leads to larger values of "reduced RSS" in global fits i.e. poorer overall agreement. The reduced RSS was calculated as RSS_0/DF where DF = # points - # parameters. The stars correspond to the "reduced RSS" values obtained for the true experimental datasets. These define the maximum average mismatch between the folding thermodynamics of trapped mutants and corresponding wild-type GR isomers (gray shaded boxes). For thermodynamic perturbations larger than 1.1-1.5 kJ mol^{-1} (DSC) or 0.2-0.5 kJ mol^{-1} (UV-Vis), the RSS is greater than what we observe experimentally. See Section 2.7.8 for details of the simulations.

The thermodynamic parameters extracted from dG>dT and dG>dI trapped mutants are quite similar (Supplementary Table 2.3 and Supplementary Table 2.4), indicating that the stabilities of the GQs are not very sensitive to the identities of residues in the targeted loop positions, whether these be dT, dI, or presumably dG. In the case of c-myc Pu18, the extracted folding enthalpies, populations of the GR isomers, and the melting temperatures based on dG>dT and dG>dI mutations correlate very well with correlation coefficients (R) varying between 0.80 and 0.99 with a mean of 0.94±0.09, for both DSC and UV-Vis measurements (Supplementary Figure 2.16a-f). For PIM1, the T_ms of the individual trapped mutants and the populations of the GR isomers also agree well. The ΔH values of GR isomer folding are less well reproduced by dG>dT and dG>dI substitutions (Supplementary Figure 2.16g-i). This is likely because the PIM1 trapped mutants harbor more substitutions, giving larger thermodynamic perturbations (Figure 2.5c). Nevertheless, the fact that we extracted similar GR isomer populations with dG>dT or dG>dI substitutions for both c-myc Pu18 and PIM1 gives us confidence that we can use these values to better understand GQ function.

2.4.6. Entropy effects and correlated motions in GR exchange

One strength of our approach is that we are able to directly compute a portion of the entropic stabilization of the folded state due to internal dynamics. The entropy of a folded GQ exchanging among N GR isomers is given by

$$S_F = \sum_{i=1}^{N} p_i S_{F,i} - R \sum_{i=1}^{N} p_i \ln p_i \qquad (2.1)$$

where p_i is the fraction of folded chains adopting the i^{th} GR isomer, and $S_{F,i}$ is the molar entropy of i^{th} folded isomer. The first term in Equation 2.1 is simply the population-weighted average

entropy over all the isomers and the second is the entropy due to interconversion between multiple states. This can be calculated directly, as the set of p_i is obtained in our global analysis. It is typically difficult to separate solvent and conformational contributions in experimental measurements of entropy changes, for biomolecules in aqueous solution[51]. In contrast, our global fitting approach gives us direct access to a well-defined component of conformational entropy, i.e. that due to GR exchange. The resulting contributions to the entropy of the folded state, increases in the folding equilibrium constant, and thermal upshifts in T_m (ΔG_F=0) are listed in Table 2.2 for the three GQs studied here. Interestingly, due to the skewed populations of the GR isomers, the actual entropic stabilization is less than the maximum that would be obtained with equal populations. Not surprisingly, the largest stabilization is seen for wild-type PIM1, which has 12 GR isomers and whose melting temperature is increased by 7.3±1.6 °C compared to that of the most stable GR isomer. Even the VEGFA GQ with only 2 GR isomers exhibits a 1.8±0.1-fold increase in the population of the folded state and a 3.1±0.3 °C upshift in melting temperature.

Table 2.2 Entropic stabilization of folded GQs by GR exchange**.** Parameters given here are the average of the results from dT and dI trapped mutant datasets where applicable (DSC and UV-Vis for c-myc Pu18, UV-Vis for PIM1).

Parameters	VEGFA	c-myc Pu18	PIM1
Number of GR isomers (*N*)	2	4	12
$\Delta\Delta S_{theo}=R\ln(N)$ J mol^{-1} K^{-1} [a]	5.8	11.5	20.7
$\Delta\Delta S_{exp}$ J mol^{-1} K^{-1} [b]	5.20±0.01	6.2±0.9	14.3±1.6
$K_{F,exch}/K_{F,indiv}$ [c]	1.8±0.1	2.1±0.5	4.5±1.9
ΔT_m °C [d]	3.1±0.3	3.4±1.2	7.3±1.6

[a]Maximum conformational entropy contribution for exchange among *N* isomers.

[b]Actual conformation entropy contribution calculated using trapped mutant populations at 25 °C.

[c]Ratios of wild-type and most-populated trapped mutant folding equilibrium constants evaluated at the wild-type T_m.

[d]Wild-type thermal upshifts relative to the most-populated GR isomer. Errors were calculated as the standard deviations of the values extracted from global fits of DSC and UV-Vis data according to variance/co-variance method (see Section 2.7.12).

In addition, our results give a direct measure of concerted dynamics via a comparison of the populations of the different GR isomers. We find that the position of one G-tract influences the sliding motions of the adjacent G-tracts. For instance, in c-myc Pu18, the 4th G-tract populates the 3'-shifted register over the 5'-shifted register with a ~10:1 ratio when the 2nd G-tract is in the 5'-shifted register (P_{53}/P_{55}=12.8, Figure 2.3e, Supplementary Table 2.5). When the 2nd G-tract is 3'-shifted, this ratio becomes ~30:1 (P_{33}/P_{35}=33.5). Cooperative motions are more pronounced in PIM1 (~50-fold differences in shifting, Supplementary Figure 2.17, Supplementary Table 2.6).

2.4.7. GR exchange in the human genome

In order to better understand the relevance of GR exchange to biological GQs, we conducted a bioinformatic analysis of promoter regions of the human genome[52], as GQ formation in these regions has been linked to biological function[53-55]. We found 28,377 potential GQ-forming sequences following the pattern 5'-$G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}$-3' (N = dA, dC, dT, or dG), in both strands of the 23,322 human promoter regions from the EPD (see Section 2.7.9). The number of retrieved sequences implies that, on average, each 600 base pair promoter region contained one or two putative GQ-forming sequences. The number of possible GR isomers was calculated for each putative GQ sequence (Section 2.7.10), and the resulting histogram is shown in Figure 2.6. Notably, 89% of GQ sequences following this pattern can potentially adopt 2 or more folded GR isomers, with 19% potentially exchanging among 12 or more. A sizeable fraction of potential GQ sequences have 20 or more different GR isomers which could stabilize the folded state by as much as ~17 °C (see Section 2.7.11). Within the selected sequences, 52, 35 and 13% of G-tracts contain 3, 4, and 5 residues, respectively. Longer G-tracts were not included in the analysis, as these can lead to more complicated dynamics that are beyond the scope of this work. Interestingly the distribution of the number of GR isomers closely matches that obtained if the lengths of adjacent G-tracts are uncorrelated. For example, the probability that a given G-tract contains 4 dG residues is independent of the lengths of the other three G-tracts in the sequence. It should be noted that since G-tracts longer than 5 dG residues were excluded, certain numbers of GR isomers do not appear in the analysis, explaining the gaps at 5, 7, 10, etc. in Figure 2.6. Furthermore, it has been recently shown that small-molecule G derivatives can "fill in" G-tetrads in GQs with G-vacancies, i.e. those lacking a full complement of dG residues[56]. This raises the interesting possibility that even GQs with the canonical four G_3 tracts might be able to undergo GR exchange, with G

derivatives filling in the vacancies left by the shifted strands. Similar arguments apply to GQs with G-tracts of unequal length, further increasing the availability of potential GR isomers.

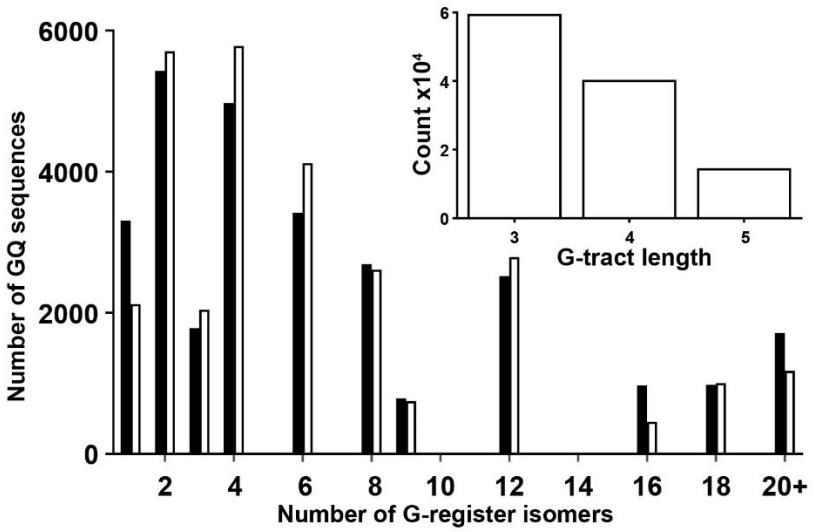


Figure 2.6. Occurrence of GR exchange in predicted human GQ sequences. Black bars show the number of putative GQ sequences identified in a database of human promoter regions[52] (28,377 putative GQ sequences in total) with a given GR isomer multiplicity. White bars show the number of sequences that would be expected to have a given GR isomer multiplicity if G-tracts containing 3, 4, and 5 consecutive dG residues are distributed randomly within GQ sequences. The inset shows the number G-tracts from the putative GQ sequences that contain 3, 4, and 5 consecutive dG residues.

2.5. Discussion

Our results clearly show that GQs with G-tracts of different lengths adopt multiple folded isomers with different core dG residues. Previous NMR spectroscopy and DMS modification studies on the c-myc Pu18 GQ have demonstrated that all possible G-register isomers are populated at equilibrium and interconvert rapidly[19-21]. Our NMR results are in agreement. The spectra of the wild-type GQs correspond closely to the population-weighted average spectra of the trapped mutants, strongly suggesting that all three wild-type GQs populate multiple GR isomers

at equilibrium. Furthermore, DSC and UV-Vis thermal analyses indicated that all three wild-type GQs are entropically stabilized with respect to the most stable of the trapped mutants. This is consistent with the wild-type GQs undergoing relatively rapid exchange among multiple GR isomers. These dynamics likely impact GQ function. GR exchange stabilizes the folded state, and there is a well-established link between GQ stability and gene expression[20, 57]. In addition, we find that GR exchange is coupled to topological rearrangements and can alter binding motifs presented by GQs. The relationship between GR isomerism and topology switching has been previously noted in a GQ from the human telomerase promoter region which has two well-populated GR isomers, one of which favours a parallel configuration, while the other is mixed (3+1) parallel/antiparallel[25]. The differing topological preferences were explained in terms of the different loop interactions in the GR isomers. For instance, in the mixed topology GR isomer, strands 2 and 3 are anti-parallel and the intervening loop is two nucleotides long, stabilized by hydrogen bonding with the third loop. In the parallel GR isomer, strands 2 and 3 are parallel and the intervening loop contains three nucleotides which pack against the GQ core[25]. For PIM1, the situation is more complicated as it exchanges among 12 GR isomers, each with distinct topological preferences, likely governed by the drive to simultaneously optimize hydrogen bonding, base stacking, and the formation of stable features such as single-residue double chain reversal loops[58, 59]. Nevertheless, these results suggest that coupling between GR exchange and topological interconversion may be a general feature of GQ dynamics. The fact that motions on different sides of the GQ are correlated is significant, as coupled conformational changes at distal sites in biological macromolecules has been identified as a prerequisite for the existence of allostery[60]. Many synthetic ligands that target GQs bind with stoichiometries greater than one, and GQs are known to interact with each other and with proteins[61-63]. Furthermore, it has been shown that

different conformational states of the same GQ can have different protein binding competencies[64]. Coupled motions of the G-tracts could thus modulate the molecular recognition of GQs *in vivo* and influence their interactions with drugs.

It must be noted that we performed *in vitro* measurements on large populations of molecules. In a living cell, a given GQ-forming sequence is present only once, or at most in a few copies. To extrapolate our results to the cellular level, one may invoke the ergodic hypothesis[65] which states that the population of a given conformational state in a large ensemble is proportional to the length of time spent by a single molecule in that state, evaluated over a long period. In other words, our finding that GR exchange increases the stability of the folded state by factors of up to 4.5 implies that an isolated G-rich DNA sequence would spend 4.5-fold more of its time in the folded rather than the unfolded state, as a direct result of GR exchange. Since folded GQs can block polymerase read-through[66], this could directly modulate transcriptional control by an individual G-rich sequence. Furthermore, the stabilities that we have determined for the different GR isomers are directly related to the total amounts of time that an individual GQ-forming sequence spends in each state, which modulates the probability that an encounter with a GQ-binding protein[62] will result in complex formation.

The internal dynamics of nucleic acids are increasingly recognized as being essential to their biological activity. Conformational transitions in RNA molecules are involved in metabolite and temperature sensing by riboswitches[67, 68], catalysis by ribozymes[69], and the assembly of RNA-protein complexes[70], among other processes. DNA duplex dynamics have been linked to recognition by transcription factors[71], DNA damage[72], and ligand binding[73]. DNA GQs are well known to be flexible in the native state[74] and can populate equilibrium folding intermediates[37, 75]. Some G-rich sequences assemble into multimeric structures[4] while others contain more than four

G-tracts, with the ability to substitute the extra "spare tire" tracts into the four-stranded core[76]. Thus the GQ dynamical repertoire is complex and varied. Studies have characterized nucleic acid motion mechanically, defining the ranges of motion of RNA hinges and relating these to molecular recognition[77]. Others have determined the thermodynamic and kinetic parameters governing excursions to weakly-populated excited states[78]. What sets the current work apart from these studies is that we have determined quantitative folding parameters for each member of large conformational ensembles containing as many as 12 discrete isomers. To our knowledge, these are the largest biomolecular conformational ensembles to be characterized at this level of detail. Our data provide a rare opportunity to precisely define the highly complex internal motions of nucleic acids.

2.6. Conclusions

We have identified the sliding of G-tracts with respect to each other as a prevalent form of internal dynamics in GQs. GR exchange likely extends to G-vacant, “spare-tire”, multimeric, and RNA GQs[4, 11, 56, 76, 79], where the GQ core could undergo similar conformational rearrangements. This reinforces the idea that GQs must be considered as dynamic ensembles with differentially populated structural forms. We have shown GR exchange contributes directly to thermodynamic stability and it could potentially modulate higher-order interactions and biological function. We demonstrate that mutant GQs harbouring dG>dT and dG>dI substitutions that quench these sliding motions are good structural and thermodynamic mimics of the corresponding wild-type GR isomers. Globally fitting thermal denaturation data for complete sets of wild-type GQs and trapped mutants thus yields the populations of all GR isomers as a function of temperature. We observed melting point elevations of up to 7.3±1.6 °C for a GQ undergoing 12-state exchange relative to the

single most stable GR isomer. A bioinformatic analysis of human promoter sequences revealed that great majority of naturally-occurring GQ sequences can potentially undergo GR exchange with many having 20 or more different GR isomers. Thus a full description of GR exchange is critical for understanding GQ stability and function. The global fitting method we introduce here is inexpensive, rapid (particularly for UV-Vis data), and yields exquisitely detailed information on the conformational sub-states comprising the native structural ensemble of GQs. We believe it represents a new and powerful tool for elucidating structure-dynamics-function relationships for this important class of molecule.

2.7. Materials and Methods

2.7.1. Sample preparation

DNA samples were produced with a MerMade6 (Bioautomation, USA) synthesizer or were ordered from Alpha DNA (Canada) with the 5'-dimethoxytrityl group on. Syntheses utilized a high loading CPG (GlenResearch, USA) to boost yields. Oligonucleotides were removed from the CPG and deprotected using ammonium hydroxide and methylamine (1:1). Samples were purified using GlenPak columns (Glen Research, USA) and purities were determined by LC-mass spectrometry using a Bruker Maxis Impact (QTOF ESI negative mode, Bruker USA) mass spectrometer. All samples were lyophilized, resuspended in buffers mentioned below and dialyzed against the same buffer for 24 hours. Concentrations were determined by measuring the A_{260} and using nearest neighbor extinction coefficients[80]. Samples were denatured at 90 °C for five minutes and then annealed in an ice bath to promote intramolecular GQ formation prior to characterization.

2.7.2. CD spectroscopy

CD experiments were performed using a JASCO J-810 (JASCO, USA) spectropolarimeter with a cell path length of 0.1 mm. Spectra were recorded with 10 μM samples in 2.5 mM K_2HPO_4, 2.5 mM KH_2PO_4 buffer, pH 7.2. The samples were scanned six times from 330 to 230 nm at 25 °C.

2.7.3. NMR spectroscopy

1D ^1H NMR spectra were collected at 25 °C using an Agilent INOVA spectrometer operating at 500 MHz proton Larmor frequency. Sample concentrations were 0.2-0.3 mM in 2.5 mM K_2HPO_4, 2.5 mM KH_2PO_4 buffer, pH 7.2, with 10% D_2O and referenced to 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS). Water suppression was achieved with double pulsed field gradient spin echos (DPFGSE) using a sweep width of 25 ppm. Each spectrum was recorded using 256 transients. The imino proton region of the spectrum was selected as the fingerprint for GR exchange. Spectra were processed and analyzed using MESTRENOVA NMR software.

2.7.4. Experimental DSC

The data were collected using a NanoDSC-III microcalorimeter (TA Instruments, USA). DNA samples were 150 μM in 2.5 mM K_2HPO_4, 2.5 mM KH_2PO_4 buffer, pH 7.2. Low ionic strength was used to reduce the melting temperatures to experimentally accessible values. Samples were scanned from 5 to 90 °C fifteen times using a scan rate of 1 °C per minute.

2.7.5. Experimental UV-Vis spectroscopy

Spectra were recorded using a Cary Win-UV spectrophotometer (Agilent Technologies, USA). Samples were scanned eighteen times at 260 and 295 nm from 15 to 95 °C. Data were collected using sample concentrations of 5 or 10 μM in 2.5 mM K_2HPO_4, 2.5 mM KH_2PO_4 buffer, pH 7.2. Experiments performed at higher K^+ concentrations (130 mM) showed identical trends to those performed with 7.5 mM K^+, but the denaturation curves were shifted to higher and less experimentally-accessible temperatures (Supplementary Figure 2.18).

2.7.6. DSC global fitting

The first step of DSC thermal melt analysis is the subtraction of data obtained for the buffer alone from the sample data (Supplementary Figure 2.19). The resulting raw C_p profile is then baseline-corrected, which accounts for the temperature dependence of the heat capacity, to first order. In the case of DSC data for GQs, baselines are typically polynomial (quadratic or cubic) in temperature [37, 48, 81]. We employed a quadratic baseline such that the corrected heat capacity is given by

$$C_p(T) = C_p^{raw}(T) - a - bT - cT^2, \tag{2.2}$$

where the coefficients *a*, *b*, and *c* were optimized in the global fits (Supplementary Figure 2.20) as described below. The trapped mutants were assumed to fold in a two-state manner (see Section 2.4.3), such that the thermogram of each mutant was uniquely defined by its folding enthalpy and entropy at a reference temperature (T_0) and the heat capacity difference between the folded and unfolded states (ΔC_p) according to

$$\Delta H^{mut}(T) = H_F^{mut}(T) - H_U^{mut}(T) = \Delta H^{mut}(T_0) + \Delta C_p(T - T_0) \tag{2.3}$$

$$\Delta S^{mut}(T) = S_F^{mut}(T) - S_U^{mut}(T) = \Delta S^{mut}(T_0) + \Delta C_p \ln\left\{\frac{T}{T_0}\right\}. \quad (2.4)$$

It is frequently assumed that $\Delta C_p = 0$ in thermal analyses of GQs[29, 82], although some studies have found that ΔC_p is small and positive for GQ unfolding[37, 83]. We performed analyses using both $\Delta C_p = 0$ and the value of $\Delta C_p = 1.3$ kJ mol^{-1} K^{-1} that was previously measured for the human telomere GQ[37] and obtained essentially the same folding parameters and populations (Supplementary Table 2.1, Supplementary Figure 2.21). We also optimized the value of ΔC_p as an adjustable parameter in the fits, yielding $\Delta C_p \approx 0.24$ kJ mol^{-1} K^{-1}. We note that our data are incompatible with ΔC_p values larger than about ~2 kJ mol^{-1} K^{-1}, as at that point the quality of the fits begins to degrade substantially. Nevertheless, virtually identical folding parameters are obtained even with ΔC_p fixed at 2.1 kJ mol^{-1} K^{-1}. Since our results are insensitive to the choice of physically realistic ΔC_p values, we performed our analysis with $\Delta C_p = 0$, for the sake of simplicity. With this assumption ΔH^{mut} and ΔS^{mut} are temperature-independent and define the folding equilibrium constant (K^{mut}) according to

$$K^{mut}(T) = \frac{[F](T)}{[U](T)} = \exp\left(\frac{-\left(\Delta H^{mut} - T\Delta S^{mut}\right)}{RT}\right), \quad (2.5)$$

which, in turn, defines the relative population of the folded state (P^{mut}) according to

$$P^{mut}(T) = \frac{K^{mut}(T)}{1 + K^{mut}(T)}. \quad (2.6)$$

Finally, the excess heat capacity thermogram for each trapped mutant is given by

$$C_p^{calc}(T) = \Delta H^{mut} \times \frac{d}{dT} P^{mut}(T). \quad (2.7)$$

In the case of a wild-type GQ undergoing GR exchange the situation is more complicated. In general, the folding enthalpy and entropy of each GR isomer can be different, such that

$$\Delta H_i^{WT} = H_{F,i}^{WT} - H_U^{WT} \tag{2.8}$$

$$\Delta S_i^{WT} = S_{F,i}^{WT} - S_U^{WT}, \tag{2.9}$$

where ΔH_i^{WT} and ΔS_i^{WT} are the folding enthalpy and entropy of the i^{th} GR isomer. Note that there is only one unfolded state for a GQ that undergoes GR exchange in the folded state. The folding equilibrium constant for the i^{th} GR isomer is defined according to

$$K_i^{WT}(T) = \frac{[F]_i^{WT}(T)}{[U]^{WT}(T)} = \exp\left(\frac{-\left(\Delta H_i^{WT} - T\Delta S_i^{WT}\right)}{RT}\right). \tag{2.10}$$

The fraction of molecules populating the i^{th} folded GR isomer is given by

$$P_i^{WT}(T) = \frac{K_i^{WT}(T)}{1 + \sum_i K_i^{WT}(T)} \tag{2.11}$$

where the sum runs over all GR isomers. The excess heat capacity thermogram for the wild-type GQ is then given by

$$C_p^{calc}(T) = \sum_i \Delta H_i^{WT} \times \frac{d}{dT} P_i^{WT}(T). \tag{2.12}$$

Provided that the trapping mutations do not perturb stability beyond restricting the GQ to a single GR isomer, the folding thermodynamics of the wild-type and trapped mutants are related according to

$$\Delta H_i^{WT} = \Delta H^{mut} \tag{2.13}$$

$$\Delta S_i^{WT} = \Delta S^{mut} \tag{2.14}$$

where ΔH_i^{WT} and ΔS_i^{WT} are the folding enthalpy and entropy of the i^{th} wild-type GR isomer and ΔH^{mut} and ΔS^{mut} are the folding enthalpy and entropy of the corresponding trapped mutant.

In the global fits, the DSC data for all trapped mutants and the wild-type GQ were analyzed simultaneously. The thermogram for each trapped mutant was calculated according to Equations 2.3-2.7 with ΔH^{mut} and ΔS^{mut} defined independently for each mutant. The thermogram for the wild-type was calculated according to Equations 2.8-2.12, using the thermodynamic parameters from the corresponding trapped mutants according to Equations 2.13 and 2.14. For a GQ with N GR isomers, the dataset comprised N+1 DSC thermograms (N mutants and the wild-type) and the global fit contained N values of ΔH and ΔS and one second-order polynomial baseline shared between all datasets, for a total of $2N$+3 adjustable parameters. The parameters were varied to minimize the residual sum of squares (RSS), which is the sum of squared differences between the experimental data points and the values calculated using the global thermodynamic parameters,

$$RSS = \sum_{j=1}^{N} \sum_{k} \left(C_{p,j}^{\exp}(T_k, \lambda) - C_{p,j}^{calc}(T_k, \xi_j) \right)^2 + \sum_{k} \left(C_{p,WT}^{\exp}(T_k, \lambda) - C_{p,WT}^{calc}(T_k, \xi_{1..N}) \right)^2 \qquad (2.15)$$

where T_k is the k^{th} experimental temperature, $C_{p,j}^{exp}(T)$ and $C_{p,j}^{calc}(T)$ are the experimental and calculated excess heat capacity values of the j^{th} trapped mutant, $C_{p,WT}^{exp}(T)$ and $C_{p,WT}^{calc}(T)$ are the experimental and calculated excess heat capacity values of the wild-type, ξ_j=[ΔH_j^{mut}, ΔS_j^{mut}] are the folding parameters of the j^{th} trapped mutant, and λ_j=[a, b, c] are the coefficients of the quadratic baseline. UV-Vis unfolding traces were modeled similarly (see Section 2.7.7).

2.7.7. UV-Vis spectroscopy global fitting

The total absorbance $A(T)$ for each trapped mutant was calculated as

$$A^{calc}(T) = P^{mut}(T) \times A_F(T) + \left(1 - P^{mut}(T)\right) \times A_U(T) \qquad (2.16)$$

where A_F and A_U are the folded and unfolded absorbance baselines, respectively, and the relative population of the folded GQ, $P^{mut}(T)$, was determined as described in Section 2.7.6 (Equations 2.3-2.6). The absorbance baselines were assumed to depend linearly on temperature giving

$$A_F(T) = m_F T + b_F \tag{2.17}$$

and

$$A_U(T) = m_U T + b_U \,. \tag{2.18}$$

For the wild-type GQs undergoing GR exchange, the total absorbance, *A(T)* was calculated as

$$A^{calc}(T) = \left(\sum_i P_i^{WT}(T) \right) \times A_F(T) + \left(1 - \left(\sum_i P_i^{WT}(T) \right) \right) \times A_U(T) \tag{2.19}$$

where P_i^{WT} is the fraction of molecules populating the i^{th} folded GR isomer. In the global fits, the UV-Vis data for all trapped mutants and the wild-type GQ were analyzed simultaneously. The absorbance profile for each trapped mutant was calculated according to Equations 2.3-2.6 and 2.16-2.18 with ΔH^{mut} and ΔS^{mut} defined independently for each mutant. The absorbance profile for the wild-type was calculated according to Equations 2.10, 2.11, and 2.19, using the thermodynamic parameters from the corresponding trapped mutants according to Equations 2.13 and 2.14. The folded baselines for all trapped mutants and the wild-type were optimized independently. The unfolded baseline intercept (b_U) was also optimized independently for each trapped mutant and wild-type, while the unfolded slope (m_U) was constrained to be identical for all absorbance profiles in the global fit. For a GQ with N GR isomers, the dataset comprised N+1 absorbance profiles (N mutants and the wild-type) and the global fit contained N values of ΔH and ΔS, N+1 values of m_F, b_F, and b_U, and a single value of m_U for a total of 5N+4 adjustable parameters. The parameters were varied to minimize the RSS

$$RSS = \sum_{j=1}^{N} \sum_k \left(A_j^{\exp}(T_k) - A_j^{calc}(T_k, \xi_j, \lambda_j) \right)^2 + \sum_k \left(A_{WT}^{\exp}(T_k) - A_{WT}^{calc}(T_k, \xi_{1..N}, \lambda_{WT}) \right)^2 \tag{2.20}$$

where T_k is the k^{th} experimental temperature, $A_j^{exp}(T)$ and $A_j^{calc}(T)$ are the experimental and calculated absorbance profiles of the j^{th} trapped mutant, $A_{WT}^{exp}(T)$ and $A_{WT}^{calc}(T)$ are the experimental and calculated absorbance profiles of the wild-type, $\xi_j=[\Delta H_j^{mut}, \Delta S_j^{mut}]$ are the folding parameters of the j^{th} trapped mutant, and $\lambda_j=[b_{F,j}, m_{F,j}, b_{U,j}, m_U]$ and $\lambda_{WT}=[b_{F,WT}, m_{F,WT}, b_{U,WT}, m_U]$ are the coefficients of the linear folded and unfolded baselines of the trapped mutants and wild-type.

2.7.8. Assessing thermodynamic perturbations in trapping GR isomers

Monte Carlo computer simulations[84] of mutation-induced thermodynamic perturbations were performed as follows. A complete set of N ΔH^{WT} and ΔS^{WT} enthalpy and entropy values were selected as the "unperturbed" folding parameters of the N wild-type GR isomers, according to the optimized values extracted from the global fits. Simulated wild-type thermal denaturation data were then generated according to

$$A^{WT}\left(T_j\right) = f\left(\Delta H_{1..N}^{WT}, \Delta S_{1..N}^{WT}, T_j\right) + \sigma \times \varepsilon_j \tag{2.21}$$

where T_j is the j^{th} temperature point, ε_j are random numbers with a mean of zero and standard deviation of 1, and $f(x)$ represents Equations 2.10-2.12 (DSC) or 2.10, 2.11 and 2.19 (UV-Vis). σ is the experimental error in each point, estimated from the average reduced RSS of the individual fits, i.e.

$$\sigma = \sqrt{\left\langle \frac{RSS^{ind}}{DF} \right\rangle} \tag{2.22}$$

where DF is the degrees of freedom of each individual trapped mutant fit and angle brackets indicate the average. Perturbed trapped mutant denaturation data were then generated according to

$$A_i^{pert}\left(T_j\right) = f\left(\left(\Delta H_i^{WT} + \Delta\Delta E \times \varepsilon_{i,H}\right), \left(\Delta S_i^{WT} + \frac{\Delta\Delta E}{T_{ref}}\varepsilon_{i,S}\right), T_j\right) + \sigma \times \varepsilon_{i,j} \quad (2.23)$$

where $A_i^{pert}(T_j)$ is the data point at the j^{th} temperature for the i^{th} trapped mutant, ΔH_i^{WT} and ΔS_i^{WT} are the folding parameters for the corresponding wild-type GR isomer, $\Delta\Delta E$ governs the size of thermodynamic perturbations with units of kJ mol^{-1}, T_{ref}=298 K is a reference temperature, $\varepsilon_{i,H}$, $\varepsilon_{i,S}$, and $\varepsilon_{i,j}$ are random numbers with means of zero and standard deviations of 1, and $f(x)$ represents Equations 2.10-2.12 (DSC) or 2.10, 2.11, and 2.19 (UV-Vis). The simulated data for the wild-type and complete set of trapped mutants were then globally fitted, iterating 10^3 times for each value of $\Delta\Delta E$ in the case of c-myc-Pu18 and 25 times in the case of PIM1. $\Delta\Delta E$ was varied from 0 to 5 kJ mol^{-1}, taking the average RSS from all iterations as the corresponding residual sum of squared differences. The sign of $\Delta\Delta E$ added to the unperturbed ΔH and ΔS was kept the same for each iteration, i.e. a mutation that was enthalpically destabilizing was also entropically stabilizing. In order to evaluate the precision of the extracted parameters, identical Monte Carlo calculations were performed with $\Delta\Delta E=0$. The precision of the measurement was taken to be the reciprocal of the standard deviation of values obtained in all iterations.

2.7.9. Identification of GR exchange in GQ sequences from the Eukaryotic Promoter Database

GQs are formed by sequences containing four G_n-tracts, of n contiguous dG residues per tract, separated by loop sequences N_m of any composition, where m is the number of loop bases. Biological GQs are thought to form according to the rule devised by Balasubramanian *et al.* which identifies 5'-$G_{\geq3}N_{1\text{-}7}G_{\geq3}N_{1\text{-}7}G_{\geq3}N_{1\text{-}7}G_{\geq3}$-3' as the consensus sequence for stable folding[85]. GQ stability decreases with increasing loop length and stable biological GQs tend to contain short

loops, thus the folding rule loop length is defined with $1 \leq m \leq 7$. GQs formed from G-tract lengths n = 1-2 are known, yet these are typically unstable relative to other conformational states[86] and so are not likely to be well populated under biological conditions. When n = 3 for all four G_n-tracts and the loops are free of dG residues, three G-tetrads are formed and every dG participates in the GQ core. Interestingly, Fox *et al.* demonstrated GQ sequences with all G_4 or G_5 tracts did not form more than three G-tetrads[12], suggesting multiple GR isomers are possible for these sequences with different subsets of dG residues participating in the formation of three G-tetrads. GQs with all G_6 or G_7 tracts did form >3 G-tetrads. However, such GQs are biologically rare, their dynamics are likely complicated, and they were not included in our analysis, i.e. we only considered sequences with $n \leq 5$. We therefore used a query sequence, 5'-$G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}$-3', where N = dA, dC, dT, or dG to search the Eukaryotic Promoter Database[52] for GQ sequences that could potentially undergo GR exchange. No more than two consecutive dG residues were allowed in any loop sequence, nor were dG residues allowed immediately adjacent to a flanking G-tract. We examined both coding and complementary strands from the -499 to +100 position of 23 322 human promoter sequences. All sequences matching the query were analyzed to determine the number of possible GR isomers (see Section 2.7.10).

2.7.10. Calculation of GR isomer numbers for GQ sequences from the Eukaryotic Promoter Database

We computed the number of possible GR isomers for each retrieved sequence as follows: The i^{th} G-tract (i=1...4) comprising n_i consecutive dG residues can adopt R_i different registers with respect to a GQ core of three G-tetrads according to

$$R_i = n_i - 2 \quad . \tag{2.24}$$

For example, a G_3-tract has one possible register relative to the GQ core, a G_4-tract has two registers, with either the first or last dG occupying a loop position, and a G_5-tract has three registers. The total number of GR isomers for each retrieved GQ sequence was then computed as

$$R_T = \prod_{i=1}^{4} n_i \,. \tag{2.25}$$

For example, the putative GQ-forming sequence 5'-$G_4N_mG_3N_mG_4N_mG_5$-3' has twelve possible GR isomers, $R_T = 2\times1\times2\times3 = 12$. Due to the maximum G-tract length of 5, the only numbers of GR isomers allowed in this analysis follow the rule $R_T=2^a\times3^b$, where a and b are integers, $0\leq(a,b)\leq4$ and $(a+b)\leq4$.

2.7.11. Predicting thermal upshifts

Assuming that the folding enthalpies of different GR isomers are approximately equal, the $T_m{}^{WT}$ of the wild-type ensemble is related to the T_ms of the individual GR isomers according to

$$T_m^{WT} = T_m \frac{\Delta S_{UF}}{(\Delta S_{UF} - \Delta\Delta S_{GR})} \tag{2.26}$$

where ΔS_{UF} is the unfolding entropy of a single GR isomer and the entropic contribution of exchanging among N GR isomers is given by $\Delta\Delta S_{GR}=R\ln(N)$.

2.7.12. Statistical analysis of errors

Errors in the group fitting parameters were calculated using the variance-covariance matrix[87] given by

$$\hat{V} = \frac{RSS}{DF}(\hat{X}\hat{W}\hat{X}^T)^{-1} \tag{2.27}$$

where *RSS* is the optimized value from the global fit, *DF* is the degrees of freedom of the fit (*N* data points minus Φ parameters of the global fit) and $\hat{W}$ is a diagonal matrix of fitting weights, in this case all taken to be identically 1. $\hat{X}$ is a matrix of the first derivatives of the differences between the experimental and calculated data points (A^{exp} and A^{calc}), with respect to increments in each of the adjustable parameters (Φ_i). The element corresponding to the i^{th} adjustable parameter and j^{th} data point is thus

$$X_{ij} = \frac{\partial\left(A_j^{\exp} - A_j^{calc}\right)}{\partial\Phi_i} \equiv \frac{\partial\alpha_j}{\partial\Phi_i} \tag{2.28}$$

where A_j^{calc} is evaluated at the optimized set of parameters, Φ. The elements were evaluated numerically according to

$$X_{ij} \approx \frac{A_j^{calc}\left(-\Delta\right) - A_j^{calc}\left(+\Delta\right)}{2\Delta} \tag{2.29}$$

where $A_j^{calc}(\pm\Delta)$ is the j^{th} data point calculated with all adjustable parameters set to their optimized values except, for the i^{th} parameter, which is incremented by $\pm\Delta$. For a global fit with *N* data points and *M* adjustable parameters this gives

$$\hat{X} = \begin{bmatrix} \frac{\partial\alpha_1}{\partial\Phi_1} & \cdots & \hat{} \\ \vdots & \ddots & \vdots \\ \frac{\partial\alpha_1}{\partial\Phi_M} & \cdots & \hat{} \end{bmatrix}. \tag{2.30}$$

The diagonal elements in $\hat{V}$ are the variances of the optimized fit parameters, while the off-diagonal elements are the covariances between the errors of the optimized parameters. The errors in group fitting populations were computed from the covariance matrix using a Monte Carlo

approach[84]. 10^4 sets of N thermodynamic parameters with random errors were generated according to

$$\hat{V}' = \hat{L}\hat{L}^T \tag{2.31}$$

$$\begin{bmatrix} \Phi_1^{MC} \\ \vdots \\ \Phi_N^{MC} \end{bmatrix} \begin{bmatrix} \Phi_1^{\exp} \\ \vdots \\ \Phi_N^{\exp} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{2.32}$$

where $\Phi^{exp}{}_{1...N}$ are the thermodynamic and baseline parameters extracted from a global fit, $\Phi^{MC}{}_{1...N}$ are the randomized Monte Carlo parameters for a single iteration, $\hat{V}'$ is the portion of the covariance matrix corresponding to the variances and covariances in the thermodynamic parameters, $\varepsilon_{1...N}$ are random numbers with means of 0 and standard deviation 1, $\hat{L}$ is the lower triangular matrix from Cholesky decomposition satisfying Equation 2.31. The 10^4 sets of thermodynamic parameters were used to generate 10^4 sets of populations. The standard deviations of the sets of populations were taken as their experimental errors.

2.7.13. Wild-type PIM1 thermal CD correction

In the case of PIM1, we found that the CD spectrum of the wild-type GQ contained a stronger signal at around 265 nm than did those of the mutants (Supplementary Figure 2.5a). This is unlikely to be due to intermolecular association of the DNA strands as it has previously been shown by gel electrophoresis that PIM1 forms a monomeric GQ at concentrations as high as 30 μM[29] while the CD analysis was performed at 10 μM. We collected CD spectra at 95 °C, a temperature at which both mutant and wild-type GQs are completely unfolded. The CD spectrum for the wild-type sequence retained a strong maximum at 265 nm, while those of the mutants were

essentially flat with values of approximately zero (Supplementary Figure 2.5b). CD spectra of poly-dG/poly-dC duplex DNA also contain strong maxima at 265 nm while those of poly-dGdC do not[88]. Therefore it appears that DNA strands with ≥5 consecutive dG residues produce this spectral signature in duplex, single-stranded, and GQ forms. To a first approximation, we corrected for this effect by subtracting the spectrum obtained at 95 °C from that obtained at 25 °C for the wild-type PIM1 GQ (Figure 2.2c,d and Supplementary Figure 2.2d-f).

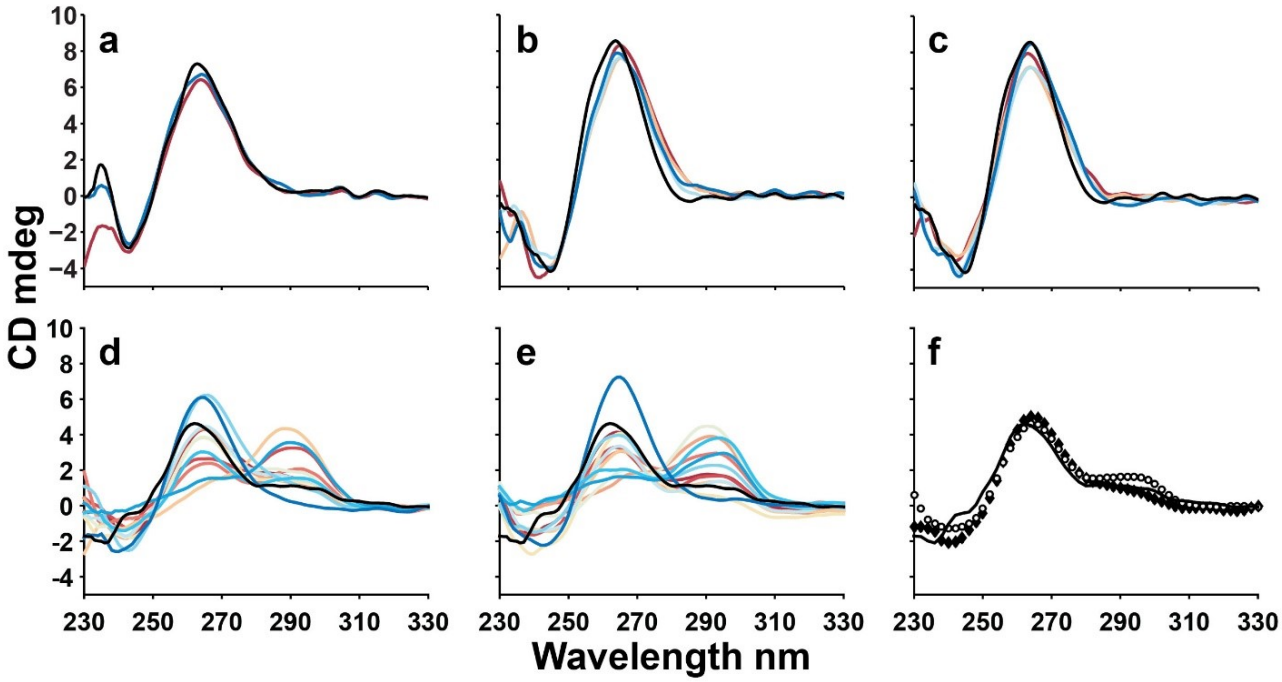
2.7.14. Monte Carlo simulations of thermodynamic perturbations in PIM1 trapped mutants

We found that the global fit of PIM1 data is less sensitive to perturbation (Figure 2.5c and Supplementary Figure 2.15) than that of c-myc. Applying the same Monte Carlo procedure, we obtained a maximum of perturbation of 2.4 kJ mol^{-1}, which is about 1.3% of the total folding enthalpy and 3.3% of the $\Delta\Delta H_F$ between most and least stable mutants. This result is not surprising, given that the global fit of PIM1 data involves many more GR isomers, such that opposing thermodynamic perturbations in different trapped mutants can essentially compensate for each other.

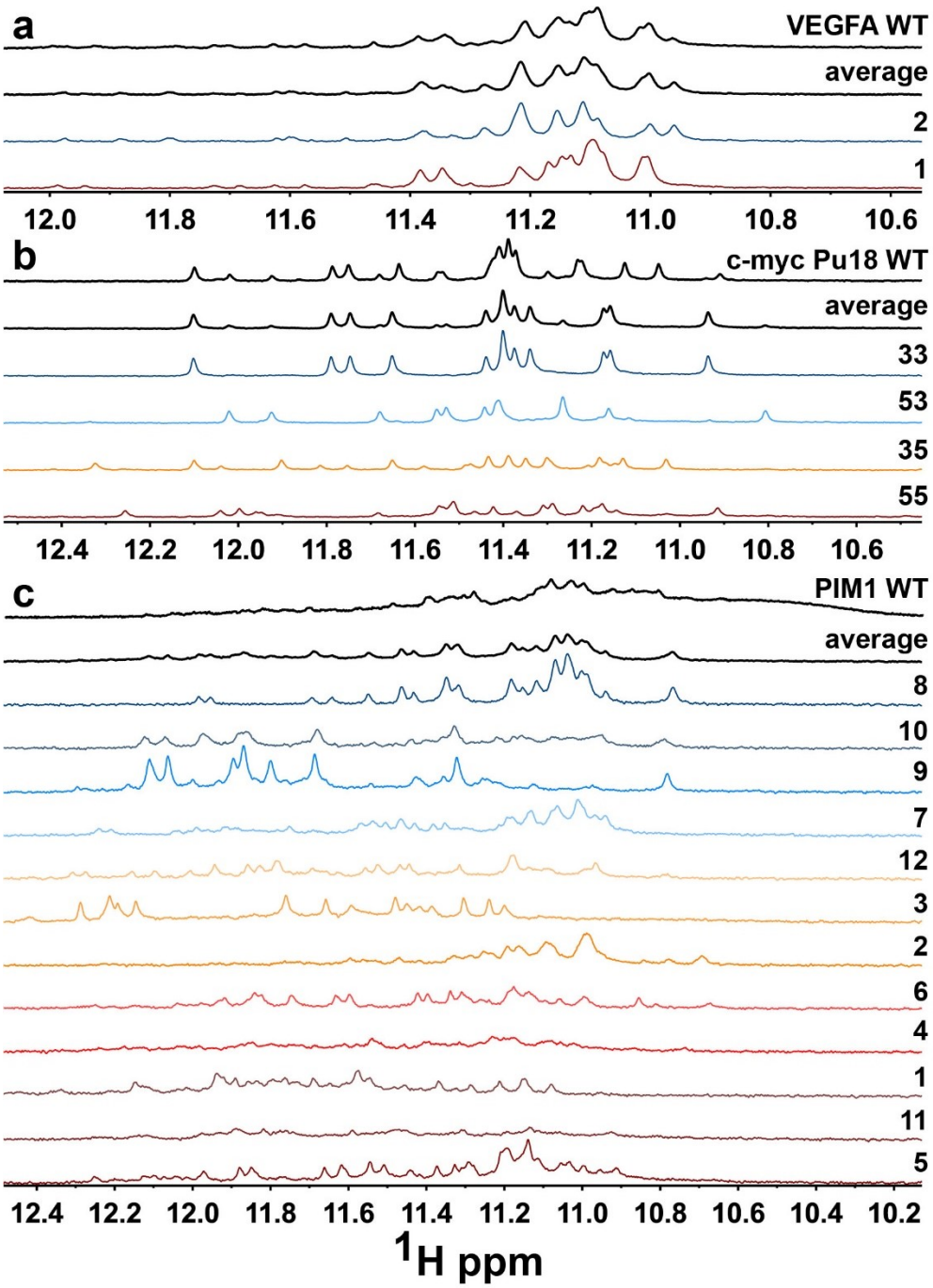
2.8. Supplementary Figures

c-myc Pu18	WT	5′-AGGGTGGGGAGGGTGGGG-3′		
	55dT	5′-AGGGTTGGGAGGGTTGGG-3′	55dI	5′-AGGGTIGGGAGGGTIGGG-3′
	35dT	5′-AGGGTGGGTAGGGTTGGG-3′	35dI	5′-AGGGTGGGIAGGGTIGGG-3′
	53dT	5′-AGGGTTGGGAGGGTGGGT-3′	53dI	5′-AGGGTIGGGAGGGTGGGI-3′
	33dT	5′-AGGGTGGGTAGGGTGGGT-3′	33dI	5′-AGGGTGGGIAGGGTGGGI-3′
VEGFA			WT	5′-GGGAGGGTTGGGGTGGG-3′
			1dI	5′-GGGAGGGTTIGGGTGGG-3′
			2dI	5′-GGGAGGGTTGGGITGGG-3′
PIM1	WT	5′-GGGCGGGGCGGGGGCGGGG-3′		
	1dT	5′-GGGCTGGGCTTGGGCTGGG-3′	1dI	5′-GGGCIGGGCIIGGGCIGGG-3′
	2dT	5′-GGGCTGGGCTTGGGCGGGT-3′	2dI	5′-GGGCIGGGCIIGGGCGGGI-3′
	3dT	5′-GGGCTGGGCTGGGTCTGGG-3′	3dI	5′-GGGCIGGGCIGGGICIGGG-3′
	4dT	5′-GGGCTGGGCTGGGTCGGGT-3′	4dI	5′-GGGCIGGGCIGGGICGGGI-3′
	5dT	5′-GGGCTGGGCGGGTTCTGGG-3′	5dI	5′-GGGCIGGGCGGGIICIGGG-3′
	6dT	5′-GGGCTGGGCGGGTTCGGGT-3′	6dI	5′-GGGCIGGGCGGGIICGGGI-3′
	7dT	5′-GGGCGGGTCTTGGGCTGGG-3′	7dI	5′-GGGCGGGICIIGGGCIGGG-3′
	8dT	5′-GGGCGGGTCTTGGGCGGGT-3′	8dI	5′-GGGCGGGICIIGGGCGGGI-3′
	9dT	5′-GGGCGGGTCTGGGTCTGGG-3′	9dI	5′-GGGCGGGICIGGGICIGGG-3′
	10dT	5′-GGGCGGGTCTGGGTCGGGT-3′	10dI	5′-GGGCGGGICIGGGICGGGI-3′
	11dT	5′-GGGCGGGTCGGGTTCTGGG-3′	11dI	5′-GGGCGGGICGGGIICIGGG-3′
	12dT	5′-GGGCGGGTCGGGTTCGGGT-3′	12dI	5′-GGGCGGGICGGGIICGGGI-3′

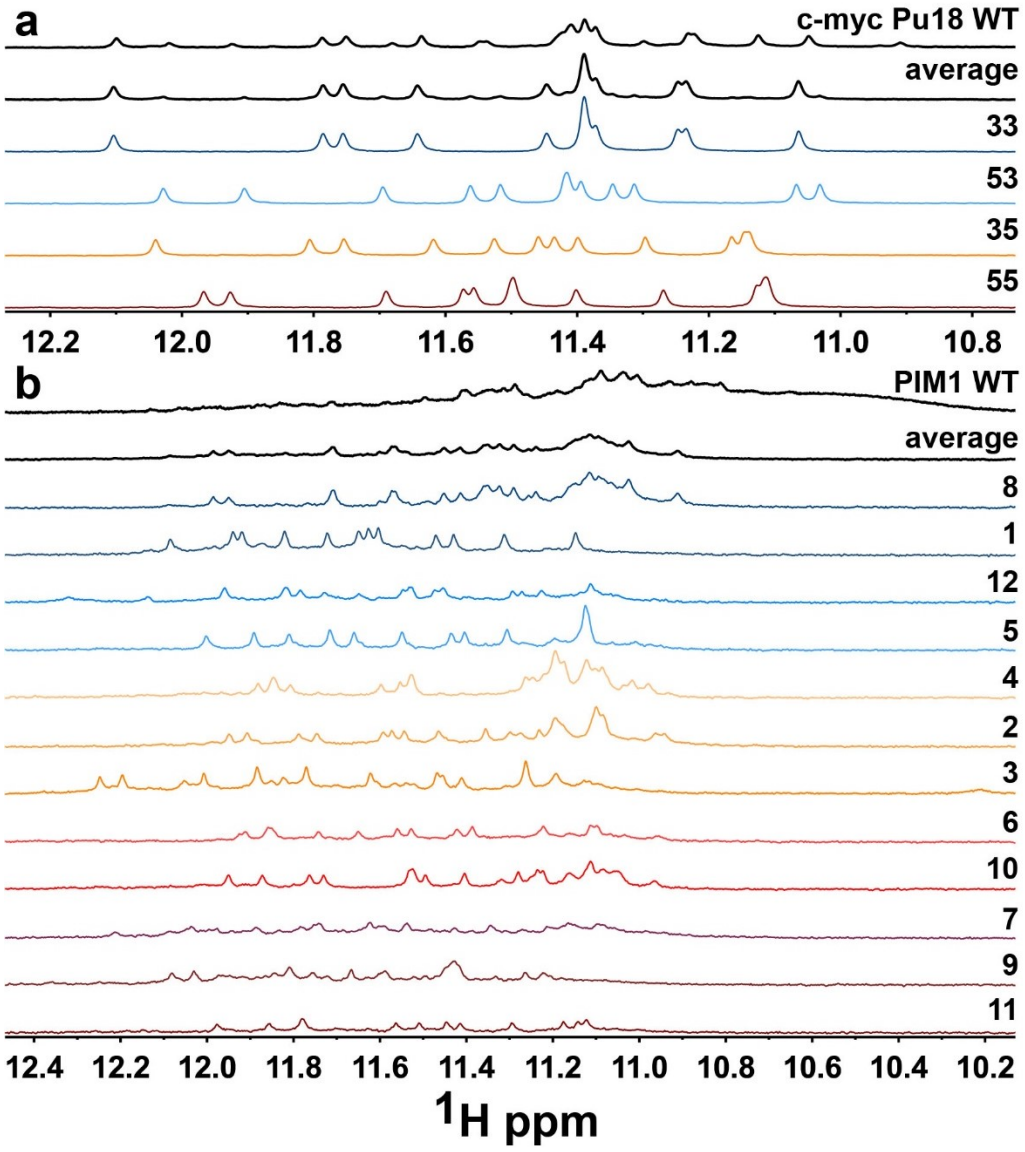
Supplementary Figure 2.1. GQ sequences investigated in this work. Red letters indicate bases that were mutated from dG>dX where dX = dT or dI.



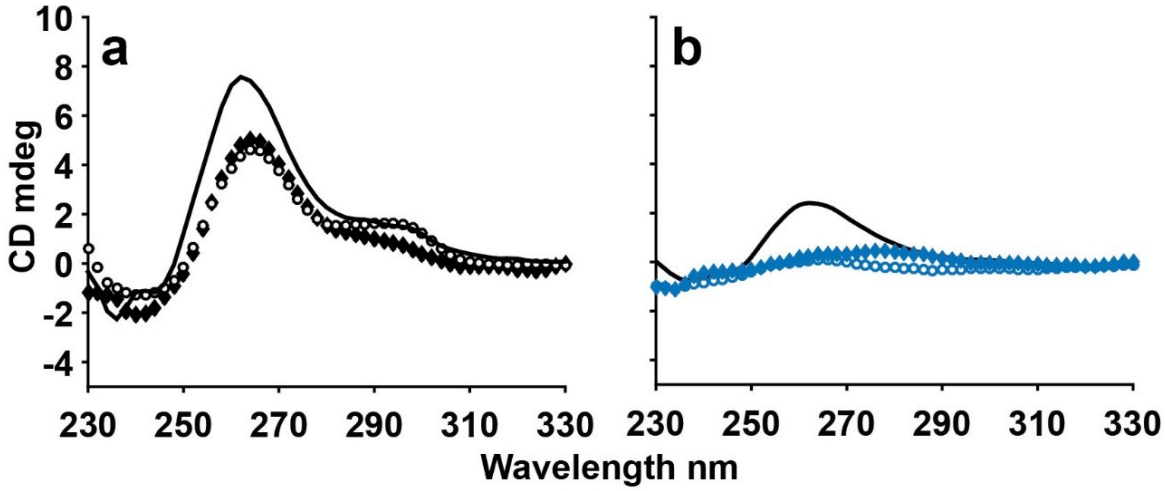
Supplementary Figure 2.2. GQ CD spectra. CD spectra for wild-type (black) and trapped mutant (colored) VEGFA dG>dI (a), c-myc Pu18 dG>dT (b), c-myc Pu18 dG>dI (c), PIM1 dG>dT (d), and PIM1 dG>dI GQs (e). The color coding of trapped mutant data in (a-e) is the same as in Figure 2.4. In (d-f), the solid line corresponds to the corrected wild-type PIM1 CD spectrum. In (f) population-weighted average spectra of dG>dT and dG>dI trapped mutants are shown with filled diamonds and empty circles, respectively.



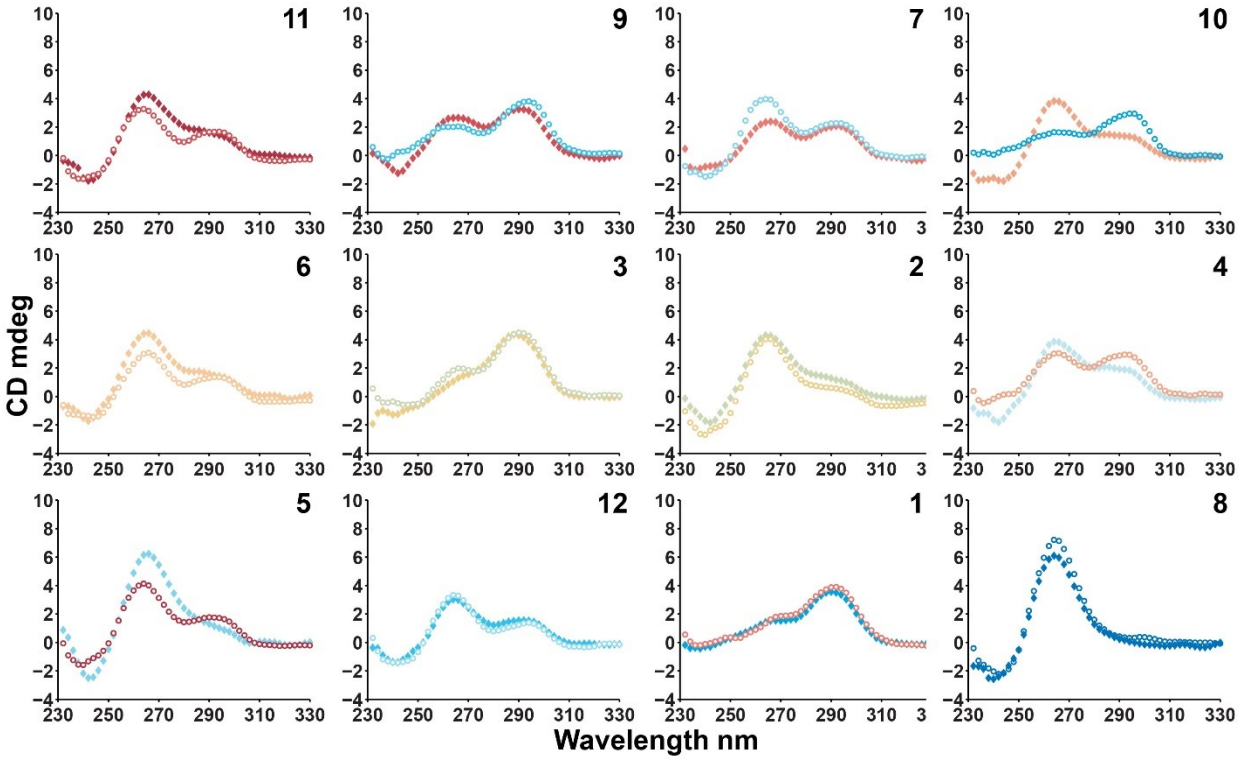
Supplementary Figure 2.3. 1D ^{1}H NMR spectra of wild-type and trapped mutant GQs. Imino proton regions of ^{1}H NMR spectra for VEGFA (a), c-myc Pu18 (b), and PIM1 (c) wild-type and dI trapped mutant GQs. Mutant spectra are ordered according to stability with more stable trapped mutants shown above less stable trapped mutants.



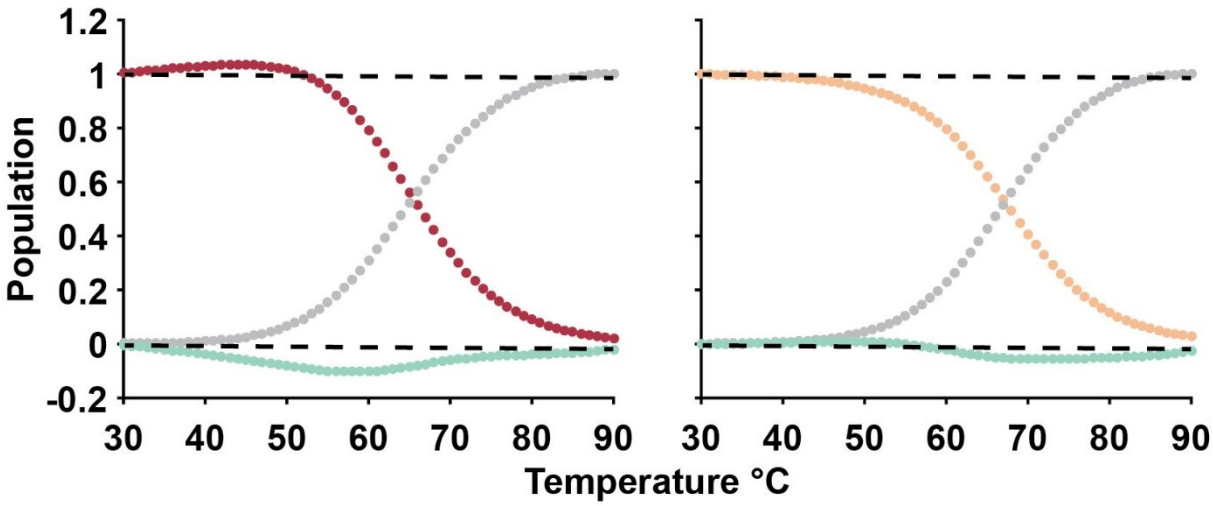
Supplementary Figure 2.4. ^{1}H NMR spectra for the wild-type and trapped dT mutant GQs. Imino proton regions of ^{1}H NMR spectra for c-myc Pu18 (a), and PIM1 (b) wild-type and dT trapped mutant GQs. Mutant spectra are ordered according to stability with more stable trapped mutants shown above less stable trapped mutants.



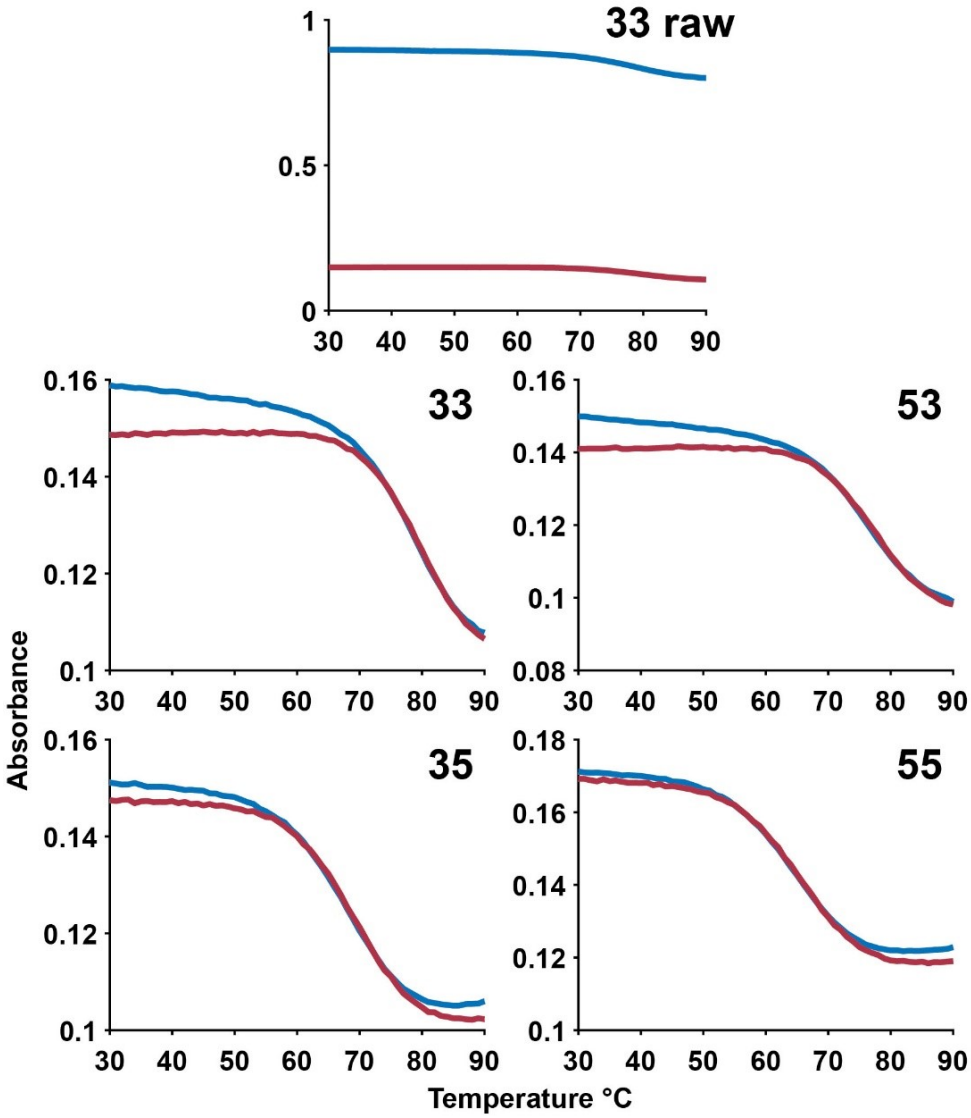
Supplementary Figure 2.5. Wild-type PIM1 CD correction. (a) CD spectrum of the wild-type PIM1 GQ (black curve) and the population-weighted average CD spectra of the dT (filled symbols) and dI (open symbols) trapped mutants at 25 °C. (b) CD spectra of the wild-type PIM1 GQ (black curve) and the most populated dT (filled symbols) and dI (open symbols) trapped mutants at 95 °C.



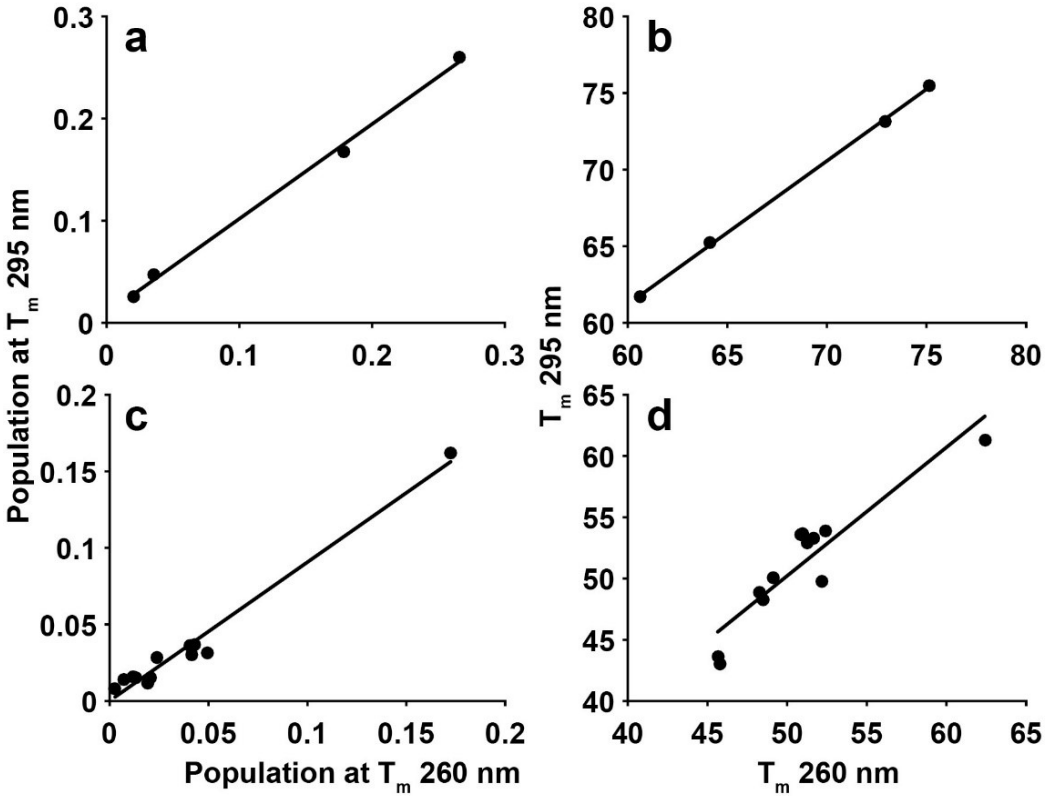
Supplementary Figure 2.6. PIM1 CD spectra. CD spectra of dG>dT (filled symbols) and dG>dI (open symbols) trapped mutants, indicated by the number in the upper right of each panel. The color coding matches that of Supplementary Figure 2.2.



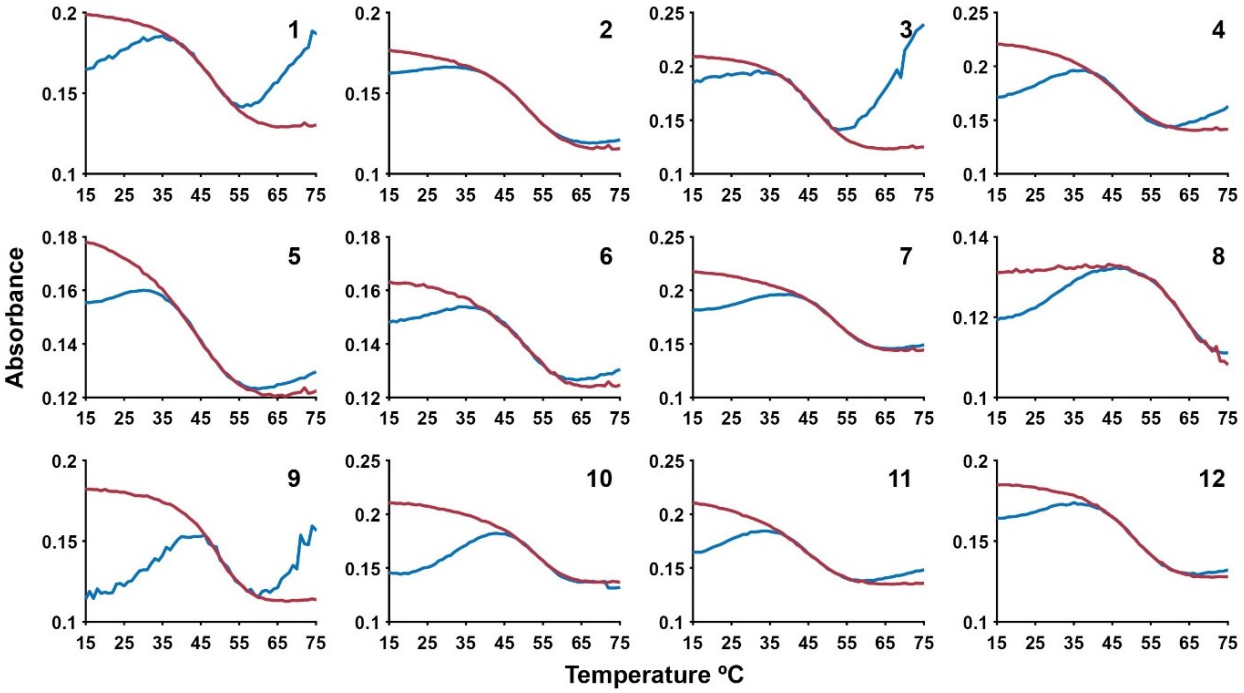
Supplementary Figure 2.7. Model-free deconvolution of experimental DSC data. The c-myc Pu18 55 and 35 dI trapped mutant DSC data (left and right panels respectively) were processed according to the Freire and Biltonen deconvolution method[36, 49]. Folded and unfolded populations (F and F_0) are shown as colored and grey filled circles respectively. 1-(F+F_0) is shown as turquoise filled circles. Dashed black lines indicate the 1 and 0 population limits.



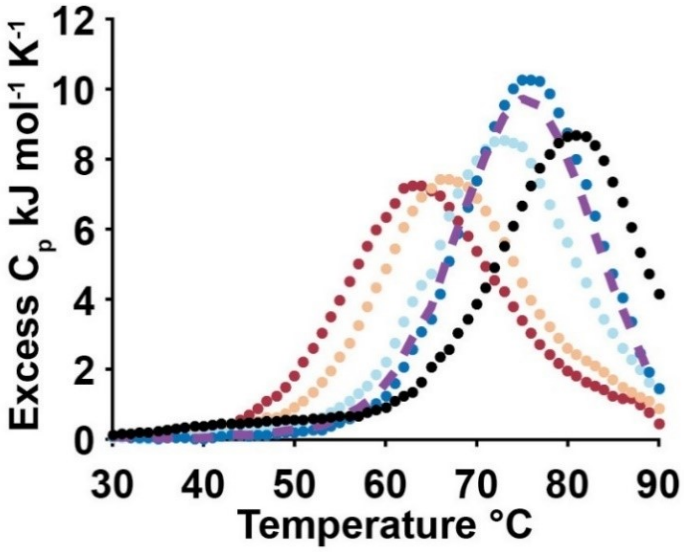
Supplementary Figure 2.8. Dual-wavelength absorbance melting for c-myc Pu18 trapped mutants. Raw absorbance data (top panel, 33 trapped mutant) at 260 nm have been vertically offset and scaled for comparison with the 295 nm data (lower four panels). Melting data at 260 and 295 nm are shown as blue and red curves respectively. Trapped dI mutant numbers are indicated in the top right corners. Experiments were performed using 5 μM strand concentrations.



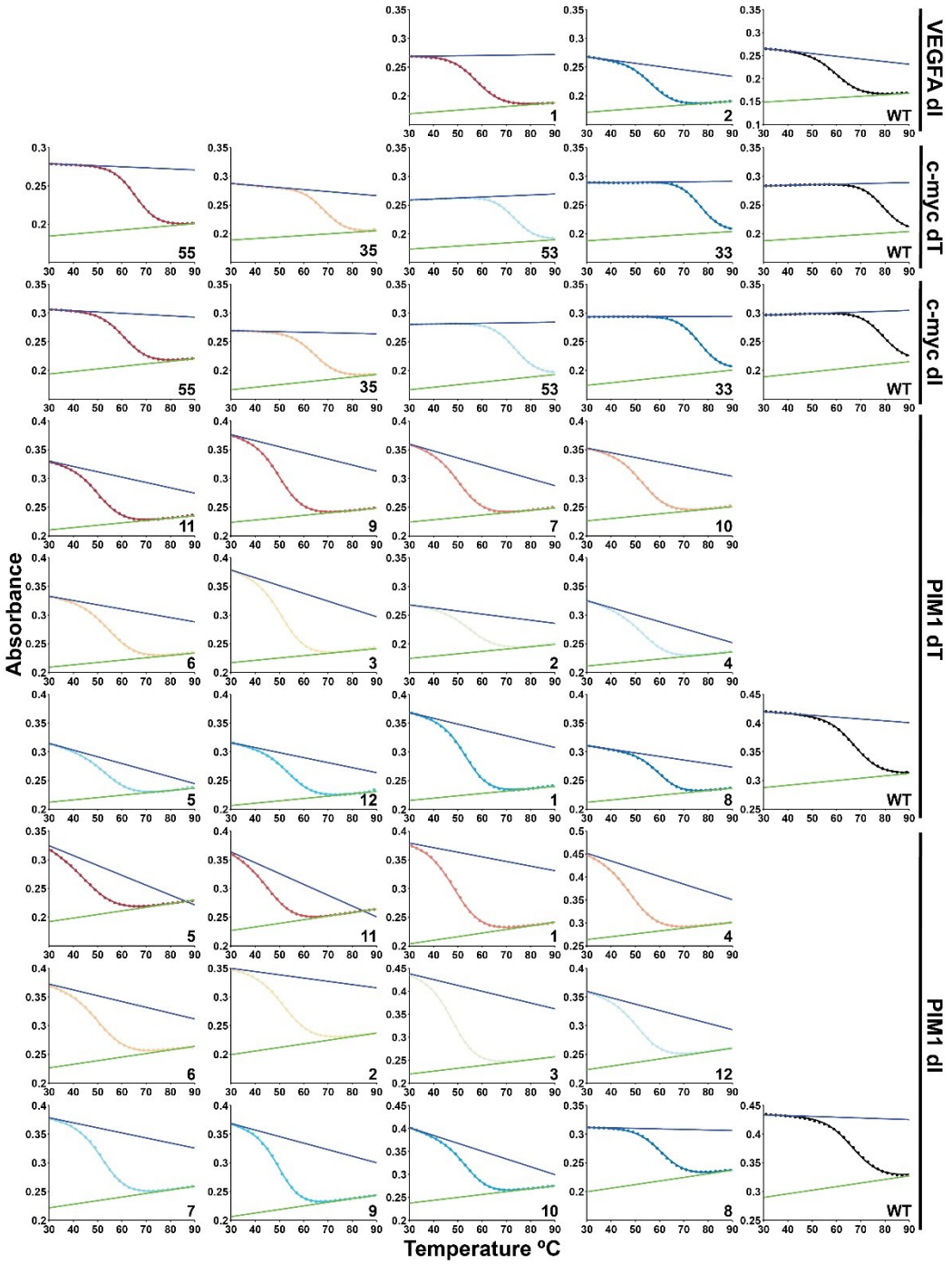
Supplementary Figure 2.9. Correlation of global fit parameters from 260 and 295 nm datasets. (a,b) The c-myc Pu18 trapped mutant populations at the T_m and T_ms at both wavelengths are strongly correlated (R=1.00 and 1.00 respectively). (c,d) The PIM1 trapped mutant populations at the T_m and T_ms at both wavelengths are strongly correlated (R=0.99 and 0.92 respectively). The dI trapped mutant data are shown here for both c-myc Pu18 and PIM1.



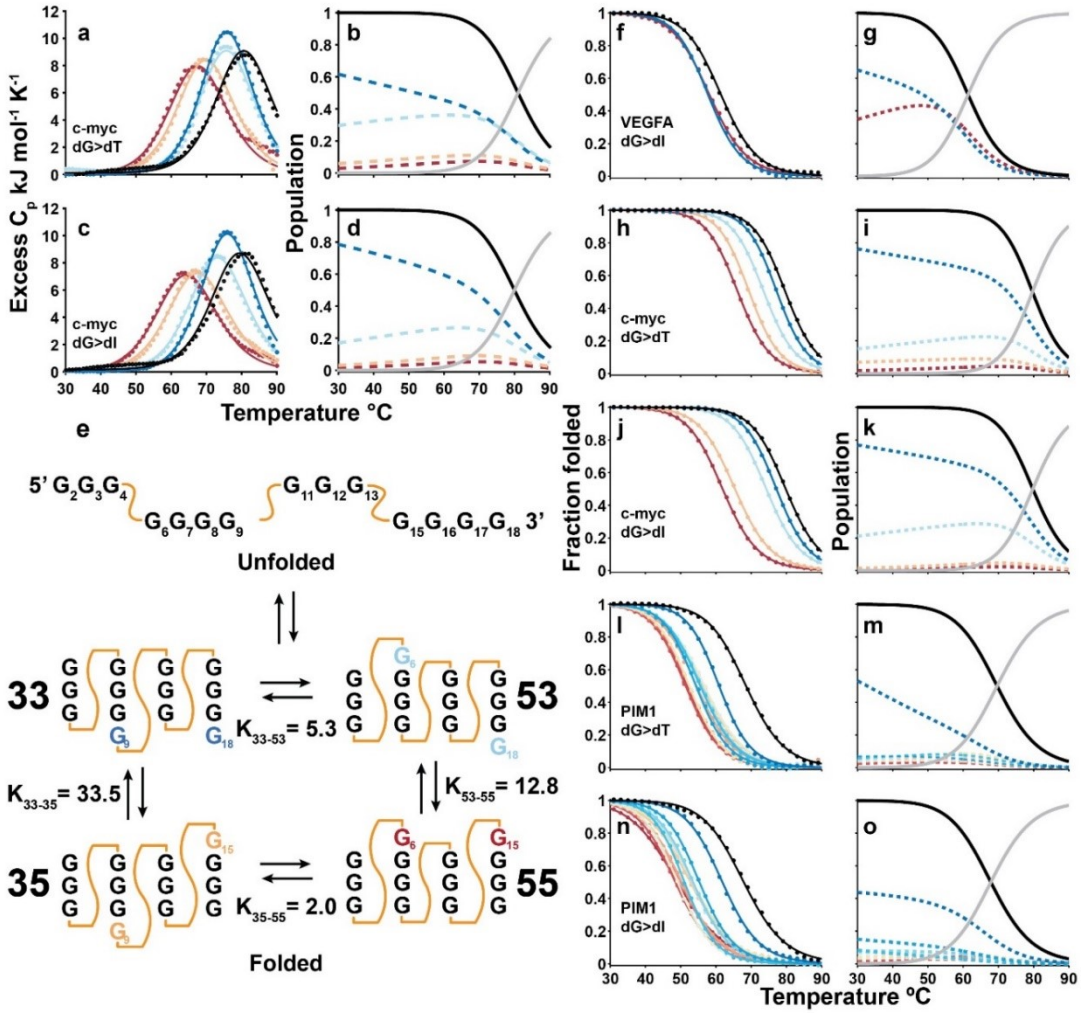
Supplementary Figure 2.10. Dual-wavelength absorbance melting for PIM1 trapped mutants. Absorbance data at 260 and 295 nm are shown as blue and red curves respectively. Raw absorbance data at 260 nm have been vertically scaled and offset, as above, for comparison with the 295 nm data. Trapped dI mutant numbers are indicated in the top right corners. Experiments were performed using 5 μM strand concentrations.



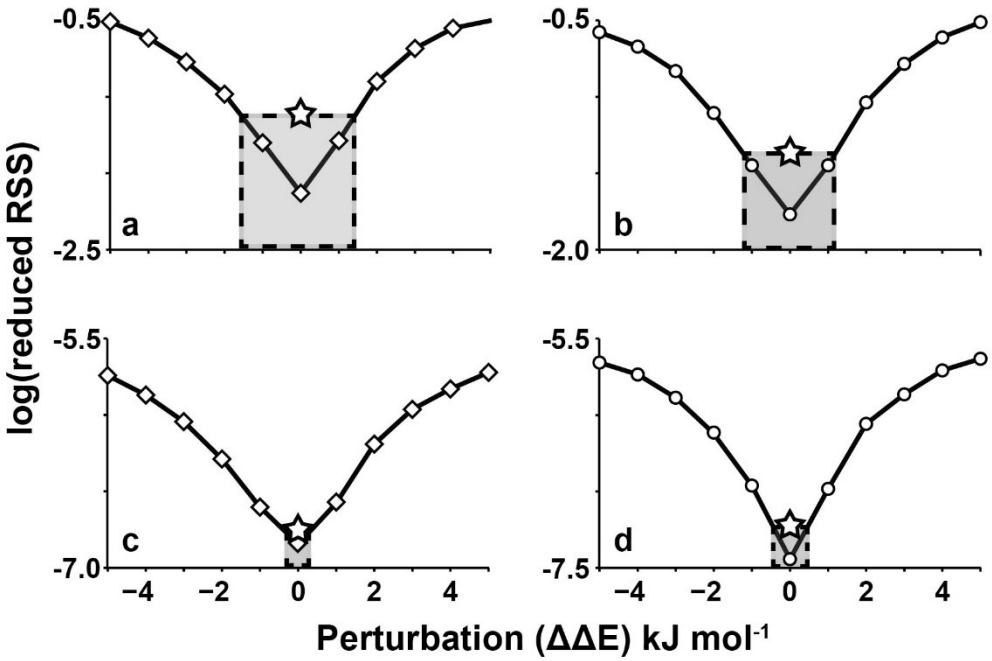
Supplementary Figure 2.11. Very slow timescale GR exchange would produce a thermal downshift. In the case of GR exchange that occurs slowly compared to the scan rate of the DSC, the thermogram of the wild-type GQ would be the population-weighted average of the individual thermograms of the GR isomers. To visualize this, the population-weighted average of the trapped mutant thermograms (c-myc Pu18 dG>dI) is shown by the purple dashed line (P_{33}=0.810, P_{53}=0.154, P_{35}=0.024, P_{55}=0.012 at 25 °C). The colored points correspond to the thermograms of the trapped mutants, and the black points correspond to the data for the wild-type GQ.



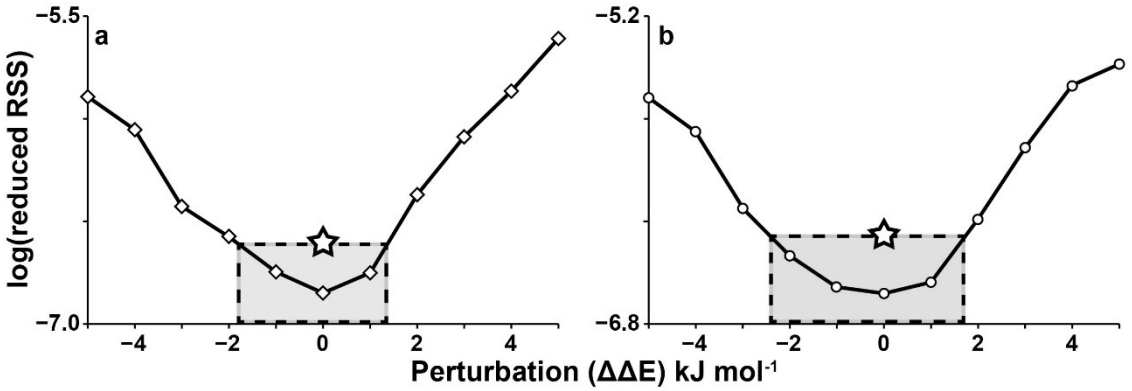
Supplementary Figure 2.12. Raw absorbance melting curves of the wild-type and trapped mutant c-myc Pu18, VEGFA, and PIM1 GQs. Color coding matches that of Figure 2.2 and Figure 2.3. The fitted folded (blue line) and unfolded (green line) baselines for each melt are indicated.



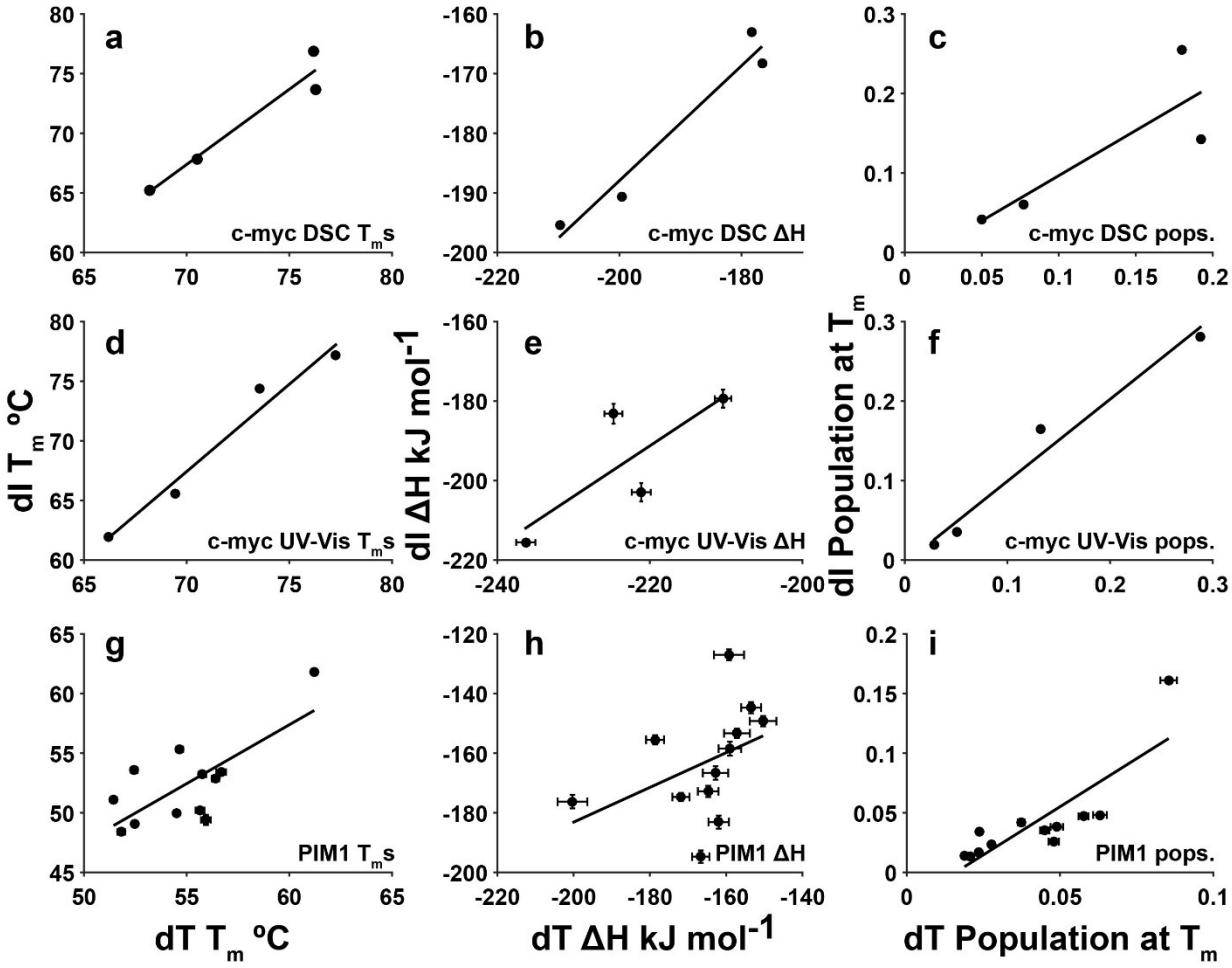
Supplementary Figure 2.13. Global fits of GQ thermal denaturation data. Thermograms (points) and global best-fit lines (lines) obtained by DSC (a,c) and UV-Vis spectroscopy (f,h,j,l,n) for a complete set of dG>dT (a,h,l) and dG>dI (c,f,j,n) trapped mutant and wild-type GQs for c-myc Pu18 (a,c,h,j), VEGFA (f), and PIM1 (l,n). The GR isomer populations extracted from fits are plotted in the panels immediately to the right of the thermograms (b,d,g,i,k,m,o). The color scheme relates to the GR isomer population order extracted from the global fits, where dark blue to dark red indicates most to least populated GR isomer respectively. Black corresponds to data for the wild-type GQ in (a,c,f,h,j,l,n) and the sum of all folded isomer populations in (b,d,g,i,k,m,o). In (b,d,g,i,k,m,o), grey curves correspond to the population of the unfolded state. (e) Cartoon of the c-myc Pu18 GQ undergoing exchange between 4 folded GR isomers and the unfolded state. $K_{X\text{-}Y}=P_X/P_Y$ is the equilibrium constant for isomers X and Y, shown as the values obtained from the trapped dI mutant fits. Errors are shown in Supplementary Table 2.5.



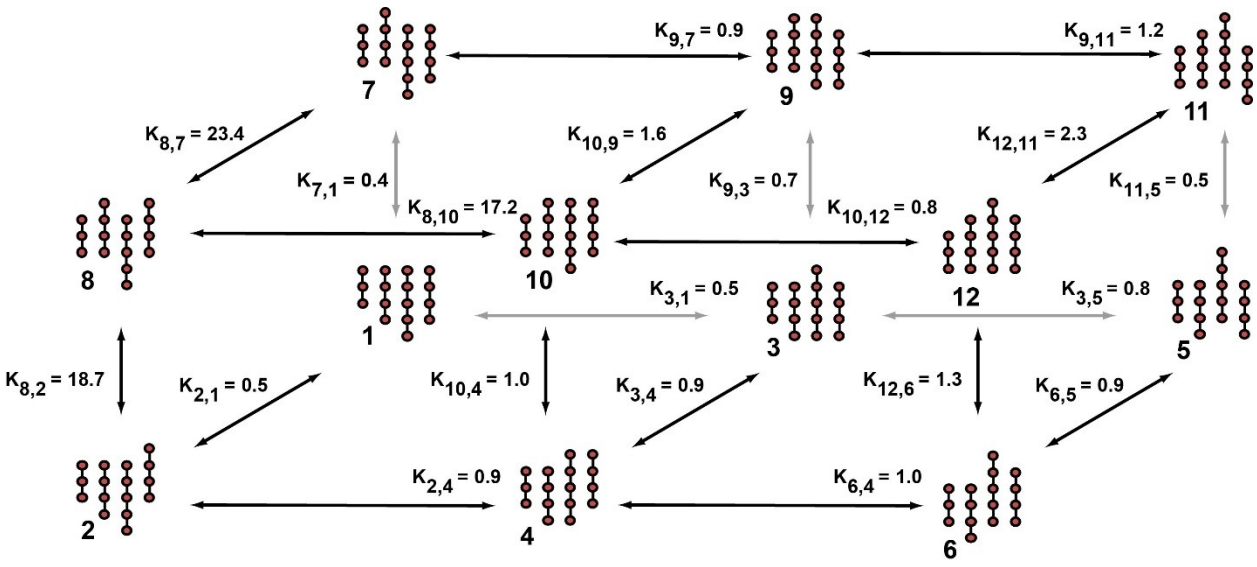
Supplementary Figure 2.14. Sensitivity of the c-myc Pu18 global fit to thermodynamic perturbations. Plot of average reduced RSS (RSS/DF) obtained for global fits of simulated c-myc Pu18 GQ (a,b) DSC and (c,d) UV-Vis data in which the folding ΔH and $T_0\Delta S$ of trapped mutants differ from those of the corresponding wild-type GR isomer by random perturbations with a mean of zero and standard deviation of $\Delta\Delta E$ kJ mol^{-1} (1000 iterations). The reduced RSS is the residual sum of squared difference between simulated data and the fits, divided by the number of degrees of freedom of the fit. The stars correspond to the experimental reduced RSS_{exp} values, while the set of $\Delta\Delta E$ giving $\langle RSS \rangle \leq RSS_{exp}$ gives the ranges of thermodynamic perturbations consistent with our data. Simulations for the dG>dT and dG>dI trapped mutant datasets are in panels (a,c) and (b,d), respectively.



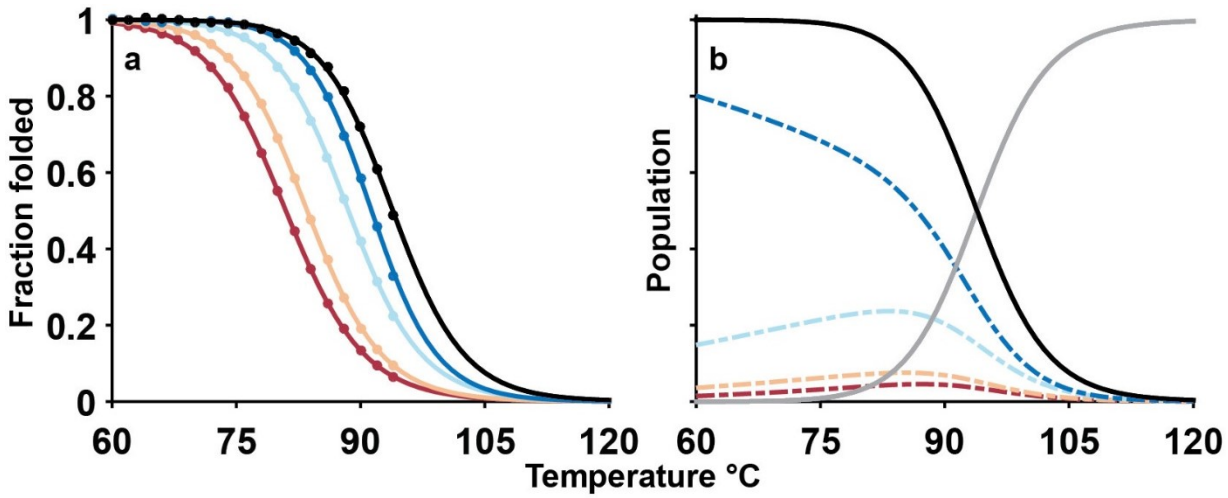
Supplementary Figure 2.15. Sensitivity of the PIM1 global fit to thermodynamic perturbations. Plot of average reduced RSS (RSS/DF) obtained for global fits of simulated PIM1 GQ UV-Vis data in which the folding ΔH and $T_0\Delta S$ of trapped mutants differ from those of the corresponding wild-type GR isomer by random perturbations with a mean of zero and standard deviation of $\Delta\Delta E$ kJ mol^{-1} (25 iterations). The stars correspond to the experimental reduced RSS_{exp} values, while the set of $\Delta\Delta E$ giving $\langle RSS \rangle \leq RSS_{exp}$ gives the ranges of thermodynamic perturbations consistent with our data. Simulations for the dG>dT (a) or dG>dI (b) trapped mutant datasets are shown in (a) and (b), respectively.



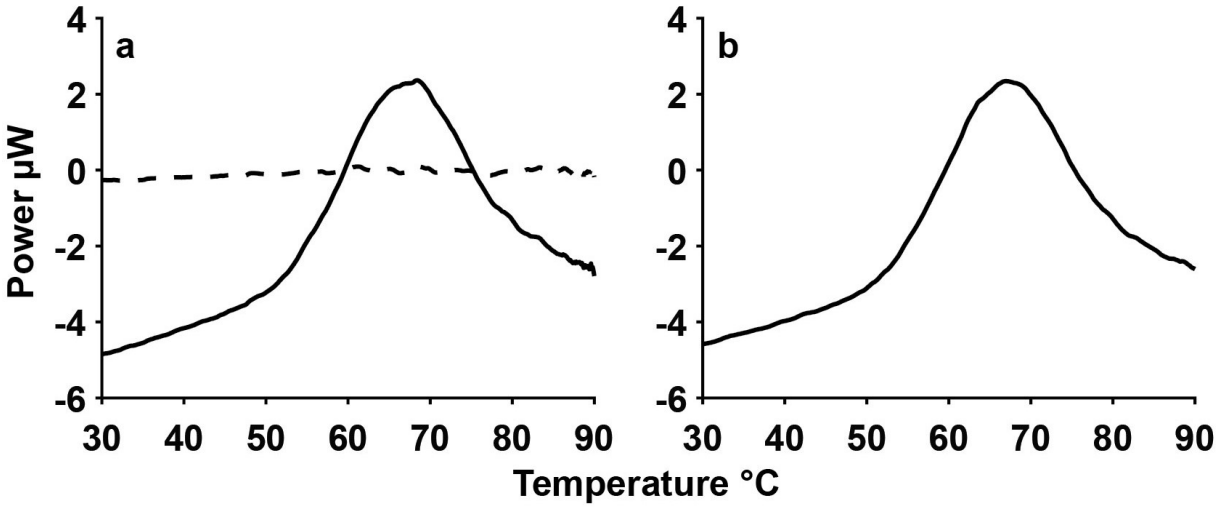
Supplementary Figure 2.16. Correlation of folding parameters extracted from global fits of dT and dI trapped mutants of the c-myc Pu18 and PIM1 GQs. Thermodynamic parameters from the global fits of c-myc Pu18 DSC data are strongly correlated (R=0.98, 0.99, 0.90 in panels a-c). Thermodynamic parameters from global fits of the c-myc Pu18 UV-Vis data are similarly well-correlated (R=0.99, 0.80, 0.99 in panels d-f). The parameters extracted from PIM1 UV-Vis data are reasonably well-correlated (R=0.73, 0.42, 0.83 in panels g-i). Errors are smaller than the symbols in some plots.



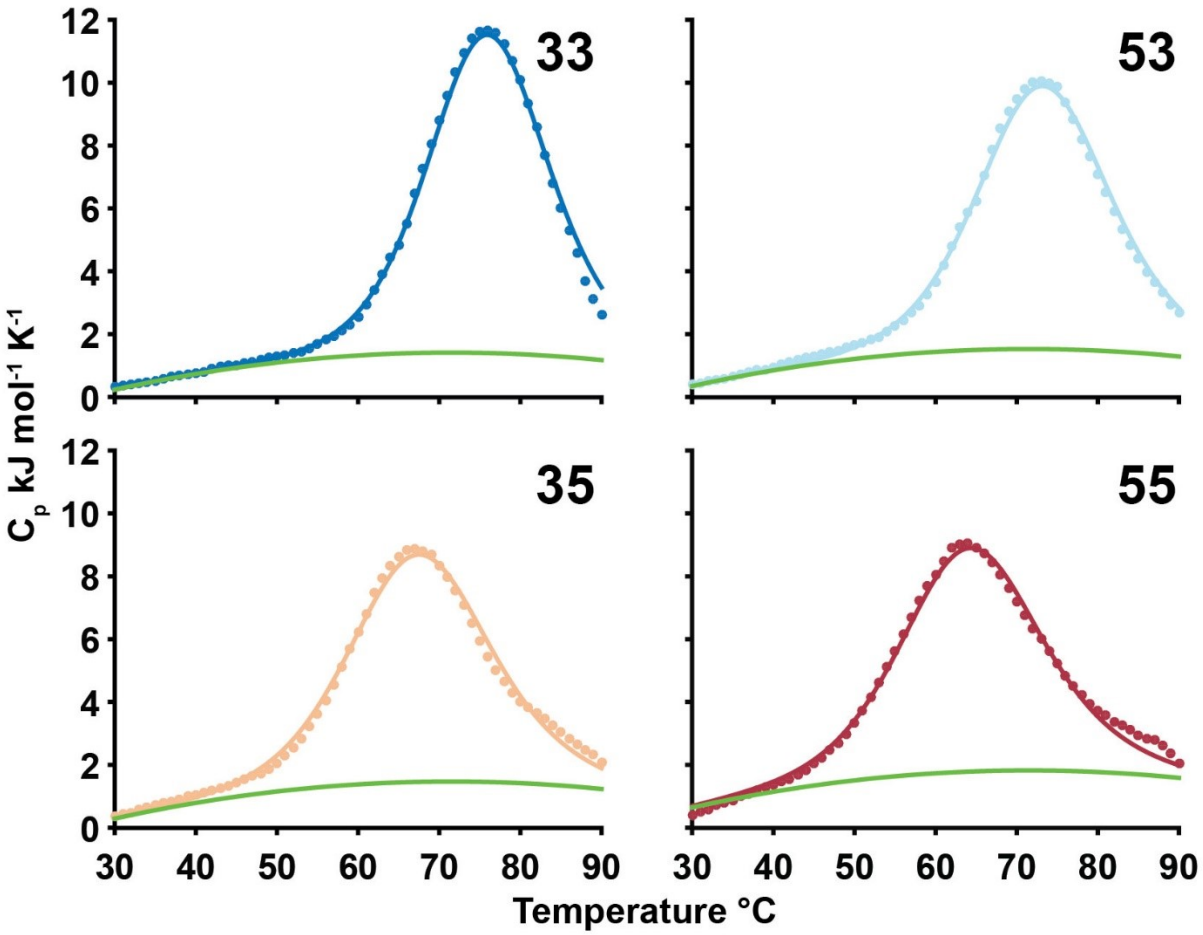
Supplementary Figure 2.17. Coupled GR exchange in the PIM1 GQ. dG residues are depicted as filled red circles and loop residues have been omitted for clarity. K_{ex} were calculated as the ratios of GR isomer populations. For example, $K_{X,Y}$ = [X]/[Y].



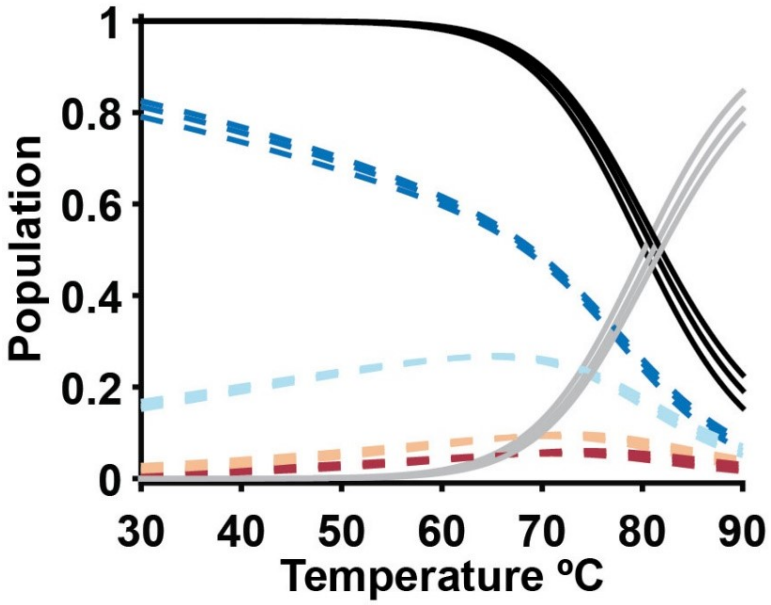
Supplementary Figure 2.18. UV-Vis thermal denaturation data for c-myc Pu18 wild-type and dT trapped mutant GQs. (a) Fraction of the folded state for the wild-type GQ and trapped mutants and (b) GR isomer populations extracted from a global fit of UV-Vis absorbance spectrophotometric data obtained with 130 mM K^+. The thermodynamic stabilities of the trapped mutants and the populations of the corresponding GR isomers rank in the same order as they do in the presence of lower K^+ concentrations. In (a), experimental data (points) and best fits (curves) are black (wild-type) and colored (trapped mutants). In (b), populations of the GR isomers are indicated by are colored dashed curves, the sum of folded isomer populations is indicated by the black curve, and population of the unfolded state is shown as a grey curve. Note that data could only be collected to 95 ºC. The extension of the curves in (a) and (b) to higher temperatures represents an experimentally inaccessible extrapolation.



Supplementary Figure 2.19. DSC buffer baseline subtraction. In (a) the buffer baseline (black dashed line) is subtracted from the sample scan (c-myc Pu18 35 dG>dI trapped mutant, black line) to yield the buffer subtracted sample curve in (b)**.**



Supplementary Figure 2.20. Heat capacity curves from global analysis of DSC data. Experimental and fitted data are shown as filled colored circles and colored lines respectively. The second-order polynomial baseline is shown as green solid line. The four c-myc Pu18 dI trapped mutant thermograms are shown in order of decreasing stability.



Supplementary Figure 2.21. Effect of ΔC_p on global fit populations. Global fits were performed with ΔC_p=0.24 (fit), 1.3 (telomere GQ), and 2.1 kJ mol^{-1} K^{-1} respectively. Populations of the GR isomers are indicated by colored dashed curves, the sum of folded isomer populations are indicated by the black curves, and populations of the unfolded state are shown as grey curves.

2.9. Supplementary Tables

Supplementary Table 2.1. Effect of ΔC_p on DSC global fit thermodynamics**.** Errors were calculated according to the variance-covariance method (Section 2.7.12).

Sequence[a]	ΔC_p (J mol^{-1} K^{-1})	ΔH^a (kJ mol^{-1})	ΔS^a (J mol^{-1} K^{-1})
c-myc Pu18 55	240[b]	-165.0±0.1	-488.1±0.5
	1300[c]	-173.0±0.2	-510.4±0.5
	2100	-182.1±0.2	-536.7±0.6
c-myc Pu18 35	240[b]	-167.9±0.1	-492.3±0.4
	1300[c]	-176.7±0.2	-516.8±0.5
	2100	-186.7±0.2	-545.6±0.5
c-myc Pu18 53	240[b]	-183.5±0.2	-529.3±0.4
	1300[c]	-192.9±0.2	-555.4±0.5
	2100	-204.0±0.2	-586.9±0.5
c-myc Pu18 33	240[b]	-203.1±0.2	-581.4±0.4
	1300[c]	-211.9±0.2	-605.6±0.5
	2100	-222.5±0.2	-635.7±0.5

[a]Global fit results using dI trapped mutants. ΔH and ΔS at T_m of each trapped mutant.

[b]ΔC_p optimized in global fit.

[c]ΔC_p reported for human telomere GQ[37].

Supplementary Table 2.2. Thermodynamic parameters obtained from two-state models and model-free analysis. Errors were calculated using the variance covariance method (global fit) or as the standard deviations of 1×10^4 Monte Carlo iterations (calorimetric and van 't Hoff parameters).

Sequence	c-myc Pu18 33 dG>dI	c-myc Pu18 53 dG>dI	c-myc Pu18 35 dG>dI	c-myc Pu18 55 dG>dI
$\Delta H^{global,a}$ (kJ mol^{-1})	-202.5±0.1	-182.7±0.2	-167.1±0.2	-164.1±0.2
$\Delta H^{cal,b}$ (kJ mol^{-1})	-207.2±3.2	-186.9±3.0	-162.6±2.8	-161.4±2.8
$\Delta H^{VH,c}$ (kJ mol^{-1})	-207.8±0.3	-186.3±0.3	-169.7±0.4	-165.8±0.5
$\Delta S^{global,d}$ (J mol^{-1} K^{-1})	-580.0±1.8	-527.3±1.8	-490.2±1.7	-486.1±1.7
$\Delta S^{cal,e}$ (J mol^{-1} K^{-1})	-605.6±9.8	-551.3±9.1	-477.9±8.0	-476.8±8.2
$\Delta S^{VH,f}$ (J mol^{-1} K^{-1})	-595.3±0.7	-538.1±0.9	-498.1±1.0	-490.7±1.5
$\Delta H^{cal}/\Delta H^{VH}$	1.00±0.02	1.00±0.02	0.96±0.02	0.97±0.02
$\Delta S^{cal}/\Delta S^{VH}$	1.02±0.02	1.03±0.02	0.96±0.02	0.97±0.02

[a]Enthalpy of folding extracted from the global analysis of DSC data.

[b]Calorimetric enthalpy calculated as area under the excess C_p curve. For the more stable 33 and 53 mutants, melting is incomplete at the maximum temperature used, thus ΔH^{cal} was calculated as twice the area under the lower-T half of the C_p curve, calculated from 25 °C to the T_m.

[c]Van 't Hoff enthalpy calculated from the slope of the progress excess C_p curve[89].

[d]Entropy of folding extracted from the global analysis of DSC data.

[e]Calorimetric entropy calculated as area under the excess C_p/T curve. For the 33 and 53 mutants, ΔS^{cal} was calculated as twice the area under the lower-T half of the C_p/T curve, calculated from 25 °C to the T_m.

[f]Van 't Hoff entropy calculated from the slope of the progress excess C_p/T curve[89].

Supplementary Table 2.3. Thermodynamic parameters from the DSC global fitting of the c-myc Pu18 GQ. Errors were calculated as stated in Section 2.7.12.

Sequence	**dT** ΔH	**dI** ΔH	**dT** ΔS	**dI** ΔS	**dT** T_m	**dI** T_m
	kJ mol^{-1}	kJ mol^{-1}	J mol^{-1} K^{-1}	J mol^{-1} K^{-1}	°C	°C
c-myc Pu18 55	-175.3±0.2	-164.1±0.2	-514.3±0.6	-486.1±1.7	67.90±0.02	64.70±0.02
c-myc Pu18 35	-181.2±0.2	-167.1±0.2	-528.0±0.4	-490.2±1.7	70.20±0.02	67.90±0.02
c-myc Pu18 53	-191.9±0.2	-182.7±0.2	-550.0±0.4	-527.3±1.8	76.00±0.03	73.50±0.02
c-myc Pu18 33	-205.9±0.2	-202.5±0.1	-590.0±0.5	-580.0±1.8	76.00±0.01	76.20±0.02

Supplementary Table 2.4. Thermodynamic parameters extracted from the global fit of UV-Vis data for the c-myc Pu18, VEGFA, and PIM1 GQs. Errors were calculated as stated in Section 2.7.12.

Sequence	**dT** ΔH	**dI** ΔH	**dT** ΔS	**dI** ΔS	**dT** T_m	**dI** T_m
	kJ mol^{-1}	kJ mol^{-1}	J mol^{-1} K^{-1}	J mol^{-1} K^{-1}	°C	°C
c-myc Pu18 55	-210.4±2.3	-179.4±1.1	-619.9±6.4	-539.6±2.9	66.2±0.1	62.00±0.04
c-myc Pu18 35	-224.8±2.5	-183.2±1.2	-656.2±6.9	-543.1±3.2	69.4±0.1	65.60±0.04
c-myc Pu18 53	-221.1±2.3	-202.9±1.2	-638.0±6.4	-588.0±3.1	73.6±0.1	74.40±0.04
c-myc Pu18 33	-236.3±2.1	-215.6±1.3	-674.6±5.9	-616.0±3.5	77.3±0.1	77.20±0.03
VEGFA-1	-	-173.2±2.2	-	-520.7±5.8	-	58.1±0.1
VEGFA-2	-	-193.3±2.9	-	-584.8±7.8	-	57.7±0.1
PIM1-1	-178.6±2.4	-155.5±1.5	-547.8±7.2	-482.4±4.5	54.5±0.1	50.0±0.1
PIM1-2	-150.3±3.5	-149.2±1.7	-455.8±10.3	-457.2±4.9	56.7±0.2	53.4±0.2
PIM1-3	-171.9±2.3	-174.7±1.4	-530.1±6.9	-542.5±4.2	52.5±0.1	49.1±0.1
PIM1-4	-157.2±3.4	-153.3±1.5	-478.0±10.1	-475.3±4.4	55.7±0.2	50.2±0.2
PIM1-5	-159.3±4.0	-127.0±1.8	-492.5±11.9	-392.5±5.1	55.9±0.2	49.4±0.4
PIM1-6	-153.5±2.6	-144.7±1.9	-461.8±7.6	-444.6±5.6	56.4±0.2	52.9±0.2
PIM1-7	-164.7±2.7	-172.8±1.9	-510.5±8.1	-528.5±5.7	52.4±0.1	53.6±0.1
PIM1-8	-200.3±3.9	-176.2±2.3	-596.0±11.5	-526.8±6.7	61.2±0.1	61.8±0.1
PIM1-9	-166.7±2.3	-194.7±2.1	-513.7±6.8	-599.8±6.3	51.4±0.2	51.1±0.1
PIM1-10	-162.0±2.7	-183.1±2.2	-489.7±7.9	-558.2±6.5	54.7±0.2	55.3±0.1
PIM1-11	-159.0±3.0	-158.5±2.3	-488.4±8.9	-492.6±6.7	51.8±0.2	48.4±0.3
PIM1-12	-162.8±3.3	-166.6±2.3	-494.7±9.3	-511.2±6.7	55.8±0.5	53.3±0.2

Supplementary Table 2.5. GR exchange equilibrium constants for the c-myc Pu18 dT and dI trapped mutants calculated from the DSC and UV-Vis global fitting parameters. The exchange equilibria follow the cycle given in the main text. K_{ex} were calculated at 25 °C from population ratios extracted from the DSC and UV-Vis global fitting, e.g. $K_{33\text{-}53}=P_{33}/P_{53}$. All K_{ex} have been expressed as >1 for ease of comparison. Errors are smaller than one decimal place for certain K_{ex}. Errors were calculated as stated in Section 2.7.12.

Equilibrium	**DSC dT K_{ex}**	**DSC dI K_{ex}**	**UV-Vis dT K_{ex}**	**UV-Vis dI K_{ex}**
33-53	2.3±0.0	5.3±0.1	5.6±1.1	3.8±0.4
53-55	11.2±0.2	12.8±0.3	9.0±1.5	38.5±3.3
35-55	2.1±0.0	2.0±0.0	4.3±0.7	2.4±0.2
33-35	12.4±0.2	33.5±0.4	11.7±2.3	60.6±5.2

Supplementary Table 2.6. GR exchange equilibrium constants for the PIM1 dT and dI trapped mutants extracted from the extracted UV-Vis global fits. K_{ex} were calculated as the ratios of GR isomer populations at 25 °C. For example, $K_{X,Y}$ = [X]/[Y]. Errors were calculated as stated in Section 2.7.12.

Equilibrium	dT K_{ex}	dI K_{ex}
8,2	18.7±2.7	13.2±1.9
2,4	0.9±0.1	1.5±0.2
10,4	1.0±0.1	7.4±0.9
8,10	17.2±2.3	2.7±0.4
6,4	1.0±0.1	1.2±0.1
12,6	1.3±0.2	2.3±0.3
10,12	0.8±0.1	2.7±0.4
7,1	0.4±0.0	3.5±0.4
3,1	0.5±0.1	1.5±0.1
9,3	0.7±0.1	2.9±0.3
9,7	0.9±0.1	1.3±0.2
3,5	0.8±0.1	4.0±0.5
11,5	0.5±0.1	2.2±0.3
9,11	1.2±0.1	5.3±0.7
2,1	0.5±0.1	1.5±0.2
3,4	0.9±0.1	1.6±0.1
6,5	0.9±0.1	3.1±0.4
8,7	23.4±2.7	5.6±0.8
10,9	1.6±0.1	1.6±0.2
12,11	2.3±0.3	3.2±0.5

2.10. References

1. Bochman, M.L., Paeschke, K. & Zakian, V.A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-780 (2012).
2. Hu, L., Lim, K.W., Bouaziz, S. & Phan, A.T. Giardia telomeric sequence d(TAGGG)4 forms two intramolecular G-quadruplexes in K+ solution: effect of loop length and sequence on the folding topology. *J Am Chem Soc* **131**, 16824-16831 (2009).
3. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res* **34**, 5402-5415 (2006).
4. Chaires, J.B. et al. An improved model for the hTERT promoter quadruplex. *PLoS One* **9**, e115580 (2014).
5. Paeschke, K. et al. Telomerase recruitment by the telomere end binding protein-beta facilitates G-quadruplex DNA unfolding in ciliates. *Nat Struct Mol Biol* **15**, 598-604 (2008).
6. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. & Hurley, L.H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* **99**, 11593-11598 (2002).
7. Murat, P. et al. G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. *Nat. Chem. Biol.* **10**, 358-364 (2014).
8. Arora, A. & Suess, B. An RNA G-quadruplex in the 3' UTR of the proto-oncogene PIM1 represses translation. *RNA Biol* **8**, 802-805 (2011).
9. Pennarun, G. et al. Apoptosis related to telomere instability and cell cycle alterations in human glioma cells treated by new highly selective G-quadruplex ligands. *Oncogene* **24**, 2917-2928 (2005).
10. Endoh, T., Kawasaki, Y. & Sugimoto, N. Suppression of gene expression by G-quadruplexes in open reading frames depends on G-quadruplex stability. *Angew Chem Int Ed Engl* **52**, 5522-5526 (2013).
11. Endoh, T., Kawasaki, Y. & Sugimoto, N. Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor alpha. *Nucleic Acids Res.* **41**, 6222-6231 (2013).
12. Rachwal, P.A., Brown, T. & Fox, K.R. Effect of G-tract length on the topology and stability of intramolecular DNA quadruplexes. *Biochemistry* **46**, 3036-3044 (2007).
13. Pandey, S., Agarwala, P. & Maiti, S. Effect of loops and G-quartets on the stability of RNA G-quadruplexes. *J Phys Chem B* **117**, 6896-6905 (2013).
14. Guedin, A., Gros, J., Alberti, P. & Mergny, J.L. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res* **38**, 7858-7868 (2010).
15. Phan, A.T., Kuryavyi, V., Burge, S., Neidle, S. & Patel, D.J. Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Am. Chem. Soc.* **129**, 4386-4392 (2007).

16. Agrawal, P., Hatzakis, E., Guo, K., Carver, M. & Yang, D. Solution structure of the major G-quadruplex formed in the human VEGF promoter in K+: insights into loop interactions of the parallel G-quadruplexes. *Nucleic Acids Res* **41**, 10584-10592 (2013).
17. Yang, D. & Hurley, L.H. Structure of the biologically relevant G-quadruplex in the c-MYC promoter. *Nucleosides Nucleotides Nucleic Acids* **25**, 951-968 (2006).
18. Agrawal, P., Lin, C., Mathad, R.I., Carver, M. & Yang, D. The major G-quadruplex formed in the human BCL-2 proximal promoter adopts a parallel structure with a 13-nt loop in K+ solution. *J Am Chem Soc* **136**, 1750-1753 (2014).
19. Phan, A.T., Modi, Y.S. & Patel, D.J. Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter. *J. Am. Chem. Soc.* **126**, 8710-8716 (2004).
20. Seenisamy, J. et al. The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. *J. Am. Chem. Soc.* **126**, 8702-8709 (2004).
21. Ambrus, A., Chen, D., Dai, J., Jones, R.A. & Yang, D. Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry* **44**, 2048-2058 (2005).
22. Hatzakis, E., Okamoto, K. & Yang, D. Thermodynamic stability and folding kinetics of the major G-quadruplex and its loop isomers formed in the nuclease hypersensitive element in the human c-Myc promoter: effect of loops and flanking segments on the stability of parallel-stranded intramolecular G-quadruplexes. *Biochemistry* **49**, 9152-9160 (2010).
23. Smith, F.W. & Feigon, J. Strand orientation in the DNA quadruplex formed from the Oxytricha telomere repeat oligonucleotide d(G4T4G4) in solution. *Biochemistry* **32**, 8682-8692 (1993).
24. Adrian, M., Heddi, B. & Phan, A.T. NMR spectroscopy of G-quadruplexes. *Methods* **57**, 11-24 (2012).
25. Lim, K.W. et al. Coexistence of two distinct G-quadruplex conformations in the hTERT promoter. *J. Am. Chem. Soc.* **132**, 12331-12342 (2010).
26. Zuckerman, D.M. Statistical Physics of Biomolecules: An Introduction. (CRC Press, 2010).
27. Kettani, A. et al. A dimeric DNA interface stabilized by stacked A.(G.G.G.G).A hexads and coordinated monovalent cations. *J Mol Biol* **297**, 627-644 (2000).
28. Zhang, N. et al. V-shaped scaffold: a new architectural motif identified in an A x (G x G x G x G) pentad-containing dimeric DNA quadruplex involving stacked G(anti) x G(anti) x G(anti) x G(syn) tetrads. *J Mol Biol* **311**, 1063-1079 (2001).
29. Kumar, N. & Maiti, S. A thermodynamic overview of naturally occurring intramolecular DNA quadruplexes. *Nucleic Acids Res* **36**, 5610-5622 (2008).
30. Risitano, A. & Fox, K.R. Inosine substitutions demonstrate that intramolecular DNA quadruplexes adopt different conformations in the presence of sodium and potassium. *Bioorg Med Chem Lett* **15**, 2047-2050 (2005).
31. Szatylowicz, H. & Sadlej-Sosnowska, N. Characterizing the strength of individual hydrogen bonds in DNA base pairs. *J Chem Inf Model* **50**, 2151-2161 (2010).

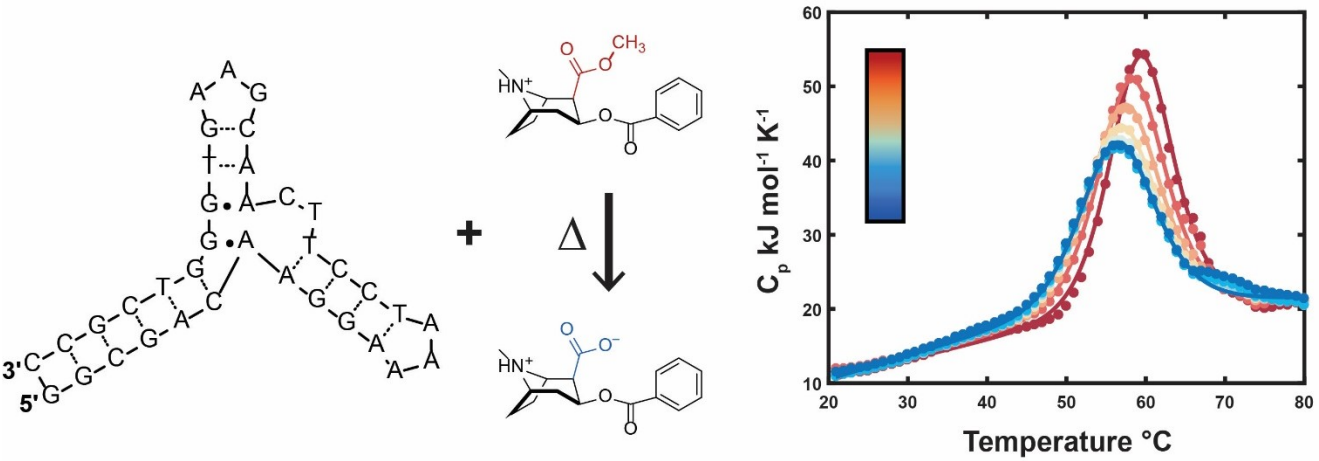
32. Karsisiotis, A.I. et al. Topological characterization of nucleic acid G-quadruplexes by UV absorption and circular dichroism. *Angew Chem Int Ed Engl* **50**, 10645-10648 (2011).
33. Hsu, S.T. et al. A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics. *J. Am. Chem. Soc.* **131**, 13399-13409 (2009).
34. Ambrus, A. et al. Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res.* **34**, 2723-2735 (2006).
35. Mergny, J.L. & Lacroix, L. UV Melting of G-Quadruplexes. *Curr Protoc Nucleic Acid Chem* **Chapter 17**, Unit 17 11 (2009).
36. Spink, C.H. The deconvolution of differential scanning calorimetry unfolding transitions. *Methods* **76**, 78-86 (2015).
37. Boncina, M., Lah, J., Prislan, I. & Vesnaver, G. Energetic basis of human telomeric DNA folding into G-quadruplex structures. *J. Am. Chem. Soc.* **134**, 9657-9663 (2012).
38. Gray, R.D., Buscaglia, R. & Chaires, J.B. Populated intermediates in the thermal unfolding of the human telomeric quadruplex. *J. Am. Chem. Soc.* **134**, 16834-16844 (2012).
39. Gray, R.D., Trent, J.O. & Chaires, J.B. Folding and unfolding pathways of the human telomeric G-quadruplex. *J. Mol. Biol.* **426**, 1629-1650 (2014).
40. Limongelli, V. et al. The G-triplex DNA. *Angew Chem Int Ed Engl* **52**, 2269-2273 (2013).
41. Mergny, J.L. & Lacroix, L. Analysis of thermal melting curves. *Oligonucleotides* **13**, 515-537 (2003).
42. Payet, L. & Huppert, J.L. Stability and structure of long intramolecular G-quadruplexes. *Biochemistry* **51**, 3154-3161 (2012).
43. Buscaglia, R., Gray, R.D. & Chaires, J.B. Thermodynamic characterization of human telomere quadruplex unfolding. *Biopolymers* **99**, 1006-1018 (2013).
44. Gray, R.D. & Chaires, J.B. Analysis of multidimensional G-quadruplex melting curves. *Curr Protoc Nucleic Acid Chem* **Chapter 17**, Unit17 14 (2011).
45. Gray, R.D. & Chaires, J.B. Kinetics and mechanism of K+- and Na+-induced folding of models of human telomeric DNA into G-quadruplex structures. *Nucleic Acids Res* **36**, 4191-4203 (2008).
46. Haq, I., Chowdhry, B.Z. & Jenkins, T.C. Calorimetric techniques in the study of high-order DNA-drug interactions. *Methods Enzymol* **340**, 109-149 (2001).
47. Wintrode, P.L., Griko, Y.V. & Privalov, P.L. Structural energetics of barstar studied by differential scanning microcalorimetry. *Protein Sci* **4**, 1528-1534 (1995).
48. Spink, C.H. Differential scanning calorimetry. *Methods Cell Biol* **84**, 115-141 (2008).
49. Freire, E., Biltonen, R. L. Statistical mechanical deconvolution of thermal transitions in macromolecules. I. Theory and application to homogeneous systems. *Biopolymers* **17**, 463-479 (1978).

50. Phan, A.T. & Patel, D.J. Two-repeat human telomeric d(TAGGGTTAGGGT) sequence forms interconverting parallel and antiparallel G-quadruplexes in solution: distinct topologies, thermodynamic properties, and folding/unfolding kinetics. *J Am Chem Soc* **125**, 15021-15027 (2003).
51. Fitter, J. A measure of conformational entropy change during thermal protein unfolding using neutron spectroscopy. *Biophys J* **84**, 3924-3930 (2003).
52. Cavin Perier, R., Junier, T. & Bucher, P. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res* **26**, 353-357 (1998).
53. Brooks, T.A. & Hurley, L.H. Targeting MYC Expression through G-Quadruplexes. *Genes Cancer* **1**, 641-649 (2010).
54. Zhou, W. et al. Possible regulatory roles of promoter g-quadruplexes in cardiac function-related genes - human TnIc as a model. *PLoS One* **8**, e53137 (2013).
55. Balasubramanian, S., Hurley, L.H. & Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat Rev Drug Discov* **10**, 261-275 (2011).
56. Li, X.M. et al. Guanine-vacancy-bearing G-quadruplexes responsive to guanine derivatives. *Proc Natl Acad Sci U S A* **112**, 14581-14586 (2015).
57. Ou, T.M. et al. G-quadruplexes: targets in anticancer drug design. *ChemMedChem* **3**, 690-713 (2008).
58. Olsen, C.M., Gmeiner, W.H. & Marky, L.A. Unfolding of G-quadruplexes: energetic, and ion and water contributions of G-quartet stacking. *J Phys Chem B* **110**, 6962-6969 (2006).
59. Lim, K.W. et al. Structure of the human telomere in K+ solution: a stable basket-type G-quadruplex with only two G-tetrad layers. *J Am Chem Soc* **131**, 4301-4309 (2009).
60. Motlagh, H.N., Wrabl, J.O., Li, J. & Hilser, V.J. The ensemble nature of allostery. *Nature* **508**, 331-339 (2014).
61. Miyoshi, D., Nakao, A. & Sugimoto, N. Structural transition from antiparallel to parallel G-quadruplex of d(G4T4G4) induced by Ca2+. *Nucleic Acids Res* **31**, 1156-1163 (2003).
62. Brazda, V., Haronikova, L., Liao, J.C. & Fojta, M. DNA and RNA quadruplex-binding proteins. *Int J Mol Sci* **15**, 17493-17517 (2014).
63. Dai, J., Carver, M., Hurley, L.H. & Yang, D. Solution structure of a 2:1 quindoline-c-MYC G-quadruplex: insights into G-quadruplex-interactive small molecule drug design. *J Am Chem Soc* **133**, 17673-17680 (2011).
64. Moye, A.L. et al. Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat Commun* **6**, 7643 (2015).
65. Moore, C.C. Ergodic theorem, ergodic theory, and statistical mechanics. *Proc Natl Acad Sci U S A* **112**, 1907-1911 (2015).
66. Broxson, C., Beckett, J. & Tornaletti, S. Transcription arrest by a G quadruplex forming-trinucleotide repeat sequence from the human c-myb gene. *Biochemistry* **50**, 4162-4172 (2011).
67. Nudler, E. Flipping riboswitches. *Cell* **126**, 19-22 (2006).

68. Narayan, S. et al. Site-specific fluorescence dynamics in an RNA 'thermometer' reveals the role of ribosome binding in its temperature-sensitive switch function. *Nucleic Acids Res* **43**, 493-503 (2015).
69. Lee, T.H. et al. Measuring the folding transition time of single RNA molecules. *Biophys J* **92**, 3275-3283 (2007).
70. Stone, M.D. et al. Stepwise protein-mediated RNA folding directs assembly of telomerase ribonucleoprotein. *Nature* **446**, 458-461 (2007).
71. Hogan, M.E. & Austin, R.H. Importance of DNA stiffness in protein-DNA binding specificity. *Nature* **329**, 263-266 (1987).
72. Bohnuud, T. et al. Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde. *Nucleic Acids Res.* **40**, 7644-7652 (2012).
73. Harris, S.A., Gavathiotis, E., Searle, M.S., Orozco, M. & Laughton, C.A. Cooperativity in drug-DNA recognition: a molecular dynamics study. *J Am Chem Soc* **123**, 12658-12663 (2001).
74. Tippana, R., Xiao, W. & Myong, S. G-quadruplex conformation and dynamics are determined by loop length and sequence. *Nucleic Acids Res* **42**, 8106-8114 (2014).
75. Stadlbauer, P., Krepl, M., Cheatham, T.E., 3rd, Koca, J. & Sponer, J. Structural dynamics of possible late-stage intermediates in folding of quadruplex DNA studied by molecular simulations. *Nucleic Acids Res.* **41**, 7128-7143 (2013).
76. Fleming, A.M., Zhou, J., Wallace, S.S. & Burrows, C.J. A Role for the Fifth G-Track in G-Quadruplex Forming Oncogene Promoter Sequences during Oxidative Stress: Do These "Spare Tires" Have an Evolved Function? *ACS Cent Sci* **1**, 226-233 (2015).
77. Reblova, K., Sponer, J. & Lankas, F. Structure and mechanical properties of the ribosomal L1 stalk three-way junction. *Nucleic Acids Res* **40**, 6290-6303 (2012).
78. Nikolova, E.N. et al. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470**, 498-502 (2011).
79. Heddi, B., Martin-Pintado, N., Serimbetov, Z., Kari, T.M. & Phan, A.T. G-quadruplexes with (4n - 1) guanines in the G-tetrad core: formation of a G-triad.water complex and implication for small-molecule binding. *Nucleic Acids Res* **44**, 910-916 (2016).
80. Kallansrud, G. & Ward, B. A comparison of measured and calculated single- and double-stranded oligodeoxynucleotide extinction coefficients. *Anal Biochem* **236**, 134-138 (1996).
81. Pagano, B. et al. Differential scanning calorimetry to investigate G-quadruplexes structural stability. *Methods* **64**, 43-51 (2013).
82. Dettler, J.M., Buscaglia, R., Le, V.H. & Lewis, E.A. DSC deconvolution of the structural complexity of c-MYC P1 promoter G-quadruplexes. *Biophys. J.* **100**, 1517-1525 (2011).
83. Majhi, P.R., Qi, J., Tang, C.F. & Shafer, R.H. Heat capacity changes associated with guanine quadruplex formation: an isothermal titration calorimetry study. *Biopolymers* **89**, 302-309 (2008).

84. Tellinghuisen, J. A Monte Carlo study of precision, bias, inconsistency, and non-Gaussian distributions in nonlinear least squares. *J. Phys. Chem. A* **104**, 2834-2844 (2000).
85. Huppert, J.L. & Balasubramanian, S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35**, 406-413 (2007).
86. Zhang, Z., Dai, J., Veliath, E., Jones, R.A. & Yang, D. Structure of a two-G-tetrad intramolecular G-quadruplex formed by a variant human telomeric sequence in K+ solution: insights into the interconversion of human telomeric G-quadruplex structures. *Nucleic Acids Res* **38**, 1009-1021 (2010).
87. Tellinghuisen, J. Statistical error propagation. *J. Phys. Chem. A* **105**, 3917-3921 (2001).
88. Kypr, J., Kejnovska, I., Renciuk, D. & Vorlickova, M. Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res* **37**, 1713-1725 (2009).
89. Saboury, A.A. & Moosavi-Movahedi, A.A. Clarification of calorimetric and van't hoff enthalpies for evaluation of protein transition states. *Biochemical Education* **22**, 210-211 (1994).

Chapter 3: Rapid characterization of biomolecular folding and binding interactions with thermolabile ligands by DSC



3.1. Preface

The previous chapter emphasized the functional effects of nucleic acid folding dynamics. Nucleic acid binding interactions are equally as important. Together, the folding and binding interactions of nucleic acids form the basis for their biological function and applications. For example, the rational design of aptamers for enhanced stability and sensitivity in drug detection is predicated on a quantitative understanding of aptamer folding and ligand binding. Yet, standard techniques for simultaneously analyzing folding and binding such as DSC are slow and low-throughput. In this chapter, the use of thermolabile ligands in DSC experiments combined with a new global fitting analysis is shown to drastically reduce the amount of time and sample required to obtain the physical parameters governing nucleic acid folding and ligand binding. The method is applied to two DNA aptamers commonly used to detect cocaine and quinine. Importantly, the binding parameters extracted from the global fitting analysis are in agreement with those derived from ITC. In addition, the approach can be used to extract information on the kinetics of the ligand conversion process. As an extension of the primary global fitting analysis given here, more complicated scenarios involving thermolabile ligands in DSC experiments are explored by computer simulation.

3.2. Abstract

DSC is a powerful technique for quantifying thermodynamic parameters governing biomolecular folding and binding interactions. This information is critical in the design of new pharmaceutical compounds. However, many pharmaceutically relevant ligands are chemically unstable at the high temperatures used in DSC analyses. Thus, measuring binding interactions is challenging because the concentrations of ligands and thermally-converted products are constantly

changing within the calorimeter cell. Here, we present a method using thermolabile ligands and DSC for rapidly obtaining thermodynamic and kinetic information on the folding, binding, and ligand conversion processes. We have applied our method to the DNA aptamers MN4 and MN19 that bind to the thermolabile ligand cocaine. Using a new global fitting analysis that accounts for thermolabile ligand conversion, the complete set of folding and binding parameters are obtained from a pair of DSC experiments. In addition, we show that the rate constant for thermolabile ligand conversion may be obtained with only one supplementary DSC dataset. The guidelines for identifying and analyzing data from several more complicated scenarios are presented, including irreversible aggregation of the biomolecule, slow folding, slow binding, and rapid depletion of the thermolabile ligand.

3.3. Introduction

DSC is a powerful method for quantitating biomolecular binding and folding interactions[1-3]. The strengths of DSC include its ability to elucidate binding and folding mechanisms, and to yield the corresponding thermodynamic parameters[2, 3]. Furthermore, DSC can be performed in solution under near-physiological conditions and does not require labeling of the biomolecule or ligand, e.g., with fluorophores, spin-labels or nuclear isotopes[4]. The instrument scans in temperature, measuring the amount of heat required to denature the biomolecule in the presence and absence of ligand. The resulting thermograms are used to extract the thermodynamic parameters governing the ligand binding and folding processes. The information provided by DSC or other thermodynamic techniques is critical to guiding the design of drugs targeting biomolecules[1, 5-8]. However, the repeated scanning to high temperatures (~60–100 °C) can be problematic. For example, many pharmaceutically important compounds undergo rearrangement

or decomposition upon sustained exposure to high temperatures[9-11], i.e., they are thermolabile. Examination of binding interactions by DSC typically requires multiple forward and reverse scans in order to verify the reproducibility of the thermogram for thermodynamic analyses[12]. Thermal conversion of an initial ligand to a secondary form with altered binding characteristics leads to pronounced differences in the shape and position of successive thermograms, since the concentration of the initial ligand decreases with each scan while the thermal conversion products accumulate. These datasets are not amenable to traditional analyses.

We have developed an experimental and global fitting method to generate and analyze thermolabile ligand DSC datasets that yield the complete set of thermodynamic parameters governing the biomolecular folding and binding interactions from a single ligand-bound experiment referenced to the requisite thermogram for the free biomolecule. The analysis reduces the experimental time and sample required by ~10-fold compared to standard DSC approaches. We have accounted for ligand thermal conversion by assuming this happens during the high temperature portion of each scan where the thermogram does not depend on ligand concentration. Therefore, the ligand concentration is a constant within the portion of the thermogram that is used to extract thermodynamic parameters. We additionally demonstrated how the rate constant for ligand thermal conversion can be obtained by performing one supplementary experiment with a longer high temperature equilibration period. For systems where ligand thermal conversion is less temperature-dependent (i.e., occurring appreciably at all temperatures), the analysis can be modified to include variable ligand concentrations. Here we demonstrate this procedure for the DNA aptamers MN4 and MN19 (Figure 3.1a) in the presence of the thermolabile ligand cocaine, which rapidly converts to benzoylecgonine at high temperatures (>60 °C). Quinine is used as a negative control for ligand thermolability since it does not undergo conversion at these

experimental temperatures and also binds to the aptamers (Figure 3.1b). We describe the acquisition of thermolabile ligand DSC datasets and their analysis yielding thermodynamic and kinetic parameters of the folding, binding, and ligand conversion processes. Furthermore, we present simulations of several non-equilibrium (kinetically controlled) folding and binding scenarios in the presence of thermolabile ligands as guidelines for analyzing these more complicated outcomes.

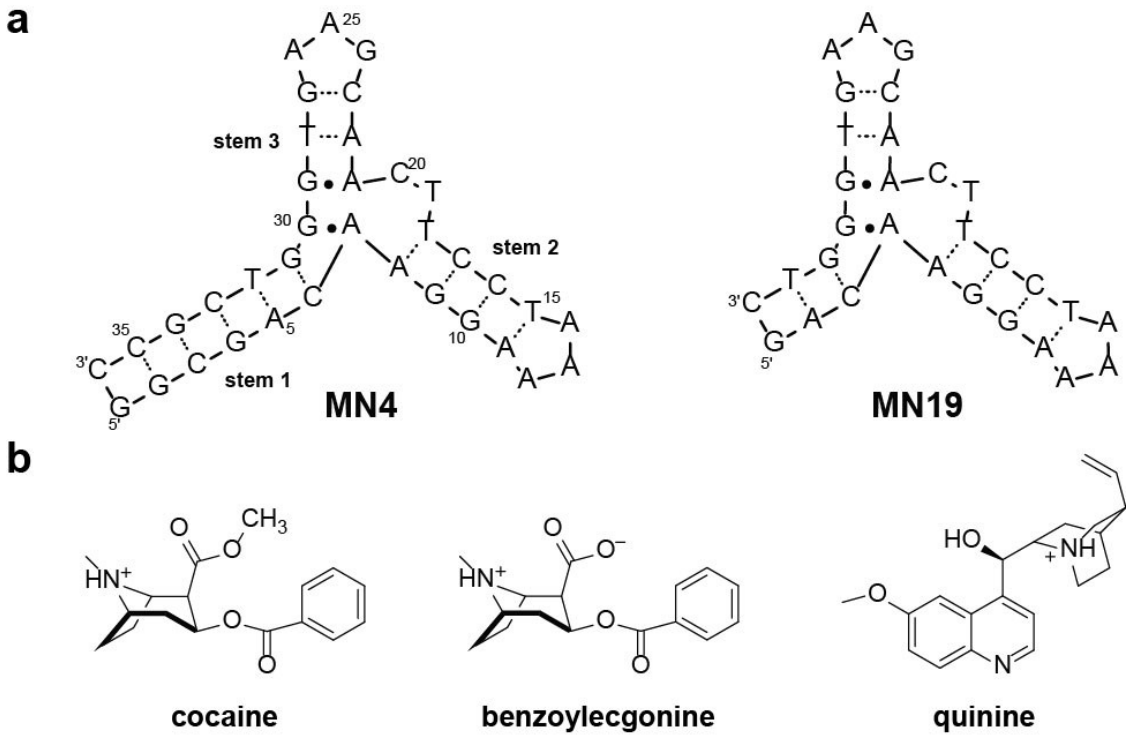


Figure 3.1. Cocaine binding aptamers, thermolabile ligand, thermal conversion product, and thermostable control. (a) Aptamers with base pair hydrogen bonds shown as dashed lines. (b) Chemical structures of ligands investigated by DSC.

3.4. Results

3.4.1. DSC with thermolabile ligands

Thermolabile ligands gradually convert from an initial to a secondary form when exposed to elevated temperatures. For example, cocaine spontaneously converts to benzoylecgonine[10] at higher temperatures (70-80 °C). We exploited this property in characterizing the interactions of MN4 and MN19 with cocaine by DSC (Figure 3.2). In a previous investigation we found by ITC that MN4 and MN19 have moderate affinities for cocaine (K_D = 7 and 27 μM respectively), and MN4 has undetectable affinity for benzoylecgonine[13, 14]. The aptamers have stronger affinities for quinine (K_D = 0.23 and 0.70 μM for MN4 and MN19 respectively)[15]. The series of replicate DSC thermograms obtained for MN4 and MN19 with cocaine are shown in Figure 3.2a,b. Each successive DSC denaturation profile shifts towards lower temperatures and smaller heights. We attribute this to the progressive conversion of cocaine to the more-weakly binding benzoylecgonine. After a large number of scans (roughly 7-10), the apparent melting temperature stabilizes at a new lower value, which we interpret as 100% conversion of cocaine to benzoylecgonine. These asymptotic scans indicate that both MN4 and MN19 bind benzoylecgonine; in the case of MN4, a slight thermal upshift and increase in peak height is apparent compared to the thermogram of the free aptamer (Figure 3.2a,b, Supplementary Figure 3.1a, Supplementary Figure 3.2). In the case of MN19, the asymptotic scans exhibit clear unfolding peaks, while an almost non-existent unfolding peak was observed for the free MN19 molecule (Supplementary Figure 3.1c). Notably, repeat scans for both aptamers in the presence of the thermostable quinine ligand are superimposable (Figure 3.2c,d, Supplementary Figure 3.1b,d).

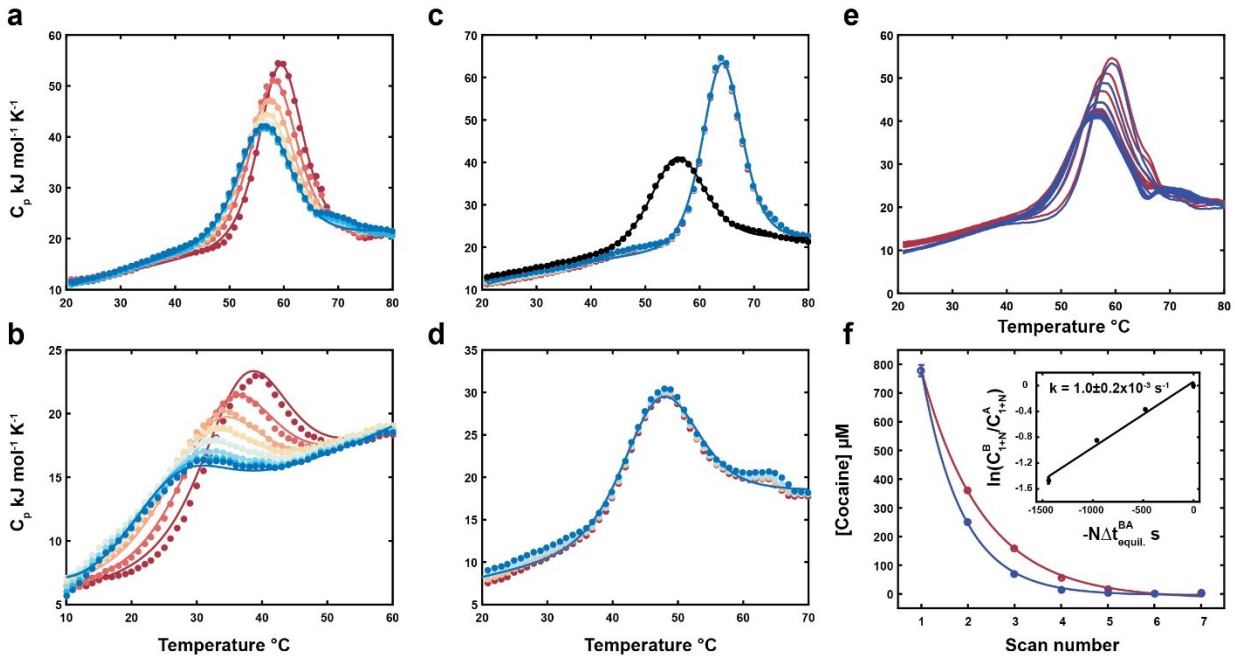


Figure 3.2. Rapid characterization of folding and binding thermodynamics using thermolabile ligands. (a) DSC heat capacity profiles for MN4 bound to cocaine and benzoylecgonine. (b) DSC heat capacity profiles for MN19 bound to cocaine and benzoylecgonine. (c) DSC heat capacity profiles for MN4 both free and bound to quinine. (d) DSC heat capacity profiles for MN19 bound to quinine. Ligand-bound experimental data points are shown as colored filled circles, fits are shown as colored lines. The first and last scans are red and blue respectively. Experimental and fitted data for free MN4 are shown as black circles and lines respectively. (e) Sets of DSC profiles for MN4 bound to cocaine and benzoylecgonine. The red profiles have equilibration times of 120 seconds at 80 °C between scans, the blue profiles have 600 second equilibration times between scans. (f) The cocaine concentrations from global analysis of datasets in (e) as a function of scan number. Experimental points and fits are shown as colored empty circles and lines respectively. Exponential fits were performed according to [cocaine] = a + $[cocaine]_0 \exp(-b*scan\ number)$. The inset shows a linear fit to Equation 3.18 for the first 4 forward scans of the 120 (A) and 600 second (B) equilibration time datasets.

3.4.2. Global analysis of thermolabile-ligand binding DSC series

We have developed a global analysis method for DSC data obtained with thermolabile ligands that yields folding and binding parameters for the initial and thermally converted ligand

(Figure 3.3, Section 3.7.3) from a pair of experiments performed with and without added ligand. This represents considerable savings in material and experiment time, as DSC-based ligand binding assays typically involve repeating multiple (~7-10) experiments over a range of ligand concentrations[3].

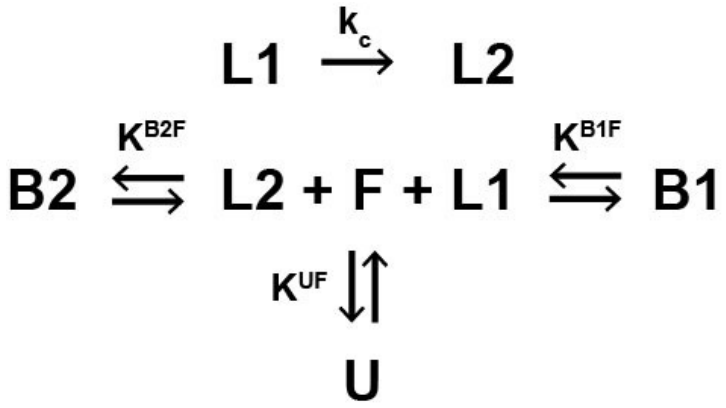


Figure 3.3. Equilibrium binding and unfolding model for a biomolecule in the presence of a thermolabile ligand during a DSC experiment. The model assumes the biomolecule (F) can unfold (U) or bind to two different ligands (L1 and L2) to form two different bound states (B1 and B2). L1 converts to L2 during the experiment with rate constant k_c.

The global fitting analysis yielded the enthalpy, ΔH, and entropy, ΔS, of the folding, cocaine-binding, and benzoylecgonine-binding reactions (Table 3.1), as well as the extent of ligand thermal conversion in each scan (Supplementary Table 3.1). The folding thermodynamic parameters obtained for MN4 with both cocaine and quinine are equal within experimental uncertainties, as expected since the stability of the free biomolecule should not depend on the identity of dilute co-solutes. The global binding parameters for both aptamers with cocaine and quinine (Table 3.1, B1F parameters) are in good agreement with ITC-derived parameters[13-15], despite differences in buffer, providing proof-of principle for this method. Interestingly, the DSC parameters show that the preference of MN4 for quinine over cocaine is driven by a much more

favourable binding enthalpy. Conversely, the preference of MN19 for quinine over cocaine is driven by a much less unfavourable binding entropy term. The analysis also yielded the binding parameters for benzoylecgonine (Table 3.1, B2F parameters), with K_D = 604 μM and 5.1 mM, for MN4 and MN19 respectively. This demonstrates the sensitivity of DSC in measuring very weak binding interactions, as benzoylecgonine binding to MN4 and similar aptamers was previously undetected by ITC, absorbance, and fluorescence spectroscopy[14, 16, 17]. Similar to what was observed for quinine, the preference of MN4 for cocaine over benzoylecgonine is due to a more favourable binding enthalpy, while in the case of MN19 it is due to a less unfavourable binding entropy. This points to a common energetic mechanism underlying the selectivity of aptamer binding and highlights the importance of obtaining thermodynamic information for understanding molecular interactions.

Table 3.1. Thermodynamic parameters extracted from global analysis of DSC data using thermolabile and thermostable ligands. ΔH and ΔG are expressed in kJ mol^{-1}, ΔS is expressed in J mol^{-1} K^{-1}and ΔC_p is expressed in kJ mol^{-1} K^{-1}. Errors were calculated according to the variance/co-variance method[18].

	MN4		[b]MN19	
Fit parameters	**Cocaine added**	**Quinine added**	**Cocaine added**	**Quinine added**
ΔH^{UF}	271.3±1.8	272.5±4.0	-	-
ΔS^{UF}	824.4±5.1	827.9±10.9	-	-
[a]ΔG^{UF}	21.6±0.2	21.6±0.9	-	-
[a]ΔH^{B1F}	-75.2±1.6	-101.0±4.0	-148.1±1.4	-105.9±10.4
[a]ΔS^{B1F}	-154.2±5.0	-213.7±12.0	-418.2±7.9	-264.9±34.1
ΔC_p^{B1F}	-1.5±0.1	-1.2±0.1	-5.2±0.1	-7.0±0.3
[a]ΔG^{B1F}	-28.5±0.2	-36.2±0.7	-21.4±0.1	-25.6±0.2
[a]ΔH^{B2F}	-33.7±1.8	-	-155.7±2.4	-
[a]ΔS^{B2F}	-49.9±5.2	-	-469.8±8.0	-
ΔC_p^{B2F}	-2.2±0.1	-	-8.2±0.2	-
[a]ΔG^{B2F}	-18.6±0.3	-	-13.3±0.1	-

[a]Parameters were calculated at 30 °C. B1F refers to cocaine- or quinine-bound folded states and B2F refers to the benzoylecgonine-bound folded state.

[b]MN19 was assumed to be only folded when bound to ligand, the parameters listed here are for unfolding of the bound folded state.

As expected, the thermal conversion of cocaine proceeds further with each successive thermogram, following a single exponential decay as a function of scan number (Figure 3.2e,f, Supplementary Table 3.1). In actuality, thermally labile ligands convert to their secondary products

continuously throughout each DSC scan, with the rate accelerating as the temperature increases, according to the activation enthalpy[19]. We made the simplifying assumption in the global analysis that the concentration is constant during each scan, but varies scan-to-scan. This assumption depends both on the rate of thermal conversion and the temperature scan rate of the calorimeter. In order to test the effects of these parameters and identify optimal ranges, we performed computer simulations with different ligand conversion kinetics and scan rates (see Section 3.7.4) in Supplementary Figure 3.3 and Supplementary Figure 3.4. Importantly, thermograms generated with fixed and varying ligand concentrations at 1 °C min^{-1} scan rate are superimposable (Supplementary Figure 3.3a) indicating that the ligand concentration can indeed be treated as constant in each scan. This makes sense as the conversion rate is ~10 000-fold faster at 80 relative to 0 °C at pH 6.8. The simulations imply that, at least in this case, ligand conversion occurs almost entirely during the high temperature portion of the scans where the thermograms are not dependent on ligand concentration. It must be noted that in cases where ligand conversion is less temperature dependent or when the biomolecular melting temperature is much higher, this assumption might not be expected to hold. We find that when the ratio of the scan rate (°C min^{-1}) to the rate constant for ligand conversion at the apparent T_m of the first forward scan (min^{-1}) is ~≤20 °C, the assumption breaks down. It would in principle, be possible to fit DSC data with continuously-varying ligand concentrations (essentially an extension of the simulations above), however this is unnecessary for the data at hand.

When the thermal conversion products bind less tightly than the original ligand, the apparent melting temperatures decrease in successive DSC scans, as observed for MN4 and MN19 interacting with cocaine. In the limit that the thermal conversion product does not bind at all, the unfolding thermogram of the ligand-free biomolecule is eventually obtained after a sufficient

number of scans. However, it is not possible to determine from these endpoint scans alone whether the thermal conversion product binds weakly or not at all. For that reason, it is important to jointly analyze the thermogram of the free biomolecule, as a reference. Conversely, if the thermal conversion product binds more tightly than the original ligand, then apparent melting temperatures increase with successive DSC scans, reaching a maximum when full conversion of the ligand is achieved. In order to illustrate possible scenarios (conversion products with no, weaker, and tighter binding) we performed simulations shown in Supplementary Figure 3.5.

3.4.3. Measuring the rate constant for ligand conversion

In addition to the thermodynamic parameters describing the folding and binding processes, the rate constant for thermal conversion is also of interest. This can be obtained in a straightforward manner by performing one additional biomolecule/ligand DSC experiment with a different high-temperature equilibration time. When a longer equilibration time is chosen, the scan-to-scan changes in ligand concentrations are greater, with a concomitant increase in the differences between successive thermograms. The ratios of successive ligand concentrations can be fit to yield the rate constant for conversion at the equilibration temperature (Section 3.7.5). We performed two sets of MN4/cocaine DSC experiments with either 120- or 600-second equilibration times between repeat scans (Supplementary Figure 3.6). The scan-to-scan decrease in ligand concentration is far more pronounced for the 600-second dataset compared to the 120-second dataset (Figure 3.2e), as anticipated. Conversion of cocaine to benzoylecgonine follows pseudo-first order kinetics and from the cocaine concentrations extracted from the global fits, we fit Equation 3.18 to obtain a rate constant for cocaine conversion of $1.0\pm0.2\times10^{-3}$ s^{-1} (Figure 3.2f

inset), in close agreement to the value previously determined at 80 °C ($1.7\pm0.3\times10^{-3}$ s^{-1}, citric acid-phosphate buffer pH 7.65)[10].

3.5. Discussion

Here, we outline practical aspects of performing and analyzing DSC binding experiments with thermolabile ligands. One of the most important experimental procedures to consider is dialysis or exchange of the biomolecule and ligand into identical working buffer solutions. Buffer mismatch between the ligand and biomolecule solutions can lead to large artifacts in the baseline and sample scans which completely obscure the relevant folding data. Additionally, it is essential that the power reading stabilizes before the DSC is pressurized so that it can be monitored during pressurization. If the power reading changes by more than ~10 μW during pressurization, bubbles have likely formed in the capillaries and can cause large artifacts in the data. In this case, the solutions need to be degassed more thoroughly. Note that a DSC baseline obtained for the thermolabile ligand alone is subtracted from the ligand + biomolecule dataset, effectively cancelling out the heat released or absorbed by the thermal conversion process itself. The standard thermolabile ligand global fitting analysis (Figure 3.2) assumes that the system is at thermodynamic equilibrium throughout the temperature scan and that the thermolabile ligand concentration is constant throughout each thermogram, decreasing exclusively during the high temperature equilibration period. We have shown that this assumption applies to cocaine-bound MN4 and is expected to hold for any thermolabile ligand/biomolecule system with kinetics similar to these.

There are, however, some situations in which the system cannot be assumed to be at thermodynamic equilibrium and/or the concentration of ligand cannot be considered constant

throughout a single scan. These include i) when the ligand thermally converts rapidly relative to the temperature scan rate, ii) when the biomolecule undergoes irreversible aggregation at high temperature, iii) when the folding/unfolding rates are slow compared to the scan rate, and iv) when the ligand association/dissociation rates are slow compared to the scan rate. In these cases, the system is under kinetic rather than thermodynamic control and equilibrium thermodynamic models cannot strictly be applied. Data may be simulated quantitatively following Figure 3.4, as described in Section 3.7.6. In principle, these kinetics-based calculations could be used to fit non-equilibrium DSC data, potentially yielding both kinetic and thermodynamic data, however this analysis is beyond the scope of this thesis. Instead, we present some representative simulated DSC data to assist the reader in identifying non-equilibrium situations.

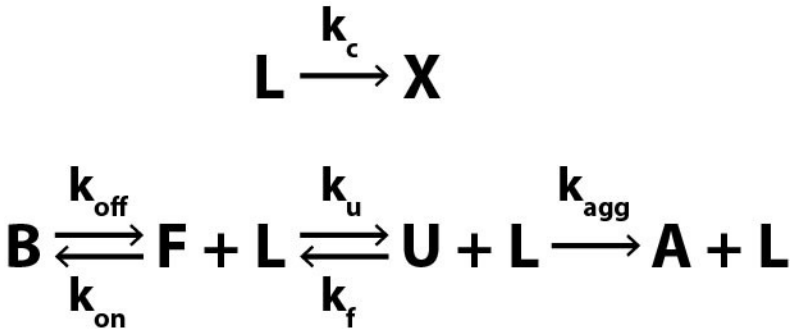


Figure 3.4 Biomolecular folding, binding to a thermolabile ligand, and irreversible aggregation. The thermolabile ligand (L) converts to product (X) with a rate constant k_c. X has no affinity for the biomolecule. The bound state (B) of the biomolecule exchanges with the free folded state (F) with rate constants k_{off} and k_{on}. F exchanges with the unfolded state (U) with rate constants k_u and k_f. U irreversibly converts to the aggregated state (A) with the rate constant k_{agg}.

An ideal example of thermodynamic control is shown in Figure 3.5a,b. DSC thermograms of the free biomolecule are superimposable (Figure 3.5a) and scans with the thermolabile ligand do not show hysteresis, such that the melting temperature observed on the up-scan matches the

folding temperature of the previous down-scan (Figure 3.5b). When the thermolabile ligand converts rapidly compared to the scan rate, large distortions appear in the thermogram and the thermodynamic equations do not account for the peak shape, as shown in Figure 3.5c,d. This can be alleviated somewhat by increasing the scan rate. When the biomolecule aggregates in a temperature-dependent manner, DSC traces for the free biomolecule show successive decreases in magnitude (Figure 3.5e), while addition of the thermolabile ligand produces a pattern of decreasing thermal upshifts similar to the ideal case, but scaled by the decreasing biomolecule concentration (Figure 3.5f). When folding/unfolding kinetics are slow compared to the scan rate, hysteresis is apparent in DSC traces of the free biomolecule such that the apparent denaturation temperature on the up-scan is higher than the apparent renaturation temperature on the down-scan (Figure 3.5g). Addition of a thermolabile ligand leads to the familiar pattern of decreasing thermal upshifts, particularly for the up-scans (Figure 3.5h). Finally, systems with rapid folding and slow binding produce hysteresis-free DSC thermograms for the free biomolecule (Figure 3.5i), however data with the thermolabile ligand show hysteresis where the apparent melting temperature of the up-scan is higher than the apparent folding temperature of the previous down-scan (Figure 3.5j). Nevertheless, the typical pattern of decreasing thermal upshifts are apparent in both up-scans and down-scans. Non-equilibrium behavior in the case of slow folding or binding kinetics can be alleviated somewhat by decreasing the scan rate, although this runs the risk of non-negligible ligand thermal conversion occurring throughout the scan. In practice, the scan rate and upper equilibration temperature can be adjusted manually to obtain data resembling Figure 3.5a,b.

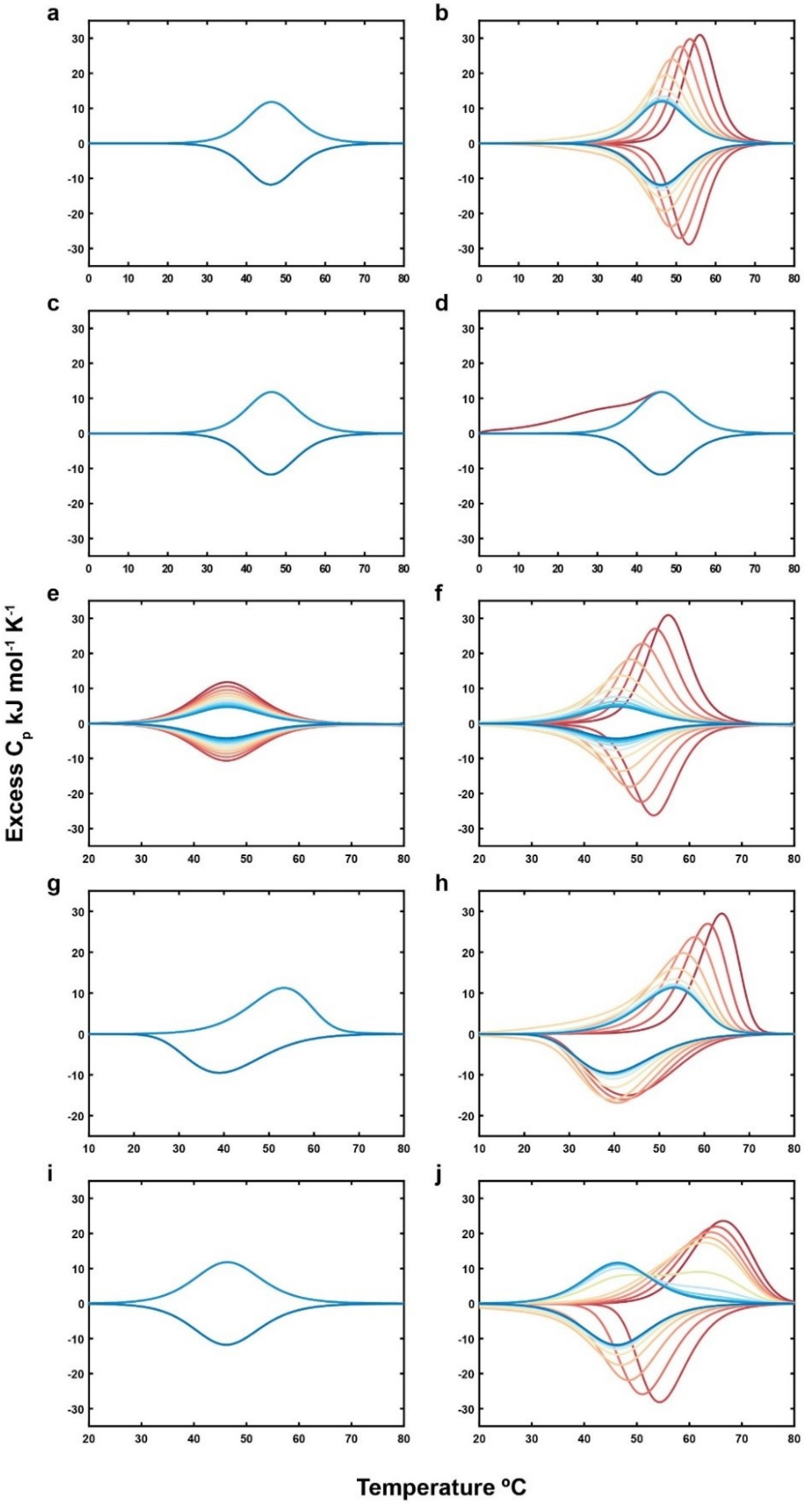

a
b
c
d
e
f
g
h
i
j
Excess C_p kJ mol^{-1} K^{-1}
Temperature ºC

Figure 3.5. Computer simulation of equilibrium and kinetically-controlled DSC experiments in the absence and presence of a thermolabile ligand. (a) Equilibrium biomolecular folding. (b) Equilibrium folding, thermolabile ligand binding, and slow thermolabile ligand conversion. (c) Equilibrium biomolecular folding. (d) Equilibrium folding, thermolabile ligand binding, and fast thermolabile ligand conversion. (e) Equilibrium biomolecular folding and slow irreversible aggregation. (f) Equilibrium folding, binding, slow thermolabile ligand conversion, and slow irreversible aggregation. (g) Slow biomolecular folding. (h) Slow folding, equilibrium binding, and slow thermolabile ligand conversion. (i) Equilibrium biomolecular folding. (j) Equilibrium folding, slow thermolabile ligand binding, and slow thermolabile ligand conversion. In all panels, the first and last simulated scans are dark red and dark blue, respectively. Panels that show only light and dark blue thermograms indicate that all simulated scans overlay, and only the last two are visible in the plot. The parameters for performing these simulations are given in Section 3.7.6.

Our thermodynamic analysis for DSC binding experiments with thermolabile ligands requires that the folding and binding processes are relatively rapid and that thermolabile ligand conversion is slow prior to the high temperature portion of each scan. When the lifetime of the folded and/or bound state is greater than about 30 s (k_{off}, k_u < 0.03 s^{-1}), hysteresis becomes discernable in scans performed at 1 °C min^{-1}. Additionally, when the ligand conversion rate constant is above approximately $k_c = 10^{-4}$ s^{-1} before the denaturation transition, there can be significant depletion of the ligand during the course of a single scan. Application of our analysis is also inappropriate when irreversible aggregation occurs. In these cases, more advanced modeling could be applied to the data. No affinity information is available if ligand conversion is so rapid that it reaches completion prior to the first denaturation transition.

3.6. Conclusions

DSC is a powerful approach for characterizing biomolecule/ligand interactions, with many applications in drug development[1, 5-7]. It is particularly well-suited to very tight interactions that are difficult to study directly by titration methods such as ITC[20, 21]. Our method for the first time allows DSC to be used to measure the binding thermodynamics of high affinity, thermolabile ligands. By performing a global simultaneous analysis of all scans, thermodynamic parameters are extracted with high accuracy[22]. An additional benefit is that the full dataset can be collected in as little as one experiment if the thermal conversion product has no affinity for the free biomolecule. In contrast, producing a typical experimental DSC series for a non-thermolabile ligand requires ~7–10 total experiments. We find here that DSC is also highly effective at measuring very weak binding interactions (high μM to mM) that may be undetectable by other techniques. DSC has the additional advantage of simultaneously providing information on both folding and binding reactions. However, the thermal lability of many known pharmaceuticals and potential drug leads can lead to DSC data with large scan-to-scan variations that are not interpretable using standard methods. Our global fitting method exploits these variations to yield folding and binding parameters in a fraction of the time and sample needed for thermally-stable compounds analyzed with conventional DSC approaches. Furthermore, just one additional DSC experiment gives the rate constant for thermal conversion. This method therefore opens the door to using DSC to characterize a class of hitherto inaccessible biomolecule/ligand interactions with high precision. This approach has direct applications to characterizing tight, thermolabile inhibitors in drug discovery campaigns. Several therapeutic compounds such as antibiotics and benzodiazepines are known to be thermolabile, undergoing rapid hydrolysis at or near physiological pH and temperatures of ~60–70 °C[11]. This DSC method is well positioned to identify and characterize

many more. As well, modification of the fitting protocol to account for systems under kinetic rather than thermodynamic control, as discussed above, has the potential to open the door to many more systems of biological relevance.

3.7. Materials and Methods

3.7.1. Sample Preparation

Oligonucleotide samples were purchased pre-purified from Integrated DNA Technologies (Iowa, USA). Samples were dissolved in 20 mM sodium phosphate buffer, 140 mM NaCl, pH 7.4. DNA concentrations were 83 and 88 μM for MN4 and MN19. Initial ligand concentrations were 778 and 880 μM for cocaine and quinine respectively.

3.7.2. Experimental DSC

Each experiment consisted of 10 melting and 10 annealing scans. Scan rates were 1.0 °C min^{-1}. Samples were scanned from 0-80 °C with equilibration times of 60 seconds between scans, except for the cocaine kinetics experiments which used 120 and 600 second equilibration times.

3.7.3. DSC global fitting

Heat capacity profiles were analyzed assuming unfolded (U), folded (F), folded cocaine-bound (B1), and folded benzoylecgonine-bound (B2) states in equilibrium where L1 and L2 are cocaine and benzoylecgonine ligands respectively, and cocaine is assumed to convert to benzoylecgonine with rate constant k_c (Figure 3.3). The heat capacity profiles were fit using temperature dependent thermodynamic parameters for the ligand binding processes

$$\Delta H^{B1F}(T) = \Delta H^{B1F}(T_0) + \Delta C_p^{B1F}(T - T_0) \tag{3.1}$$

$$\Delta S^{B1F}(T) = \Delta S^{B1F}(T_0) + \Delta C_p^{B1F} \ln\left\{\frac{T}{T_0}\right\} \tag{3.2}$$

where $\Delta H^{B1F}(T)$, $\Delta S^{B1F}(T)$, and $\Delta C_p{}^{B1F}$ are the changes in enthalpies, entropies, and heat capacities for the folded to cocaine-bound (B1F) processes respectively and T_0 is the reference temperature. Folded to benzoylecgonine-bound (B2F) parameters were fit using the same equations as the B1F equilibrium but these have been omitted here for the sake of brevity. The change in heat capacity for unfolding, $\Delta C_p{}^{UF}$, was set equal to zero as it was found to be negligibly small when included in the global fits. Equilibrium constants for the folding and binding processes were calculated according to

$$K^{UF}(T) = \frac{[U](T)}{[F](T)} = \exp\left(\frac{-\left(\Delta H^{UF} - T\Delta S^{UF}\right)}{RT}\right) \tag{3.3}$$

$$K^{B1F}(T) = \frac{[B1](T)}{[F](T)[L1](T)} = \exp\left(\frac{-\left(\Delta H^{B1F}(T) - T\Delta S^{B1F}(T)\right)}{RT}\right) \tag{3.4}$$

and combined in the partition function assuming folded, cocaine-bound, benzoylecgonine-bound, and unfolded states according to

$$Q(T) = 1 + K^{UF}(T) + K^{B1F}(T)[L1](T) + K^{B2F}(T)[L2](T) \tag{3.5}$$

where 1 is the relative population of the folded state. The concentration of folded state as a function of temperature (or similarly the unfolded state for MN19 as $[F](T) = 0$) was obtained by numerically solving the real, positive root of

$$a[F](T)^3 + b[F](T)^2 + c[F](T) - C_T = 0 \tag{3.6}$$

where $a = K^{UF}(T)K^{B1F}(T)K^{B2F}(T) + K^{B1F}(T)K^{B2F}(T)$, $b = K^{UF}(T)K^{B1F}(T) + K^{UF}(T)K^{B2F}(T) + K^{B1F}(T) + K^{B2F}(T) - C_TK^{B1F}(T)K^{B2F}(T) + [L2]_TK^{B1F}(T)K^{B2F}(T) + [L1]_TK^{B1F}(T)K^{B2F}(T)$, and $c = 1 + K^{UF} -$

$C_T K^{B1F}$ - $C_T K^{B2F}$ + $[L1]_T K^{B1F}(T)$ + $[L2]_T K^{B2F}(T)$, and C_T, $[L1]_T$, and $[L2]_T$ are the total concentrations of aptamer, cocaine, and benzoylecgonine respectively. Free ligand concentrations $[L1](T)$ and $[L2](T)$ were calculated using

$$[L1](T) = \frac{[L1]_T}{1 + K^{B1F}[F]} \tag{3.7}$$

where the equation for calculating $[L2](T)$ is analogous to that for $[L1](T)$. The populations of each state were computed as

$$P^U(T) = \frac{K^{UF}(T)}{Q(T)} \tag{3.8}$$

and

$$P^{B1}(T) = \frac{K^{B1F}(T)[L1](T)}{Q(T)} \tag{3.9}$$

Where we have omitted the calculation of P^{B2} for brevity. The DSC thermogram (C_p) profiles were calculated using

$$\begin{aligned} C_p^{calc}(T) &= C_p^F(T) + \Delta H^{UF} \times \frac{d}{dT} P^U(T) + \Delta H^{B1F}(T) \times \frac{d}{dT} P^{B1}(T) \\ &+ \Delta H^{B2F}(T) \times \frac{d}{dT} P^{B2}(T) + P^{B1}(T) \times \Delta C_p^{B1F} + P^{B2}(T) \times \Delta C_p^{B2F} \end{aligned} \tag{3.10}$$

where $C_p^F(T)$ is the heat capacity of the reference folded state, calculated as a second order polynomial in temperature. Note that since ΔC_p^{UF} is zero, the heat capacity of the unfolded state is identical. The calculated thermograms were globally fit to the experimental DSC profiles by minimizing the RSS

$$RSS(\xi) = \sum_{i=1}^{M} \sum_{j=1}^{N} (C_p^{\exp}(i,j) - C_p^{calc}(\xi,i,j))^2 \tag{3.11}$$

where M is the number of replicate DSC scans, N is the number of data points in each scan, and ξ = [ΔH^{UF}, ΔH^{B1F}, ΔH^{B2F} ΔS^{UF}, ΔS^{B1F}, ΔS^{B2F}, ΔC_p^{B1F}, ΔC_p^{B2F}, $[L1]_{T,i}$, a_f, b_f, c_f]. (ΔH^{UF}, ΔH^{B1F}, ΔH^{B2F}

ΔS^{UF}, ΔS^{B1F}, ΔS^{B2F}, ΔC_p^{B1F}, ΔC_p^{B2F}) are the thermodynamic parameters for folding and binding, (a_f, b_f, c_f) define the C_p^F baseline, and $[L1]_{T,i}$ is the total cocaine concentration of the ith DSC profile. $[L1]_{T,1}$ was fixed at the initial ligand concentration and $[L1]_{T,2-M}$ were optimized in the fits as adjustable parameters. The total concentration of benzoylecgonine in each scan was calculated as $[L2]_{T,i} = [L1]_{T,0} - [L1]_{T,i}$, and the final scan of the cocaine-bound DSC manifold assumed cocaine conversion was 100% complete, i.e. $[L2]_{T,M} = [L1]_{T,0}$. The quinine-bound MN4 global fits were constrained by including the unbound dataset as i=0, with $[L]_{T,0}$=0. The code for performing these global fits is freely available at http://www.rsc.org/suppdata/c6/cc/c6cc05576a/c6cc05576a2.pdf.

3.7.4. Testing the high temperature ligand conversion assumption

In our global fits we used constant total ligand concentrations for each thermogram, based on the assumption that ligand thermal conversion predominantly occurs during the high temperature equilibration period. However, this assumption depends on both the temperature scan rate and the rate of ligand conversion. It is possible for the assumption to be violated with very slow temperature scanning or rapid ligand conversion. Slow scan rates lengthen the amount of time the ligand spends at each temperature and rapid ligand conversion causes the initial ligand to be depleted within the first scan. In our case, cocaine conversion is strongly temperature dependent and scanning at 1 °C min^{-1} does not violate our assumption that most of the conversion happens at high temperatures. Simulations with continuously-varying cocaine concentrations at scan rates of 1 °C min^{-1} are superimposable with those using fixed concentrations (Supplementary Figure 3.3a). We have simulated experiments where we varied the scan rate or the ligand conversion kinetics (by modifying the activation energy) in order to provide visual evidence of when the assumption

breaks down (Supplementary Figure 3.3). The rate constants for ligand conversion were calculated every 0.1 °C from 20-80 °C using

$$k(T) = Ae^{-\frac{E_a}{RT}} \tag{3.12}$$

Where A (7.51×10^{10} s^{-1}) and E_a (95.9 kJ mol^{-1}) are the pre-exponential factor and activation energy for cocaine conversion respectively, and R is the ideal gas constant. Ligand concentrations at every 0.1 °C were computed with

$$[L]_t = [L]_{t-1}e^{-k(T)\times t} \tag{3.13}$$

where $[L]_t$ is the new ligand concentration after converting for a time t (set by the scan rate, for example the average time per 0.1 °C at 1 °C min^{-1} = 6 seconds) at temperature T, and $[L]_{t-1}$ is the ligand concentration at the previous temperature.

Clear distortions of the DSC peak shape occur when the ligand is depleted in the early portion of the thermogram, either when scanning extremely slowly or when rapid conversion occurs at lower temperatures. We note that the 0.005 °C min^{-1} scan rate is unfeasible in practice as the DSC signal to noise becomes poor below scan rates of 0.1 °C min^{-1}. We found that a conversion ratio (CR, °C) defined as the scan rate (°C min^{-1}) divided by the rate constant for ligand conversion at the apparent T_m of the first forward scan (min^{-1}) gives a measure of when our ligand conversion assumption is violated. For conversion ratios below ~20 °C, substantial depletion of the ligand occurs before and during the thermogram leading to distortions of the transition shape. This indicates continuously-varying ligand concentrations must be applied in the fit. By increasing the scan rate, one may adjust the conversion ratio for a thermolabile ligand DSC experiment in order to obtain data that can be fit with constant ligand concentrations. The scan rate must however remain slow enough to avoid TH. This can be tested with an experiment on the free biomolecule.

Ligand conversion can also be modulated by protection of the ligand in the biomolecular binding pocket. The experiments were performed with ligand concentrations in excess of the biomolecule (10:1), i.e. the total amount of available ligand is largely unbound. Therefore, the overall ligand conversion is dominated by that of the free ligand molecules. Assuming the biomolecule can protect the ligand and the ligand is in excess, the apparent rate constant is given by $k^{app.}(T) = k(T) \times P_{free}(T)$, where $P_{free}(T)$ is the population of free ligand as a function of temperature. This assumes the equilibration between free and bound ligand is rapid relative to the conversion rate. Ligand concentrations at each temperature are accordingly calculated with Equation 3.13. We have simulated the DSC profiles where protection of the ligand occurs at scan rates of 1 °C min^{-1} and 0.005 °C min^{-1}, overlaying these with the case where ligand is not protected by the biomolecule and continuously varies at the corresponding scan rates (Supplementary Figure 3.4). We find that protection of the ligand does not appreciably modify the result, even at 0.005 °C min^{-1} scan rate where it can play a greater role.

3.7.5. Calculation of the rate constant for conversion of cocaine to benzoylecgonine

The concentration of cocaine in any scan, C_N, is related to the concentration remaining in the subsequent scan, C_{N+1} (Supplementary Figure 3.6), according to

$$C_{N+1} = f * C_N e^{-kt_{equil.}} \tag{3.14}$$

where t_{equil} is the length of the high-temperature (80 °C) equilibration period separating each heating scan from the following cooling scan, k is the rate constant for thermal conversion, and f (<1) accounts for the thermal conversion occurring during a cooling scan and subsequent heating scan. The factor f cancels out when a comparison is made between the ratio of two N scans and two N+1 scans obtained with different equilibration times (A and B), as follows:

$$\frac{C_{N+1}^{B}}{C_{N+1}^{A}} = \frac{f * C_{N}^{B} e^{-kt_{equil.}^{B}}}{f * C_{N}^{A} e^{-kt_{equil.}^{A}}} = \frac{C_{N}^{B}}{C_{N}^{A}} e^{-k\Delta t_{equil.}^{BA}} \tag{3.15}$$

and

$$k = -\ln\left(\frac{\left(\frac{C_{N+1}^{B}}{C_{N+1}^{A}}\right)}{\left(\frac{C_{N}^{B}}{C_{N}^{A}}\right)}\right) \Delta t_{equil.}^{BA\ -1}. \tag{3.16}$$

Extending this to the general case for the initial forward scan (number *1*) and the N^{th} later scan (number *1+N*) gives

$$\frac{C_{1+N}^{B}}{C_{1+N}^{A}} = \frac{f^{N} * C_{1}^{B} e^{-kNt_{equil.}^{B}}}{f^{N} * C_{1}^{A} e^{-kNt_{equil.}^{A}}} = \frac{C_{1}^{B}}{C_{1}^{A}} e^{-kN\Delta t_{equil.}^{BA}} \tag{3.17}$$

and the rate constant for ligand thermal conversion can be obtained from fitting

$$\ln\left(\frac{\left(\frac{C_{1+N}^{B}}{C_{1+N}^{A}}\right)}{\left(\frac{C_{1}^{B}}{C_{1}^{A}}\right)}\right) = -kN\Delta t_{equil.}^{BA}. \tag{3.18}$$

When the initial concentrations in each experiment are equal, $C_1^B / C_1^A = 1$, as was the case here. A plot of $\ln\left(\frac{C_{1+N}^{B}}{C_{1+N}^{A}}\right)$ versus $-N\Delta t_{equil.}^{BA}$ yields a straight line with a slope of k = 1.0±0.2x10^{-3} s^{-1} (Figure 3.2f inset). Note that it is necessary to use the earlier scans (1-4 here, this depends on ligand conversion kinetics) in the experiment where the concentrations of ligand can be more accurately fit, as the relative error in determining ligand concentrations is high for later scans.

Similarly, the rate constant can be calculated from a minimum number of 4 forward scans (Supplementary Figure 3.6). A two-forward scan experiment with a short equilibration time at high

temperature is performed first (Experiment A), and, in the case of thermolabile ligands like cocaine, C_1^A can be assumed to be what was loaded into the calorimeter and the C_2^A can be extracted from the second forward scan. An additional two-scan experiment (Experiment B) with a longer equilibration time at the highest temperature gives C_1^B and C_2^B, from which the rate constant can be calculated using the $\Delta t_{equil.}^{BA}$ and Equation 3.18.

3.7.6. Characterizing non-equilibrium biomolecular folding and binding interactions with thermolabile ligands by DSC

Non-equilibrium biomolecular folding and binding processes are identified by the presence of TH, or a lack of reproducibility of the scan signatures. These types of thermograms can be considered to be kinetically controlled and may be generally analyzed with rate equations describing the rates of change of concentrations for each of the species involved[23, 24]. This approach can be extended to DSC binding experiments with thermolabile ligands by including an equation governing the ligand concentration as a function of time and temperature in the DSC experiment. Here we consider the model in Figure 3.4 for a biomolecule undergoing thermolabile ligand binding, folding, and aggregation processes.

We simulated scenarios using the model in Figure 3.4 where aggregation occurs or biomolecular folding and ligand binding are kinetically controlled. The simulation temperatures were defined according to

$$T_n = T_0 + (T_f - T_0) n \frac{\Delta t_{\text{int}}}{t_{tot}} \tag{3.19}$$

where T_0 and T_f are the initial and final temperatures for each simulated DSC scan, n is the increment number, Δt_{int} is the integration time increment (s) (see below), and t_{tot} (s) is the total length of time for the DSC scan calculated by

$$t_{tot} = \frac{(T_f - T_0)}{\frac{dT}{dt}} \tag{3.20}$$

where dT/dt is the scan rate (°C s^{-1}). The simulations were bounded by the total number of increments n_{tot} computed with

$$n_{tot} = \frac{(T_f - T_0)}{\Delta t_{\text{int}} \left(\frac{dT}{dt} \right)} \tag{3.21}$$

such that when $n = n_{tot}$, $T_n = T_f$. Considering Figure 3.4, the changes in each species concentration with respect to time are given by

$$\begin{aligned} &\frac{d}{dt}[L](T_n,t_n) = k_{off}(T_n)[B](T_n,t_n) - k_{on}(T_n)[F](T_n,t_n)[L](T_n,t_n) \\ &- k_c(T_n)[L](T_n,t_n) \end{aligned} \tag{3.22}$$

$$\frac{d}{dt}[B](T_n,t_n) = k_{on}(T_n)[F](T_n,t_n)[L](T_n,t_n) - k_{off}(T_n)[B](T_n,t_n) \tag{3.23}$$

$$\begin{aligned} &\frac{d}{dt}[F](T_n,t_n) = k_{off}(T_n)[B](T_n,t_n) - k_{on}(T_n)[F](T_n,t_n)[L](T_n,t_n) \\ &+ k_{fold}(T_n)[U](T_n,t_n) - k_{unfold}(T_n)[F](T_n,t_n) \end{aligned} \tag{3.24}$$

$$\frac{d}{dt}[U](T_n,t_n) = k_{unfold}(T_n)[F](T_n,t_n) - k_{fold}(T_n)[U](T_n,t_n) - k_{agg}(T_n)[U](T_n,t_n) \tag{3.25}$$

$$\frac{d}{dt}[A](T_n,t_n) = k_{agg}(T_n)[U](T_n,t_n) \tag{3.26}$$

and the rate constants are given by

$$k(T_n) = Ae^{\frac{-E_a}{RT_n}}. \tag{3.27}$$

The concentrations of each species can be computed numerically starting from an initial condition using equations of the form

$$[L](T_{n+1},t_{n+1}) = [L](T_n,t_n) + \begin{pmatrix} k_{off}(T_n)[B](T_n,t_n) - k_{on}(T_n)[F](T_n,t_n)[L](T_n,t_n) \\ -k_c(T_n)[L](T_n,t_n) \end{pmatrix} \Delta t_{\text{int}} \quad (3.28)$$

where for brevity, we have demonstrated the integration with just the ligand concentration. We simulated DSC profiles with a scan rate of 1 °C min^{-1}. The populations of each biomolecular state and their temperature derivatives are calculated according to

$$P_B(T_n,t_n) = \frac{[B](T_n,t_n)}{C_T} \quad (3.29)$$

$$\frac{d}{dT}P_B(T_n,t_n) = \frac{dt}{dT}\frac{d}{dt}[B](T_n,t_n)\frac{1}{C_T} \quad (3.30)$$

where dt/dT is the inverse scan rate. We have shown the calculation for just the bound state population for brevity. The excess heat capacity function for Figure 3.4 is given by

$$C_p^{excess}(T_n,t_n) = \frac{d}{dT}P_B(T_n,t_n)\Delta H_{BF} + \frac{d}{dT}P_U(T_n,t_n)\Delta H_{UF} + \frac{d}{dT}P_A(T_n,t_n)\Delta H_{AF} \quad (3.31)$$

calculated relative to the reference folded state. Here we have chosen temperature-independent changes in enthalpy relative to the folded state for simplicity. If required, the temperature dependences of ΔH are accounted for by including the ΔC_p parameter. The changes in enthalpy for the reversible folding and binding steps were calculated according to

$$\Delta H_{UF} = E_a^{unfold} - E_a^{fold} \quad (3.32)$$

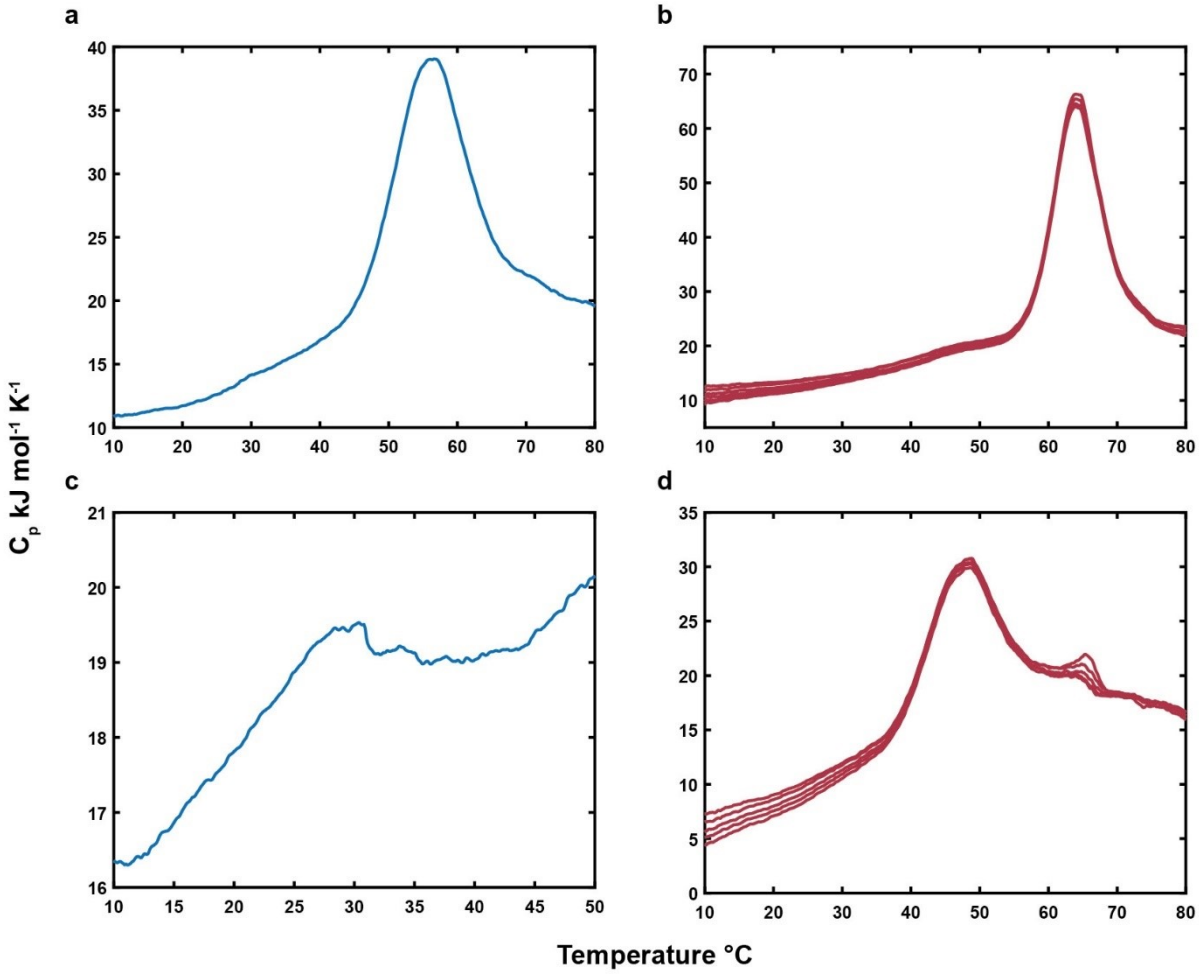
and

$$\Delta H_{BF} = E_a^{on} - E_a^{off} \quad (3.33)$$

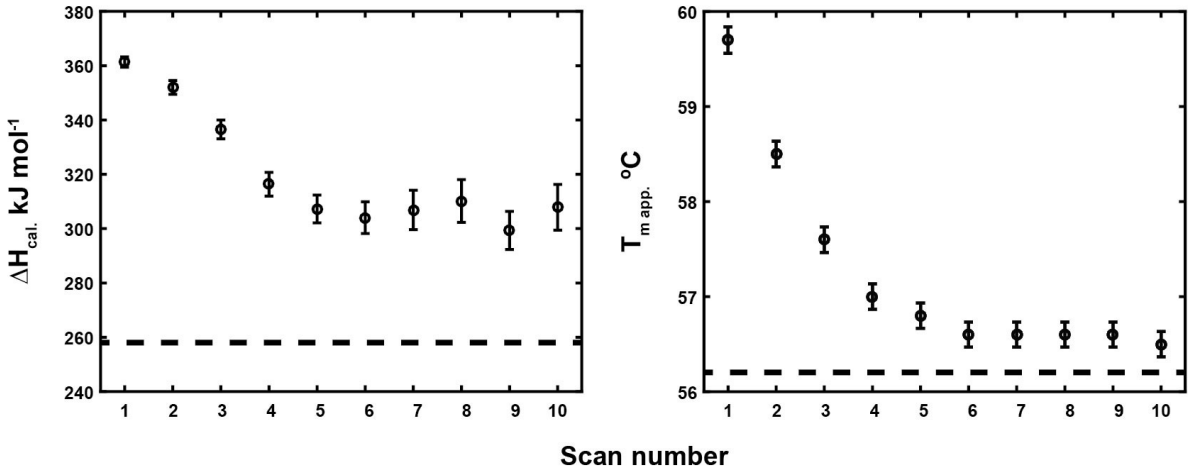
with simulation values of 200 and -140 kJ mol^{-1} respectively. The change in enthalpy for aggregation relative to the folded state was chosen as 50 kJ mol^{-1}. Arrhenius parameters for

equilibrium binding and folding were A_{on} = 5×10^{-1} M^{-1} s^{-1}, A_{off} = 1×10^{19} s^{-1}, E_a^{on} = -20 kJ mol^{-1}, E_a^{off} = 120 kJ mol^{-1}, A_{fold} = 1×10^{-14} s^{-1}, A_{unfold} = 5×10^{18}, E_d^{fold} = -80 kJ mol^{-1}, and E_a^{unfold} = 120 kJ mol^{-1}. Arrhenius parameters for kinetically-controlled binding and folding were A_{on} = 5×10^{-3} M^{-1} s^{-1}, A_{off} = 1×10^{16} s^{-1}, E_a^{on} = -20 kJ mol^{-1}, E_a^{off} = 120 kJ mol^{-1}, A_{fold} = 1×10^{-16} s^{-1}, A_{unfold} = 5×10^{16}, E_d^{fold} = -80 kJ mol^{-1}, and E_a^{unfold} = 120 kJ mol^{-1}. Arrhenius parameters for slow and rapid thermolabile ligand conversion were A_{slow} = 7.509×10^{10} s^{-1}, E_a^{slow} = 94.65 kJ mol^{-1}, and A_{fast} = 1 s^{-1}, E_d^{fast} = 10 kJ mol^{-1}. Arrhenius parameters for slow irreversible aggregation were $A_{agg.}$ =5×10^{7} s^{-1} and $E_a^{agg.}$ = 80 kJ mol^{-1}. The simulations were performed with lower and upper temperatures of 0 and 80 °C and a scan rate of 1 °C min^{-1}. 20 scans (10 melting and 10 annealing) were simulated with total biomolecule and ligand concentrations of 200 μM and 10 mM respectively. The concentrations of all species were allowed to equilibrate at 0 °C for ten minutes to simulate the pre-scan equilibration time in the calorimeter. After each scan, the concentrations of each species were allowed to equilibrate for 60 s. The MATLAB code for generating kinetically controlled DSC datasets with thermolabile ligands is available at https://www.jove.com/video/55959/measuring-biomolecular-dsc-profiles-with-thermolabile-ligands-to.

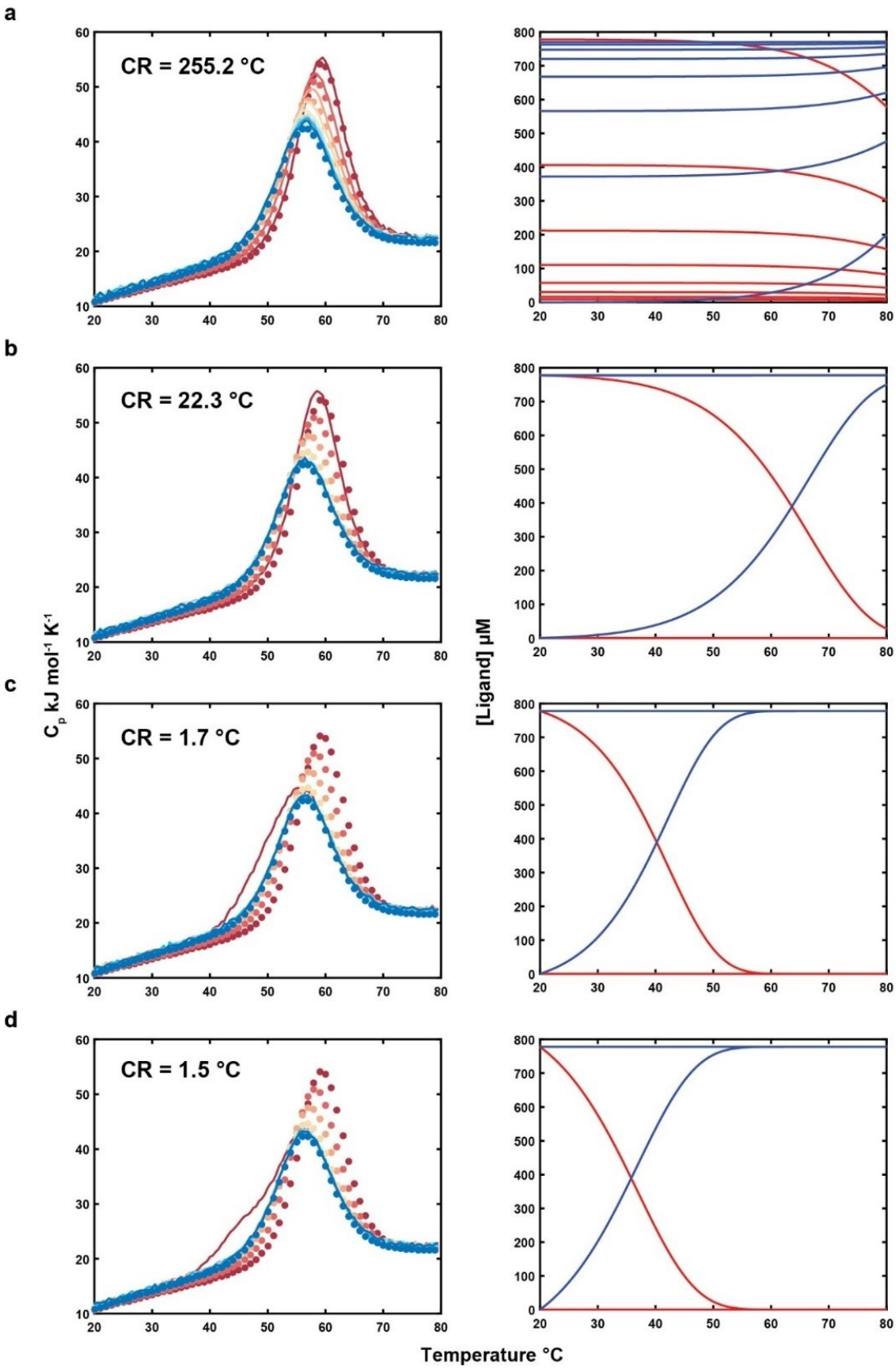
3.8. Supplementary Figures



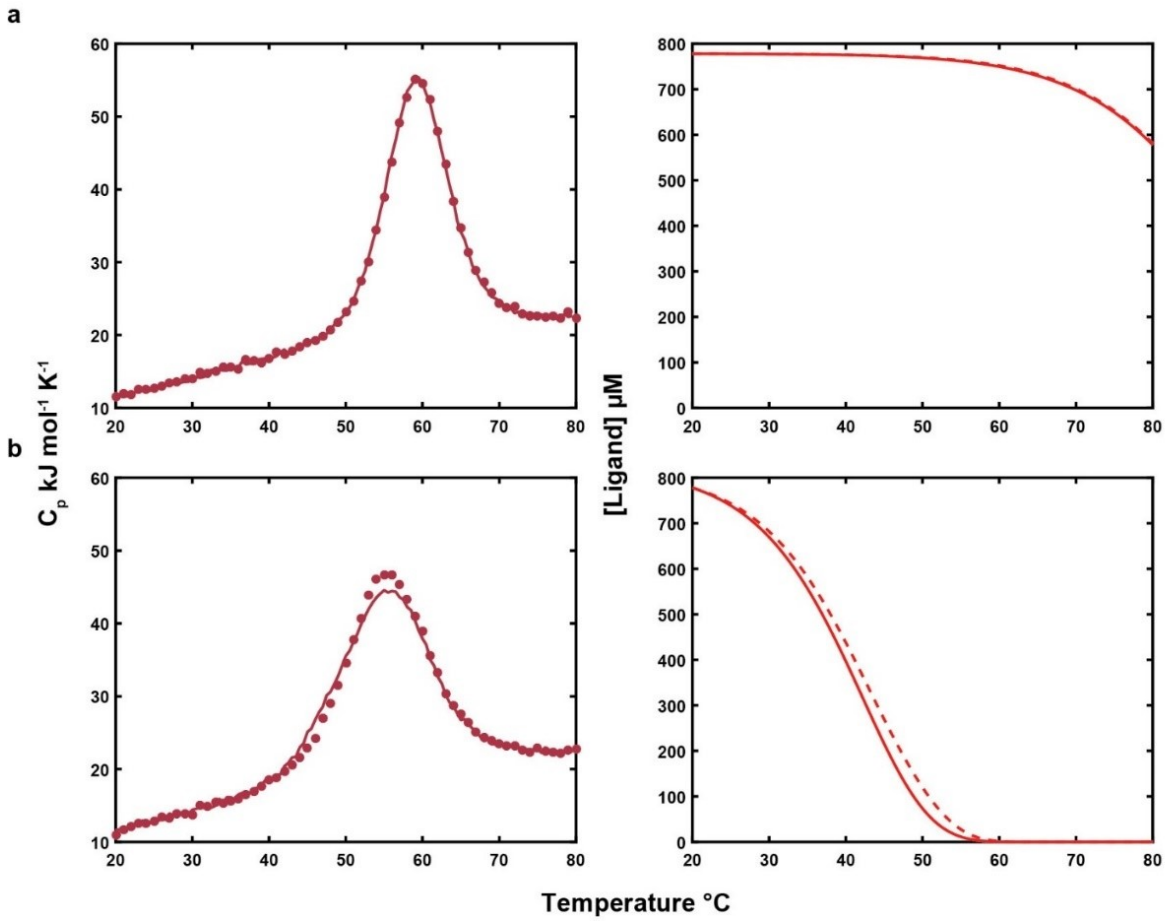
Supplementary Figure 3.1. DSC profiles for free and quinine-bound aptamers. (a) DSC profile for free MN4. (b) DSC profiles for MN4 bound to quinine. Successive scans show no change in profile. (c) DSC profile for free MN19. MN19 is largely unstructured in its free state. (d) DSC profiles for MN19 bound to quinine. Successive scans show no change in profile.



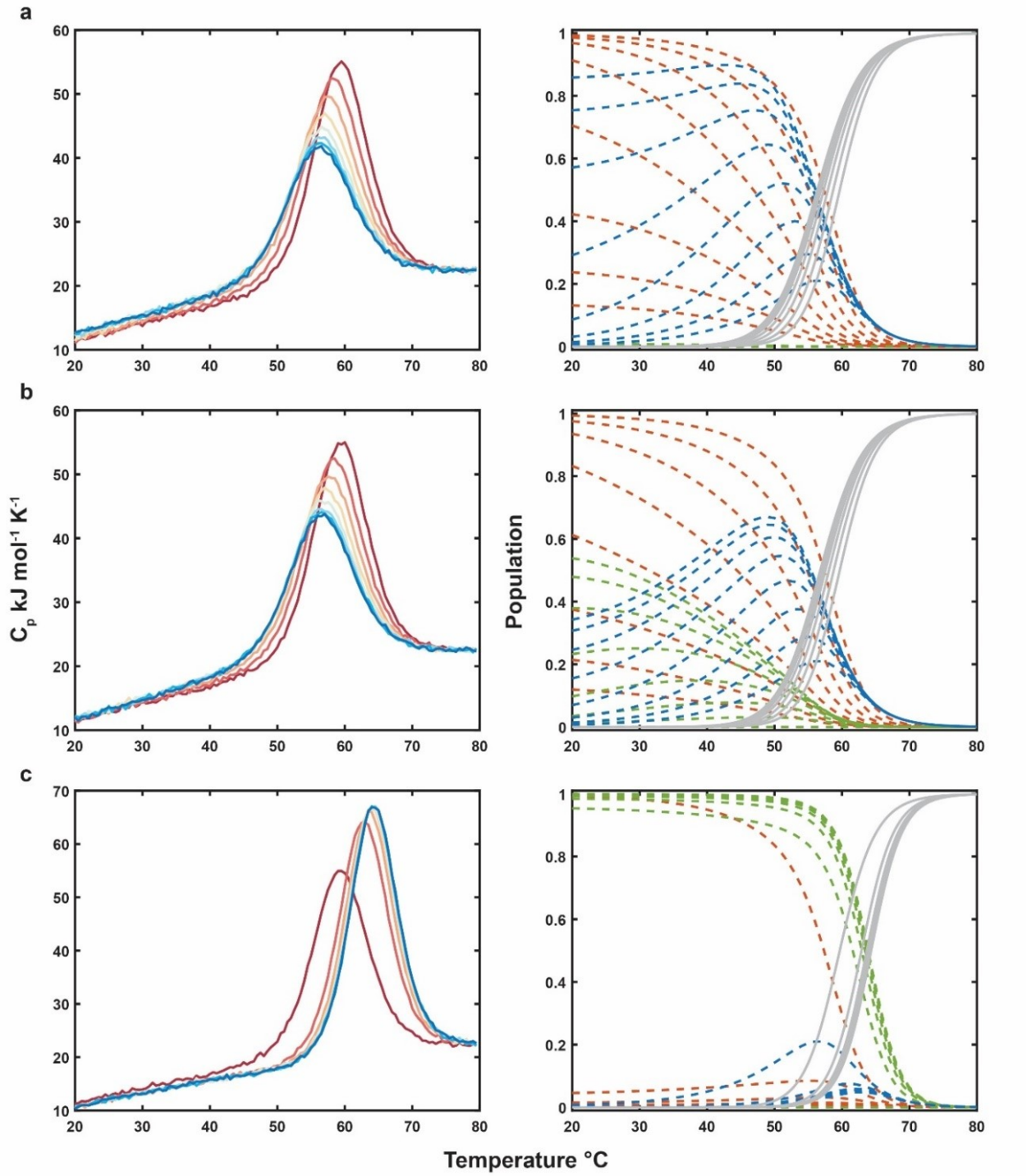
Supplementary Figure 3.2. Evidence for benzoylecgonine binding. Calorimetric enthalpy and apparent melting temperature versus forward scan number for the cocaine-bound MN4 dataset. Dashed black lines indicate the calorimetric enthalpy and T_m for free MN4 respectively.


a
CR = 255.2 °C
b
CR = 22.3 °C
c
CR = 1.7 °C
d
CR = 1.5 °C
Cp kJ mol-1 K-1
[Ligand] µM
Temperature °C

Supplementary Figure 3.3. Effects of scan rate and continuously-varying ligand conversion kinetics on thermolabile ligand DSC profiles. (a) Simulated DSC profiles with 1 °C min^{-1} scan rate and slow ligand conversion at low temperatures (E_a = 95.9 kJ mol^{-1}). (b) Simulated DSC profiles with 1 °C min^{-1} scan rate and moderate ligand conversion at low temperatures (E_a = 89.0 kJ mol^{-1}). (c) 0.005 °C min^{-1} scan rate and slow ligand conversion at low temperatures (E_a = 95.9 kJ mol^{-1}). (d) 1 °C min^{-1} scan rate and rapid ligand conversion at low temperatures (E_a = 81.0 kJ mol^{-1}). In (a-d), simulated DSC profiles where the concentration of cocaine is fixed (i.e. using the optimized parameters from fits of experimental data to our model) through each transition are shown as colored circles while data simulated with continuously-varying ligand concentrations are shown as solid curves. First and last scans are shown in dark red and dark blue. Gaussian noise was added to the simulated scans using the standard deviation of the high temperature experimental baseline. Concentrations of ligand 1 and 2 as a function of temperature are shown as red and blue lines respectively in the right-hand panels. First order kinetics in (a-d) were calculated with a pre-exponential factor estimated from literature rate constants ($A = 7.51\times10^{10}$ s^{-1})[10]. The conversion ratios (CR) given in the upper left corners of (a-d) were calculated as CR = scan rate/$k(T_{m,app.})$. Values of CR below ~20 °C give distortions of the first scan's shape and violate the assumption that most of the ligand conversion happens at high temperatures. Subsequent scans superimpose as the initial ligand is depleted entirely within the first scan.

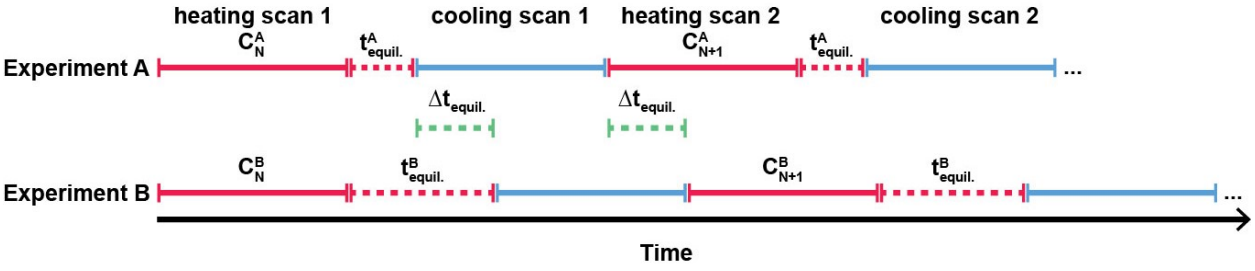


Supplementary Figure 3.4. Protection of the ligand by the biomolecule. (a) Simulated DSC profiles at 1 °C min^{-1} scan rate for continuously-varying ligand conversion in the absence (dark red line) and presence (dark red circle) of protection of the ligand in the biomolecule binding pocket. (b) DSC profiles simulated at 0.005 °C min^{-1} scan rate for continuously-varying ligand conversion in the absence (dark red line) and presence (dark red circle) of protection of the ligand in the biomolecule binding pocket. If the ligand can be protected by the biomolecule and the ligand is in excess, the rate constant for ligand conversion is given by $k_{app.}(T) = k(T)P_{free}(T)$. Ligand concentrations for the protected (dashed red line) and unprotected (red line) cases in (a) and (b) are shown in the panels immediately to the right. Rate constants for ligand conversion were calculated using E_a = 95.9 kJ mol^{-1} and $A = 7.51\times10^{10}$ s^{-1}.



Supplementary Figure 3.5. Simulation of thermolabile ligand binding scenarios**.** (a) Conversion of initial bound ligand to a form with negligible affinity for the aptamer (i.e. decrease in total bound ligand concentration). ΔH^{B1F} = -61.4 kJ mol^{-1}, ΔS^{B1F} = -108.1 J mol^{-1} K^{-1}. ΔH^{B2F} = -2.0 kJ mol^{-1}, ΔS^{B2F} = 16.0 J mol^{-1} K^{-1}. (b) Conversion of the initial ligand to a weaker binding form. ΔH^{B1F} = -61.4 kJ mol^{-1}, ΔS^{B1F} = -108.1 J mol^{-1} K^{-1}. ΔH^{B2F} = -14.0 kJ mol^{-1}, ΔS^{B2F} = 16.0 J mol^{-1}

K^{-1}. (c) Conversion of the initial ligand to a tighter binding form. ΔH^{B1F} = -61.4 kJ mol^{-1}, ΔS^{B1F} = -108.1 J mol^{-1} K^{-1}. ΔH^{B2F} = -68.4 kJ mol^{-1}, ΔS^{B2F} = -103.0 J mol^{-1} K^{-1}. In each simulated profile, the [aptamer] = 83 µM, $[ligand]_{initial}$ = 778 µM, ΔH^{UF} = 271.3 kJ mol^{-1}, ΔS^{UF} = 824.2 J mol^{-1} K^{-1}, $\Delta C_p^{UF} = 0$, ΔC_p^{B1F} = -1.5 kJ mol^{-1} K^{-1}, ΔC_p^{B2F} = -2.2 kJ mol^{-1} K^{-1}, and $k_{conversion} = 5\times10^{-3}$ s^{-1}. Ligand was assumed to convert as a first order process at high temperature for 120 seconds. Heat capacity baselines were calculated as $12.8 + 0.292(T - T_0) - 0.0022(T - T_0)^2$. Gaussian noise was added to the simulated profiles using the standard deviation of horizontal high temperature experimental baselines. Simulated DSC scans are shown as colored lines, where dark red and dark blue indicate first and last scans respectively. Populations for each ligand binding scenario are shown in the panels immediately to the right. The populations of initial ligand bound, converted ligand bound, and folded states are shown as orange, green, and blue dashed lines respectively. Populations of the unfolded state are shown as grey solid lines.



Supplementary Figure 3.6. Time evolution of two DSC experiments with different high temperature equilibration periods. Heating and cooling scans are shown by solid red and blue increments respectively. Experiment A and B have shorter and longer high temperature equilibration times respectively, shown as dashed red increments. The difference in equilibration times is indicated by dashed green increments.

3.9. Supplementary Tables

Supplementary Table 3.1. Cocaine concentrations extracted from global analysis of the cocaine-added MN4 datasets assuming benzoylecgonine can bind the aptamer.

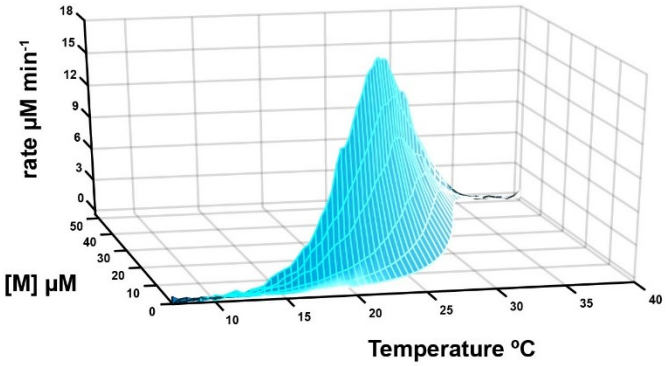
Scan number	[Cocaine] µM 120 second equilibrations	[Cocaine] µM 600 second equilibrations
1	778.0±19.8	778.0±19.8
2	361.3±4.0	250.3±4.8
3	159.7±2.7	68.3±3.3
4	57.4±2.1	13.3±2.8
5	16.3±1.8	3.1±2.5
6	1.3±1.2	2.1±2.3
7	0.9±1.0	3.2±2.5

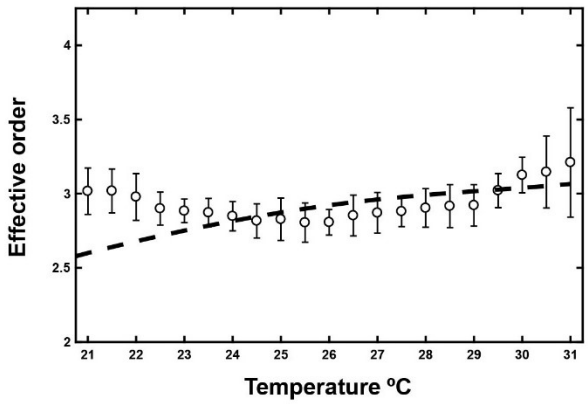
3.10. References

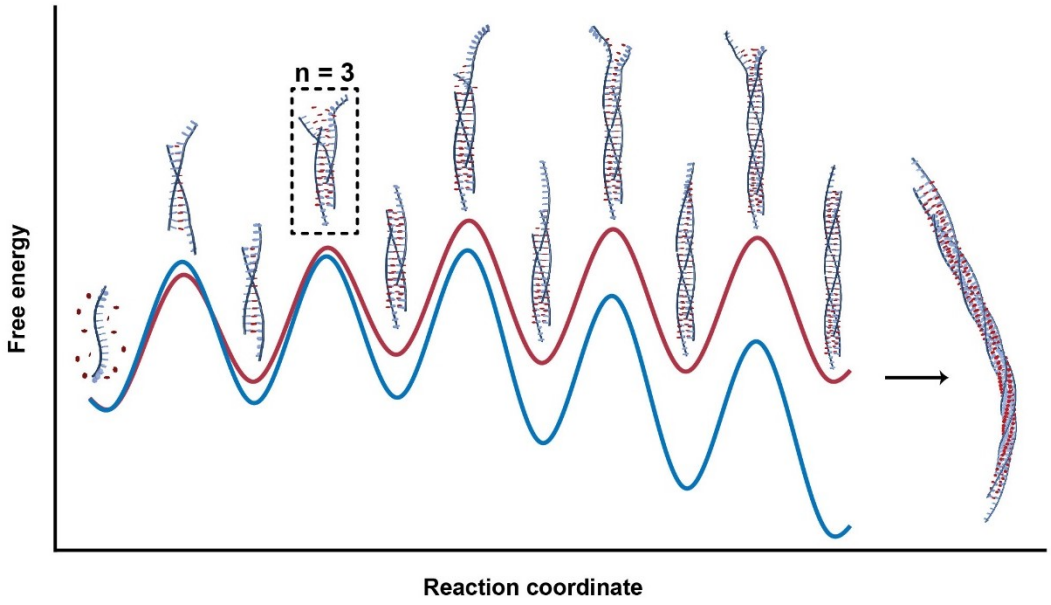
1. Bruylants, G., Wouters, J. & Michaux, C. Differential scanning calorimetry in life science: thermodynamics, stability, molecular recognition and application in drug design. *Curr Med Chem* **12**, 2011-2020 (2005).
2. Privalov, P.L. & Dragan, A.I. Microcalorimetry of biological macromolecules. *Biophys Chem* **126**, 16-24 (2007).
3. Brandts, J.F. & Lin, L.N. Study of strong to ultratight protein interactions using differential scanning calorimetry. *Biochemistry* **29**, 6927-6940 (1990).
4. Harkness, R.W., Slavkovic, S., Johnson, P.E. & Mittermaier, A.K. Rapid characterization of folding and binding interactions with thermolabile ligands by DSC. *Chem Commun* **52**, 13471-13474 (2016).
5. Garbett, N.C. & Chaires, J.B. Thermodynamic studies for drug design and screening. *Expert Opin Drug Dis* **7**, 299-314 (2012).
6. Holdgate, G.A. & Ward, W.H.J. Measurements of binding thermodynamics in drug discovery. *Drug Discov Today* **10**, 1543-1550 (2005).
7. Plotnikov, V. et al. An autosampling differential scanning calorimeter instrument for studying molecular interactions. *Assay Drug Dev Techn* **1**, 83-90 (2002).
8. Schon, A., Lam, S.Y. & Freire, E. Thermodynamics-based drug design: strategies for inhibiting protein-protein interactions. *Future Med Chem* **3**, 1129-1137 (2011).
9. L. Periánez Parraga, A.G.-L., I. Gamón Runnenberg, R. Seco Melantuche, O. Delgado Sánchez, F. Puigventós Latorre Thermolabile Drugs. Operating Procedure In the Event of Cold Chain Failure. *Farmacia Hospitalaria* **35**, 1-28 (2011).
10. Murray, J.B. & Alshora, H.I. Stability of Cocaine in Aqueous-Solution. *J Clin Pharmacy* **3**, 1-6 (1978).
11. Waterman, K.C. et al. Hydrolysis in pharmaceutical formulations. *Pharm Dev Technol* **7**, 113-146 (2002).
12. Mergny, J.L. & Lacroix, L. Analysis of thermal melting curves. *Oligonucleotides* **13**, 515-537 (2003).
13. Neves, M.A., Reinstein, O. & Johnson, P.E. Defining a stem length-dependent binding mechanism for the cocaine-binding aptamer. A combined NMR and calorimetry study. *Biochemistry* **49**, 8478-8487 (2010).
14. Slavkovic, S., Altunisik, M., Reinstein, O. & Johnson, P.E. Structure-affinity relationship of the cocaine-binding aptamer with quinine derivatives. *Bioorg. Med. Chem.* **23**, 2593-2597 (2015).
15. Reinstein, O. et al. Quinine binding by the cocaine-binding aptamer. Thermodynamic and hydrodynamic analysis of high-affinity binding of an off-target ligand. *Biochemistry* **52**, 8652-8662 (2013).

16. Stojanovic, M.N., de Prada, P. & Landry, D.W. Fluorescent sensors based on aptamer self-assembly. *J Am Chem Soc* **122**, 11547-11548 (2000).
17. Stojanovic, M.N. & Landry, D.W. Aptamer-based colorimetric probe for cocaine. *J Am Chem Soc* **124**, 9678-9679 (2002).
18. Tellinghuisen, J. Statistical error propagation. *J Phys Chem A* **105**, 3917-3921 (2001).
19. Peter Atkins, J.d.P. Atkins' Physical Chemistry, Edn. 8th. (Oxford University Press, Great Britain; 2006).
20. Wiseman, T., Williston, S., Brandts, J.F. & Lin, L.N. Rapid Measurement of Binding Constants and Heats of Binding Using a New Titration Calorimeter. *Anal Biochem* **179**, 131-137 (1989).
21. Velazquez-Campoy, A. & Freire, E. Isothermal titration calorimetry to determine association constants for high-affinity ligands. *Nat Protoc* **1**, 186-191 (2006).
22. Drobnak, I., Vesnaver, G. & Lah, J. Model-based thermodynamic analysis of reversible unfolding processes. *J Phys Chem B* **114**, 8713-8722 (2010).
23. Prislan, I., Lah, J. & Vesnaver, G. Diverse polymorphism of G-quadruplexes as a kinetic phenomenon. *J Am Chem Soc* **130**, 14161-14169 (2008).
24. Toledo-Nunez, C., Vera-Robles, L.I., Arroyo-Maya, I.J. & Hernandez-Arana, A. Deconvolution of complex differential scanning calorimetry profiles for protein transitions under kinetic control. *Anal Biochem* **509**, 104-110 (2016).

Chapter 4: Mapping the energy landscapes of supramolecular assembly by thermal hysteresis







4.1. Preface

Although nucleic acids frequently adopt their folded structure in a one-step intramolecular process (as discussed in previous chapters), many nucleic acids in biology and biotechnology applications assemble from multiple component strands via transient, partly-structured intermediates. The complexity of supramolecular nucleic acids means they often have slow assembly and disassembly kinetics and acquiring detailed information on their assembly mechanisms frequently amount to months of experimental input. The rational design of novel biomaterials based on these structures, and a robust understanding of supramolecular nucleic acid assembly in biological function is hindered by this experimental bottleneck. This chapter discusses a combined model-free and global fitting method developed to harness TH measurements for mapping the assembly pathways of supramolecular nucleic acids in as little as one day. To demonstrate the generality of this approach in analyzing the assembly of supramolecular nucleic acids, the method is applied to the formation of two considerably different structures: a tetramolecular GQ and poly(A)-CA fibers. Importantly, the method can be used to identify rate-determining steps in an assembly pathway, and provides information on the size of the nucleus structure in cooperative supramolecular polymerizations.

4.2. Abstract

Understanding how biological macromolecules assemble into higher-order structures is critical to explaining their function in living organisms and engineered biomaterials. Transient, partly-structured intermediates are essential in many assembly processes but are challenging to characterize. Here we present a simple thermal hysteresis method based on rapid, non-equilibrium melting and annealing measurements that maps the rate of supramolecular assembly as a function

of temperature and concentration. A straightforward analysis of these surfaces provides detailed information on the natures of assembly pathways, offering temperature resolution beyond that accessible with conventional techniques. Validating the approach using a tetrameric GQ, we obtained strikingly good agreement with previous kinetics measurements and revealed temperature-dependent changes to the assembly pathway. In an application to the recently-discovered co-assembly of poly(A) and CA, we show that fiber elongation is initiated when an unstable complex containing three poly(A) monomers acquires a fourth strand.

4.3. Introduction

The non-covalent assembly of biological or biomimetic subunits into large supramolecular structures is critical to the function of living organisms and the creation of novel biomaterials. Supramolecular assemblies are validated drug targets[1, 2], and contain tightly controlled internal structure on the nanometer scale, offering new opportunities in the bottom-up design of functional materials[3, 4]. In general, these large supramolecular structures are too complicated to form in a single step, and instead follow multi-step assembly pathways comprising multiple transient, partly-assembled, intermediate states. The nature of these intermediates and the factors governing their interconversion are critical to understanding biological supramolecular self-assembly and yet remain poorly understood[5, 6]. Assembly intermediates are often unstable and short-lived, thus direct detection is not always possible. Nevertheless, detailed information on assembly pathways and the intermediate states that comprise them can be obtained through careful study of how the overall reaction rate varies with environmental conditions, particularly monomer concentration, $[M]$, and temperature. In particular, for homomeric assembly, an effective reaction order, n, implies that the reaction rate varies as $[M]^n$, and can be related to the molecularity of the rate-determining

transition state[7], or the critical nucleus size for polymerization[8, 9], depending on the system. We find that measuring the effective reaction order as a function of temperature allows one to map the energy landscapes of supramolecular assembly with remarkable detail.

Measurements of self-assembly kinetics typically involve triggering the reaction by rapid mixing[10], temperature jump[11], or flash photolysis[12] and monitoring the accumulation of the assembled product as a function of time, while the temperature is held constant. In order to obtain robust measurements of the reaction order, these experiments must be repeated for different initial values of $[M]$[7]. This series of experiments must then be repeated for each temperature of interest. The costs in time and material are high for this type of analysis, which consequently has been performed on just a handful of systems with only modest temperature sampling[7]. Motivated by the need for new methods to efficiently characterize the pathways of supramolecular assembly, we turned to spectroscopic TH, a simple and rapid technique that had previously been used mainly to measure two-state folding and unfolding rates of biomolecules[13, 14]. This experiment entails measuring a spectroscopic signature of folding or assembly (such as absorbance or ellipticity) while raising and lowering the temperature to cause melting and annealing. The temperature scan rate is chosen to be rapid compared to the length of time needed for the system to relax to equilibrium, such that both folding and unfolding occur out of equilibrium. The populations effectively lag behind the rapidly changing temperature such that the 50% folding point (i.e. the apparent T_m) is reached at a higher temperature than the true T_m on the up-scan and at a lower temperature than the true T_m on the down-scan. The folding and unfolding rates can then be calculated as a function of temperature based on the size of the lag[13]. TH has been widely used to measure unimolecular folding and unfolding[15]. A small number of TH studies have examined multimeric assembly, but in these cases the reaction mechanism was assumed *a priori*[16, 17]. Here

we show that TH experiments have a great and largely untapped potential for *de novo* elucidation of complex supramolecular assembly pathways.

We have developed a novel TH method in which data obtained from several (6 in our study) different scan rates are combined to create a 3D map of reaction rate as a function of both *[M]* and temperature. These surfaces are then analyzed in a two-step procedure. In the first step, effective assembly and disassembly reaction orders (*n* and *m* respectively) are extracted as a function of temperature in a model-free manner. A single set of experiments yields reaction orders across the entire thermal transition, spanning in our case up to 40 degrees, sampled in increments of 0.5 degrees, delivering a level of kinetic detail that is unattainable by conventional methods. In the second step, explicit mechanistic models are constructed, based on the observed reaction orders, and are globally fit to the TH datasets, simultaneously yielding the kinetic and thermodynamic parameters that quantify the supramolecular assembly pathway in terms of interconversion between partly-assembled intermediates. The combined model-free and global fitting approach can be applied to self-assembling systems that follow widely divergent mechanisms, as illustrated below.

4.4. Results

4.4.1. Model-free analysis of TH profiles

Our model-free analysis is based on measuring sets of thermal melting and annealing spectrophotometric profiles with different temperature scan rates. The datasets contain low (assembled) and high (disassembled) temperature baselines bridged by transition regions where assembly and disassembly occur. Faster heating rates push the melting curves to successively higher temperatures as the populations lag further behind their equilibrium values. Conversely,

faster cooling rates push the reannealing curves to successively lower temperatures. Each trace is then used to estimate the fraction of subunits that are dissociated (monomeric), θ_U, as a function of temperature and scan rate (see Section 4.7.6). The rate of monomer release or consumption ($d[M]/dt$) can then be calculated from the slopes of the curves ($d\theta_U/dT$), the temperature scan rate (dT/dt), and the total concentration of subunits (C_T) according to the simple expression[13]

$$\frac{d}{dt}[M] = C_T \frac{d}{dT}\theta_U \frac{dT}{dt}. \tag{4.1}$$

Our method relies on the fact that the set of curves obtained with different scan rates sample multiple $[M]$ and $d[M]/dt$ values at any given temperature within the transition region (Figure 4.1a). These measurements yield the effective reaction orders as follows: If the assembled structure contains N subunits and assembly and disassembly occur with effective reaction orders of n and m, respectively, then the rate of monomer formation (or consumption) is given by

$$\frac{d}{dt}[M] = -k_{on}[M]^n + k_{off} N \left(\frac{C_T - [M]}{N} \right)^m. \tag{4.2}$$

When the degree of hysteresis is small, melting and annealing curves lie close together, both terms on the right-hand side of Equation 4.2 are similar in magnitude, and the values of N, n, and m must be known *a priori* in order to extract values of k_{on} and k_{off}[16, 17]. The situation is considerably simpler when the degree of hysteresis is large. In this case, the first term dominates during the annealing scan and

$$\frac{d}{dt}[M] \approx -k_{on}[M]^n \tag{4.3}$$

a plot of $\log(-d[M]/dt)$ versus $\log([M])$ is therefore linear with a slope of n and y-intercept of $\log(k_{on})$. During the melting scan, the second term dominates and

$$\frac{d}{dt}[M] \approx k_{off} N \left(\frac{C_T - [M]}{N} \right)^m \tag{4.4}$$

a plot of $\log(d[M]/dt)$ versus $\log(C_T - [M])$ is therefore linear with a slope of m and y-intercept of $(1-m)\log(N) + \log(k_{off})$. The values of n and m thus obtained provide model-free estimates of the reaction orders at each temperature throughout the transitions, while k_{on} and k_{off} are essentially phenomenological constants describing the rate of the reaction. The magnitude of hysteresis required for these approximations may be judged by comparing the slopes of the melting and annealing curves at any given temperature. We consider a ratio of slopes of roughly 3-fold or more between the slowest annealing and melting scan rates in the middle of the annealing transition to be sufficient, although a more rigorous examination of this approximation can be achieved by numerical simulation. Fortunately, the degree of hysteresis can be tuned by changing the scan rates and C_T. Increasing the scan rate and lowering the total concentration of subunits tend to increase hysteresis. Therefore, this approach can be applied to a wide variety of supramolecular assembly processes with careful selection of the experimental conditions.

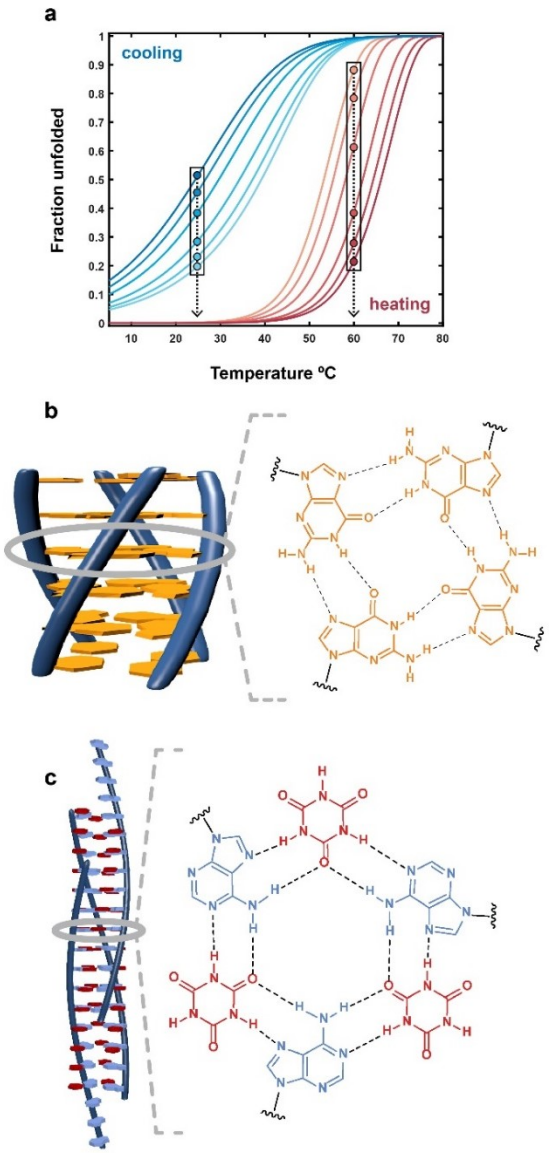


Figure 4.1. Supramolecular assemblies and model-free analysis of multi-scan rate TH datasets. (a) A multi-scan rate TH dataset for generating a 3D supramolecular assembly map. The dashed arrows indicate temperature slices at which monomer concentrations and reaction rates are calculated at each temperature scan rate (colored circles in the black boxes). Light to dark blue and orange to dark red indicate slow to fast scan rates respectively. (b) Tetrameric GQ structure formed by TG_4T. The tetramolecular structure (left) contains stacked, HG-hydrogen bonded G-tetrads (right). (c) Fiber structure formed by co-assembly of CA and poly(A) strands (left). CA brings about the growth of nanofibers from poly(A) strands by participating in hexameric rosette arrays with A residues (right).

4.4.2. Assembly of a tetrameric DNA GQ

To test the TH method, we applied it to the well-studied tetrameric DNA GQ TG_4T (Figure 4.1b), which is believed to fold via a pathway involving small populations of partly assembled intermediates[7]. We used spectroscopic absorbance measurements to determine the fraction of unfolded DNA strands as the temperature was raised and lowered at rates varying from 0.2 to 2 °C min^{-1}. (Figure 4.2a, Supplementary Figure 4.1, Supplementary Figure 4.2, Supplementary Figure 4.3a,b, and Sections 4.7.5 and 4.7.6 for details of baseline and temperature correction). Equation 4.1 was then applied to map assembly (cooling) and disassembly (heating) rates as a function of temperature and monomer concentration (Figure 4.2b). Slices through this landscape perpendicular to the temperature axis yield reaction rates as a function of [*M*] at constant temperature. Log-log plots were constructed (Figure 4.2c,d) yielding linear correlations, as predicted by Equations 4.3 and 4.4. The level of agreement is remarkable, as each point in the graph is obtained from a separate melt with a different temperature scan rate. To our knowledge, this is the first time such an analysis method is applied to TH data, and the high degree of linearity gives us confidence in the analysis that follows. The slopes of the plots correspond to the effective reaction orders of assembly and disassembly sampled as a function of temperature (Figure 4.2e). Assembly reaction orders, *n*, are roughly 2.75 at 5 °C and gradually rise to about 3.5 at 45 °C, while disassembly orders, *m*, are steady near 1.2 from 65 down to ~55 °C. Perfectly simultaneous assembly of all four strands would produce assembly reaction orders of exactly 4. These results therefore imply that the dominant energy barriers for assembly of this GQ are crossed by intermediates with fewer than four strands. A similar temperature-dependent increase in *n* from roughly 3 to 4 was previously reported[7], which is notable as that study employed a series of variable-concentration isothermal folding reactions monitored by NMR spectroscopy. Our results

are thus in excellent agreement with those obtained from a completely orthogonal methodology. Furthermore, it must be emphasized that the total experiment time of the previous study was on the order of months and sampled only 6 temperatures, while our data set was obtained in 24 hours and sampled orders at 80 different temperatures.

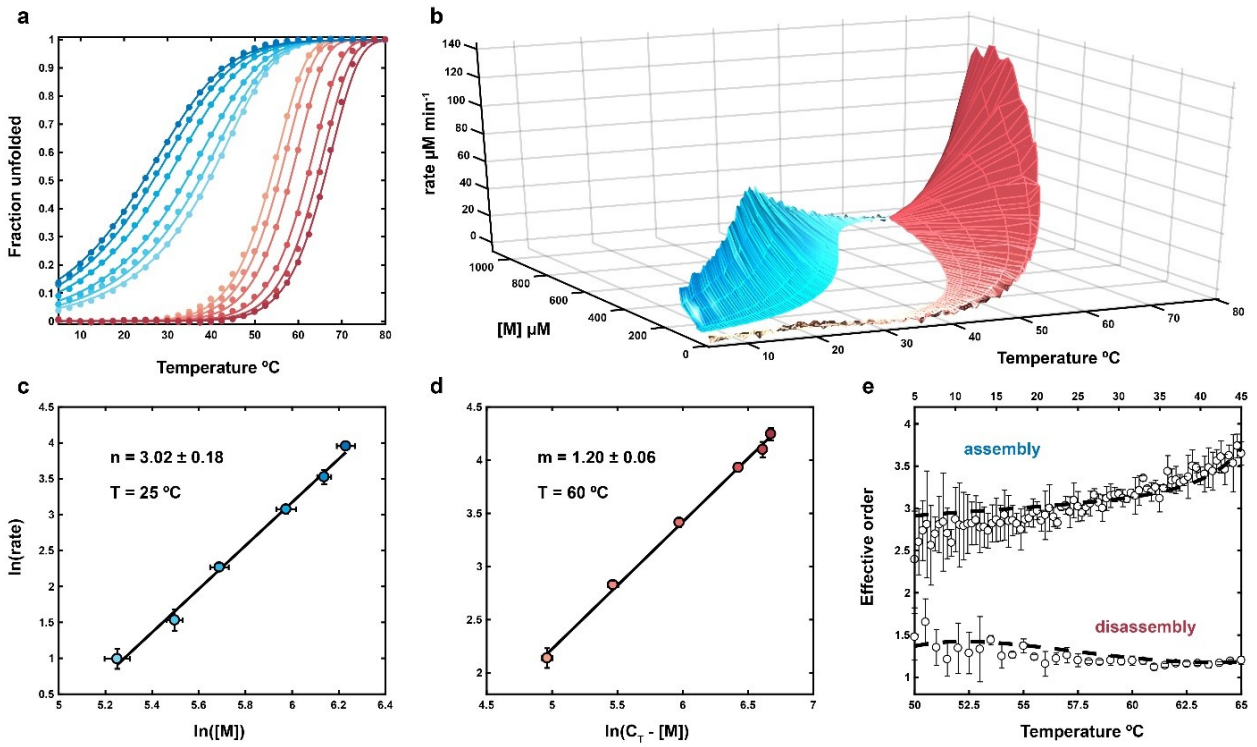


Figure 4.2. Mapping the energy landscape of TG_4T assembly by TH. (a) Multi-scan rate fraction unfolded TH profiles for TG_4T assembly and disassembly. Colored points and lines are experimental data and the globally-fitted step-wise monomer association model respectively. Only every 5th experimental point is shown for clarity. (b) 3D temperature-concentration-rate supramolecular assembly map calculated from the experimental data in (a). The monomer reaction rates (shown as absolute values) increase with faster scanning. (c) Isothermal slices from (b) at 25 °C plotted as ln(assembly rate) vs. ln([*M*]). The effective assembly order *n* is the slope of the line. (d) Isothermal slices from (b) at 60 °C plotted as ln(disassembly rate) vs. ln(C_T-[*M*]). The effective disassembly order *m* is the slope of the line. (e) Effective TG_4T assembly and disassembly reaction orders from model-free analysis of the surface in (b) as a function of temperature through the annealing and melting transitions. The top and bottom temperature axes are for assembly and disassembly respectively. White circles and dashed black lines correspond to effective orders from experimental data and the globally fit step-wise monomer association model respectively. In (a-d), light to dark blue and orange to dark red corresponds to slow to fast annealing and melting scan rates respectively. In (c-e), error bars are the standard deviations of the values obtained from analysis of three replicate TH datasets.

We suspected that the observed variation of the assembly reaction order is likely due to temperature-dependent shifts in the energetic barriers along the assembly pathway. To test whether this hypothesis is consistent with the data, we simulated TH curves using kinetic models with explicit assembly intermediates and examined the extent to which they could reproduce the experimental datasets and effective reaction orders (see Section 4.7.7, Figure 4.3). A one-step *monomer* ↔ *tetramer* model gave very poor agreement with the TH data (Supplementary Figure 4.4a), as expected from the extracted experimental values of $n < 4$. The assembly of TG_4T and other similar tetramolecular GQs has been proposed to follow either step-wise *(monomer* ↔ *dimer* ↔ *trimer* ↔ *tetramer*)[7] or dimer-of-dimers type (*monomer* ↔ *dimer* ↔ *tetramer*)[18] mechanisms. Both models gave generally good agreement with the raw data and their corresponding intermediate populations never reached more than ~5%, consistent with the effectively two-state assembly previously observed for TG_4T (Supplementary Figure 4.4b,c). However, the step-wise model fit substantially better than the dimer-of-dimers model (roughly 1.4-fold in terms of residual sum of squares). According to the Akaike Information Criterion[19], the relative likelihood of the dimer-of-dimers model being correct is <0.01% and therefore the step-wise model is preferred. The simulated melting/annealing curves and reaction orders for the step-wise model are shown in (Figure 4.2a,e) and agree closely with both experimental TH data and extracted reaction orders. The extracted rate constants are physically reasonable: the dimer intermediate is highly unstable at 45 °C, with an equilibrium dissociation constant, K_D, of ~17 M, and forms very slowly with a kinetic association constant of only 300 M^{-1} min^{-1}. Addition of a third strand is more favourable and occurs more rapidly, with a K_D of ~2 μM and association rate constant of ~2×10^5 M^{-1} min^{-1}, similar to the association kinetics of intermolecular triplex DNA[20]. A simple and realistic physical kinetic model is thus consistent with the temperature dependence of the reaction orders obtained

from our model-free analysis, and the values of the extracted step-wise rate constants given in Table 4.1 provide quantitative insight into the nature of the assembly pathway.

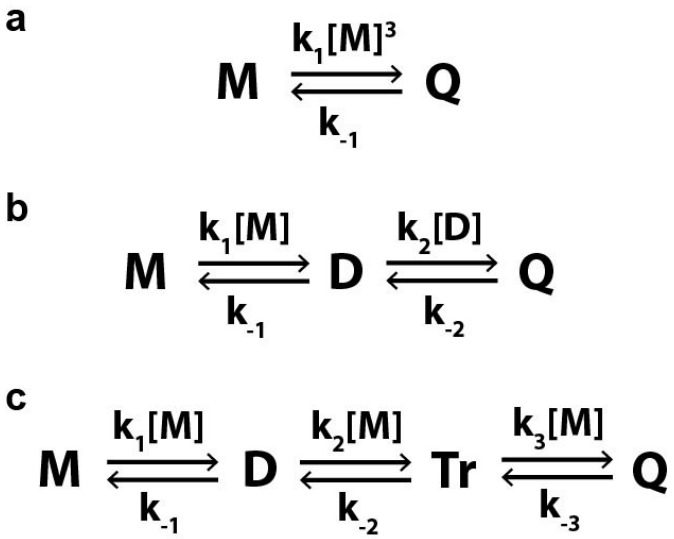


Figure 4.3. TG_4T assembly models. (a) One-step assembly. (b) Dimer-of-dimers assembly. (c) Step-wise monomer association. M, D, Tr, and Q correspond to monomer, dimer, trimer, and tetrameric GQ respectively.

Table 4.1. TH global fit parameters for TG_4T assembly with the step-wise monomer association model. Activation energies are given in kcal mol^{-1}. Rate constants are given at the reference temperature of 45 °C and in M^{-1} min^{-1} and min^{-1} for forward and reverse steps respectively. Errors were calculated according to the variance-covariance method[21].

	Activation energies		**Rate constants**
E_1	-5.4±0.8	k_1	(3.0±0.3) × 10^2
E_{-1}	14.4±0.4	k_{-1}	(5.0±1.0) × 10^3
E_2	-4.0±1.2	k_2	(1.6±0.2) × 10^5
E_{-2}	15.9±0.4	k_{-2}	(3.1±0.2) × 10^{-1}
E_3	-3.8±0.7	k_3	(8.2±0.5) × 10^2
E_{-3}	37.4±0.2	k_{-3}	(8.4±0.1) × 10^{-3}

4.4.3. Co-polymerization of poly(A) and CA

It was recently discovered that short poly(A) chains co-assemble with CA to form long fibers[22]. A cross-section of the proposed structure (Figure 4.1c) shows three A residues from different DNA strands hydrogen bonded to three CA molecules and forming stacked, planar, hexameric rosettes perpendicular to the fiber axis. This system provides a different type of challenge for the TH method than does the assembly of the tetrameric GQ. Rather than identifying specific intermediates formed on route to a well-defined final structure, characterizing poly(A) fiber formation corresponds to elucidating the supramolecular polymerization mechanism i.e. determining how individual poly(A) chains initiate and add to indefinitely growing fibers. We performed TH measurements of poly(A) fiber formation in the presence of excess CA. Heating scans produced identical curves regardless of the scan rate, indicating that dissociation occurs too rapidly at these temperatures to characterize using this method. In contrast, the cooling scans showed a pronounced scan rate dependence (Figure 4.4a, Supplementary Figure 4.3c,d) and were subjected to further analysis. The rate of unfolded poly(A) consumption was calculated as a function of temperature and concentration (Figure 4.4b) and the resulting log-log plots (Figure 4.4c) were linear. The apparent reaction orders for assembly were calculated as the slopes of the plots, yielding values close to 3 (Figure 4.4d). It must be noted that apparent reaction orders for polymerization reactions do not reflect a single rate-limiting barrier as monomers are consumed by adding to an ensemble of nascent fibers of different lengths[8]. Effective reaction orders can nevertheless yield mechanistic insight and can be related to the molecularity of the nucleus[8, 9], as discussed below (see Section 4.7.9 and 4.7.10).

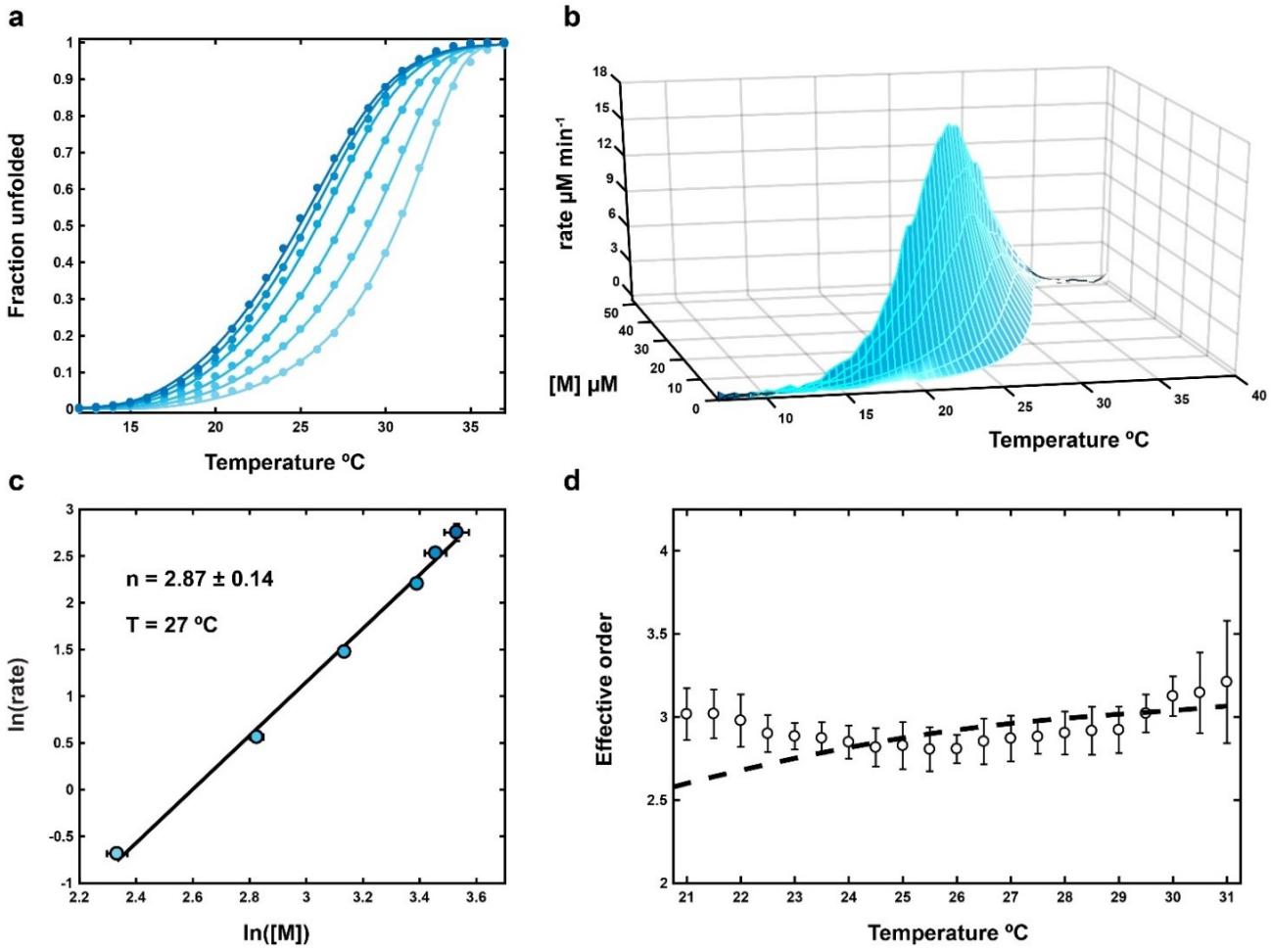


Figure 4.4. Mapping the energy landscape of poly(A) fiber assembly by TH. (a) Fraction unfolded TH profiles for poly(A) fiber assembly as a function of temperature scan rate. Colored points are the experimental data and colored lines correspond to the globally fit Goldstein-Stryer model for cooperative supramolecular assembly assuming a nucleus size of 3. Only every 2nd experimental point is shown for clarity. (b) 3D supramolecular assembly map for poly(A) fiber assembly. (c) Model-free analysis of the map in (b) at 27 °C. The effective assembly order *n* is the slope of the line. (d) Effective poly(A) fiber assembly orders as a function of temperature through the annealing transition. White circles and dashed black lines are the effective orders obtained from model-free analysis of the experimental and globally fitted data respectively. In (a-c), light to dark blue corresponds to slowest and fastest annealing scan rates respectively. In (d), error bars for the experimental points were taken as the standard deviation of the values from analysis of three replicate TH experiments.

We tested whether a simple kinetic model could account for the temperature dependent annealing curves and reaction orders by fitting the Goldstein-Stryer assembly model[8] directly to the experimental data (Figure 4.5). This model explicitly tracks the populations of all oligomers up to a certain number of monomer units (100 in Figure 4.4a, see Section 4.7.8), while populations of longer fibers were accounted for using the approximation of Korevaar *et al*[10, 23]. Association and dissociation of monomers and short oligomers less than the critical nucleus size, s (where s is the number of poly(A) strands in our case), were described by the nucleation rate constants k_{n+} and k_{n-} respectively, while oligomers larger than s were described with the elongation rate constants k_{e+} and k_{e-}. For the critical nucleus itself, the association rate constant was taken as k_{e+}, while dissociation was taken as k_{n-}. We held the forward rate constants as equal, $k_{n+} = k_{e+}$, a simplification previously applied in other systems[8-10, 24], and allowed the assembly activation enthalpies to vary with temperature (i.e. $\Delta C_p \neq 0$)[25].

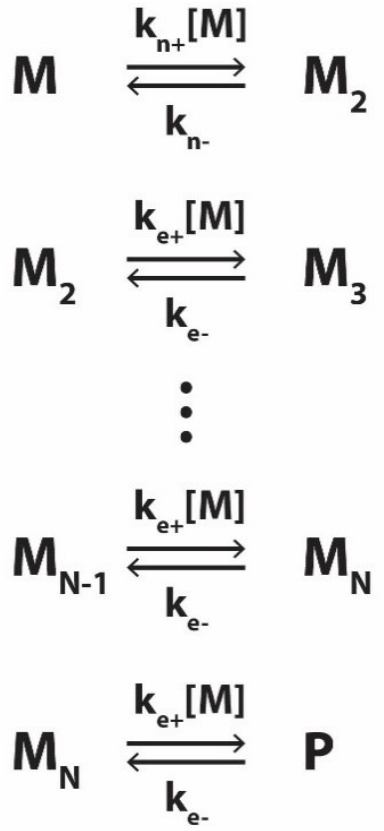


Figure 4.5. The Goldstein-Stryer model for cooperative self assembly. The monomer (M) associates in a step-wise manner to form a nucleus of a defined size which then elongates to give large assemblies. The two regimes are defined by nucleation and elongation rate constants. The case for a nucleus size of 2 is shown here. We allowed post-nucleus oligomers to elongate up to an explicitly described size of *N*, beyond which they are treated as a fibril pool (P) according to the approximation by Korevaar *et al*[10, 23].

We applied the model and systematically varied the value of *s* from 1 to 7 (a nucleus of 1 corresponds to non-cooperative assembly) to find the optimal nucleus size (Supplementary Figure 4.5). Excellent fits were obtained with nucleus sizes of 2-4, with substantial worsening of the fit quality below or above these nucleus sizes. While nucleus sizes of 2-4 are all physically realistic for poly(A) fiber assembly, the best fit was obtained with a nucleus of 3 and therefore this is our preferred nucleus size. The forward rate constants of $k_{n+} = k_{e+} \approx 7\times10^4$ M $^{-1}$ min^{-1} (Table 4.2) are

somewhat slower than those of duplex DNA, which is to be expected, given that each poly(A) strand joining the growing fiber must simultaneously organize a column of CA molecules along the interface. The trimeric nucleus is relatively unstable at 25 °C, with $K_D \approx 100$ μM, compared to a total poly(A) concentration of 50 μM, meaning that it is never more populated than the monomeric state under these conditions. The dissociation equilibrium constant for each subsequent poly(A) strand is much more favourable ($K_D \approx 1$ μM), meaning that fibers spontaneously elongate at poly(A) concentrations above this value. Thus, the full set of TH data for poly(A) fiber assembly agrees quantitatively with a simple, realistic, model of supramolecular assembly.

Table 4.2. TH global fit parameters for poly(A) fiber assembly using the Goldstein-Stryer model with a nucleus size of 3. Activation energies are given at the reference temperature of 25 °C in kcal mol^{-1}. Activation heat capacities are given in kcal mol^{-1} K^{-1}. Rate constants are given at the reference temperature of 25 °C and in M^{-1} min^{-1} and min^{-1} for forward and reverse steps respectively. Errors were calculated according to the variance-covariance method[21].

Activation energies		**Rate constants**		**Heat capacities**	
$E_{n+} = E_{e+}$	14.5±1.1	$k_{n+} = k_{e+}$	$(6.7\pm0.2) \times 10^4$	$\Delta C_p^{\ddagger}{}_{n+} = \Delta C_p^{\ddagger}{}_{e+}$	5.2±0.3
E_{n-}	96.4±2.0	k_{n-}	6.8±0.3	$\Delta C_p^{\ddagger}{}_{n-}$	5.5±0.5
E_{e-}	133.0±0.8	k_{e-}	$(5.8\pm0.4) \times 10^{-2}$	$\Delta C_p^{\ddagger}{}_{e-}$	0.5±0.2

4.5. Discussion

We have shown that a simple analysis of multiple-scan rate TH data yields reaction orders for supramolecular assembly over a broad range of temperatures. This approach is model-free in the sense that no assumptions regarding the populations and interconversion rates of partly-assembled intermediate forms are necessary. A multi-scan rate TH dataset can be collected in a few hours with a small amount of material and the extraction of reaction orders is straightforward

and can be achieved with standard spreadsheet software. This protocol thus brings kinetic information into easy reach at a level of detail that is not readily obtainable from existing methods.

Likely due to the current scarcity of reaction order data for supramolecular assembly, there has not been much discussion of how values of *n* and *m* relate to the underlying pathways. It is therefore useful to examine in more detail how the model-free reaction orders extracted for the tetrameric GQ and poly(A) fibers relate to the energy surfaces predicted by direct model fitting to the TH curves. In the case of GQ assembly, the experimental values of *n* are approximately 2.75 at 5 °C, rising to over 3.5 at 45 °C. The reaction energy diagrams predicted by the step-wise model for 5 and 45 °C are shown in Figure 4.6a. The heights of the energy barriers correspond directly to the transition probabilities, where larger barriers indicate fewer molecules traversing the barrier in a given direction per unit time. At 5 °C, the first (*monomer* ↔ *dimer*) barrier is larger than the second (*dimer* ↔ *trimer*) and third (*trimer* ↔ *tetramer*) barriers, and the experimental reaction order (2.75) is closer to the molecularity of first transition state (2) than it is to that of the third (4). At 45 °C, the second and third barriers are slightly higher than the first and the observed reaction order (3.5) moves closer to the molecularity of the third transition state. We note that differences in effective order are also somewhat influenced by differences in monomer concentration at different temperatures, but changing barrier heights are the main factor (see Section 4.7.11, Supplementary Figure 4.6). In the case of disassembly reaction orders, *m*, the values are all ~1, since the dominant barrier is located at the *tetramer* → *trimer* transition and the corresponding transition state has the same molecularity as the fully folded GQ. The temperature-dependent effective reaction orders thus reveal detailed information on the locations and sizes of energetic barriers along the reaction pathway.

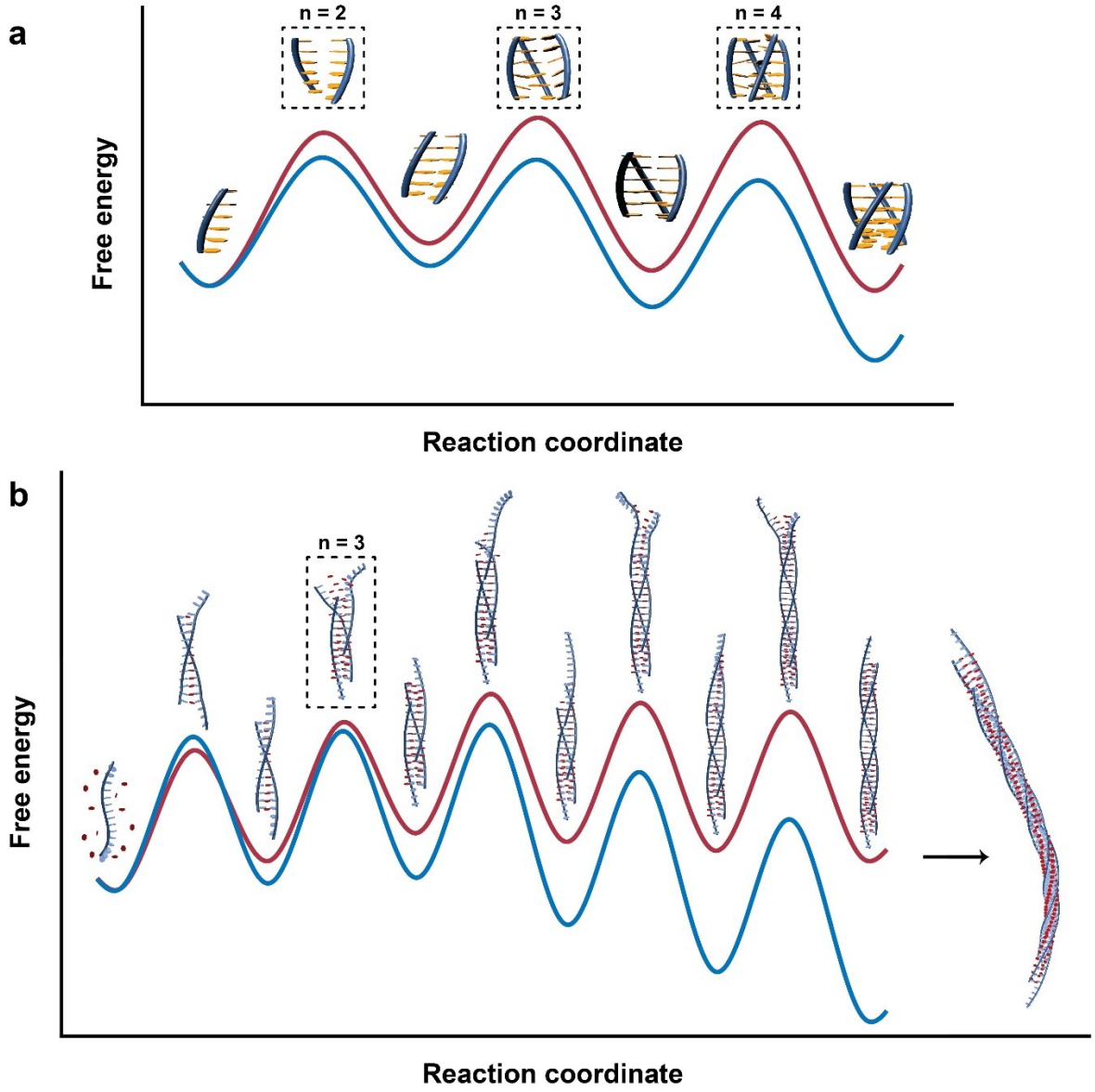


Figure 4.6. Quantitative free energy diagrams for supramolecular assembly by TH. (a) TG_4T assembly at 45 (red) and 5 (blue) °C. At 5 °C (experimental $n \approx 2.75$), the dimer barrier dominates. The trimer and tetramer barriers become dominant at 45 °C (experimental $n \approx 3.5$). The limiting cases of n = 2, 3, and 4 are indicated by the dimeric, trimeric, and tetrameric transition state structures enclosed in dashed black boxes. (b) Poly(A) fiber formation at 30 (red) and 20 (blue) °C assuming a nucleus size of 3 poly(A) strands. Elongation is driven by addition of a fourth poly(A) monomer. The reaction coordinate is truncated at the hexamer. Longer fibers form from step-wise association of monomers as indicated by the black arrow.

For poly(A) fiber formation, the reaction energy diagram predicted by the Goldstein-Stryer assembly model with a nucleus size of 3 is shown in Figure 4.6b. The least stable state is the trimeric nucleus, while the addition of each subsequent poly(A) chain produces a successively more stable oligomer. This matches the proposed structure of the fiber, which requires a minimum of three strands to complete the rosette arrangement, and suggests that the addition of a fourth strand effectively stabilizes the nascent fiber in an arrangement that is primed to bind to additional chains. We note that, while in this case, the effective order of the reaction (≈ 3) matches the molecularity of the trimeric nucleus, this relationship does not necessarily hold for polymerization reactions in general. Monomers are consumed at each step of the assembly process and elongation continues indefinitely so no single energy barrier is completely rate-determining. Nevertheless, the effective order of a polymerization reaction can be quantitatively interpreted in terms of the fluxes of the individual steps. The net rate at which the *N*-mer oligomer binds monomers to produce (*N*+1)-mers is given by

$$\Phi_N = k_{on} c_1 c_N - k_{off} c_{N+1} \tag{4.5}$$

where c_N is the concentration of the *N*-mer, and k_{on} and k_{off} are the appropriate association and dissociation rate constants. The total rate of monomer consumption, *R*, is then

$$R = -\frac{\partial}{\partial t} c_1 = \sum_{N=1}^{\infty} \Phi_N \,. \tag{4.6}$$

It can be shown (see Section 4.7.9) that the effective order, *n*, of monomer consumption is given by

$$n = \frac{\partial \ln(R)}{\partial \ln(c_1)} = \sum_{N=1}^{\infty} \frac{\Phi_N}{R} \frac{\partial \ln(\Phi_N)}{\partial \ln(c_1)} \,. \tag{4.7}$$

In other words, the effective order of the polymerization reaction is given by the weighted average of the orders of the individual fluxes, where the weight of each term is simply the relative contribution of the individual flux to the total rate, Φ_N/R. According to the Goldstein-Stryer model applied to poly(A) fiber assembly, the major fluxes for monomer consumption involve dimers up to about 10-mers and have orders ranging from about 2 to 5, with a weighted average of roughly 3. The order of each flux $\partial\ln(\Phi_N)/\partial\ln(c_1)$ is largely governed by how the steady-state concentration of the N-mer varies with monomer concentration, and on the magnitude of the depolymerization rate $k_{off}c_{N+1}$ (see Section 4.7.9).

We have simulated sequential polymerization reactions according to the Goldstein-Stryer model and find that, without changing rate constants, larger nuclei produce larger effective reaction orders (Supplementary Figure 4.7). Thus the effective reaction orders provide information on the size of the assembly nucleus, s. This is particularly true for canonical nucleated assembly, where the concentration of fibers scales as $[M]_0^{(s+1)/2}$ and the rate scales as $[M]_0^{(s+3)/2}$, giving an apparent reaction order of (s+3)/2, with respect to the initial monomer concentration, $[M]_0$, in isothermal annealing reactions[8]. For the sake of comparison, we have simulated TH data for canonical nucleated assembly. TH experiments are quite different from isothermal annealing reactions since the temperature varies throughout the measurement leading to fiber accumulation that varies with scan rate. Nevertheless, we find empirically similar relationships such that the concentration of fibers scales approximately as $[M]^{(s-1)/2}$ and the polymerization rate scales as $[M]^{(s+1)/2}$, giving an apparent reaction order of (s+1)/2, where in this case $[M]$ is the actual (instantaneous) monomer concentration (see Section 4.7.10, Supplementary Figure 4.8). We note that for poly(A) assembly this empirical relationship would predict an effective order of 2 while we observe orders closer to 3, but poly(A) assembly does not meet the criteria for a canonical nucleated mechanism so the lack

of agreement is unsurprising. While we find that effective reaction orders obtained from TH data are information rich and closely linked to the sizes of critical nuclei for self-assembly, the precise relationship is complex and would be an interesting area for further theoretical study.

Our TH-based approach is applicable to many different types of supramolecular self-assembly systems and thus represents a general approach for studying these complex processes. The main requirements are that the reaction is reversible and temperature-controlled, and that the degree of self-assembly correlates with a real-time observable such as spectroscopic absorbance or ellipticity. Nucleic acids are particularly amenable, as illustrated by the results presented here. There is growing interest in understanding nucleic acid self-assembly in molecular biology[26] and biotechnology[3], providing many interesting opportunities for application of this method. Furthermore, biological and biomimetic systems such as collagen fibers[16, 27], SNARE proteins[28], ganglioside micelles[29], viral capsids[30], peptide amphiphiles[31], elastin-mimetic peptides[32, 33], and DNA ribbons[34], along with many others[35] exhibit TH in temperature-driven assembly and represent excellent candidates for TH-based analysis of their assembly mechanisms. In addition, this approach is equally well applicable to non-biological assembly processes, such as rod formation by trisurea disks[36]. A complete TH dataset can be acquired very rapidly, in as little as a single day, compared to weeks or months for comparable existing methods[7, 8]. Interestingly, when melting/annealing kinetics are slow, a TH dataset can be measured in much less time than an equilibrium melting experiment, as there is no need to allow the system to fully equilibrate at each temperature. Extracting the model-free effective reaction orders and rate constants (Equations 4.1-4.4) is easily accomplished without specialized software. The reaction orders can then be interpreted in terms of the molecularities of the highest energy transition states for discrete assembly, or in terms of nucleus size for polymerization reactions. Subsequently, direct fitting of

explicit mechanistic models to the TH data provides detailed insight into the stabilities and interconversion rates of individual assembly intermediates, even those that are very weakly populated and short-lived.

4.6. Conclusions

Information on partly-formed intermediates is critical to understanding, and ultimately controlling, macromolecular assembly processes. The growing interest in this challenging problem has led to the development of a variety of biophysical approaches. Many of these focus on direct observation of assembly intermediates. When assembly is extremely slow, intermediates may be sufficiently long-lived for direct structural analysis. For instance, Aβ oligomeric precursors to amyloid fibril formation represent the dominant species after several days of incubation for some variants, and were recently characterized by solid-state NMR and IR spectroscopy[37]. Alternatively, partly-assembled intermediates can be distinguished from monomers and fully-formed structures by single-molecule methods. For instance cryo-EM and AFM were used to identify and characterize partly assembled viral capsids on the basis of shape, and to track their abundance as a function of time[38]. Single-molecule microscopy and spectroscopy can identify individual intermediates on the bases of FRET intensity or diffusion rates[39], and yield information on their populations and lifetimes[40]. NMR spectroscopy is highly sensitive to millisecond association kinetics and was recently used to dissect a dimer-of-dimers association pathway, giving overall kinetic parameters and identifying interaction surfaces[41]. In contrast with these techniques which require costly specialized equipment, extensive user expertise, and lengthy analysis, multi-scan rate TH analysis can be performed with only a thermally-controlled spectrophotometer and rapidly yields quantitative information on the assembly process in a straightforward manner, in the form

of temperature-dependent reaction orders. Similar data are not readily accessible via existing approaches and are highly complementary to those of the more specialized techniques listed above. TH experiments focus on the relative sizes of the kinetic barriers along the assembly pathway rather than on the properties of individual assembly intermediates which may or may not be on-pathway or kinetically relevant. Used in combination with the techniques mentioned above, they can help to identify which of the intermediates are involved in the rate limiting step(s) of assembly. Furthermore, given the simplicity, speed, and low cost of multi-scan rate TH analysis, it is highly suitable as an initial screening method for optimizing samples and assembly conditions for more detailed study. The approach laid out here thus represents a powerful new tool for better understanding supramolecular assembly.

4.7. Materials and Methods

4.7.1. Materials

CA, tris(hydroxymethyl)aminomethane (Tris), magnesium chloride hexahydrate ($MgCl_2 \cdot 6 H_2O$), sodium cacodylate (NaCaco), sodium chloride (NaCl), glacial acetic acid and urea were used as purchased from Sigma-Aldrich. Boric acid was obtained from Fisher Scientific and used as supplied. Acrylamide/bis-acrylamide (40% 19:1) solution, ammonium persulfate and tetramethylethylenediamine (TEMED) were used as purchased from BioShop Canada Inc. Sephadex G-25 (super fine, DNA grade) was purchased from Glen Research.

Desalted $d(A_{15})$ and $d(TG_4T)$ oligonucleotides were purchased from Integrated DNA Technologies (IDT). $d(A_{15})$ was purified by denaturing polyacrylamide gel electrophoresis (PAGE) (8 M urea, 1xTBE running buffer) and desalted with Sephadex G-25. $d(TG_4T)$ was used without further purification.

1X TBE (Tris-boric acid-EDTA) buffer was composed of 45 mM Tris, 45 mM boric acid and 2 mM EDTA at pH 8.3. 1X AcMg buffer was composed of 40 mM acetic acid, 7.6 mM $MgCl_2 \cdot 6\ H_2O$, with pH adjusted to 4.5. 1X NaCaco buffer was composed of 10 mM NaCaco and 100 mM NaCl (110 mM total Na^+) at pH 7.2. Buffers and samples were prepared with Milli-Q water.

4.7.2. Instrumentation

UV-Vis absorbance-based quantification of $d(A_{15})$ was performed on a Nanodrop Lite spectrophotometer from Thermo Scientific. Quantification of $d(TG_4T)$ was performed on an Agilent Cary 300 UV-Vis spectrometer at 95 °C using a 10 mm path length quartz cuvette. DNA purification by PAGE was carried out on a 20×20 cm vertical acrylamide Hoefer 600 electrophoresis unit.

UV-Vis absorbance studies were performed using a 1 mm path length quartz cuvette on a Jasco-810 spectropolarimeter equipped with a Peltier temperature control unit and a water recirculator. Temperature verification on the instrument was performed with a handheld digital thermometer (Oakton) equipped with a fine gage thermocouple (Omega).

4.7.3. Acquisition of $d(TG_4T)$ TH profiles

Samples contained 1 mM $d(TG_4T)$ in 1X NaCaco buffer. The absorbance signals for annealing and melting were monitored at 295 nm over the 0 °C to 85 °C temperature range at different scan rates (0.2, 0.3, 0.5, 1, 1.5, and 2 °C/min). The rates were selected to ensure good separation between the curves. Samples were maintained at 85 °C for 10 minutes before annealing and at 0 °C for 10 minutes before melting. A layer of silicon oil was applied on top of the sample

solution to minimize evaporation. A stream of nitrogen gas was supplied to the sample chamber to prevent condensation on the cuvette. Curves were obtained in triplicate.

4.7.4. Acquisition of d(A_{15}) TH profiles

Samples contained 50 μM d(A_{15}) and 15 mM CA in 1X AcMg pH 4.5. The absorbance signals for annealing and melting were monitored at 252 nm over the 2 °C to 65 °C temperature range at different rates of temperature change (0.2, 0.5, 1, 2, 3, and 4 °C/min). The rates were selected to ensure good separation between the curves. Samples were maintained at 65 °C for 5 minutes before annealing and at 2 °C for 5 minutes before melting. A layer of silicon oil was applied on top of the sample solution to minimize evaporation. A stream of nitrogen gas was supplied to the sample chamber to prevent condensation on the cuvette with heating. Curves were obtained in triplicate.

4.7.5. Temperature correction

Thermal melting and annealing experiments can be subject to differences between the temperature of the solution in the experimental cuvette and the sample block temperature recorded by the instrument. Furthermore, this temperature difference changes as a function of experimental scan rate and additionally depends on the scanning direction (heating or cooling). Therefore, we measured the cuvette solution temperature with a digital thermocouple during heating and cooling scans as a function of scan rate, finding strongly linear correlations between the solution and block temperatures at all tested rates (Supplementary Figure 4.1). We corrected for the block temperature offset at each scan rate to a first approximation using

$$T_{solution} = mT_{block} + b \tag{4.8}$$

where m is the slope of the temperature correlation, found to be ~0.985 at all tested scan rates (Supplementary Figure 4.2a) and b is the temperature offset, i.e. the solution temperature when the block temperature is equal to 0 °C. We found that b varied linearly with the scan rate (dT/dt) (Supplementary Figure 4.2b), following the empirical relationship

$$b = -0.5578\frac{dT}{dt} + 1.9421. \tag{4.9}$$

TH profiles were subsequently corrected and resampled with linear interpolation (Supplementary Figure 4.2c,d) so that temperature points were identical for all scan rates (5-80 °C for TG_4T and 7-65 °C for poly(A) fibers in in 0.5 °C increments). The temperature corrected, resampled data were used for all analyses herein.

4.7.6. Model-free analysis of TH datasets for generating 3D assembly maps

Scan-rate dependent TH profiles were fit with linear baselines for the assembled (A_F) and monomeric, unfolded (A_U) signals according to

$$A_F(T) = m_F T + b_F \tag{4.10}$$

and

$$A_U(T) = m_U T + b_U \tag{4.11}$$

where m_U, m_F, b_U, and b_F are the unfolded and assembled baseline slopes and intercepts respectively. Using the baselines, the TH profiles were converted to fraction unfolded (θ_U)

$$\theta_U(T) = \frac{A(T) - A_F(T)}{A_U(T) - A_F(T)} \tag{4.12}$$

where *A(T)* are the experimental thermal melting and annealing data. At low and high temperatures $\theta_U(T)$ takes limiting values of 0 and 1 respectively, corresponding to the completely assembled (=0) or completely monomeric (=1) states. The total concentration of nucleic acid in the experimental cuvette C_T is related to the concentrations of the monomeric and assembled states at each temperature by

$$C_T = [M](T) + N[F](T) \tag{4.13}$$

where N accounts for the number of monomers that reside within a folded assembly. The concentration of free monomers at each scan rate were calculated from the fraction unfolded assuming

$$[M](T) = \theta_U C_T . \tag{4.14}$$

The slopes of the monomer concentration with respect to temperature $d[M]/dT$ were calculated numerically using rolling window regression where the derivative of a third-order polynomial fit to the calculated $[M](T)$ in a centered five point moving window is used with the experimental temperature increment of 0.5 °C to calculate the local slope (the movingslope function in MATLAB, https://www.mathworks.com/matlabcentral/fileexchange/16997-movingslope). The rates of change of the monomer concentration were then obtained from the slopes and the scan rate:

$$\frac{d}{dt}[M](T) = \frac{dT}{dt}\frac{d}{dT}[M](T) . \tag{4.15}$$

The choice of polynomial order and window size was not found to dramatically influence the calculated values of $d[M]/dt$. Note that the scan rate is positive in the heating direction and

negative in the cooling direction, leading to positive and negative $d[M]/dt$ in the heating and cooling directions respectively. The sets of scan rate dependent $d[M]/dt$ and $[M]$ from the annealing and melting portions of the experiment provide access to the temperature/reaction rate supramolecular assembly maps. These surfaces have larger reaction rates at faster temperature scan rates, as expected. The middle of the surface appears as a valley between the assembly and disassembly portions of the experiment and corresponds to concentrations close to their equilibrium values.

The surfaces are sliced with respect to temperature and a log-log analysis is performed according to Equations 4.1-4.4 in Section 4.4.1. The intercepts correspond to effective rate constants for assembly and disassembly, however these contain contributions from a number of processes and are not meaningful for supramolecular pathway analysis. As guidelines for extraction of effective reaction orders using the model-free analysis presented here, we find the method requires (i) that there is adequate separation of the TH profiles for a given assembly or disassembly process as a function of temperature scan rate, e.g. for the assembly process, profiles collected at different scan rates must differ from each other in the transition region by substantially more than the scatter due to experimental noise. (ii) The assembly and disassembly portions of TH data occur independently of each other. We suggest a roughly 3-fold difference in calculated annealing and melting rates in the middle of the annealing transition to ensure the observed orders reflect the pure assembly or disassembly processes. (iii) The concentration-rate plots are not performed using experimental data near or within the baseline regions. We recommend restricting the analysis to the ~10-90% fraction unfolded regions in order to obtain accurate orders.

4.7.7. Global analysis of TG4T TH profiles

The TH profiles for TG4T were globally fit assuming a model where GQ assembly proceeds via step-wise association of monomers[7] (Figure 4.2, Figure 4.3, and Supplementary Figure 4.4c). The changes in concentration with respect to temperature are

$$\frac{d}{dT}[M] = \left(2k_{-1}[D] - 2k_1[M]^2 - k_2[M][D] + k_{-2}[Tr] - k_3[M][Tr] + k_{-3}[Q]\right)\frac{dt}{dT} \tag{4.16}$$

$$\frac{d}{dT}[D] = \left(k_1[M]^2 - k_{-1}[D] + k_{-2}[Tr] - k_2[D][Tr]\right)\frac{dt}{dT} \tag{4.17}$$

$$\frac{d}{dT}[Tr] = \left(k_2[M][D] - k_{-2}[Tr] + k_{-3}[Q] - k_3[D][Tr]\right)\frac{dt}{dT} \tag{4.18}$$

$$\frac{d}{dT}[Q] = \left(k_3[Tr][M] - k_{-3}[Q]\right)\frac{dt}{dT} \tag{4.19}$$

where *dt/dT* is the inverse temperature scan rate. In what follows, the rate constants are assumed to be functions of temperature, and the concentrations of each species are assumed to be functions of temperature and scan rate, but we omit this notation for clarity. The temperature dependences of the rate constants are given by

$$k(T) = k_0 e^{\frac{E_a}{R}\left(\frac{1}{T_{ref}} - \frac{1}{T}\right)} \tag{4.20}$$

where k_0 is the rate constant at the reference temperature T_{ref} and E_a is the activation energy. In the global fit of the TG4T TH profiles, the set of TG4T assembly Equations 4.16-4.19 were numerically integrated using the ordinary differential equation (ODE) solvers in MATLAB (with ten minute pre-scan equilibrations) to obtain the concentrations of monomer, dimer, trimer, and

tetramer as a function of temperature. The concentrations were converted to fraction unfolded and folded respectively using

$$C_T = [M] + 2[D] + 3[Tr] + 4[Q] \tag{4.21}$$

$$\theta_U = \frac{[M]}{C_T} \tag{4.22}$$

$$\theta_F = \frac{4[Q] + 3[Tr] + 2[D]}{C_T} \tag{4.23}$$

which permitted calculation of the thermal absorbance profiles as

$$A(T) = A_F(T)\theta_F(T) + A_U(T)\theta_U(T) \tag{4.24}$$

where $A_F(T)$ and $A_U(T)$ are the linear folded and unfolded absorbance baselines calculated according to Equations 4.10 and 4.11. The sets of TH profiles were fit by varying the kinetic parameters to minimize the RSS between the experimental and fitted absorbance data according to

$$RSS = \sum_{j=1}^{N} \sum_{k} \left(A_j^{\exp}(T_k) - A_j^{calc}(T_k, \xi) \right)^2 \tag{4.25}$$

where $A_j^{\exp}(T_k)$ and $A_j^{calc}(T_k)$ are the j^{th} experimental and fitted absorbance profiles respectively, T_k is the k^{th} experimental temperature, and ξ = *[k_1, k_{-1}, k_2, k_{-2}, k_3, k_{-3}, E_1, E_{-1}, E_2, E_{-2},* E_3, E_{-3}*]* are the rate constants at the reference temperature and activation energies governing assembly and disassembly of the tetramer.

4.7.8. Global analysis of TH profiles for CA-mediated poly(A) fiber formation

The TH profiles for CA-mediated poly(A) fiber formation were globally fit with the Goldstein-Stryer model for cooperative self-assembly[8] (Figure 4.5). The model assumes reversible, cooperative stepwise association of monomers (M) to form nuclei (M_s), which then elongate to form fibers (M_N). The model has two distinct phases, where the pre-nucleus equilibria are governed by the nucleation rate constants k_{n+} and k_{n-}, and post-nucleus equilibria are governed by the elongation rate constants k_{e+} and k_{e-}. In order to limit the number of equations that must be numerically integrated, only fibers up to size N are explicitly described. Korevaar *et al.* showed that by treating all structures larger than the explicitly described size of N as a reversibly-formed fibril pool, increased numerical accuracy is obtained in solving this system of equations compared to straight truncation at a certain fiber length N[10, 23]. The Goldstein-Stryer model including the fibril pool for reversible self-assembly by Korevaar *et al.* is described by the following rate equations

Monomer

$$\begin{aligned}\frac{d}{dt}[M] = &-k_{n+}[M]\left(2[M]+\sum_{i=2}^{s-1}[M_i]\right)-k_{e+}[M]\left(\sum_{i=s}^{N}[M_i]+[P]\right)\\ &+k_{n-}\left(2[M_2]+\sum_{i=3}^{s}[M_i]\right)+k_{e-}\left(\sum_{i=s+1}^{N}[M_i]+[P]\right)\end{aligned} \tag{4.26}$$

Pre-nucleus oligomers

$$\frac{d}{dt}[M_i] = k_{n+}[M]\left([M_{i-1}]-[M_i]\right)+k_{n-}\left([M_{i+1}]-[M_i]\right) \tag{4.27}$$

Nucleus

$$\frac{d}{dt}[M_s] = k_{n+}[M][M_{s-1}] - k_{e+}[M][M_s] + k_{e-}[M_{s+1}] - k_{n-}[M_s] \tag{4.28}$$

Post-nucleus fibers

$$\frac{d}{dt}[M_i] = k_{e+}[M]\left([M_{i-1}] - [M_i]\right) + k_{e-}\left([M_{i+1}] - [M_i]\right) \tag{4.29}$$

Fiber length N

$$\frac{d}{dt}[M_N] = k_{e+}[M]\left([M_{N-1}] - [M_N]\right) + k_{e-}\left((1-\alpha)[P] - [M_N]\right) \tag{4.30}$$

Fibril number concentration

$$\frac{d}{dt}[P] = k_{e+}[M][M_N] - k_{e-}(1-\alpha)[P] \tag{4.31}$$

Fibril mass concentration

$$\frac{d}{dt}[Z] = k_{e+}[M]\left((N+1)[M_N] + [P]\right) - k_{e-}\left([P] + N(1-\alpha)[P]\right) \tag{4.32}$$

where α is given by

$$\alpha = 1 - \left(\frac{[P]}{[Z] - N[P]}\right). \tag{4.33}$$

In our global fits, we assumed $k_{n+} = k_{e+}$[10]. Additionally, we allowed for non-zero nucleic acid folding ΔC_ps[25, 42] in the nucleation and elongation steps by including the $\Delta C_p^{\ddagger}$ parameter in calculating temperature-dependent activation energies

$$E_{n+}(T) = E_{n+}^0 + \Delta C_p^{\ddagger}(T - T_{ref}) \tag{4.34}$$

where $E_{n+}{}^{0}$ is the activation energy at the reference temperature. We have shown only the equation for the forward nucleation step for brevity. The set of differential equations for the Goldstein-Stryer model were numerically integrated as a function of temperature using the inverse scan rate *dt/dT* and the fractions of the unfolded monomer and polymerized states were calculated according to

$$\theta_U = \frac{[M]}{C_T} \tag{4.35}$$

$$\theta_F = \frac{\sum_{i=2}^{N} i[M_i] + [Z]}{C_T} \tag{4.36}$$

which permitted the calculation of the thermal absorbance profiles according to Equation 4.24. Global fits to the TH profiles for CA-mediated poly(A) fiber formation were carried out by minimizing the RSS in an identical manner to TG_4T. We varied the nucleus size to optimize the fit quality and agreement with the experimentally-determined effective assembly orders (Supplementary Figure 4.5). Out of the arrayed nucleus sizes, 2-4 gave excellent agreement with the data. While sizes of 2-4 are all physically realistic for poly(A) fiber formation, a nucleus of 3 fit the data best and therefore this is our preferred nucleus size. In addition, we varied the explicitly described fiber size *N* in order to verify that the fit results did not depend on its value. We found that annealing profiles simulated with a nucleus size of 3 and N = 50, 100, and 200 overlay, highlighting the utility of the approach developed by Korevaar *et al.*[10, 23] in global fitting of TH datasets for supramolecular systems, as well as improving numerical accuracy and reducing computational time by allowing the use of smaller *N* values.

4.7.9. General interpretation of reaction orders for step-wise polymerization

The net flux of N-mer conversion to (N+1)-mers, Φ_N, depends on the concentrations of monomer, N-mer, and (N+1)-mer, (c_1, c_N, c_{N+1}, respectively), as well as the association rate of monomers and N-mers ($k_{on,N}$) and the dissociation rate of the (N+1)-mer ($k_{off,N+1}$) according to

$$\Phi_N = k_{(on,N)} c_1 c_N - k_{(off,N+1)} c_{N+1} \tag{4.37}$$

for N>1. The flux of monomer to dimer conversion is given by

$$\Phi_1 = 2\left(k_{(on,1)} c_1^2 - k_{(off,2)} c_2\right). \tag{4.38}$$

The total rate of monomer consumption is given by

$$R = -\frac{\partial}{\partial t} c_1 = \sum_{N=1}^{\infty} \Phi_N \tag{4.39}$$

and the effective order is

$$\frac{\partial \ln(R)}{\partial \ln(c_1)} = \sum_{N=1}^{\infty} \frac{\Phi_N}{R} \frac{\partial \ln(\Phi_N)}{\partial \ln(c_1)}. \tag{4.40}$$

In other words, the effective order of monomer consumption is given by the weighted average of the orders of the individual fluxes $\partial\ln(\Phi_N)/\partial\ln(c_1)$ where the N^{th} weight Φ_N/R is the relative contribution of the N^{th} flux to the total rate. The order of each flux depends on how the populations of the N-mer and (N+1)-mer vary relative to the monomer concentration at a given temperature across the different scan rates, as well as the relative rate of depolymerization ($k_{off,N+1}$)

$$\frac{\partial \ln(\Phi_N)}{\partial \ln(c_1)} = 1 + \frac{\partial \ln(c_N)}{\partial \ln(c_1)} + \frac{k_{(off,N+1)} c_{N+1}}{\Phi_N}\left(1 + \frac{\partial \ln(c_N)}{\partial \ln(c_1)} - \frac{\partial \ln(c_{N+1})}{\partial \ln(c_1)}\right). \tag{4.41}$$

Thus if depolymerization ($k_{(off,N+1)}c_{N+1}$) is slow compared to the flux (Φ_N) and the concentration of the N-mer varies as the m^{th} power of the monomer concentration across the scan rates ($c_N \propto c_1^m$, $\partial\ln(c_N)/\partial\ln(c_1) = m$), then the apparent order of the N^{th} flux $\partial\ln(\Phi_N)/\partial\ln(c_1)$ is m+1. With faster depolymerization rates, i.e. at values of [M] approaching the critical concentration, larger apparent reaction orders are obtained. The order of the monomer-to-dimer flux, Φ_1, is given by

$$\frac{\partial \ln\left(\Phi_1\right)}{\partial \ln\left(c_1\right)} = 2\left[1 + \frac{k_{(off,2)}c_2}{\Phi_1}\left(2 - \frac{\partial \ln\left(c_2\right)}{\partial \ln\left(c_1\right)}\right)\right]. \tag{4.42}$$

4.7.10. Simulating TH profiles for classical nucleated supramolecular polymerizations

Classical nucleated polymerizations were simulated according to the assumptions that (i) the monomer concentration changes only by addition to and subtraction from polymers longer than the nucleus, therefore the nucleus and pre-nuclear oligomers have small concentrations and are in rapid equilibrium with the monomer, (ii) polymer formation is irreversible, and (iii) the polymer elongation rate becomes zero when the monomer concentration reaches the critical concentration, $[M]_{critical}$. The equations for a classical nucleated polymerization[8] are

$$[s] = K_n^{\,s-1}[M]^s \tag{4.43}$$

$$\frac{d}{dT}[M] = -k_{e+}[P]\left([M] - [M]_{critical}\right)\frac{dt}{dT} \tag{4.44}$$

$$\frac{d}{dT}[P] = k_{e+}[s]\left([M] - [M]_{critical}\right)\frac{dt}{dT} \tag{4.45}$$

Where [s] is the concentration of the nucleus of size s, K_n is the equilibrium constant for nucleation = k_{n+}/k_{n-}, [P] is the concentration of polymers larger than the nucleus, and the $[M]_{critical} = k_{e-}/k_{e+}$.

Fraction unfolded TH profiles were simulated by numerically solving the concentration of monomer with the ODE solvers in MATLAB and dividing by the total monomer concentration C_T.

4.7.11. Calculating apparent reaction orders

We calculated theoretical reaction orders for TG_4T assembly for the step-wise association of monomers model approaching thermodynamic equilibrium with negligible concentrations of dimer and trimer ($[D]=(k_1/k_{-1})[M]^2$, $[Tr]=(k_2/k_{-2})[M][D]$). The rate of tetramer conversion to monomer is thus approximately equal to the rate of monomer conversion to tetramer. This is equal to the rate of monomer conversion to dimer ($k_1[M]^2$) multiplied by the net fraction of dimers (F_{DQ}) that continue forward to tetramer versus those that disassociate back to monomers

$$F_{DQ} = \frac{\left(\frac{k_2 k_3 [M]^2 [D][Tr]}{k_{-2}[Tr] + k_3[M][Tr]}\right)}{k_{-1}[D] + \left(\frac{k_2 k_3 [M]^2 [D][Tr]}{k_{-2}[Tr] + k_3[M][Tr]}\right)} \tag{4.46}$$

where the numerator of Equation 4.46 is the net rate of dimer to tetramer transition and $k_{-1}[D]$ is the net rate of the dimer to monomer transition[43]. The forward (monomer to tetramer) rate in the dynamic equilibrium is thus given by

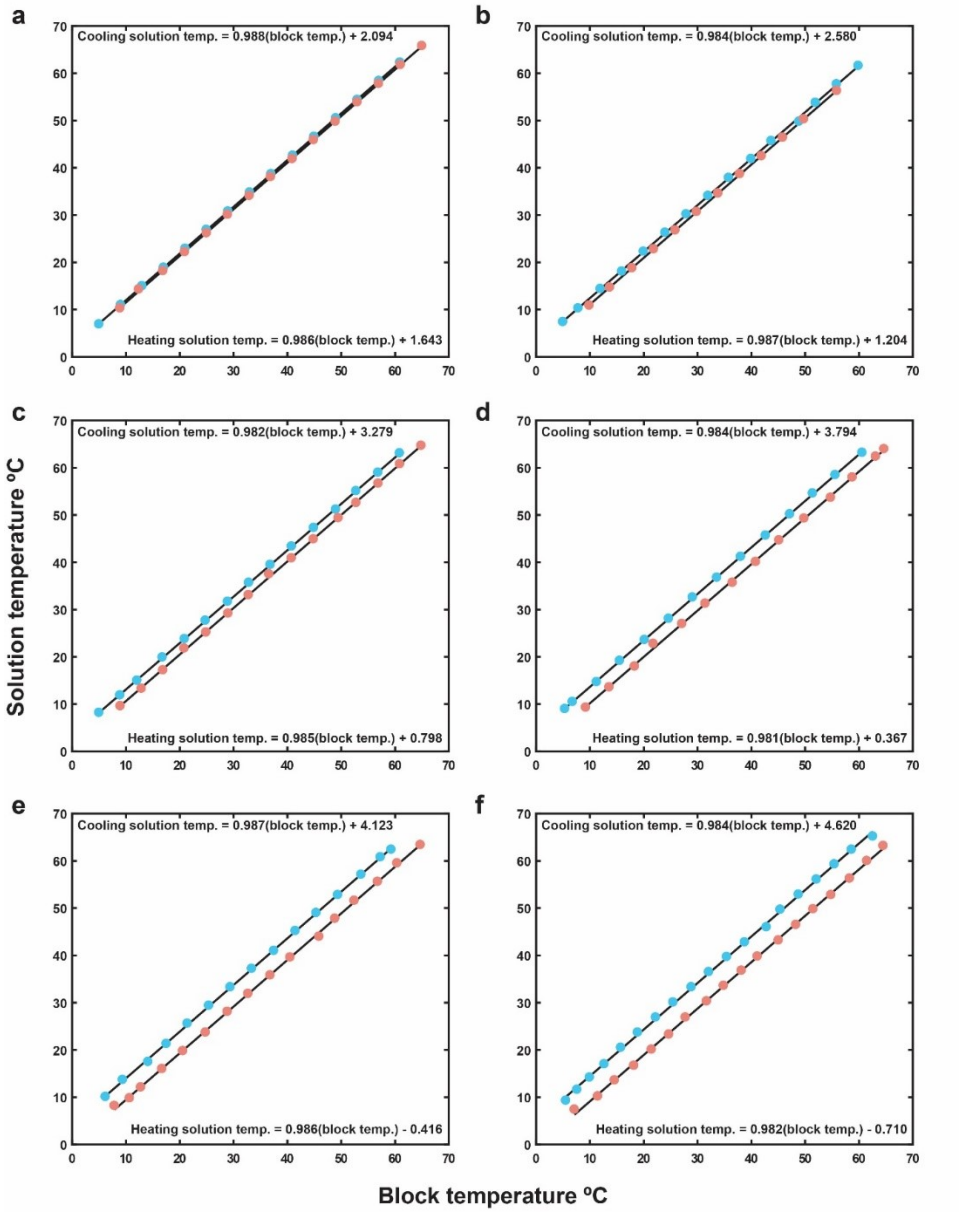
$$\begin{aligned} R &= k_1[M]^2 \left[\frac{\left(\frac{k_2 k_3 [M]^2 [D][Tr]}{k_{-2}[Tr] + k_3[M][Tr]}\right)}{k_{-1}[D] + \left(\frac{k_2 k_3 [M]^2 [D][Tr]}{k_{-2}[Tr] + k_3[M][Tr]}\right)} \right]. \\ &= \frac{k_1 k_2 k_3 [M]^4}{k_{-1}k_{-2} + k_{-1}k_3[M] + k_2 k_3 [M]^2} \end{aligned} \tag{4.47}$$

The order of R with respect to [M] is thus

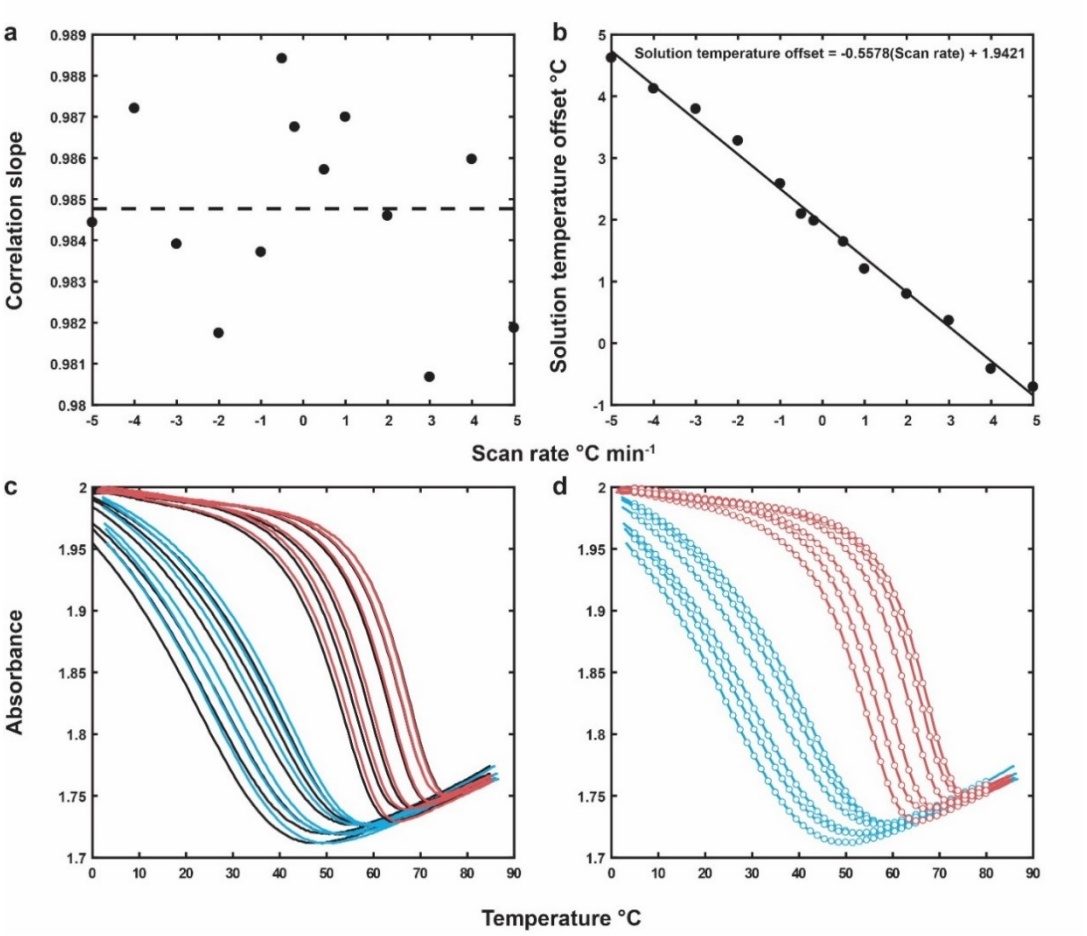
$$n_{app.} = \frac{\partial \ln(R)}{\partial \ln([M])} = \frac{[M]}{R}\frac{\partial R}{\partial [M]}. \tag{4.48}$$

To assess the effects of changes in monomer concentration on the TH orders for TG_4T assembly, we simulated the apparent orders at fixed temperature and variable monomer concentration (Supplementary Figure 4.6). The change in R with respect to [M] was calculated numerically using the movingslope function in MATLAB. The temperatures were held fixed at the lower and upper limits of 5 and 45 °C respectively (dark blue and dark red dashed lines in Supplementary Figure 4.6), while at each temperature, the monomer concentration was set at the average value used in the TH analysis at that temperature. Note that the monomer concentration was lower at low temperatures and higher at high temperatures. The theoretical orders at both limiting temperatures decrease with increasing [M] (and T), as expected. This effect is overwhelmed in the experimental data by the shift in rate-determining barrier which leads to an increase in the apparent reaction order with increasing temperature.

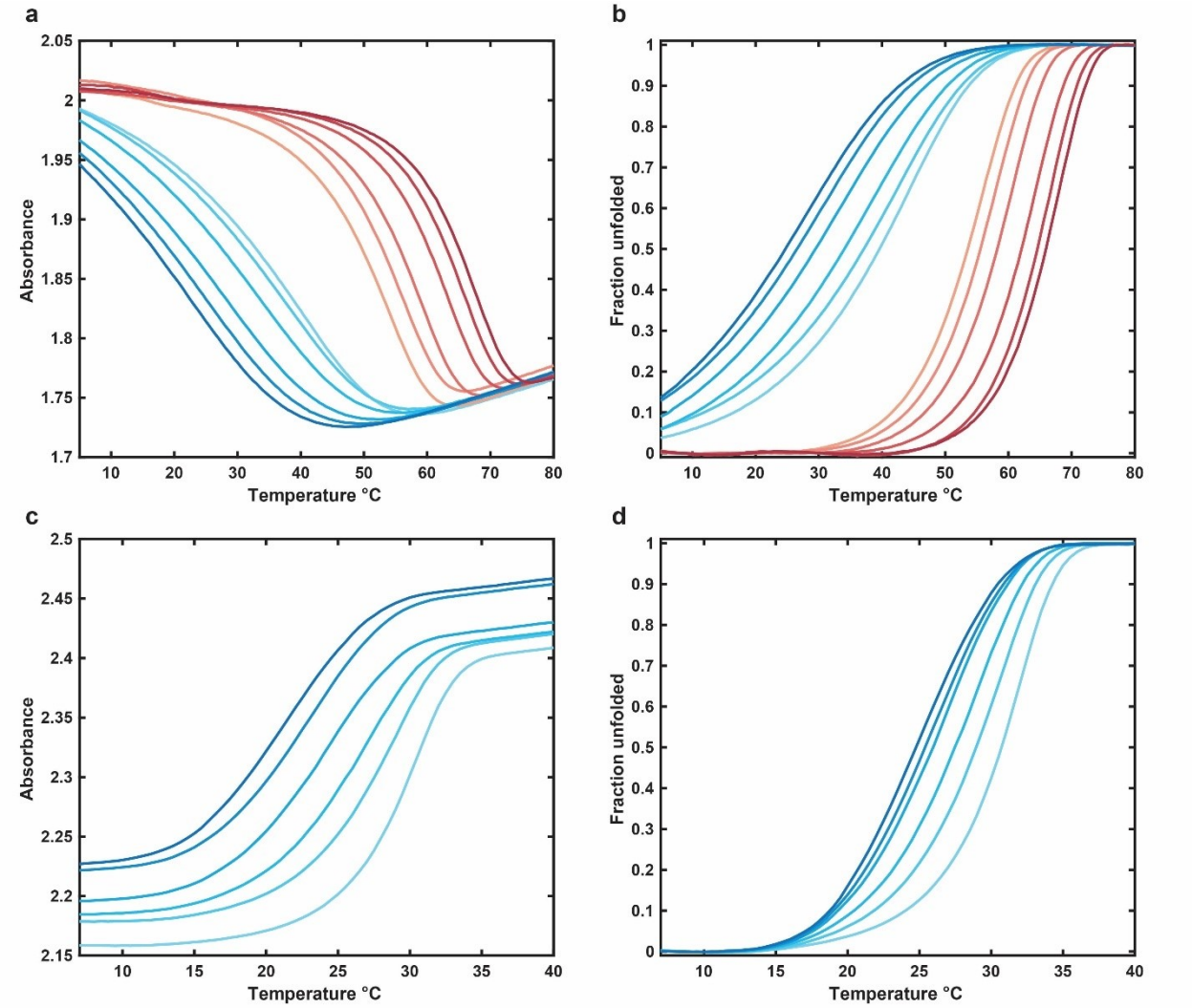
4.7.12. Supplementary Figures



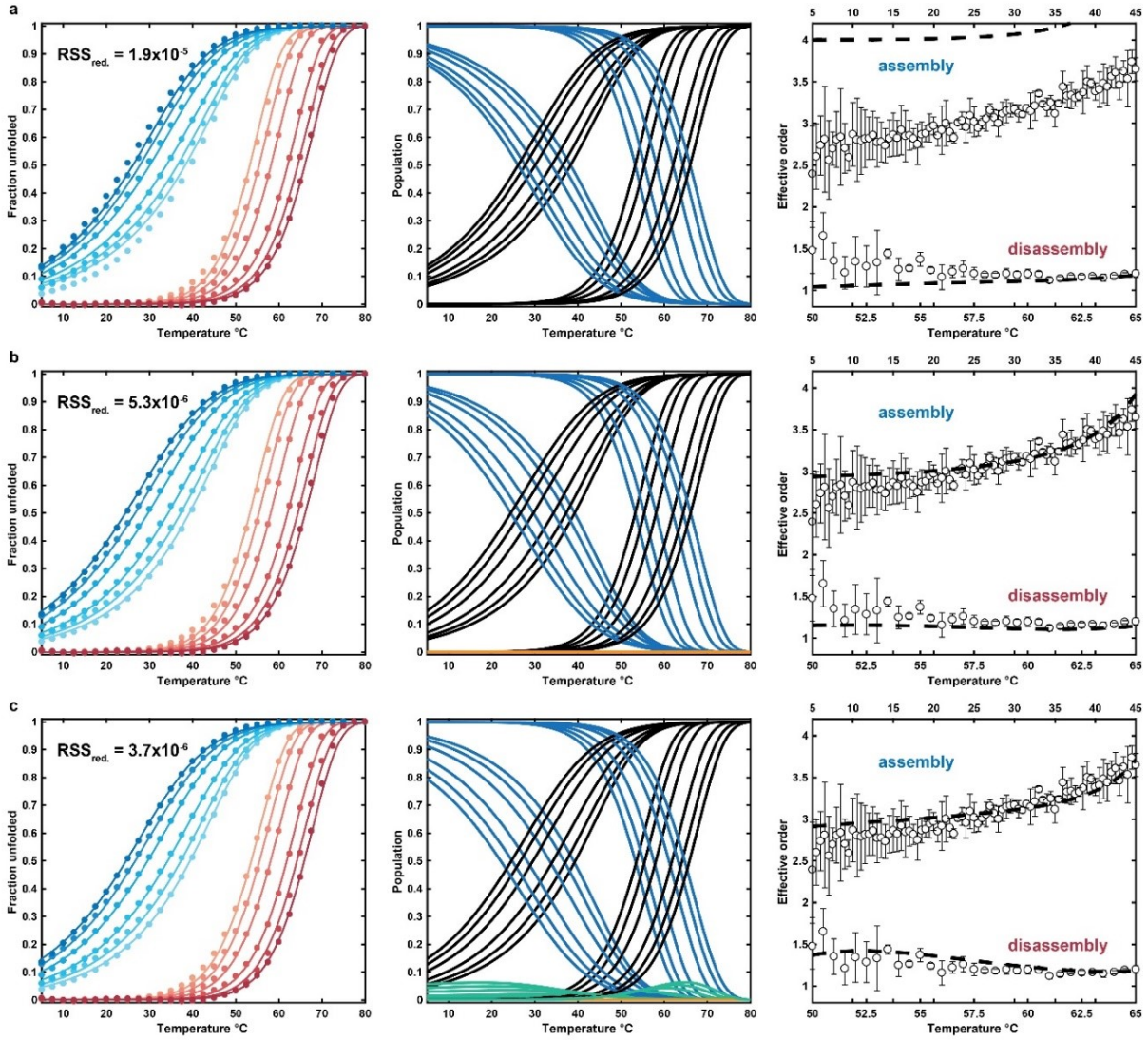
Supplementary Figure 4.1. Solution versus block temperature as a function of scan rate. (a) ±0.5 °C min^{-1} scan rates. (b) ±1 °C min^{-1} scan rates. (c) ±2 °C min^{-1} scan rates. (d) ±3 °C min^{-1} scan rates. (e) ±4 °C min^{-1} scan rates. (f) ±5 °C min^{-1} scan rates. In all panels, the cooling and heating scan temperatures are shown as blue and red circles respectively. Linear fits to the scan temperatures are shown as black lines. The parameters corresponding to the linear fits of the cooling and heating scan temperatures are shown in the top left and bottom right of each panel respectively.



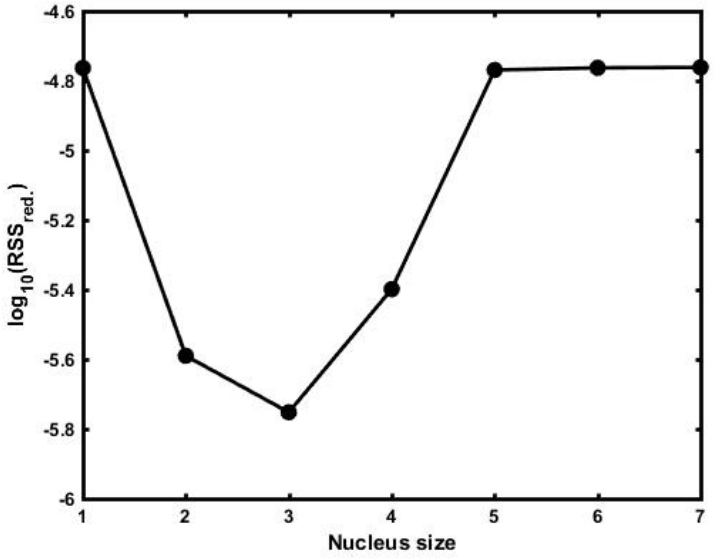
Supplementary Figure 4.2. Temperature correction of TH data. (a) Slopes from the correlations in Supplementary Figure 4.1 as a function of temperature scan rate. The dashed black line indicates the mean slope (~0.985). (b) Solution temperature offset as a function of scan rate. Offsets were taken as the intercepts from the linear fits to the correlations in Supplementary Figure 4.1. A linear fit to the solution temperature offsets as a function of scan rate is shown as a black line, with the corresponding fit parameters given in the top right of the panel. (c) Uncorrected (black lines) and temperature corrected TG_4T TH profiles (blue and red lines for annealing and melting respectively). The correction was performed using $T_{solution} = 0.985(T_{block})$ + offset where the offset was calculated from the equation given in Supplementary Figure 4.2b. (d) Linearly interpolated TH profiles (blue and red empty circles for annealing and melting respectively) overlaid with the temperature corrected profiles from Supplementary Figure 4.2c. Only every third interpolated point is shown for clarity. The interpolation was performed to place the corrected data on the same temperature domain, from 5-80 °C in 0.5 °C increments.



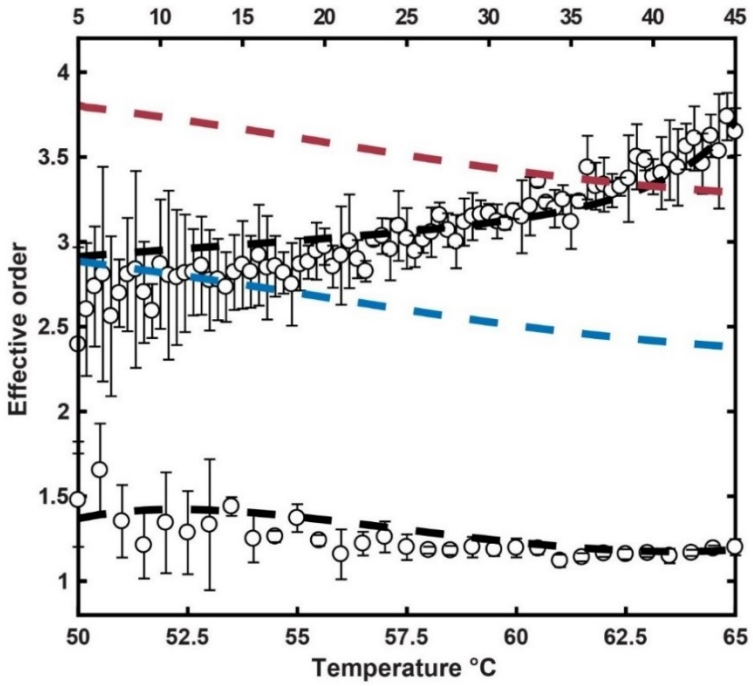
Supplementary Figure 4.3. Raw and corrected TH profiles. (a) Raw TG_4T absorbance data. (b) Baseline and temperature corrected TG_4T data used in all model-free and global fitting analyses. (c) Raw poly(A) fiber assembly TH profiles. (d) Baseline and temperature corrected poly(A) fiber assembly data used in all model-free and global fitting analyses. In all panels, dark to light blue indicate indicates fastest to slowest annealing scan rates, while dark red to light orange in (a,b) indicates fastest to slowest melting scan rates.



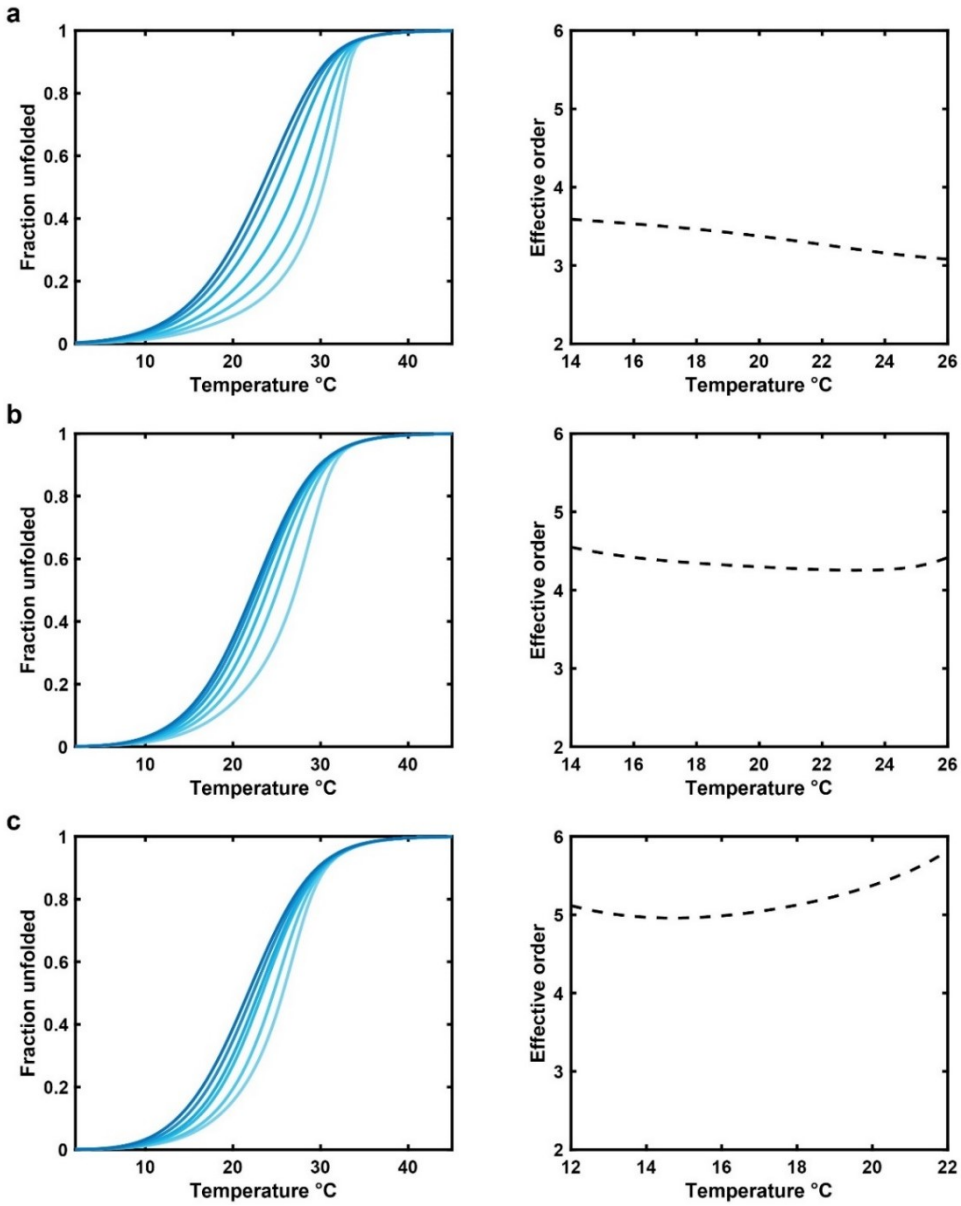
Supplementary Figure 4.4. Comparison of global fits of kinetic models to TG_4T TH profiles. (a) One-step assembly model (Figure 4.3a), $RSS_{red.} = 1.9\times10^{-5}$. (b) Dimer-of-dimers model (Figure 4.3b), $RSS_{red.} = 5.3\times10^{-6}$. (c) Step-wise monomer association model (Figure 4.3c), $RSS_{red.} = 3.7\times10^{-6}$. For (a-c), panels show: (Left) Fraction unfolded TH profiles, where fits and experimental data are shown as colored lines and circles respectively. Only every 5th experimental point is shown for clarity. Dark to light blue corresponds to fastest and slowest annealing scan rates, and dark red to light orange corresponds to fastest to slowest melting scan rates respectively. The reduced RSS, $RSS_{red.}$ was calculated as RSS_0/DF where DF = # points - # fitted parameters. (Middle) Populations, where black and dark blue lines correspond to monomer and tetramer respectively. In (b) and (c), orange lines correspond to dimer population. In (c), green lines correspond to trimer population. (Right) Effective orders obtained from model-free analysis of experimental and fitted data, shown as white circles and dashed black lines respectively. The error bars for the experimental points are the standard deviation of model-free analysis on three replicate TH experiments.



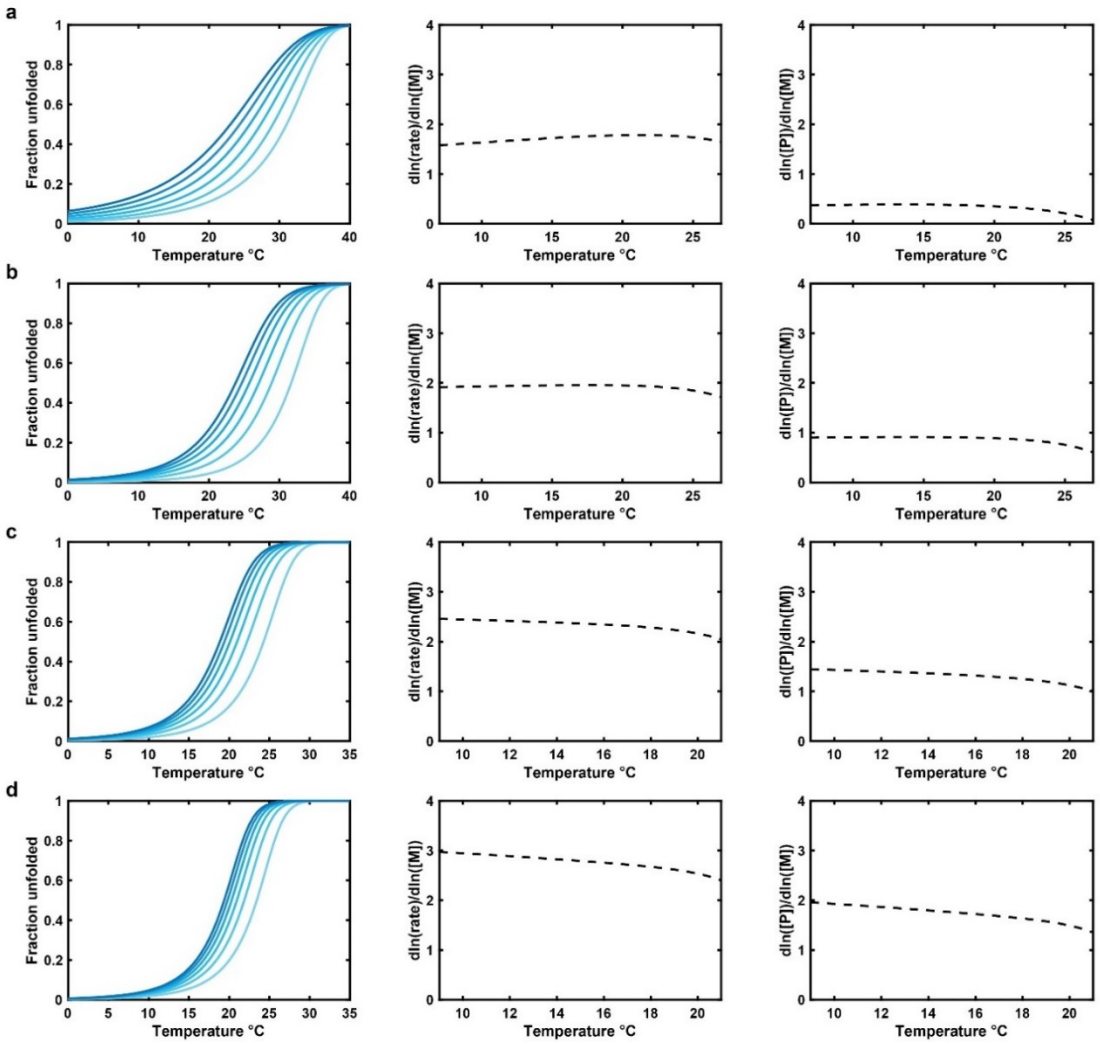
Supplementary Figure 4.5. Global fit quality as a function of nucleus size for global fits to CA-mediated poly(A) assembly TH profiles. The reduced RSS, $RSS_{red.}$, was calculated as RSS_0/DF, where DF = # points - # fit parameters. The nucleus size of 1 corresponds to a fit with an isodesmic (non-cooperative) mechanism where $k_{e+} = k_{n+}$ and $k_{e-} = k_{n-}$. The best fit was obtained with a nucleus size of 3.



Supplementary Figure 4.6. Assessing the concentration dependence of the TG_4T assembly reaction orders at low and high temperature. The dark red and blue dashed lines are the assembly orders simulated with the step-wise model at 45 and 5 °C respectively as described in Section 4.7.11.



Supplementary Figure 4.7. Simulations of Goldstein-Stryer TH profiles as a function of nucleus size with fixed kinetic parameters. (a) Nucleus size of 3. (b) Nucleus size of 4. (c) Nucleus size of 5. In all left panels, dark to light blue indicates fastest to slowest annealing scan rates respectively. In all right panels, the effective assembly reaction order is shown as dashed black lines. Simulation parameters were E_{n+} = -27, k_{n+} = 1.75×10^4, E_{n-} = 35, k_{n-} = 2.4, E_{e+} = 50, k_{e+} = 7.5×10^4, E_{e-} = 90, k_{e-} = 0.5. Activation energies are given in kcal mol^{-1} and forward and reverse rate constants are given in M^{-1} min^{-1} and min^{-1} respectively at the reference temperature of 25 °C. Simulations were performed with $\Delta C_p^{\ddagger} = 0$ for all steps. The scan rates were (a,b) 0.2, 0.5, 1, 2, 3, 4 and (c) 0.1, 0.2, 1, 2, 4, 6 °C min^{-1}.



Supplementary Figure 4.8. Simulations of TH profiles for classical nucleated polymerizations. (a) Nucleus size of 2, ΔH_n = -30, K_n = 143, E_{e+} = 5, k_{e+} = 3.5×10^4, E_{e-} = 30, k_{e-} = 0.2. (b) Nucleus size of 3, ΔH_n = -40, K_n = 2×10^3, E_{e+} = 5, k_{e+} = 1×10^5, E_{e-} = 50, k_{e-} = 0.1. (c) Nucleus size of 4, ΔH_n = -40, K_n = 2×10^3, E_{e+} = 5, k_{e+} = 1×10^5, E_{e-} = 60, k_{e-} = 0.3. (d) Nucleus size of 5, ΔH_n = -40, K_n = 2×10^3, E_{e+} = 5, k_{e+} = 3×10^5, E_{e-} = 60, k_{e-} = 1. In all left panels, dark to light blue lines indicate fastest to slowest annealing scan rates respectively. Effective monomer reaction rate orders are shown in the middle panels as dashed black lines, tracking approximately as (s+1)/2 where s is the nucleus size. Fiber concentration ([P]) orders are shown in the right panels as dashed black lines, tracking approximately as (s-1)/2. Nucleation ΔHs and elongation activation energies are given in kcal mol^{-1}, nucleation equilibrium constants are given in M^{-1}, and forward and reverse rate constants are given in M^{-1} min^{-1} and min^{-1} respectively at the reference temperature of 25 °C. The critical monomer concentration $[M]_{critical}$ was calculated as k_{e-}/k_{e+}. Simulations were performed with $\Delta C_p^{\ddagger} = 0$ for all steps. In (a), scan rates were 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7 °C min^{-1}. In (b-d), scan rates were 0.2, 0.4, 0.6, 0.8, 1, and 1.2 °C min^{-1}.

4.8. References

1. Schiff, P.B., Fant, J. & Horwitz, S.B. Promotion of microtubule assembly in vitro by taxol. *Nature* **277**, 665-667 (1979).
2. Jacques, D.A. et al. HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis. *Nature* **536**, 349-353 (2016).
3. Seeman, N.C. & Sleiman, H.F. DNA nanotechnology. *Nature Reviews Materials* **3**, 17068 (2017).
4. Aida, T., Meijer, E.W. & Stupp, S.I. Functional supramolecular polymers. *Science* **335**, 813-817 (2012).
5. Raskatov, J.A. & Teplow, D.B. Using chirality to probe the conformational dynamics and assembly of intrinsically disordered amyloid proteins. *Sci Rep* **7**, 12433 (2017).
6. Schwierz, N., Frost, C.V., Geissler, P.L. & Zacharias, M. Dynamics of Seeded Abeta40-Fibril Growth from Atomistic Molecular Dynamics Simulations: Kinetic Trapping and Reduced Water Mobility in the Locking Step. *J Am Chem Soc* **138**, 527-539 (2016).
7. Bardin, C. & Leroy, J.L. The formation pathway of tetramolecular G-quadruplexes. *Nucleic Acids Res* **36**, 477-488 (2008).
8. Goldstein, R.F. & Stryer, L. Cooperative polymerization reactions. Analytical approximations, numerical examples, and experimental strategy. *Biophys J* **50**, 583-599 (1986).
9. Powers, E.T. & Powers, D.L. The kinetics of nucleated polymerizations at high concentrations: amyloid fibril formation near and above the "supercritical concentration". *Biophys J* **91**, 122-132 (2006).
10. Korevaar, P.A. et al. Pathway complexity in supramolecular polymerization. *Nature* **481**, 492-496 (2012).
11. Hiragi, Y. et al. Dynamic mechanism of the self-assembly process of tobacco mosaic virus protein studied by rapid temperature-jump small-angle X-ray scattering using synchrotron radiation. *Journal of Molecular Biology* **213**, 495-502 (1990).
12. Marx, A., Jagla, A. & Mandelkow, E. Microtubule assembly and oscillations induced by flash photolysis of caged-GTP. *Eur Biophys J* **19**, 1-9 (1990).
13. Mergny, J.L. & Lacroix, L. Analysis of thermal melting curves. *Oligonucleotides* **13**, 515-537 (2003).
14. Hatzakis, E., Okamoto, K. & Yang, D. Thermodynamic stability and folding kinetics of the major G-quadruplex and its loop isomers formed in the nuclease hypersensitive element in the human c-Myc promoter: effect of loops and flanking segments on the stability of parallel-stranded intramolecular G-quadruplexes. *Biochemistry* **49**, 9152-9160 (2010).
15. Mergny, J.L. & Lacroix, L. Kinetics and thermodynamics of i-DNA formation: phosphodiester versus modified oligodeoxynucleotides. *Nucleic Acids Res.* **26**, 4797-4803 (1998).

16. Mizuno, K., Boudko, S.P., Engel, J. & Bachinger, H.P. Kinetic hysteresis in collagen folding. *Biophys J* **98**, 3004-3014 (2010).
17. Prislan, I., Lah, J. & Vesnaver, G. Diverse polymorphism of G-quadruplexes as a kinetic phenomenon. *J Am Chem Soc* **130**, 14161-14169 (2008).
18. Wyatt, J.R., Davis, P.W. & Freier, S.M. Kinetics of G-quartet-mediated tetramer formation. *Biochemistry* **35**, 8002-8008 (1996).
19. Motulsky, H.J., Christopoulos, A. Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting. (GraphPad Software, Inc., San Diego, CA; 2003).
20. Paes, H.M. & Fox, K.R. Kinetic studies on the formation of intermolecular triple helices. *Nucleic Acids Res* **25**, 3269-3274 (1997).
21. Tellinghuisen, J. Statistical error propagation. *Journal of Physical Chemistry A* **105**, 3917-3921 (2001).
22. Avakyan, N. et al. Reprogramming the assembly of unmodified DNA with a small molecule. *Nat Chem* **8**, 368-376 (2016).
23. van der Zwaag, D. et al. Kinetic Analysis as a Tool to Distinguish Pathway Complexity in Molecular Assembly: An Unexpected Outcome of Structures in Competition. *J Am Chem Soc* **137**, 12677-12688 (2015).
24. Frieden, C. & Goddette, D.W. Polymerization of actin and actin-like systems: evaluation of the time course of polymerization in relation to the mechanism. *Biochemistry* **22**, 5836-5843 (1983).
25. Tikhomirova, A., Taulier, N. & Chalikian, T.V. Energetics of nucleic acid stability: the effect of DeltaCP. *J Am Chem Soc* **126**, 16387-16394 (2004).
26. Bochman, M.L., Paeschke, K. & Zakian, V.A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-780 (2012).
27. Sarkar, B., O'Leary, L.E. & Hartgerink, J.D. Self-assembly of fiber-forming collagen mimetic peptides controlled by triple-helical nucleation. *J Am Chem Soc* **136**, 14417-14424 (2014).
28. Fasshauer, D., Antonin, W., Subramaniam, V. & Jahn, R. SNARE assembly and disassembly exhibit a pronounced hysteresis. *Nat Struct Biol* **9**, 144-151 (2002).
29. Cantu, L., Corti, M., Del Favero, E., Muller, E., Raudino, A., Sonnino, S. Thermal hysteresis in ganglioside micelles investigated by differential scanning calorimetry and light-scattering. *Langmuir* **15**, 4975-4980 (1999).
30. Singh, S. & Zlotnick, A. Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly. *J Biol Chem* **278**, 18249-18255 (2003).
31. Korevaar, P.A., Newcomb, C.J., Meijer, E.W. & Stupp, S.I. Pathway selection in peptide amphiphile assembly. *J Am Chem Soc* **136**, 8540-8543 (2014).
32. Ma, X., Sun, C., Huang, J. & Boutis, G.S. Thermal hysteresis in the backbone and side-chain dynamics of the elastin mimetic peptide [VPGVG]3 revealed by 2H NMR. *J Phys Chem B* **116**, 555-564 (2012).

33. Reguera, J., Lagaron, J. M., Alonso, M., Reboto, V., Calvo, B., and Rodriguez-Cabello, J. C. Thermal behaviour and kinetic analysis of the chain unfolding and refolding and of the concomitant nonpolar solvation and desolvation of two elastin-like polymers. *Macromolecules* **36**, 8470-8476 (2003).
34. Schulman, R. & Winfree, E. Synthesis of crystals with a programmable kinetic barrier to nucleation. *Proc Natl Acad Sci U S A* **104**, 15236-15241 (2007).
35. Yin, P. et al. Programming DNA Tube Circumferences. *Science* **321**, 824-826 (2008).
36. Van Gorp, J.J., Vekemans, J.A. & Meijer, E.W. C3-symmetrical supramolecular architectures: fibers and organic gels from discotic trisamides and trisureas. *J Am Chem Soc* **124**, 14759-14769 (2002).
37. Liang, C. et al. Kinetic intermediates in amyloid assembly. *J Am Chem Soc* **136**, 15146-15149 (2014).
38. Medrano, M. et al. Imaging and Quantitation of a Succession of Transient Intermediates Reveal the Reversible Self-Assembly Pathway of a Simple Icosahedral Virus Capsid. *J Am Chem Soc* **138**, 15385-15396 (2016).
39. Rajagopalan, S., Huang, F. & Fersht, A.R. Single-Molecule characterization of oligomerization kinetics and equilibria of the tumor suppressor p53. *Nucleic Acids Res* **39**, 2294-2303 (2011).
40. Chung, H.S. et al. Oligomerization of the tetramerization domain of p53 probed by two- and three-color single-molecule FRET. *Proc Natl Acad Sci U S A* **114**, E6812-E6821 (2017).
41. Rennella, E., Sekhar, A. & Kay, L.E. Self-Assembly of Human Profilin-1 Detected by Carr-Purcell-Meiboom-Gill Nuclear Magnetic Resonance (CPMG NMR) Spectroscopy. *Biochemistry* **56**, 692-703 (2017).
42. Mikulecky, P.J. & Feig, A.L. Heat capacity changes associated with DNA duplex formation: salt- and sequence-dependent effects. *Biochemistry* **45**, 604-616 (2006).
43. Fersht, A. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. (W.H. Freeman, New York; 1999).

Chapter 5: Conclusions and future directions

5.1. Preface

Conventional techniques for studying nucleic acid folding and assembly dynamics are labor-intensive, time consuming, and low throughput. The overarching goal of this thesis has been to develop methods that improve this bottleneck and offer robust physical descriptions of nucleic acid dynamics in biological and applied settings. To this end, several new global fitting analyses applied to thermal denaturation datasets for nucleic acids have been presented. These permit the rapid extraction of thermodynamic and kinetic parameters governing the folding and assembly dynamics of large nucleic acid ensembles. Furthermore, these approaches require simple instrumentation and are applicable in nearly any laboratory. The parameters obtained from these analyses aid in the understanding of nucleic acid function and the development of novel nucleic acid-inspired biomaterials. This chapter summarizes the main advances in this thesis and discusses current and future work applying the methodologies developed herein.

5.2. Conclusions and contributions to knowledge

This thesis has explored several new methods that we developed for characterizing the folding and assembly dynamics of large nucleic acid ensembles with an exquisite level of physical detail. The research presented in this thesis has made three distinct primary contributions to knowledge. These are: (i) that we have made substantial progress in the understanding of the folding and assembly dynamics of intra- and intermolecular GQ structures which are heavily implicated in regulating biological processes and also form key components of nanotechnology applications. We also performed the first literature review of GQ dynamics (a major portion of Chapter 1) where we grouped GQ conformational exchange into three main categories and highlighted the influence that each type of dynamics has on biological function. The goal of our

review was to emphasize that GQ dynamics are likely strongly linked to function in a similar manner to the dynamics of RNA and proteins. We suggested that GQ dynamics and biological function should be more thoroughly investigated and discussed several biophysical methodologies for studying conformational excursions in GQs. (ii) We have quantitatively characterized the assembly mechanism of the recently discovered fibers formed by CA and poly(A) strands, opening the door for the rational design of novel poly(A) fiber-based biomaterials. (iii) To address the complex folding and assembly dynamics of biomolecules such as GQs, aptamers, and nucleic acid fibers, we developed several new global fitting analyses that have allowed us to extract quantitative information on biomolecular ensembles featuring upwards of one hundred members.

The methods that we have developed are superior to previously existing approaches in terms of speed, cost, experimental burden (they require as little as one experiment), and level of detail that can be accessed in a relatively short amount of time (as little as one day). Furthermore, our methods are generally applicable in any laboratory, owing to their ability to be performed with data acquired on relatively simple and ubiquitous instrumentation such as the UV-Visible spectrophotometer. With the aim of making our methods accessible to the wider scientific community, we additionally performed extensive computer simulations of scenarios that may be encountered when using our methods to act as visual guides for novice users. Furthermore, the computer code to perform some of these global fitting analyses and simulations has been made freely available to the public so that it can be used as an educational tool for learning how to program and perform complex fitting routines. In conjunction with the three primary contributions to knowledge given above, the major conclusions and contributions to knowledge from each chapter in this thesis are outlined below.

5.2.1. Chapter 2: G-register exchange dynamics in guanine quadruplexes

We characterized ensembles of GQ structures from the promoter regions of human genes that undergo a form of folding dynamics we termed GR exchange. To understand these complex biomolecular ensembles, we developed a novel thermal denaturation global fitting analysis to extract the thermodynamic parameters governing transitions between individual GR isomer conformations and the unfolded state. The approach relies on making systematic mutations to wild-type GQ sequences undergoing GR exchange in order to trap their structures as mimics of individual GR isomers from the ensemble. Thermal denaturation data for the set of trapped and wild-type structures are then measured and globally fit with a model assuming the wild-type thermal profile is described by the folding parameters for the trapped mutants. The key assumption of this global fitting analysis is that the trapped mutants are thermodynamically equivalent to the corresponding GR isomers in the wild-type ensemble, which we demonstrated to be valid using Monte Carlo computer simulations. This method is particularly rapid (it needs as little as a few days of total input time) since it can be performed on a multi-sample absorbance spectrophotometer, and it requires small amounts of GQ sample (~nmol).

We applied this method to promoter GQs containing 2 (VEGFA), 4 (c-myc), and 12 (PIM1) GR isomers. To our knowledge, the PIM1 ensemble is one of the largest to be characterized to date. The GR isomer populations extracted from our global fitting analysis allowed us to compute the contributions to conformational entropy from GR exchange, which is a parameter that is typically difficult to directly measure for biomolecules. Our method shows that the increase in conformational entropy from populating multiple GR isomers in the wild-type ensemble can in theory amount to thermal stabilizations of up to ~20 °C relative to the most populated GR isomer. We additionally showed that shifting G-tracts in GQs undergoing GR exchange is a cooperative

process, meaning that the shifting of one G-tract depends on the relative positions of other exchanging G-tracts. As an example of the highly coupled GR exchange dynamics in GQs, we measured up to ~50-fold differences in G-tract shifting equilibrium constants depending on the relative positions of exchanging G-tracts in the PIM1 ensemble. Furthermore, we demonstrated that GR exchange is likely coupled to topological interconversion. Since promoter GQ stability is tied to gene expression levels, we concluded that GR exchange dynamics are a method to modulate the downregulation of genes, with the added benefit of providing multiple interaction motifs for GQ-binding proteins. A bioinformatic analysis of human promoter sequences revealed that many thousands of putative GQ-forming sequences can undergo GR exchange dynamics, highlighting how these dynamics could be a widespread, evolved regulatory mechanism for gene expression.

5.2.2. Chapter 3: Rapid characterization of biomolecular folding and binding interactions with thermolabile ligands by DSC

In Chapter 3, we developed a dual experimental and global fitting DSC technique that utilizes thermolabile ligands to rapidly characterize the folding and binding interactions of biomolecules from a minimal set of two experiments. The experimental method allows an entire DSC ligand binding series to be collected in a single experiment, in contrast to the traditional approach for studying binding by DSC which requires multiple separate experiments. The global fitting analysis enables the extraction of binding affinities to at least two ligands in competition-type binding scenarios. Our approach amounts to an order of magnitude reduction in the experimental time and sample required to characterize biomolecular folding and binding processes, with even greater levels of physical detail. Furthermore, we showed how the rate

constant for thermolabile ligand conversion can be obtained from one additional experiment performed with a longer high temperature equilibration time.

To validate our method, we applied it to two cocaine-binding aptamers, finding that their extracted cocaine and quinine binding parameters are in agreement with those derived from ITC analyses. As well, the rate constant for cocaine conversion extracted from our analysis was nearly identical to the literature value. The global analysis also revealed that the aptamers bind to the thermal conversion product benzoylecgonine with a weak affinity (roughly mM). Since no binding to this ligand was found by ITC, we concluded that our DSC method can be used to extract weakly-binding ligand affinities, in addition to the traditional use for ultra-tight binding ligands. We further performed a series of computer simulations as visual guidelines for more complicated situations that may be encountered when using thermolabile ligands in DSC experiments such as when the folding and binding kinetics are slow, or the biomolecule irreversibly aggregates at high temperature. The extension of our analysis to these non-equilibrium DSC experiments was presented and discussed. The computer code for performing our global fitting analysis and computer simulations of DSC folding and binding experiments with thermolabile ligands is freely available in the Supplementary Information PDFs at the corresponding Chemical Communications and Journal of Visualized Experiments article webpages.

5.2.3. Chapter 4: Mapping the energy landscapes of supramolecular assembly by thermal hysteresis

We developed a model-free analysis of multi-scan rate TH datasets to extract effective supramolecular assembly and disassembly reaction orders. We demonstrated how the effective orders provided by our method are exquisitely sensitive to the highest free energy barriers in the

assembly pathway, and the size of the nucleus in cooperative supramolecular polymerizations. Importantly, the model-free analysis can be easily performed with standard spreadsheet software, making it highly applicable for users who wish to obtain information on the nature of supramolecular assembly pathways as a function of temperature but do not have experience in data fitting. In combination with our model-free approach, we developed a global fitting method for TH datasets that yields the barrier energies and rate constants governing the kinetics of supramolecular assembly. The kinetic constants permit the calculation of affinities that can be used to predict assembly propensities as a function of temperature and monomer concentration. The global fitting method also enables the ability to track the distributions of polymer sizes in thermally-driven assembly. From our globally-fit kinetic models, we developed several analytical expressions for calculating effective assembly orders. These can be used to monitor the dominant energy barriers and fluxes throughout supramolecular assembly reactions. The combined model-free and global fitting of TH datasets is particularly rapid for acquiring highly detailed information on supramolecular assembly as a function of temperature, needing as little as one day of combined experimentation and analysis time. This represents an enormous reduction in user input compared to traditional methods for characterizing supramolecular assembly, which require repeating experiments over multiple separate temperatures and monomer concentrations, followed by a fitting procedure.

In application to the assembly of a tetrameric GQ, we found model-free assembly orders that suggested temperature-dependent variations in the assembly pathway. The global fitting analysis revealed that these orders are the result of an assembly pathway containing dimeric and trimeric intermediates where the dominant energy barrier shifts from being early (dimer limited) to late (trimer and tetramer limited) as the temperature is increased. We then applied our model-

free analysis to the recently discovered assembly of CA and poly(A) fibers, extracting orders that were roughly 3 through the assembly transition. We globally fit a cooperative polymerization model for supramolecular assembly, finding that poly(A) fiber elongation is driven by the favorable addition of a fourth poly(A) strand to an energetically unfavorable nucleus containing three poly(A) monomers. Importantly, this is consistent with the orders obtained from the model-free analysis. We also performed simulations of TH data for several commonly encountered supramolecular polymerizations to show that the assembly orders obtained from our analysis are closely linked to the size of the nucleus, demonstrating the generality of our model-free methodology in characterizing cooperatively assembling systems.

5.3. Future directions

5.3.1. Reconstructing parallel folding pathways in GQs by TH

We are currently engaged in a follow-up study based on GR exchange dynamics in GQs (Chapter 2). In Chapter 2, significant advances were made in the understanding the equilibrium folding dynamics of GQs from the promoter regions of human genes. We are now interested in developing a TH global fitting method to extract information on the folding kinetics of GQs that can populate multiple GR isomers. Relatively few investigations have examined the folding pathways of conformationally heterogeneous GQs from the unfolded state (i.e. upon duplex opening in the cell). Therefore, quantitative descriptions of the influence that populating multiple conformations has on the overall GQ folding rate is highly desirable for understanding putative effects on biological function. For example, folding via multiple parallel pathways can in theory accelerate the net folding rate, which in turn could enhance the ability to modulate gene expression by disrupting polymerase read through. Having the option to rapidly fold into distinct GQ

structures with different kinetic probabilities could also play a role in recruiting proteins to the promoter regions of genes for regulatory purposes[1].

The experimental and fitting work described in the following paragraphs was primarily performed by an excellent undergraduate student, Christopher Hennecker, under the supervision of myself and Dr. Anthony Mittermaier. Our study draws on the global fitting methods developed in Chapter 2 and Chapter 4 to reconstruct the parallel folding pathways of a GQ that can adopt four GR isomers. We are using a slightly longer version of the c-myc GQ sequence where have performed mutations to trap the four GR isomers (Table 5.1). In addition, we have made a set of four mutant "half-trapped" sequences, where only one of the two exchanging G-tracts has been mutated (Table 5.1). The half-trapped sequences undergo exchange between only two out of the four possible GR isomers, allowing us to obtain information on the folding kinetics and thermodynamics of GR exchange on a subset of the wild-type ensemble when one of the exchanging G-tracts is locked into different positions.

Table 5.1. The sequences being investigated in our study of the effects of parallel folding pathways on GQ folding kinetics. The L and X in the sequence names correspond to long and exchanging respectively. The red bold I indicates a dG>dI mutation.

Name	Sequence
c-myc L	5′-**TGAGGGTGGGGAGGGTGGGGAA**-3′
5X L	5′-TGAGGGT**I**GGGAGGGTGGGGAA-3′
3X L	5′-TGAGGGTGGG**I**AGGGTGGGGAA-3′
X5 L	5′-TGAGGGTGGGGAGGGT**I**GGGAA-3′
X3 L	5′-TGAGGGTGGGGAGGGTGGG**I**AA-3′
55 L	5′-TGAGGGT**I**GGGAGGGT**I**GGGAA-3′
35 L	5′-TGAGGGTGGG**I**AGGGT**I**GGGAA-3′
53 L	5′-TGAGGGT**I**GGGAGGGTGGG**I**AA-3′
33 L	5′-TGAGGGTGGG**I**AGGGTGGG**I**AA-3′

Briefly, we performed thermal denaturation experiments on the set of wild-type and trapped sequences at low [K^+] and fast temperature scan rates to induce TH. Using the global fitting approaches developed in Chapters 2 and 4, we fit the multi-scan rate TH datasets for all nine sequences simultaneously. The global fit assumes a model where the unfolded state can fold via four parallel pathways into each of the four GR isomers (Figure 5.1). The parameters from fits to the four full trapped mutants were used to calculate the TH profiles for the wild-type c-myc sequence. Concurrently, the half-trapped mutant TH profiles were fit with the parameters from their corresponding full-trapped sequences. This acts as a fit constraint because the half-trapped sequences comprise a subset of the conformations available to the wild-type c-myc sequence.

Therefore, the parameters describing the full-trapped TH datasets should describe the profiles for both the half-trapped and wild-type c-myc sequences.

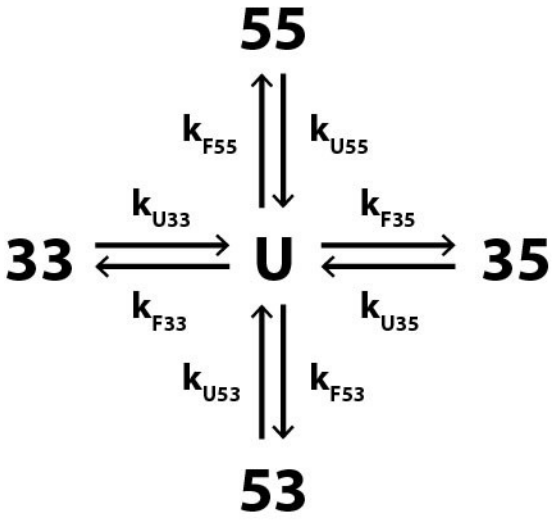


Figure 5.1. Parallel folding pathways in the extended c-myc GQ sequence. The unfolded state folds into the four-membered ensemble via four parallel pathways. The rate constants for folding (F) into and unfolding (U) from each GR isomer are indicated with the corresponding GR isomer numbers.

The global fits are in excellent agreement with the experimental data across the board (Figure 5.2a), confirming that the trapped mutants are good mimics of the corresponding wild-type GR isomers. The optimized global fit parameters (Figure 5.2b) reveal that the folding of the wild-type ensemble is accelerated by a factor of 2 relative to the fastest folding GR isomer, since folding via four parallel pathways is given by the net rate constant $k_{FWT} = k_{F33} + k_{F35} + k_{F53} + k_{F55}$. This influence on the folding rate is consistent with our previous study (Chapter 2) where we demonstrated that the wild-type folding equilibrium constant is improved by roughly 2-fold from populating four GR isomers.

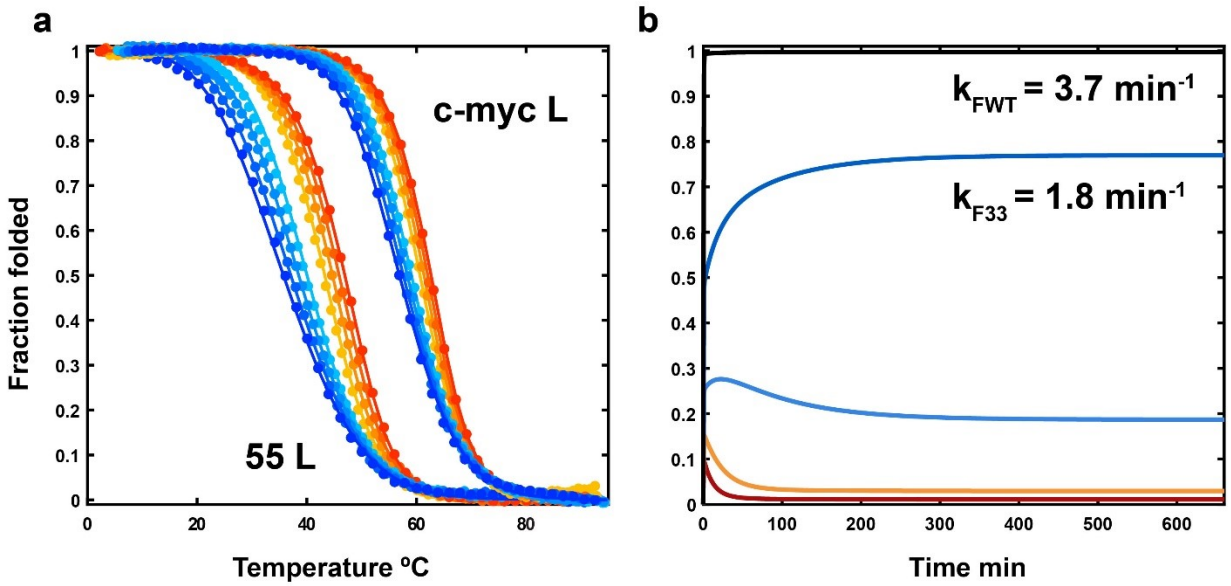


Figure 5.2. Reconstructing parallel folding pathways in GQs by TH. (a) Fraction folded TH profiles for the c-myc 55 and wild-type L GQ sequences. Experimental data and fits are shown as colored points and lines respectively. Dark to light blue and red to light orange indicate fastest to slowest cooling and heating scan rates respectively. Fits were performed with TH data for all 9 sequences in Table 5.1, however only these two datasets are shown for clarity. (b) Isothermal annealing fraction folded profiles for the four c-myc wild-type L GR isomers calculated using the global fit parameters from (a) at 35 °C. Folding rate constants for the wild-type and fastest-folding GR isomer are given in the top right of the panel.

One highly important benefit to performing global fits of TH data is that we extract kinetic parameters that permit the examination of how the GR isomer population distributions evolve isothermally over time, starting from the unfolded state in analogy to biological conditions. We performed isothermal simulations of the unfolded to folded transition for the wild-type c-myc ensemble, finding that the short timescale population distribution features substantially greater proportions of the lesser-stable GR isomers relative to their equilibrium values (Figure 5.2b). This implies that folding into lesser-stable GR isomers not only accelerates the net folding rate - it may also influence protein binding under biological conditions (i.e. during transcription) to a much

greater extent than is reflected by the equilibrium population distribution. We are currently investigating the isothermal folding kinetics of the wild-type and trapped mutant c-myc structures by NMR spectroscopy in order to provide experimental support for the kinetic intermediate populations and their role in accelerating the global folding rate. We expect to see rapid initial increases to relatively large populations of the lesser-stable GR isomers, followed by slower relaxation to the equilibrium mixture where the two most stable GR isomers are populated to nearly 100%.

5.3.2. Applications to other systems

The research presented in this thesis has laid the groundwork for applications to several other highly interesting nucleic acids implicated in biological function and employed in nanotechnology applications. Currently, a member of the Mittermaier laboratory is exploring the folding kinetics of tandem non-canonical GQ repeats from the human genome using the TH methods we have developed over the past two years. Our methods are also highly amenable to the folding dynamics of genomic i-motif sequences. For example, a future member of the Mittermaier laboratory could examine the extent of protonation at the transition state of i-motif folding using TH profiles collected as a function of solution pH in a Φ-value type analysis[2]. A separate TH analysis in molecular crowding conditions that simulates the intracellular environment, collected as a function of pH, would then reveal how the extent of i-motif protonation during folding shifts in response to the constricted milieu. This study would provide clues as to how i-motifs are stabilized inside the cell[3].

With respect to poly(A) fiber formation, the Sleiman laboratory is currently developing modified fibers which can be treated with the approach developed in Chapter 4 to elucidate the

changes to the assembly pathway relative to the system characterized herein. A related project would be to address the question of whether poly(A) fibers form *in-vivo*[4]. Biological CA-like small molecules might organize around the poly(A) tails of mRNA strands to induce fiber assembly. An experimental or computational screen of biological CA-like small molecules may reveal other candidates for poly(A) fiber formation. In theory, it is possible that poly(A) fibers act as an organizing center for mRNA translation where the 3' poly(A) tails are docked in the fiber assemblies and the 5' ends remain free in solution to be acted on by the ribosome. Interestingly, membraneless organelles are enriched with poly(A) RNA molecules[5]. The liquid protein phase inside membraneless organelles has a dielectric constant similar to acetonitrile[6] which could stabilize the hydrophobic surfaces of the fiber rosettes. Combined with the presence of CA-like biological small molecules, the assembly of poly(A) fibers may be quite favorable in these environments. A simple demonstration of poly(A) fiber assembly within simulated membraneless organelles would provide proof of principle for this concept.

Another exciting extension of our work would be to apply the model-free and global fitting analysis of supramolecular TH datasets to cooperatively-assembling systems that exhibit pathway complexity[7], i.e. assembly that occurs via multiple pathways with unique intermediate and product structures. The solution conditions and scan rate could be used to selectively drive assembly down individual pathways and the resulting datasets could be globally fit by a multi-pathway kinetic model to extract the determinants of pathway selection over a widely sampled temperature range. In conjunction with the project dedicated to developing modified poly(A) fibers described above, the resulting parameter sets may be used to develop structure-kinetics-function relationships for the rational design of novel biomaterials.

5.4. List of publications

1. Harkness, R. W., V, and Mittermaier, A. K. G-register exchange dynamics in guanine quadruplexes. *Nucleic Acids Research* **44(8)**, 3481-3494 (2016). **Cover article***.

2. Stabilization of i-motif structures by 2′-β-fluorination of DNA. Abou Assi, H., Harkness, R. W., V, Martin-Pintado, N., Wilds, C. J., Campos-Olivas, R., Mittermaier, A. K., González, C., and Damha, M. J. *Nucleic Acids Research* **44(11)**, 4998-5009 (2016).

3. Harkness, R. W., V, Slavkovic, S., Johnson, P. E., and Mittermaier, A. K. Rapid characterization of folding and binding interactions with thermolabile ligands by DSC. *Chemical Communications* **52**, 13471-13474 (2016).

4. Harkness, R. W., V, and Mittermaier, A. K. G-quadruplex dynamics. *Biochimica et Biophysica Acta – Proteins and Proteomics* **1865(11B)**, 1544-1554 (2017).

5. Harkness, R. W., V, Johnson, P. E., and Mittermaier, A. K. Measuring biomolecular DSC profiles with thermolabile ligands to rapidly characterize folding and binding interactions. *The Journal of Visualized Experiments* **129**, e55959 (2017).

6. Harkness, R. W., V, Avakyan, N., Sleiman, H. F., and Mittermaier, A. K. Mapping the energy landscapes of supramolecular assembly by thermal hysteresis. *Nature Communications*, in review (2018).

5.5. References

1. Bessi, I., Jonker, H.R., Richter, C. & Schwalbe, H. Involvement of Long-Lived Intermediate States in the Complex Folding Pathway of the Human Telomeric G-Quadruplex. *Angew Chem Int Ed Engl* **54**, 8444-8448 (2015).
2. Fersht, A.R. & Sato, S. Phi-value analysis and the nature of protein-folding transition states. *Proc Natl Acad Sci U S A* **101**, 7976-7981 (2004).
3. Dzatko, S. et al. Evaluation of the Stability of DNA i-Motifs in the Nuclei of Living Mammalian Cells. *Angew Chem Int Ed Engl* **57**, 2165-2169 (2018).
4. Avakyan, N. et al. Reprogramming the assembly of unmodified DNA with a small molecule. *Nat Chem* **8**, 368-376 (2016).
5. Visa, N., Puvion-Dutilleul, F., Harper, F., Bachellerie, J.P. & Puvion, E. Intranuclear distribution of poly(A) RNA determined by electron microscope in situ hybridization. *Exp. Cell Res.* **208**, 19-34 (1993).
6. Nott, T.J., Craggs, T.D. & Baldwin, A.J. Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nat Chem* **8**, 569-575 (2016).
7. Korevaar, P.A. et al. Pathway complexity in supramolecular polymerization. *Nature* **481**, 492-496 (2012).

The only way out is through.