New biophysical assays to study biomacromolecular folding, assembly, and function

Christopher D. Hennecker
Department of Chemistry, McGill University
Montréal, Québec, Canada
Submitted January 2024

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

Table of Contents

Abstract

Biomacromolecules are fundamental in virtually all aspects of biology. Developing a deep understanding about how these complex molecular systems behave is crucial for understanding their biological function and exploiting their physical properties for new technologies. For example, the self-assembly of these biomacromolecules can produce interesting materials which exhibit emergent physical properties such as self-healing and stimuli responsiveness. Furthermore, both enzymes and non-canonical DNA secondary structures are popular targets for small molecule therapeutics. However, in order to develop these molecules, scientists need to know how stable these structures are, how strongly they interact with their targets, and how to improve their desired properties. However, due to their highly complex nature, there are many examples of when conventional experimental techniques fail. Thus, new methods which enable quantitative characterization of these more complex systems are highly desirable.

This thesis explores novel biophysical analyses that combine common laboratory equipment with modern day computational power and mathematical modelling to address the need for rapid and cost-effective biomacromolecular characterization. Chapter 2 details a global-fitting analysis for non-equilibrium thermal denaturation experiments and its application to the folding dynamics of guanine quadruplexes (G4s), four stranded non-canonical nucleic acid structures formed from guanine rich sequences that are implicated in wide variety of cancers. We demonstrate that these sequences can fold into several different structures via parallel pathways which significantly increases their folding rate, potentially influencing their biological function.

Chapter 3 proposed the concept of G4 containing regions (G4CRs). We then developed a bioinformatic algorithm which characterizes these G4CRs based on their length and total number of G4 structures. This algorithm was applied to human promoter sequences where we found G4CRs that were up to several hundred nucleotides long and had the potential to form thousands of G4 structures. These polymorphic G4CRs were clustered directly adjacent to the transcription start site, suggesting that polymorphism has a functional role in biology.

In Chapter 4, we developed an experimental approach based off cyclic heating and cooling ramps which can measure thermodynamic information on slowly assembling supramolecular structures, information which was previously unobtainable with any other method. We used this technique to study the co-assembly of poly-adenosine strands and cyanuric acid (CA) into long supramolecular fibres. We discovered that roughly one third of CA binding sites were unoccupied, which has implications for the use of this system as a drug delivery vehicle.

Finally, Chapter 5 describes how to measure the binding kinetics of covalent enzyme inhibitors using isothermal titration calorimetry (ITC). Our new method allowed for faster and more robust characterization of these inhibitors when compared to conventional methods and removes the need for modified or spectroscopically active substrates as ITC is able to directly measure the rate of enzymatic catalysis. Together, these approaches represent powerful new additions to researchers' toolkit for rigorous characterizations of biomacromolecular folding, assembly, and function.

Resume

Les biomacromolécules ont un rôle fondamental dans pratiquement tous les aspects de la biologie. Il est essentiel de développer une compréhension profonde du comportement de ces systèmes moléculaires complexes pour comprendre leur fonction biologique et exploiter leurs propriétés physiques dans les nouvelles technologies. Par exemple, l'auto-assemblage des biomacromolécules peut mener à des matériaux intéressants avec des propriétés physiques émergentes, telles que l'autorégénération et le contrôle par stimuli. De plus, les enzymes et les structures secondaires non canoniques de l'ADN sont des cibles populaires pour les thérapies à petites molécules. Pour développer ces molécules, les scientifiques doivent savoir quelle est la stabilité de ces structures, de quelle manière elles interagissent avec leurs cibles et comment améliorer les propriétés souhaitées. Cependant, en raison de la nature extrêmement complexe de ces systèmes, il existe de nombreux exemples où les techniques expérimentales conventionnelles échouent. Par conséquent, de nouvelles méthodes permettant une caractérisation quantitative de ces systèmes plus approfondies sont en demande.

Cette thèse explore des analyses biophysiques innovantes qui combinent les équipements de laboratoire courants avec la puissance de calcul moderne et la modélisation mathématique pour adresser la nécessité d'une caractérisation biomacromoléculaire rapide et rentable. Le Chapitre 2 détaille une méthode d'analyse d'ajustement global pour les expériences de dénaturation thermique non équilibrée et son application à la dynamique de repliement des quadruplexes de guanine (G4), des structures nucléiques non canoniques à quatre brins formés à partir de séquences riches en guanine qui sont, entre autres, impliquées dans une grande variété de cancers. Nous démontrons que ces séquences peuvent se replier en plusieurs structures différentes via des voies parallèles, ce qui augmente considérablement leur vitesse de repliement, ce qui pourrait influencer leur fonction biologique.

Le Chapitre 3 propose le concept des régions contenant des G4 (G4CR). Nous avons ensuite développé un algorithme bioinformatique qui caractérise ces G4CR en fonction de leur longueur et de leur nombre total de structures G4. Cet algorithme a été appliqué à des séquences de promoteurs humains où nous avons trouvé des G4CR pouvant atteindre plusieurs centaines de nucléotides et ayant le potentiel de former des milliers de structures G4. Ces G4CR polymorphes étaient regroupés directement à côté du site de début de transcription, ce qui suggère que le polymorphisme joue un rôle fonctionnel en biologie.

Dans le Chapitre 4, nous avons développé une approche expérimentale basée sur des rampes de chauffage et de refroidissement cycliques qui permet de mesurer des informations thermodynamiques sur les structures supramoléculaires à assemblage lent, informations qui n’étaient auparavant pas possible d’obtenir avec d’autres méthodes. Nous avons utilisé cette technique pour étudier le co-assemblage de brins poly-adénosine et de l’acide cyanurique (CA) en longues fibres supramoléculaires. Nous avons découvert qu'environ un tiers des sites de liaison au CA étaient inoccupés, une découverte qui a des implications pour l'utilisation de ce système comme véhicule de livraison de médicaments.

Enfin, le Chapitre 5 décrit comment mesurer les cinétiques de liaison des inhibiteurs covalents en utilisant la calorimétrie à titration isothermique (ITC). Notre nouvelle méthode a permis une caractérisation plus rapide et plus robuste de ces inhibiteurs par rapport aux méthodes conventionnelles et élimine le besoin de substrats modifiés ou spectroscopiquement actifs, car l'ITC peut mesurer directement la vitesse de la catalyse enzymatique. Ensemble, ces approches représentent des ajouts puissants aux outils des chercheurs pour des caractérisations rigoureuses du repliement, de l'assemblage et de la fonction des biomacromolécules.

This thesis is dedicated to my grandfather – Dusty Rhodes – an inventor and musician who never had the opportunities afforded to me. Dusty passed away in 1999 at the young age of 59, but his memory lives on in my family, and my music.

Acknowledgements

Wow, am I actually done? It sure has been a long journey since I first started at McGill back in 2013, and to be honest, I can't believe I've finally finished – it's surreal. Back in fourth grade I remember answering the question of "What do you want to be when you grow up?" with "Scientist". Back in 10th grade I set my sights on getting a PhD from McGill University. Nowadays? Well, I'm running out of goals! So, I guess it's time to come up with some new ones, but before I start on that I want to take the time to acknowledge my fantastic support network which has gotten me through the past decade.

Of course, there's no way you could rank the people who have helped me along the way – that would be both ridiculous and unfair – but if you did the clear winner would be my parents, Elaine and Kim Hennecker. Throughout my life they have consistently supported and encouraged me to reach my potential. Whenever I've needed to vent, needed advice, or just wanted to chat, they have always been there to lend their ear. In a similar vein, there's no one who understands me more than my brother Aaron who's let me crash at his place when I visit home and always invites me out with his friends for some much-needed rest and relaxation. Finally, my partner Alexia Piercey (and our cat Esteban!) deserves a lot of credit for helping me get this thesis across the finish line. Her love and support throughout this process gave me the opportunity focus on this thesis, without falling into a pit of neglected cleaning duties and take-out delivery orders. Not to mention a specific quote from her from back on September 23rd 2023, which was – "You need to stop talking about your thesis, and actually start it".

Now what is a PhD student without their PhD supervisor? The answer, probably not a doctor. Professor Anthony Mittermaier is a special kind of person, not only is he smart, caring, and funny – but he's also particularly good at working in a Lord of the Rings or music reference into a conversation about science. He's cultivated an excellent research environment, and the Mittermaier lab has been full of amazing people throughout the years. I also want to the thank my committee – Prof. Matthew Harrington and Prof. Hanadi Sleiman – for their support and guidance throughout my PhD. Finally, before this section gets too long. I want to thank all of my friends, collaborators, and bandmates for putting up with me for the last five years – the past years have truly been some of the best of my life.

Contribution to original knowledge

This thesis contains several key original contributions to knowledge. Chapters 2-4 are already published in peer-reviewed journals, and Chapter 5 is currently in preparation for submission.

Chapter 2 uses a combination of mutagenesis, thermal hysteresis kinetic experiments, and global analysis to study the individual folding pathways of a conformationally heterogenous guanine quadruplex sequence located in the promoter of the c-MYC oncogene. We discovered that this quadruplex did indeed have multiple folding pathways, resulting in a net increase in the effective folding rate of the wild-type sequence. This result has important implications for the biological function of the G4, and is discussed in detail in the context of current literature.

Chapter 3 introduces a new bioinformatic tool for characterizing the structural heterogeneity of guanine quadruplex containing regions in the human promoters. The algorithm, GReg, finds long multimeric sequences and classifies them based on the number of possible guanine quadruplex structures, the number of possible guanine quadruplexes that can fold in tandem, the G-richness, and the length of the region. We made the discovery that polymorphism is much more prevalent than previously thought, with some guanine quadruplex containing regions with thousands of possible isomers. We discuss these results in the context of Chapter 2, where we saw that the presence only four distinct isomers doubles the effective folding rate. For systems with thousands of isomers, this effect may be much larger.

Chapter 4 develops the first method to obtain thermodynamic information from slowly assembling supramolecular systems. We begin by using extensive simulations to show how ignoring the effects of thermal hysteresis leads to completely incorrect characterization of these systems. We then introduce transient equilibrium mapping (TREQ) and describe best practices for setting up and analyzing these experiments. Through simulations, we show when and where the TREQ method is suitable and give two criteria which the system must follow for the TREQ experiment to work. We used TREQ to determine the small-molecule loading efficiency of polyadenosine-cyanuric acid supramolecular fibres and determine a loading efficiency of only 67%. To rationalize this, we develop a mathematical model, applicable to other multivalent systems, which we show provides a good description of the behaviour of our system.

Finally in Chapter 5, we show how to set up isothermal titration calorimetry experiments to measure the binding kinetics of two-step irreversible covalent inhibitors. We compare our method to the conventional ways of measuring this binding and show that our new method provides more robust characterization. Furthermore, while making this comparison we discovered a systematic flaw with time-dependent IC_{50} analysis and show how this flaw leads to underestimation of both binding constants. We then use our new method to measure 10 different covalent warheads and 10 different scaffolds and discuss how changes in these two parts of the molecules lead to changes in both the reactivity and affinity of the inhibitors. We made the discovery that the scaffold can have a large effect on the reactivity of the covalent warhead and discuss the implications of this in a drug-discovery effort.

Contribution of Authors

Chapter 1: Introduction and literature review of biomacromolecular structures

Chapter one was written and researched by Christopher Hennecker, with editing done by Prof. Anthony Mittermaier.

Chapter 2: Parallel reaction pathways accelerate folding of a guanine quadruplex

Adapted with permission from: Harkness, R. W.[†], Hennecker, C.[†], Grün, J. T., Blümler, A., Heckel, A., Schwalbe, H., & Mittermaier, A. K. (2021). Parallel reaction pathways accelerate folding of a guanine quadruplex. *Nucleic acids research*, 49(3), 1247-1262.

In Chapter 2, Prof. Anthony Mittermaier, Robert Harkness, and I conceptualized the research. Robert Harkness and I are co-first authors of the published manuscript. Robert Harkness and I synthesized the non-photolabile DNA sequences. I ran and analyzed all of the thermal hysteresis traces and circular dichroism spectra. Robert Harkness ran and analyzed the ^{1}H NMR spectra of each mutant. Tassilo Grün, Anja Blümler, Alexander Heckel and Prof. Harald Schwalbe, synthesized and analyzed the photolabile DNA sequences. Prof. Anthony Mittermaier, Robert Harkness, and I, discussed the results and wrote the original draft of the paper. I adapted the original manuscript to this thesis.

Chapter 3: Structural polymorphism of guanine quadruplex-containing regions in human promoters

Adapted with permission from: Hennecker, C., Yamout, L., Zhang, C., Zhao, C., Hiraki, D., Moitessier, N., & Mittermaier, A. (2022). Structural polymorphism of guanine quadruplex-containing regions in human promoters. *International Journal of Molecular Sciences*, 23(24), 16020.

In Chapter 3, Prof. Anthony Mittermaier and I conceptualized the GReg algorithm. I wrote all of the MATLAB code required to implement the algorithm on human promoters, and analyzed my results with Prof. Anthony Mittermaier. Lynn Yamout developed the equations to predict the stability of individual quadruplexes. Chuyang (Amos) Zhang translated the GReg algorithm into python, and along with David Hiraki, launched the

GReg Webserver. Chenzhi (Ian) Zhao helped in the early stages of algorithm development. Prof. Anthony Mittermaier and Prof. Nicolas Moitessier supervised the research. The original manuscript was written by Prof. Anthony Mittermaier, Lynn Yamout, and I. Adapting the original manuscript to this thesis was done by me.

Chapter 4: Using transient equilibria (TREQ) to measure the thermodynamics of slowly assembling supramolecular systems

Adapted with permission from: Hennecker, C. D., Lachance-Brais, C., Sleiman, H., & Mittermaier, A. (2022). Using transient equilibria (TREQ) to measure the thermodynamics of slowly assembling supramolecular systems. *Science Advances*, 8(14), eabm8455.

In Chapter 3, Prof. Anthony Mittermaier and Prof. Hanadi Sleiman conceived the project. All of the experimental data on the polyA-CA fibres and intermolecular quadruplex was acquired by Christophe Lachance-Brais. I acquired all the experimental data for the intramolecular guanine quadruplex along with the zinc porphyrin. I ran all the simulations and data analysis. Prof. Anthony Mittermaier and I interpreted the results. The original manuscript was written by Prof. Anthony Mittermaier, and me. Adapting the original manuscript to this thesis was done by me.

Chapter 5: Kinetic Isothermal titration calorimetry

In Chapter 5 Prof. Anthony Mittermaier and Prof. Nicolas Moitessier conceived the research. Felipe Venegas, Andres Reuda, and Guanyu (Chris) Wang expressed and purified the protein. Guanyu (Chris) Wang and Julia Stille synthesized the small molecule inhibitors and ran fluorescence experiments. Felipe Venegas ran all the isothermal titration calorimetry experiments. I analyzed all of the isothermal titration calorimetry data and used simulations to validate and compare the method. Prof. Anthony Mittermaier and I wrote the original draft of the manuscript.

List of Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables

List of Abbreviations

$[M]_c$	Critical monomer concentration
A	Adenine
A. mellifera	Apis mellifera
A. thaliana	Arabidopsis thaliana
AFM	Atomic force microscopy
ALT	Alternative lengthening mechanism
BTT	5-Benzylthio-1H-tetrazole
C	Cytosine
C. elegans	Caenorhabditis elegans
C. familiaris	Canis familliaris
CA	Cyanuric acid
CARPs	Consensus ankyrin repeat proteins
CD	Circular Dichroism
CL	Confidence level
d	deoxy
D. melanogaster	Drosophilia melanogaster
D. rerio	Danio rerio
DNA	Deoxyribonucleic acid
DOF	Degrees of Freedom
DSS	Sodium trimethylsilylpropanesulfonate
DX	Double crossover junctions
E. coli	Esherichia coli
FID	Free induction decay
G	Guanine
G. gallus	Gallus gallus
G4	Guanine Quadruplex
G4CR	Guanine quadruplex containing region
GR	G-register isomers
GReg	G4-containing region algorithm
GS	Goldstein-Stryer

HPLC	High performance liquid chromatography
HRMS	High-resolution mass spectrometry
I	Inosine
IC_{50}	Half-maximal inhibitory concentration
IDPC	Inhibitor concentration-dependent progress curve
ITC	Isothermal titration calorimetry
kb	kilobases
LNA	Locked nucleic acid
M	Monomer
M. mulatta	Macaca mulatta
M. musculus	Mus musculus
mRNA	Messenger RNA
M_s	Nucleus
NMR	Nuclear magnetic resonance
nt	Nucleotides
ODE	Ordinary differential equation
P. falciparum	Plasmodium falciparum
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PDB	Protein data bank
polyA	polydexoyadenosine
R. norvegicus	Rattus norvegius
RNA	Ribonucleic Acid
RPA	replication protein A
RSS	Residual sum-of-squares
S. cerevisiae	Saccharomyces cerevisiae
S. pombe	Schizosaccharomyces pombe
SELEX	Systematic evolution of ligands by exponential enrichment
smFRET	single-molecule Förster resonance energy transfer
T	Thymine
TBE	tris/borate/edta

TDIC$_{50}$	Time-dependent IC$_{50}$
TEMED	tetramethylethylenediamine
TH	Thermal hysteresis
TREQ	Transient equilibria
Tris	tris(hydroxymethyl)aminomethane
TSS	Transcription start site
U	Uracil
UTR	Untranslated region
UV	Ultraviolet
VNTR	Variable number tandem repeats
WT	Wild-type
Z. mays	Zea mays

Chapter 1: Introduction and literature review of biomacromolecular structures

1.1 Biomacromolecules

Biomacromolecules play fundamental roles in virtually all aspects of biology. They are large polymeric molecules made up of small monomeric units which exhibit diverse emergent physical properties based on their structure and sequence. In biology there are four main classes of biomacromolecules: nucleic acids, proteins, carbohydrates, and lipids. This thesis will focus on two of these classes: nucleic acids and proteins, which are both critical to the cell and have complex dynamics, folding and assembly pathways, and functions. This introduction will provide a background on the structural characteristics of these molecules along with their importance to biology. Three systems in particular will be highlighted: in section 1.2, guanine quadruplexes (G4s) and polyadenosine-cyanuric acid supramolecular fibres will be discussed as these systems are studied in detail in Chapters 2-4 of this thesis. In section 1.3 covalent enzyme inhibition will be introduced and common techniques to measure it will be discussed. Finally, two key biophysical analysis methods will be introduced, thermal analysis and isothermal titration calorimetry. These techniques are expanded upon throughout the chapters of this thesis, with each chapter describing the development of a new technique. The methods described in this thesis vary from incremental advances in classical methods, to the development of completely novel techniques which are able to measure properties of systems which were previously unobtainable by any other means.

1.2 Nucleic Acids

1.2.1 Nucleotides

Nucleic acids are biopolymers found in all cellular organisms and viruses[1, 2]. Their monomeric unit is referred to as a nucleotide, which consists of a nitrogenous base, a pentose sugar, and a phosphate group (*Figure 1.1a*). The pentose sugars are either ribose, which gives rise to ribonucleic acid (RNA), or deoxyribose, which give deoxyribonucleic acid (DNA). While deoxyribose is simply ribose which has been dehydrated at the C2’ position, this dehydration plays an important part in dictating the final structures and stabilities of both DNA and RNA. Both nucleic acids have substantial roles in biology, for DNA these include information storage and gene regulation, for RNA protein synthesis and signaling. The bases are planar aromatic heterocycles, either built from purine or pyrimidine scaffolds. Three bases are shared between DNA and RNA: Adenine (A), Guanine (G), and Cytosine (C). The fourth base in RNA is Uracil (U) and in DNA it is Thymine (T). As elaborated further in this section, interactions between these nucleotides give rise to a diverse range of secondary structures which leads to emergent physical properties necessary for biological function and exploitable in nanotechnology.

The non-planar nature of the pentose sugar causes the five-membered ring to "pucker", resulting in two major conformations: the C3' endo (North) and C2' endo (South) puckers. *Figure 1.1b* shows the orientations of these conformations, where the North conformation has the C3' carbon above the plane of the O4', with the opposite being true for the South pucker. These two conformations can interconvert but have large energetic barriers that have important implications for the overall structure of nucleic acids[3, 4]. The sugar pucker is influenced by interactions between the substituents at the four carbons of the sugar with hydrogen-bonding and steric hinderance playing a large part in determining the conformation.

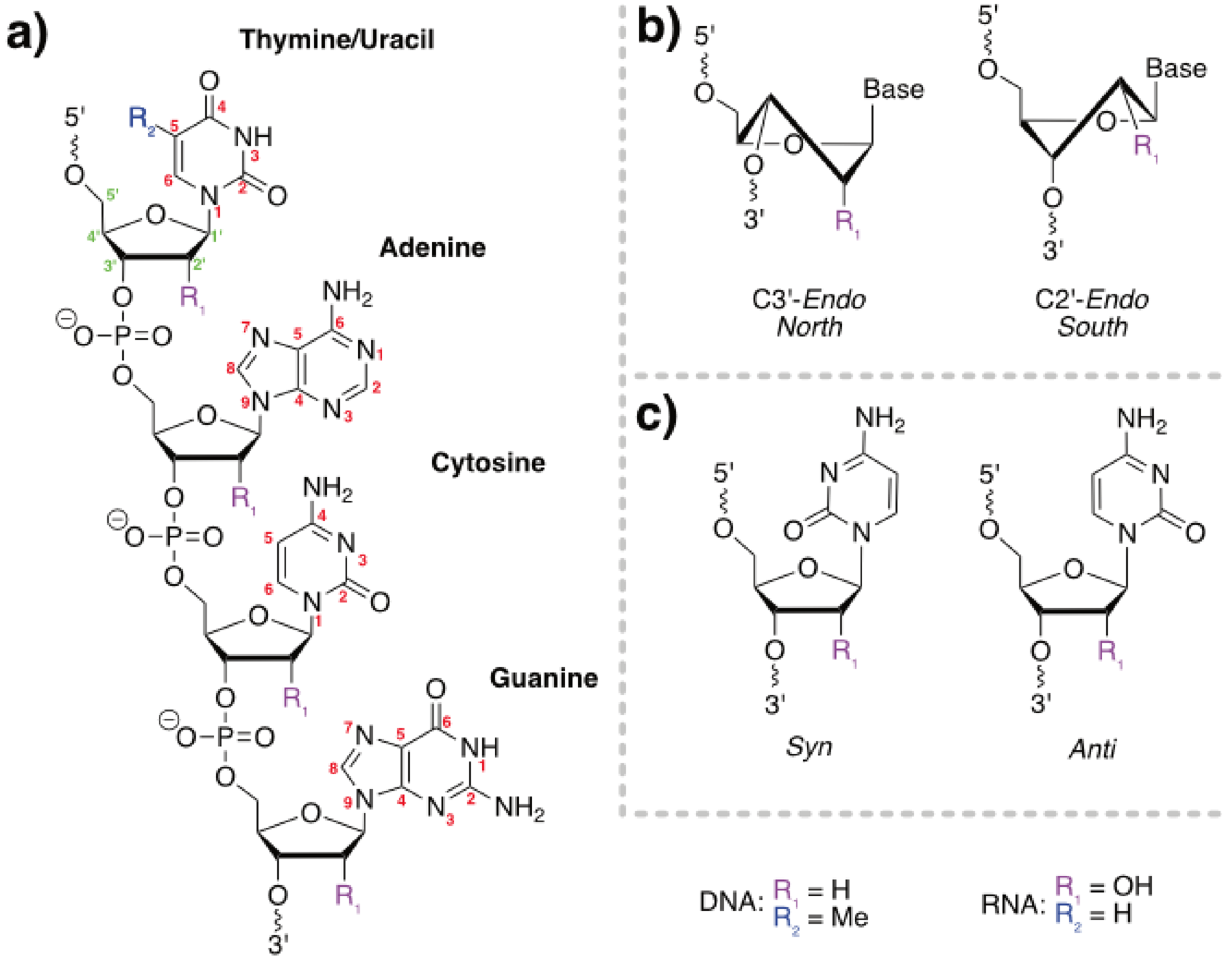


Figure 1.1: Nucleotide structures. a) a nucleic acid polymer consisting of the four basic nucleotides for DNA and RNA, glycosidic bond angles are shown in the anti-conformation. b) North (C3'-endo) and South (C2'-endo) sugar puckers. c) Syn and anti conformations shown with the Cytosine moiety.

The bases are connected to the sugars via a glycosidic bond located between the C1' of the pentose sugar and either the N9 of the purine or the N1 of pyrimidine bases. In biology this bond is in β-stereochemistry with the base located above the plane of the sugar. However, α-nucleotides can be synthesized and exhibit resistance to nuclease degradation and increased intracellular stability, making them an attractive option for nucleic acid therapeutics[5]. This glycosidic bond is able rotate giving rise to two main conformations, *anti* and *syn*. Shown in *Figure 1.1c*, the *anti*-conformation has the N1 of purines, and the N3 of pyrimidines facing away from the sugar resulting in an angle between -120° and 180°, angles in the region near -90° are described as *high anti*.

Conversely, in the *syn* conformation these atoms are angled towards the phosphate backbone and have angles between 0° and 90°. The orientation of this bond is highly sensitive to the identity of the base, pucker of the sugar, and hydrogen bonding pattern the bases are participating in[1].

The final component of nucleic acids is a phosphate group which connects two pentose rings via the C5' of the sugar below the phosphate and the C3' of the sugar above (*Figure 1.1a*). This allows polymeric nucleic acid strands to form, with a long phosphate and sugar backbone and a sequence of different bases. By convention the sequence of nucleic acid polymers is written starting with the base which has a terminal 5' phosphate and ending with the base with the terminal 3' OH. When referring to DNA, nucleotide sequences are often preceded by a small "d" denoting the deoxyribose sugar (dATCG), and RNA strands are preceded by a small "r" for ribose (rATCG). This phosphate backbone has highly negatively charged, which is generally stabilized by divalent cations such as Mg^{2+}, but monovalent cations can also lead to changes in stability and strucutre[6, 7]. DNA and RNA biomacromolecules can contain millions of monomeric units, which, as elaborated below, leads to a diverse array of secondary structures.

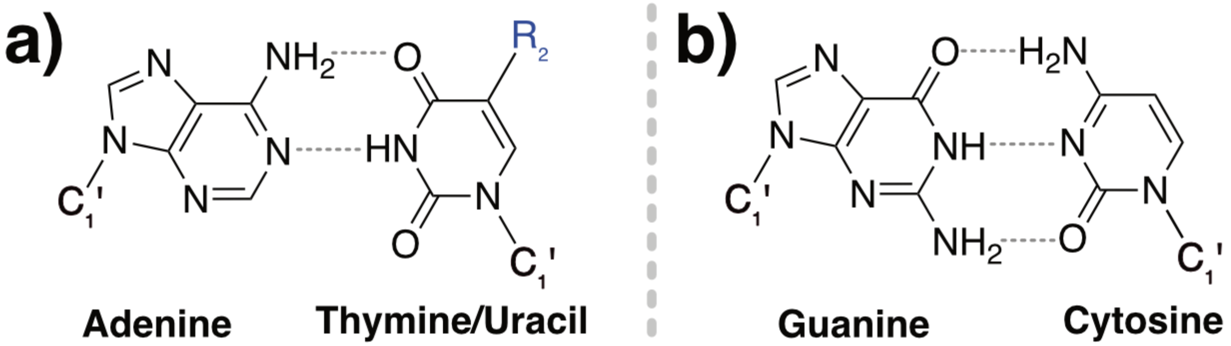


Figure 1.2: Watson-Crick Hydrogen bonding patterns. a) A-T base pairs with R_2 = Me, A-U base pairs with R_2 = H. b) G-C base pairs.

A key feature of nucleic acids is the ability of the bases to pair with each other in a well-defined pattern. This is accomplished via the hydrogen bonding patterns occurring between specific bases. In canonical base pairing, referred to as Watson-Crick base pairing, A will preferentially bind to either T or U, and C with G (*Figure 1.2*). These base pairs both contain a purine and a pyrimidine base and have similar distances between the C1' carbon of the pentose sugars, leading to a relatively constant dimensions for base pairs[8]. While the bases themselves are planar, the hydrogen bonds can have several different angles dictating the twist, buckle, roll and slide, along with others, between paired bases and stacked bases[9]. Other types of hydrogen bonding patterns taking advantage of different atoms for each base has been observed in nucleic acid structures, as well as during protein-DNA binding[10, 11].

1.2.2 Deoxyribonucleic acid structures

1.2.2.1 Duplex structures

The most common structure of deoxyribonucleic acid (DNA) is that of the double helix[12]. These structures are made up of two antiparallel nucleic acid strands which are held together by the interaction between complementary bases. Hydrogen bonding, π- π stacking, and London dispersion forces, all influence duplex formation[13, 14]. Within this family, double-helical structures can adopt several conformational forms, including the right-handed B-DNA, the left-handed Z-DNA, and the more compact right-handed A-DNA[1]. They can be formed either intermolecularly between two complementary DNA strands, or intramolecularly, when a single DNA strand has complementary regions and folds in on itself. The structure of B-DNA was first published in 1953 in a paper by James Watson and Francis Crick, which was part of a trio of papers detailing the fundaments of nucleic acid structures[15-17]. One of the key pieces of experimental evidence supporting their model was X-ray crystallography data acquired by Rosalind Franklin while under the supervision of Maurice Wilkins[18]. B-DNA is the most prevalent form under cellular conditions and is characterized by having Watson-Crick base pairing, *anti*-glycosidic bond angles and South sugar puckers *(Figure 1.3a)*. The Watson-crick base pairing in this structure has both phosphate backbones on the same side of the bases, leading to a distinct major and minor groove. These grooves play an important part in protein and small molecule binding[19, 20]. In B-DNA, the separation between the bases is nearly identical to the helical rise, causing the bases to be nearly perpendicular to the axis. Double helical nucleic acid sequences tend to be quite dynamic and are able to adopt transient Hoogsteen base pairing[21] and bulged residues[22], as well as being able to incorporate damaged bases[23].

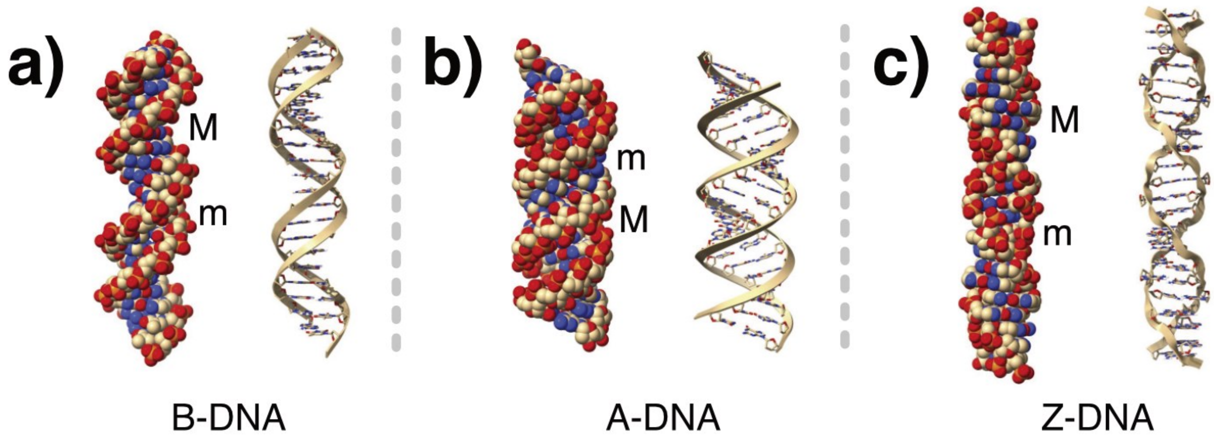


Figure 1.3: *Duplex structures of deoxyribonucleic acid. a) Space filling and cartoon representations of B-DNA. b) Space filling and cartoon representations of A-DNA. c) Space filling and cartoon representations of Z-DNA. M and m represent the major and minor grooves of all structures respectively. Adapted from Neidle and Sanderson With permissions[1].*

A-DNA is one of the other double helical structures and is a more compact right-handed DNA helix (*Figure 1.3b*). It is characterized by *anti*-glycosidic bonds and North sugar puckers. RNA tends to adopt A helices due to the 2'-OH group favoring the North sugar pucker[1]. Base pairs are displaced from the axis of the helix which leads to a hole being present through the middle of the duplex. The preference for B and A conformations is influenced by humidity, with A conformations being stable at low humidity and B conformations being stable at high humidity[1]. Z-DNA is a left-handed helix which occurs with specific nucleic acid sequences such as the repeating CGCGCG sequence, particularly under conditions of high salt (usually >2.5M NaCl) (*Figure 1.3c)*[24]. It has different sugar puckers and glycosidic angles depending on the identity of the base[25]. Purine bases have *syn*-glycosidic angles and the North sugar puckers. Pyrimidine bases have *anti*-glycosidic angles and south sugar puckers. While Z-DNA requires very specific conditions to form, it has been shown to occur near transcription start sites, and is implicated in gene expression[26].

1.2.2.2 Triplex structures

Triplex nucleic acid structures were first discovered just 3 years after the original B-DNA duplex structure was proposed[27]. They are formed from when a third nucleic acid strand binds to the major groove of a polypyrimidine and polypurine duplex (*Figure 1.4a*). The triplex-forming strand can be either an additional polypyrimidine or polypurine strand which forms Hoogsteen pairs with the purines of the duplex[28]. The hydrogen bonding patterns for the T-AT and C*-GC triplexes are shown in *Figure 1.4b/c,* in the case of the C*-GC triplex, half of the cytosine residues must be protonated*.* Triplexes can be made from a mixture of DNA and RNA strands, however the stability of these mixtures varies substantially[29]. The sugar puckers can also vary. They are generally South puckers[30], but RNA strands often take on the North pucker[31]. From a biological perspective, triplex DNA has garnered attention due to its potential role in gene regulation[32], DNA repair[33], recombination[34], and mutagenesis[35]. Triplex formation has been shown to be highly selective[36], however the close proximity of the phosphate backbones can cause interactions to be transient and make residency times for the third strand short, limiting its potential as a therapeutic target[1].

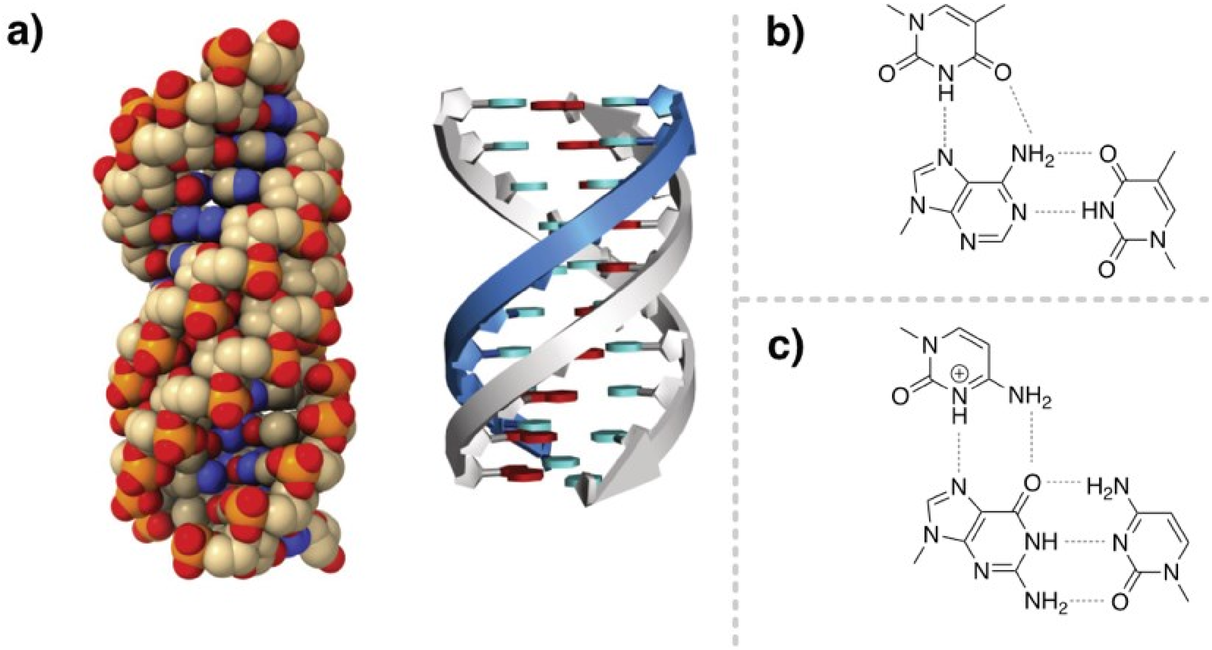


Figure 1.4: Triplex structures. a) Space filling and cartoon representations of a parallel triplex nucleic acid structure. A poly purine-pyrimidine duplex is shown in grey, with purine bases shown in red and pyrimidine bases shown in blue. It is bound to a third poly pyrimidine strand shown with a blue backbone and blue bases. b) Hydrogen bonding pattern of T-AT triplex, the Watson-Crick face of the adenine base is pointing to the right and the Hoogsteen face is pointing up. c) Hydrogen bonding pattern of C*-GC triplex, the Watson-Crick face of the guanine is pointing to the right and the Hoogsteen face is pointing up. Adapted from Neidle and Sanderson With permissions[1].

1.2.2.3 Tetrameric structures

Tetrameric DNA complexes deviate significantly from the canonical B-DNA double helix. Guanine quadruplexes, commonly referred to as G-quadruplexes or G4s, are one of these structures, forming from guanine rich nucleic acid sequences[37]. They are enriched in specific regions in the genome, such as telomeres and promoters, and have been implicated in a many cellular processes[38]. G4s can form from both DNA and RNA and are generally made up of four tracts of three guanines, referred to as G-tracts, separated by loops of one to seven nucleotides ($G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$). Structures with only two guanines in each G-tract have been shown to fold but are generally less stable[39]. Longer G-tracts of six or seven guanines have also been hypothesized[40], however experimental studies have shown that these sequences only form G4s with four tetrads[41]. In G4s, one guanine from each of the four G-tracts form planar arrangements referred to as G-quartets that are stabilized via Hoogsteen hydrogen bonding, and the coordination of cations to the O6 of each guanine (*Figure 1.5b*)[42]. Generally, this cation is monovalent such as K^+ or Na^+, Rb^+, Tl^+, NH_4^+, however under certain conditions divalent cations including Ca^{2+}, Pb^{2+}, Sr^{2+}, and Ba^{2+}, have also shown to promote quadruplex formation[43-54]. The identity of this cation has a large effect on both the stability of the G4 as well as its overall structure[42, 55 56]. Cations are also able to interact with nucleotides in the loops of quadruplexes, further stabilizing their structure[57].

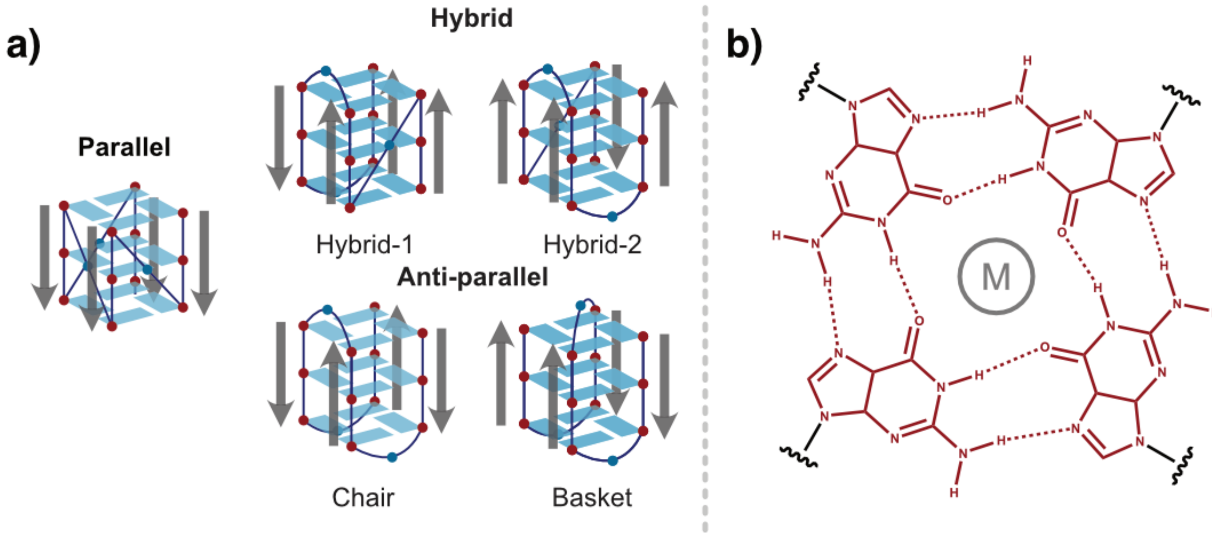


Figure 1.5: Guanine quadruplex structures. a) Possible loop arrangements for G4s. Red circles represent guanines with blue squares representing the G-quartets. Blue circles represent any nucleotide. Arrows show direction of the nucleic acid strand. b) Hydrogen bonding pattern of a G-quartet. M represents a metal ion (such as K^+ or Na^+).

There are three main types of topologies that quadruplexes can adopt. These different topologies are characterized by the directions of the four G-tracts relative to each other along with the orientations of the loops which connect them (*Figure 1.5a*)[58]. Parallel topology occurs when all four G-tracts are oriented in the same direction, anti-parallel topology occurs when only two of the four G-tracts are aligned, and hybrid topology occurs when three tracts are aligned[59]. In the parallel topology all the glycosidic bonds are either *anti* or *syn* and in both anti-parallel and hybrid there is a mixture[58]. G4s have been shown to be able to form both right and left handed structures[60, 61]. While it is still unclear why certain G4s adopt different topologies, G-tract length, loop length, and loop interactions have been shown to play roles[41, 62, 63] . Furthermore, RNA quadruplexes are exclusively found in the parallel conformation, due to their favoring the North sugar pucker[64].

The i-motif is another type of tetrameric nucleic acid structure which can form from C-rich regions of either DNA or RNA. These structures were first studied in 1963 and thought be dimers[65], however their true tetrameric structure was identified 30 years later in 1993[66]. They tend to form under slightly acidic conditions from cytosine-rich strands. Their structure consists of two parallel duplexes intercalated together in an antiparallel configuration (*Figure 1.6a*) stabilized by hemi-protonated cytosine-cytosine$^+$ base pairs, where one of the cytosines is protonated at the N3 position (*Figure 1.6b*). This unique base pairing makes i-motifs highly sensitive to pH, as values that are too high or too low cause the structure to become unstable[67]. They are the most stable near the *pKa* of the N3 of a cytosine, which is roughly 4.5, conditions under which roughly half of the cytosines are protonated[68]. Like G4s, i-motifs can adopt multiple topologies depending on the intercalation of the two duplexes, referred to as either 5'E, where the outermost CC$^+$ base pair occurs at the 5' end of the sequence, or 3'E where this base pair occurs at the 3' end of the sequence[69]. In i-motifs, the glycosidic bond tends to be *anti* and the sugar mainly adopts the North conformation[70].

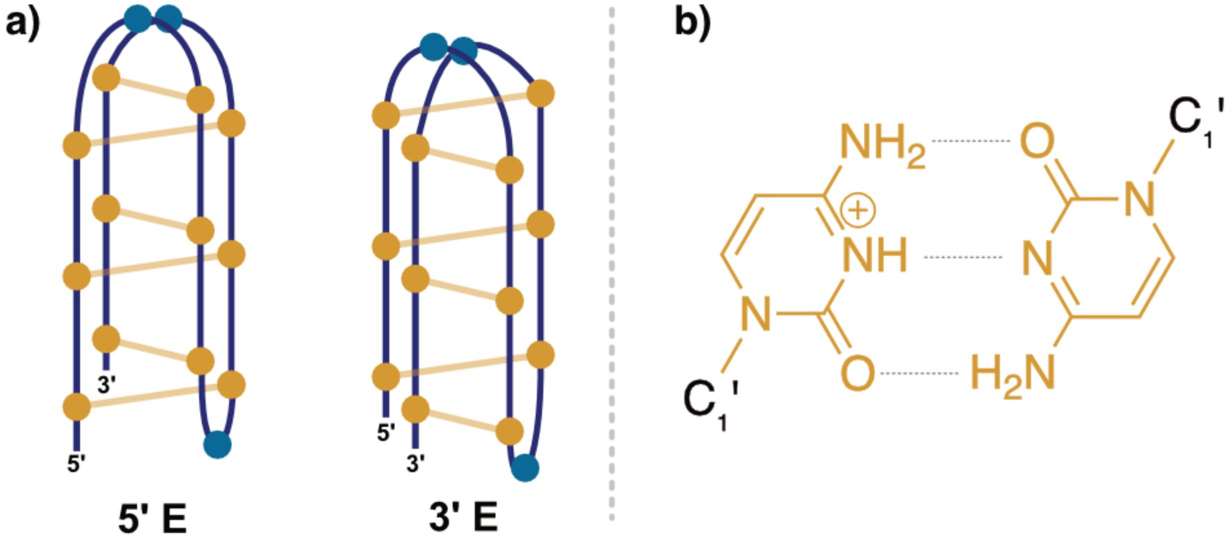


Figure 1.6: i-motif structures. a) Possible intercalation topologies. Yellow circles represent cytosines with yellow circles representing the C^+C base pairs. Blue circles represent any nucleotide. b) Hydrogen bonding pattern of a C^+C base pair.

The biological role of i-motifs is not well understood. Initially it was thought that they couldn't form under cellular conditions due to their stability being pH dependent[58]. However, it has since been shown that there are indeed sequences in the human genome that can form i-motifs at neutral pH[71]. i-motifs can be further stabilized by molecular crowding[72], secondary loop interactions[73], backbone interactions[74], and longer C-tracts[75]. Furthermore, non-natural chemical modifications such as locked nucleic acids (LNA)[76], 2'-Fana substitutions[77], and end ligation[78] can all increase i-motif stability. i-motifs have been visualized in cells using antibodies and several proteins have been found which bind i-motifs[79-81]. Like G4s they are enriched in the promoter regions of genes as well as the telomeres and have been shown to affect both transcription and telomerase activity[82, 83]. Interestingly, while G4s and i-motifs are located in the same places in the genome, several studies have shown that their formation is mutually exclusive, and suggested that this is due to steric hinderance and their stability under different conditions[84-86].

i-motifs find significant applications in biotechnology[87]. Their dynamic response to pH fluctuations makes them well-suited for pH-sensitive biosensors, enabling the detection of specific pH changes associated with biological processes or external stimuli[88]. Additionally, the pH-dependent stability of i-motifs can be leveraged in targeted drug delivery systems, allowing controlled release of therapeutic agents in response to acidic microenvironments[89]. Finally, they have been used to produce the first DNA molecular motor which is driven by pH changes[90].

1.2.2.4 Aptamers and Riboswitches

Nucleic acid aptamers are short sequences of nucleotides which bind selectively to non-nucleic acid molecules. These molecules were first described in two seminal papers in 1990[91, 92]. These papers outline well-defined procedures for developing aptamers based upon an iterative approach called systematic evolution of ligands by exponential enrichment (SELEX). The process begins with an initial library of nucleic acids which contains segments of randomized nucleotides as well as constant regions which allow for the hybridization of polymerase chain reaction (PCR) primers. Basic aptamers are created with the four natural nucleotides which causes the number of unique strands to grow as N^4, where N is the length of the randomized nucleotide sequence[93]. This initial library is then run through a selection process to find sequences which have the highest binding affinity for the target molecule. This selection process can be based on capillary electrophoresis[94], or immobilization[95] among other techniques, and may include a counterselection step to prevent nonspecific binding. The selected sequences are then amplified using error-prone PCR to introduce a few new mutations, creating a new pool of potential aptamer sequences which are then subjected to selection again. After several rounds of SELEX, aptamers can be found which are selective for the desired ligand with affinities in the nanomolar range[96]. Some groups have increased the chemical space of aptamers by using non-canonical nucleotides[97] as well as functionalizing the nucleic acids with different chemical groups[98].

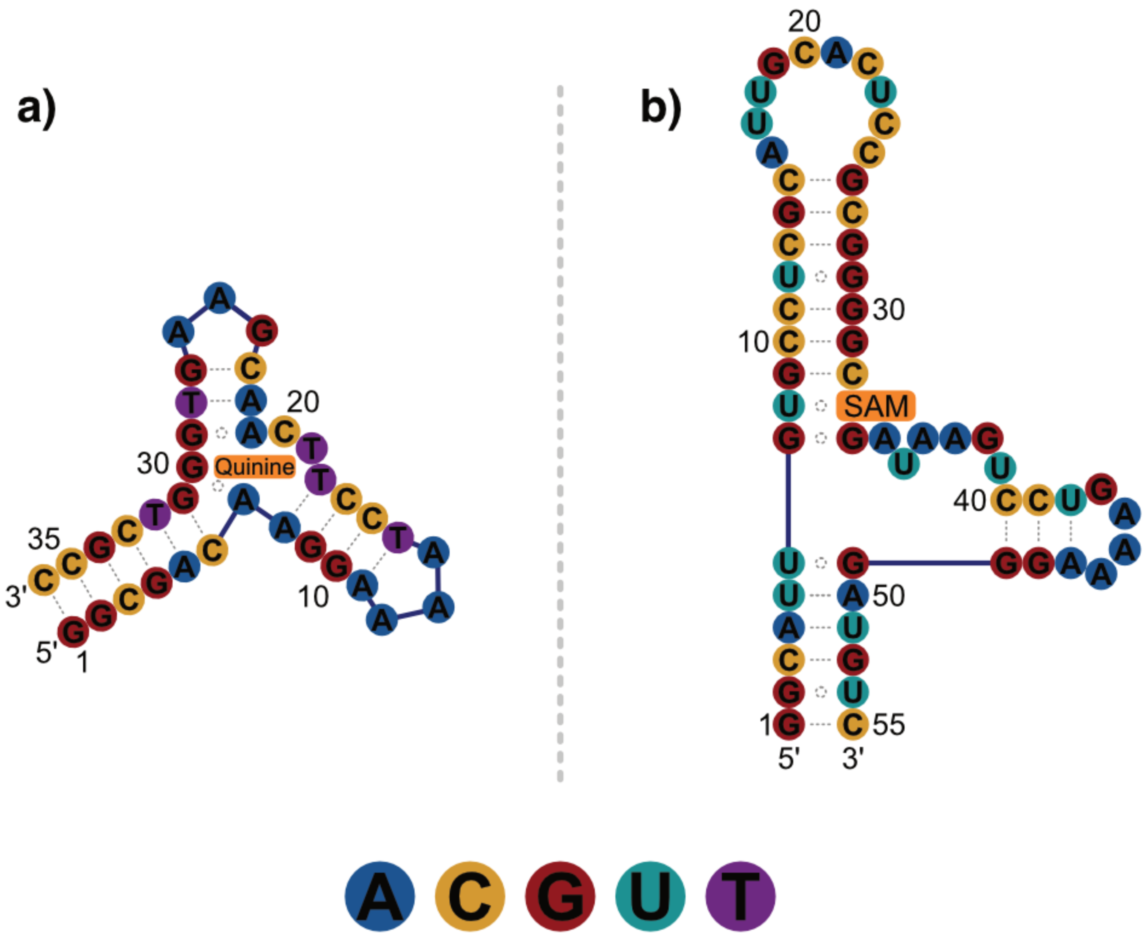


Figure 1.7: Aptamer and riboswitch structures. a) Cartoon representation of the MN4 aptamer bound to quinine[99]. b) Cartoon representation of the SAM-VI riboswitch[100].

Aptamers structures can incorporate a diverse array of nucleic acid motifs, including hairpins, guanine quadruplexes, and pseudoknots (*Figure 1.7a*)[99, 101, 102]. Their tertiary structure is critical in their ability to bind selectively to their ligands[101]. The relative ease of developing aptamers has made them attractive for biotechnological applications including sensing[103], drug delivery[104], and diagnostic tools[105]. There are cases however, where aptamers can have off-target interactions that negatively impact their use. For example, an aptamer which was initially designed to bind to cocaine has been shown to interact with quinine roughly 30- to 40-fold more tightly[99, 106]. This would have detrimental effects if this aptamer was used in a drug sensing application as quinine is an ingredient in tonic water, the common drink. Thus, properly measuring both the selectivity and binding affinity of aptamers is important before using them in real world applications[96].

Initially aptamers were believed to be completely synthetic, however in the early 2000’s it was discovered that biology has been making use of small molecule binding RNA molecules, called riboswitches[100, 107-109] (*Figure 1.7b*). These messenger RNA (mRNA) sequences selectively bind to small molecules and then undergo structural rearrangements, effectively switching structures in the presence of a specific molecule[110]. Typically, these structures occur in the 5’ untranslated region (5’-UTR) of bacterial mRNA transcripts, where they have been shown to control gene expression[111] however they have also been discovered in eukaryotes, albeit to a lesser extent[112]. Their prevalence in bacteria has made them an interesting target for novel therapeutics and antibiotics[113].

1.2.3 Guanine Quadruplex folding and function

1.2.3.1 Position in the genome

Guanine rich sequences were first identified in gene promoters[114, 115] and chromosomal telomeres[116]. Many of these sequences were later shown to fold into quadruplex structures[117]. After the consensus sequence of the human genome was published in 2001[118], interest grew in searching it for guanine rich, potential G4-forming sequences. Quadparser was the first bioinformatic algorithm designed to located and predict the formation of G4s from a sequence of DNA[119]. The authors searched the genome for segments which followed a folding rule where *“a sequence in the form* $d(G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+})$ *will fold into a quadruplex under near-physiological conditions”*[119]. They came to this sequence after investigating synthetic quadruplex sequences using a combination of biophysical techniques and molecular modelling[120]. This initial search predicted 376,000 quadruplex sequences, and they found that half of all genes had a putative G4 structure near their transcription start site. Further studies have shown G4s are not just clustered near the transcription start site of promoters[114, 115, 121-124] and the telomeres of many eukaryotes[125-128], but also at splice sites and the 5’ and 3’ UTRs of RNA[129-132127]. While this initial study provided a solid foundation for predicting the formation of G4s, it suffered from having many false negatives, where a sequence which was not predicted to form a G4 could indeed form a stable structure[133]. To overcome this weakness, the G4hunter algorithm was developed, this algorithm wasn’t a simple pattern matching algorithm like Quadparser, and instead scored sequences based on their G-richness and G-skewness[133]. This algorithm increased the accuracy of predicting if a G4 would form from a specific sequence, and concluded that the number of sequences in the human genome which could possibly form G4s was significantly higher than what was predicted with Quadparser[133]. While Quadparser and G4Hunter remain the most used algorithms for predicting G4 formation, a host of different algorithms have been developed to predict different features of G4 formation including G4s with long loops, hairpins, multimeric quadruplexes, and prediction using machine learning algorithms, with each of these algorithms having their upsides and downsides which were all recently reviewed[134].

Whereas bioinformatic searches and G4 prediction serve as a good basis for the identification of G4 structures in the genome, new sequencing methods have been developed which allow for the *in vitro* identification of both DNA (G4-seq)[135] and RNA (rG4-seq)[136] quadruplexes directly from the sequence itself. G4-seq relies on Illumina next-generation sequencing[137], and the observation that the quality of the sequencing data, quantified by Phred quality scores[138], lowers when G4's are formed (*Figure 1.8*). In G4-seq, two sequencing runs are produced, one where the sequencing buffer contains Li^+ or Na^+ cations, which do not strongly promote G4 formation, and one where the buffer contains K^+ cations which do promote G4 formation. The quality of this sequencing data is then assessed, and regions with high-quality sequencing data in the Li^+/Na^+ buffer and low-quality sequencing data in the K^+ buffer are identified as quadruplex forming regions. This study identified over 700,000 regions forming G4s in the human genome, nearly double what was predicted with the original Quadparser algorithm. Surprisingly, 70% of identified structures do not fit the canonical sequence G4s and have since been shown to adopt structures which include shorter and longer G-tracts, bulged residues, missing guanines, and longer loops[135, 139]. G4-seq has since been applied a host of different genomes[140], providing valuable information for G4 prediction, especially in the case of machine learning algorithms which have seen a rise in popularity over the last few years[141-143].

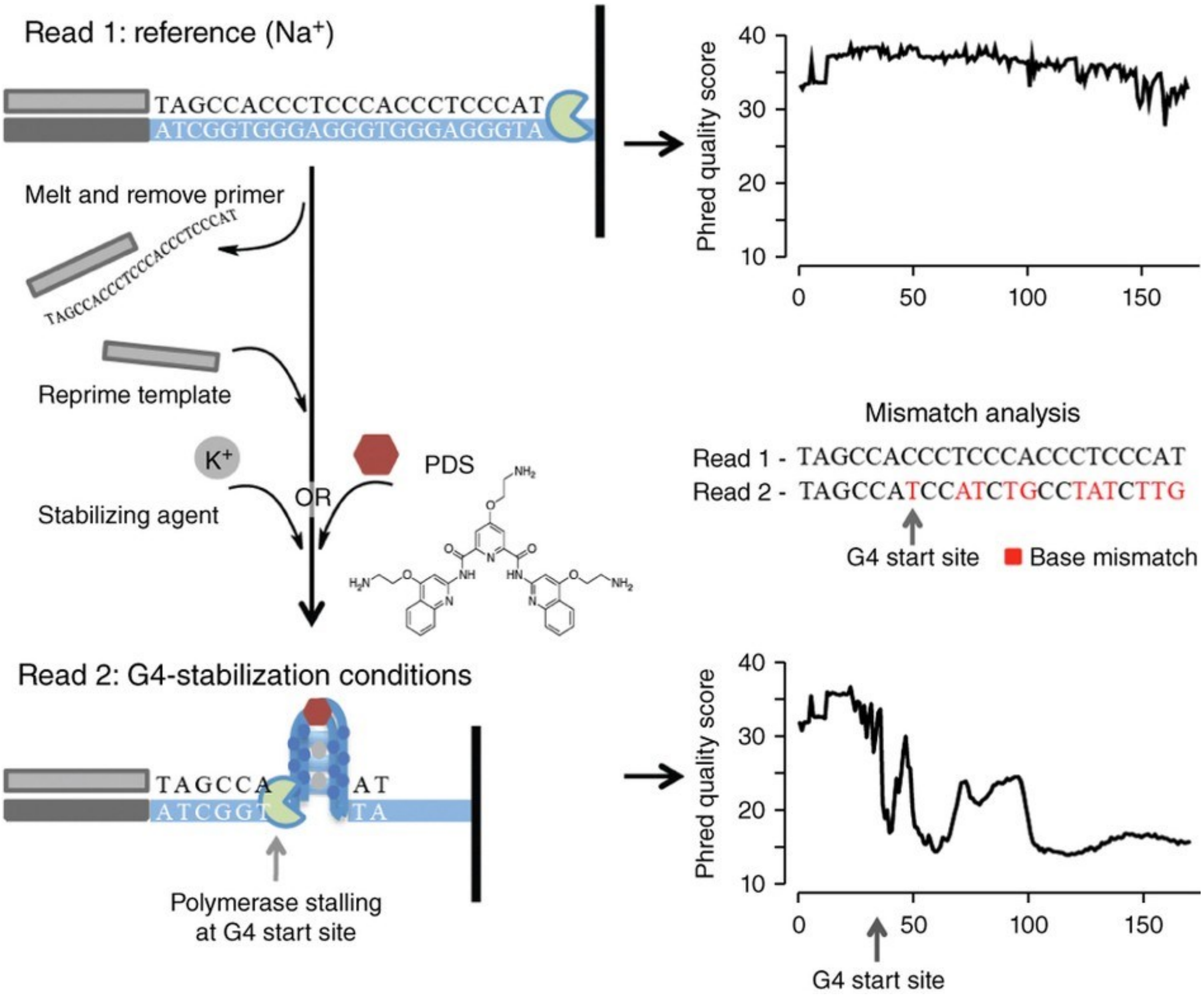


Figure 1.8: Example of a G4-seq workflow. The first read is done under conditions which do not promote G4 formation, in this case a buffer with Na⁺ which does not stabilize G4s as much as K⁺. The second read is done under G4 stabilizing conditions, in this case a buffer with K⁺. The Phred quality score is then compared to see where the G4 start site is. Adapted from Chambers et al. with permission[135].

Following the development of *in vitro* detection methods for G4s, several methods have been developed to detect G4s *in vivo*. These techniques generally use imaging to visualize quadruplexes in cells, and have used either fluorescent nanobodies[144] or small ligands[145, 146] which have been designed to selectively bind to G4s. However, these molecules can influence G4 stability, which begs the question of whether they are promoting G4 formation *in vivo*. Consequently, groups tend to try minimize the concentrations of these ligands to reduce this effect[146]. These examples show some of the first steps in detecting G4s *in vivo,* however this still remains a challenging task[147].

1.2.3.2 Biological function

Guanine quadruplexes (G4s) have been hypothesized to influence many biological processes[38]. They are implicated in telomere maintenance and structure[125, 127, 148, 149], DNA replication[150-152], chromatin strcuture[153-155], transcription control[121-123, 156, 157], translation control[132, 158, 159], and DNA damage response[160-162]. Their prevalence in regulatory elements has made them attractive targets for the development of novel cancer therapeutics[163-165], and they have been suggested as both anti microbial[166], and antiviral[167, 168] targets. However, there are currently no approved G4 targeting therapeutics, with only a few examples making it to phase 2 clinical trials[169]. Understanding how G4s influence biology as well as how to selectively target them when they are critical in so many different processes remains a large challenge for researchers[170].

Quadruplexes in telomeres

The telomeres of most eukaryotes contain quadruplex sequences which are able to form stable G4s *in vitro*[171]. They are able to cap chromosome ends[38], and recruit telomeric proteins that bind G4 structures[149, 172]. In humans, the telomeric repeat d(TTAGGG) can reach lengths of 5-25kb with a single stranded overhanging region of around 35-600 nt[173-175]. This long single stranded region theoretically allows dozens of quadruplexes to fold simultaneously, and questions remain as to how these quadruplexes interact and what structure(s) these regions adopt *in vivo.* When studied in its monomeric form, d(TTAGGG)$_4$ can fold into many different topologies. Parallel[125], hybrid[176], and antiparallel[117] structures have all been observed *in vitro*, and are all stable under different solution conditions with cation identity and molecular crowding agents promoting specific folds. Studying the extended form of the telomeric region containing multiple d(TTAGGG)$_4$ motifs is much more challenging than studying the monomer containing a single d(TTAGGG)$_4$ motif. Several structural studies have shown that longer telomeric repeats form a “beads-on-a-string” arrangement, where the maximum number of quadruplexes is formed[177, 178]. Recent work has shown that longer segments of the human telomeric repeat adopt multiple G4 topologies in the same strand[179]. However, these studies only looked at shorter sequences which could only fold 2-4 quadruplexes, which is substantially smaller than what could theoretically be formed *in vivo*.

Work from previous members of the Mittermaier lab has shown that longer telomeric regions tend to have unfolded regions, especially at the ends of the DNA sequence[180]. This is an important consideration, as the formation of G4s at the ends of telomeres has been shown to control access of telomerase, an RNA-reverse transcriptase which extends the telomeric region[181]. Indeed, telomere length has been shown to be modulated by adding G4 interacting compounds[182], as well as depleting G4-unwinding helicases[183]. Furthermore, G4s have been shown to prevent alternative lengthening mechanisms (ALT), where telomerase is not used to lengthen the telomeres[184]. This is of particular interest, as ALT is a mechanism which is activated in 15% of cancers and preliminary results have shown that ligands targeting G4s are able to kill these cells[38, 184, 185]. The fact that G4s can modulate both telomerase activity and ATL makes them attractive targets for controlling this process *in vivo*.

Quadruplexes in DNA synthesis and damage

Quadruplexes have long been known to block polymerase read through *in vitro,* and this feature is often used as evidence to support the formation of G4 in a DNA strand[186]*.* However, there has been also been mounting evidence that the formation of G4 structures impedes replication *in vivo*[187]*,* where they have been shown to prevent both lagging strand[188] and leading strand synthesis[189]. G4s have been found in the vast majority of identified origins of replication[152] where they promote replication-fork collapse, leading to double-strand breaks and genome instability[190]. Deletions caused by this collapse have been shown occur directly adjacent to quadruplexes, leaving a distinct genomic “scar” after multiple rounds of mitotic division is *C. elegans* (*Figure 1.9a*)[191]. Interestingly, when a sequence of DNA is predicted to form multiple different G4s at different positions, patterns of deletions will emerge near each of these expected structures (*Figure 1.9b*). This suggests that different quadruplexes can be formed *in vivo* and can give insight into which quadruplexes are more likely to form.

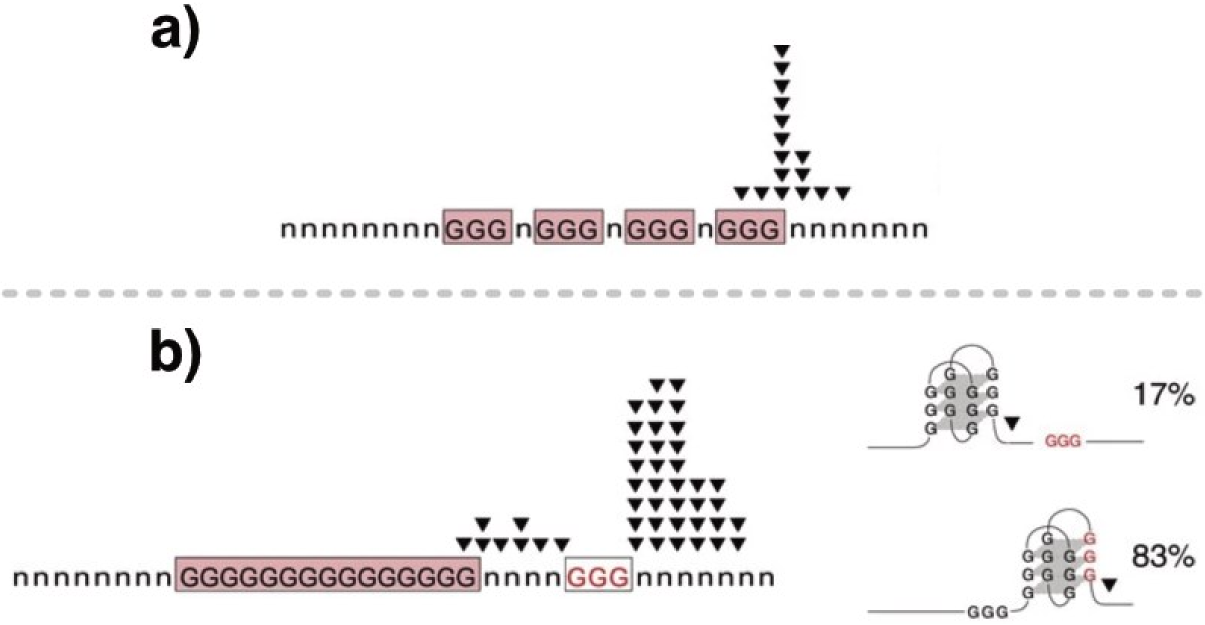

a)
nnnnnnnnGGGnGGGnGGGnGGGnnnnnnn
b)
nnnnnnnnGGGGGGGGGGGGGGGnnnnGGGnnnnnnn
17%
83%
GGG
GGG

Figure 1.9: Genomic deletions from G4 formation. G-tracts are shown in red, any nucleotide other than guanine is represented by n, and the position of deletions is shown as triangles. a) Deletions occurring near a single quadruplex. b) Deletions occurring in a sequence with two possible G4s. Adapted from Lemmens et al. with permission[191].

Another form of evidence that G4 formation can lead to DNA damage comes from looking at diseases which have mutations in G4-unwinding helicases. In these diseases, the activity of a helicase is generally negatively impacted, leading to a lower efficiency in unwinding DNA structures. Increased deletions adjacent to G4 have been observed in these diseases, along with telomere damage[192]. The Werner Syndrome Helicase (WRN) is one of these examples. This helicase has the ability to unfold G4s, and its function prevents telomere loss during the lagging strand replication[193]. Mutations in this gene causes Werner Syndrome, an autosomal recessive disorder which is characterized by premature aging[194]. Recent work looking at zebrafish as a model system has shown that the WRN helicase regulates short-stature homeobox (SHOX) expression through its G4-unwinding activity, suggesting that WRN has multiple different functions[195].

Quadruplexes in gene regulation

While quadruplexes are abundant in regulatory elements, the role they play in gene regulation is not well understood[196, 197]. Depending on their sequence and position they have been shown to upregulate or downregulate both transcription and translation[159, 198-202]. *Holder et al.* attempted to understand how G4s affected the processes by performing a systematic study on how G4s influenced both transcription and translation based on their position in a gene (*Figure 1.9*)[198]. They inserted different model G4 sequences into the promoter, 5'-UTR, and 3'-UTR of a gene encoding for the green fluorescent protein in *E. coli*. To account for potential artifacts arising from the specific promoter sequence used, they opted to run each experiment on two different promoter sequences. They measured gene expression by looking at the resulting fluorescence in *E. coli* cultures, and further studied mRNA levels to determine if transcription or translation was playing a role in the gene expression. The model G4 sequences that they measured all had different stabilities and topologies. G_2T and G_2TC were the least stable (T_m < 60°C) quadruplexes they studied, both forming only two tetrads, with either parallel or anti-parallel topologies respectively. G_3T and G_3A both form highly stable (T_m > 80°C) three-tetrad parallel structures. When located in the promoter directly adjacent to the transcription start site (TSS), G4s had no effect on transcription or translation in the coding strand, however a large decrease in gene expression was seen when there was a G4 in this region in the non-coding strand. This decrease in gene expression in the non-coding strand was correlated with the stability with more stable G4s repressing gene expression more. Furthermore, mRNA levels were also notably decreased, suggesting that a G4 in this position effected transcription and not translation. This was expected as any G4 before the transcription start site (TSS) shouldn't be transcribed into mRNA.

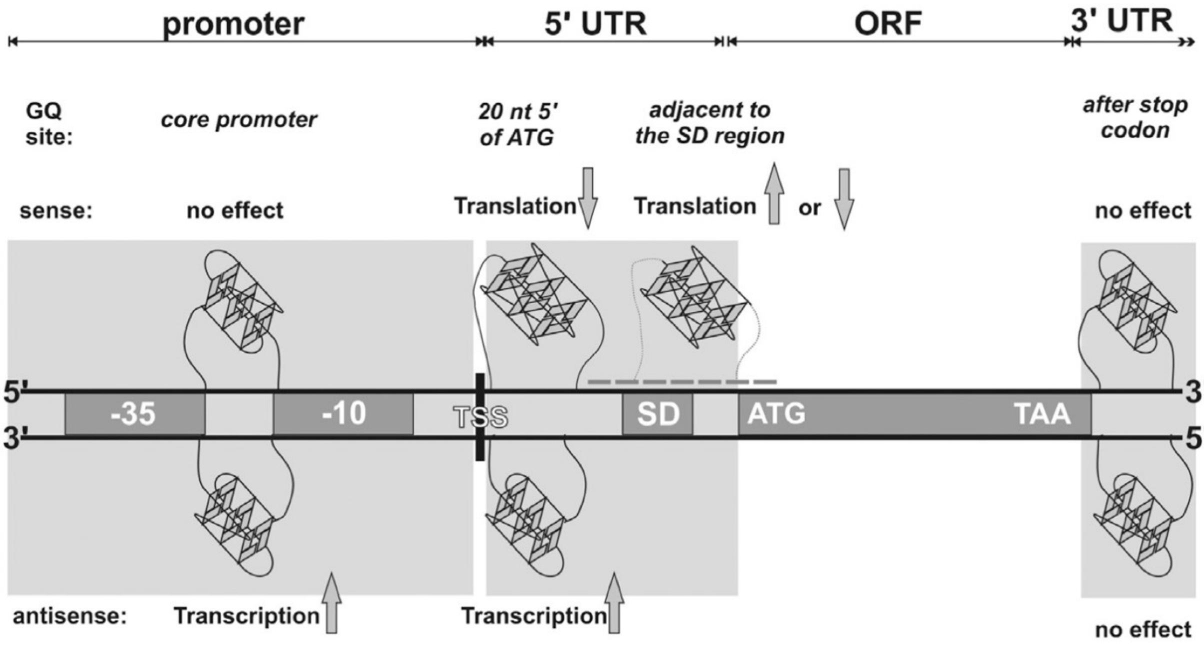


Figure 1.10: Positional dependence of G4s on gene regulation. Adapted from Holder et al with permission[198].

When a G4 was placed into the 5'-UTR of the non-coding strand gene expression was increased. This once again correlated well with G4 stability, and mRNA levels were increased relative to the gene expression, suggesting that a G4 in this position had an effect on transcription and not translation once again. G4's in the 5'-UTR of the coding strand had a much different effect, when located close to the TSS gene expression was reduced. However, there was no difference in mRNA levels, suggesting that at this position a G4 had a negative effect on translation but not transcription. When the G4 was located close to the open reading frame of the gene but still in the 5'-UTR, translation was once again affected, but this time it could either be upregulated or downregulated depending on the sequence used.

In conclusion, while quadruplexes have been shown to effect gene expression there is still no consensus on how this is accomplished. In some cases, G4s downregulate transcription which is hypothesized to occur due to blocking polymerase readthrough[195]. In other cases, they have been shown enhance transcription by recruiting transcription factors, or by blocking rehybridization of the duplex DNA[202]. We do know that the position of the G4 drastically changes its ability to alter gene expression, and that helicases can help unwind quadruplexes to influence gene expression. However, there is still a lot to learn about the impact quadruplexes have on gene expression and understanding their structure, stability, and interactions with other molecules will provide a basis to exploit them as therapeutic targets.

Guanine quadruplex therapeutics

Development of quadruplex targeting therapeutics has been ongoing for years[203-205]. However, to date, only a few G4 interacting compounds have made it into clinical trials. Two examples are Quarfloxin and Pidnarulex, which are fluoroquinolones developed to have dual topoisomerase II and G4 interactions (*Figure 1.11*)[169]. Quarfloxin was the first-in-class G4 targeting therapeutic which went through both phase 1 (NCT00955292) and phase 2 (NCT00780663) clinical trials for lymphoma and solid tumors. It was originally developed by Cylene Pharmaceutical, however it was withdrawn from clinical studies after phase II in the favor of newer derivatives.

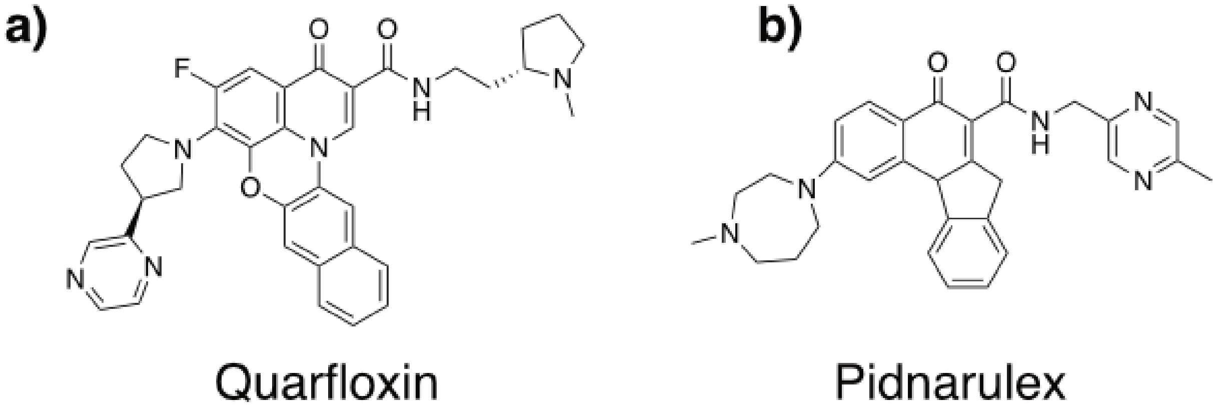


Figure 1.11: Chemical structures of G4 interacting compounds. a) Quarfloxin. b) Pidnarulex.

Pidnarulex is a much more recent derivative of quarfloxin that recently finished a phase I clinical trial (NCT02719977) for breast carcinoma and malignant solid tumors[169]. In this trial, 40 patients were treated across 10 different dose levels. Responses were observed in 14% of patients. Interestingly, patients with defective homologous recombination were more likely to show a response to treatment[206]. Homologous recombination has previously shown to be effected by G4s[207], and Pidnarulex is now in a phase I clinical trial focusing on this mechanism of action (NCT04890613). The drug was well tolerated and showed clinical activity without the characteristic toxicities of other topoisomerase inhibitors, suggesting a different mechanism compared to current therapeutics. While these two drugs are the first specifically designed to interact with G4s, some indenoisoquinolines that have entered phase 1 clinical trials have recently been shown to bind and stabilize G4s *in vitro*[208]. Important questions remain on G4 therapeutics[169]: What is the importance of selectivity in G4 interacting compounds? How do epigenetic features change quadruplex therapeutics? What are the roles other DNA structures such as i-motifs and triplexes?

1.2.3.3 Energy landscape

The energy landscape of guanine quadruplexes is quite complex. Even a simple G4 made up of exactly four runs of three guanines folds through a kinetic partitioning mechanism with many possible folds and transition states (*Figure 1.12*)[209]. The situation gets even more complex when G-tracts contain more than three guanines, which leads to different G-register isomers, where G-tracts are shifted either in the 3' or 5' direction[210]. Sequences can also contain more than four G-tracts, which leads to different spare-tire isomers[211], and when sequences contain eight or more G-tracts multiple G4s can fold together[180]. Finally, secondary interactions with loop residues and intermolecular association of other DNA or RNA G-tracts can also occur, complicating folding substantially[139].

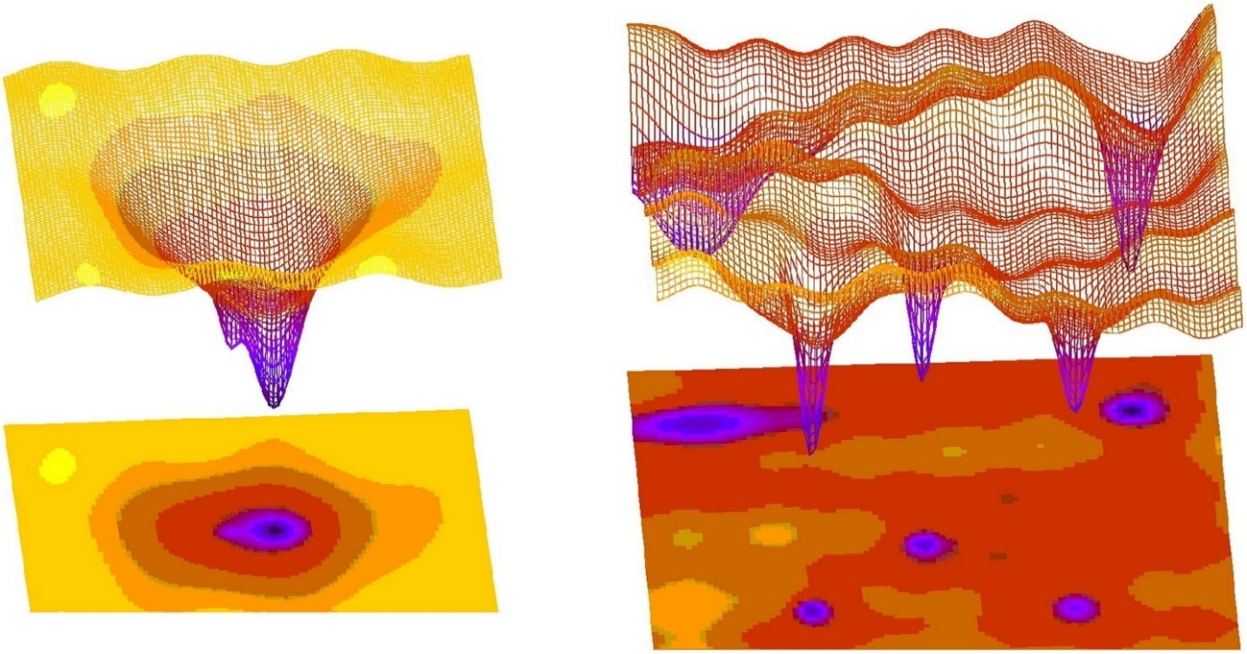


Figure 1.12: Representation of energy landscapes.Left) Funnel landscape with only a single free-energy basin. Right) Kinetic partitioning landscape with many competing free-energy basins. Adapted from Šponer et al. with permissions[209].

To start, let us take the simplest case of a G4 made up of four tracts of three guanines separated by three loops (L), which do not interact with the structure at all, GGG-L_1-GGG-L_2-GGG-L_3-GGG (*Figure 1.13a*). Upon first inspection, there is only one set of twelve guanines which can make up the core of this structure. However, each of these guanines can be adopt multiple sets of both sugar puckers and glycosidic bond angles. In fact, just through glycosidic bond angles alone, a G4 core can adopt 4096 (2^{12}) independent structures, with 2336 of these forming either two or three complete tetrads[209]. The combinations of these angles is what defines the loop arrangements and G-tract orientations, leading to either parallel, antiparallel, or hybrid G4-topologies[209]. Furthermore, the folding of this quadruplex proceeds through intermediates such as G-duplexes and G-triplexes which contain incomplete G-tetrads[212-215], and the final structures can have missing cations[56] or incorrect topologies[216]. Studying this number of both intermediate and folded states experimentally can be challenging, as most spectroscopic methods cannot differentiate between *syn* and *anti* angles, and methods that can are not sensitive enough to detect small populations of intermediates[209].

Molecular dynamics, however, can give insight into how this large number of *syn* and *anti* patterns effect the energy landscape. *Šponer et al.* have shown that this extreme complexity is what causes the exceptionally long folding times observed for G4s when compared to duplex DNA and that even this simple sequence folds through a kinetic partitioning mechanism[209].

Now let us consider the case where a sequence is made up of more than the 12 guanines required to form the core of the G4 structure. Take the sequence GGGG-L_1-GGG-L_2-GGG-L_3-GGG, which has four guanines in its first G-tract (*Figure 1.13b*). In this case, not including the loop and angle isomers described in the last paragraph, the molecule can form two distinct isomers. The isomers are created by the using a subset of the four guanines in the first tract to create the core. For example, taking the first three guanines leaves the fourth to be in the first loop, GGG-GL_1-GGG-L_2-GGG-L_3-GGG, or taking the G-GGG-L_1-GGG-L_2-GGG-L_3-GGG, resulting in two possible isomers. In the first case, the odd G in the loop is shifted to the 3' end of the sequences, and thus we refer to this as a 3' shifted isomer. In the second case the odd G is shifted to the 5' direction and is therefore the 5' isomer. In a previous study from the Mittermaier lab, *Harkness et al.* studied this type of isomer, referred to as a G-register isomer[210]. In this case they were able to study the isomers independently by mutating select guanine residues to either thymine or inosine nucleotides. They characterized the isomers and the wildtype and developed a thermodynamic model to explain this behavior, finding that the presence of these isomers increased the entropic stabilization of the folded state, leading to the full sequence being more stable than either of the two isomers themselves.

Next, consider a sequence of five tracts of three guanines separated now by four loops, where any subset of four tracts can make distinct isomers, GGG-L_1-GGG-L_2-GGG-L_3-GGG-L_4-GGG (*Figure 1.13c*). Once again, along with the isomers describe previously, this sequence can make five different isomer using different G-tracts. For example, you can make a G4 from the first four G-Tracts, GGG-L_1-GGG-L_2-GGG-L_3-GGG-L_4-GGG, leaving the last G-tract as dangling nucleotides at the 3' end of the sequence, referred to as the 1234 Isomer. You can also take all but the middle G-tract, GGG-L_1-GGG-L_2GGGL_3-GGG-L_4-GGG, where the third G-tract, along with loops two and three, now form the middle loop of the quadruplex. This quadruplex would be referred to as the 1245 quadruplex. You can continue to do this to find five different isomers, which are called spare tire isomers. *Grun et al.* dissected the folding kinetics of this type of isomer in detail, finding that the G4 forms many of these isomers, getting caught in kinetic traps along its conformational landscape[211].

Finally, consider a sequence made up of eight tracts of three guanines, now separated by seven loops, GGG-L_1-GGG-L_2-GGG-L_3-GGG-L_4-GGG-L_5-GGG-L_6-GGG-L_7-GGG (*Figure 1.13d*). Once again a single quadruplex can be formed from any subset of four G-tracts. In fact, assuming only one G-tract can be in a loop, there are twenty-eight different ways to fold a single quadruplex. However, only two of these possibilities allow for the folding of a second adjacent quadruplex. While this two G4 folded form is likely to be the most energetically favorable folded state, the formation of any one of the other twenty-six single quadruplexes will prevent two G4s from folding. This is a phenomenon referred to as frustrated folding, which was studied in detail by a previous member of the Mittermaier lab. *Carrino et al.* studied these sequences by mutating specific G-tracts into T-tracts to prevent formation of some G4s, and showed that this effect can inhibit complete formation of multiple quadruplexes, even when there are cooperative effects pushing the system to be fully formed[180]. Even at equilibrium, sequences do not always form the maximally folded structure instead leave single stranded regions, especially near the ends of the sequence.

Even after a folding event occurs, G4's still exhibit a large degree of dynamics. G-tracts that contain more than 3 guanines can slide to incorporate different guanine residues in the core, without the need for complete unfolding[216], cations can also exchange from the core of the quadruplex[217]. While originally only the thermodynamically most stable G4 was favored for study, there is growing evidence to suggest that kinetic partitioning early in quadruplex folding is more biologically relevant[139].

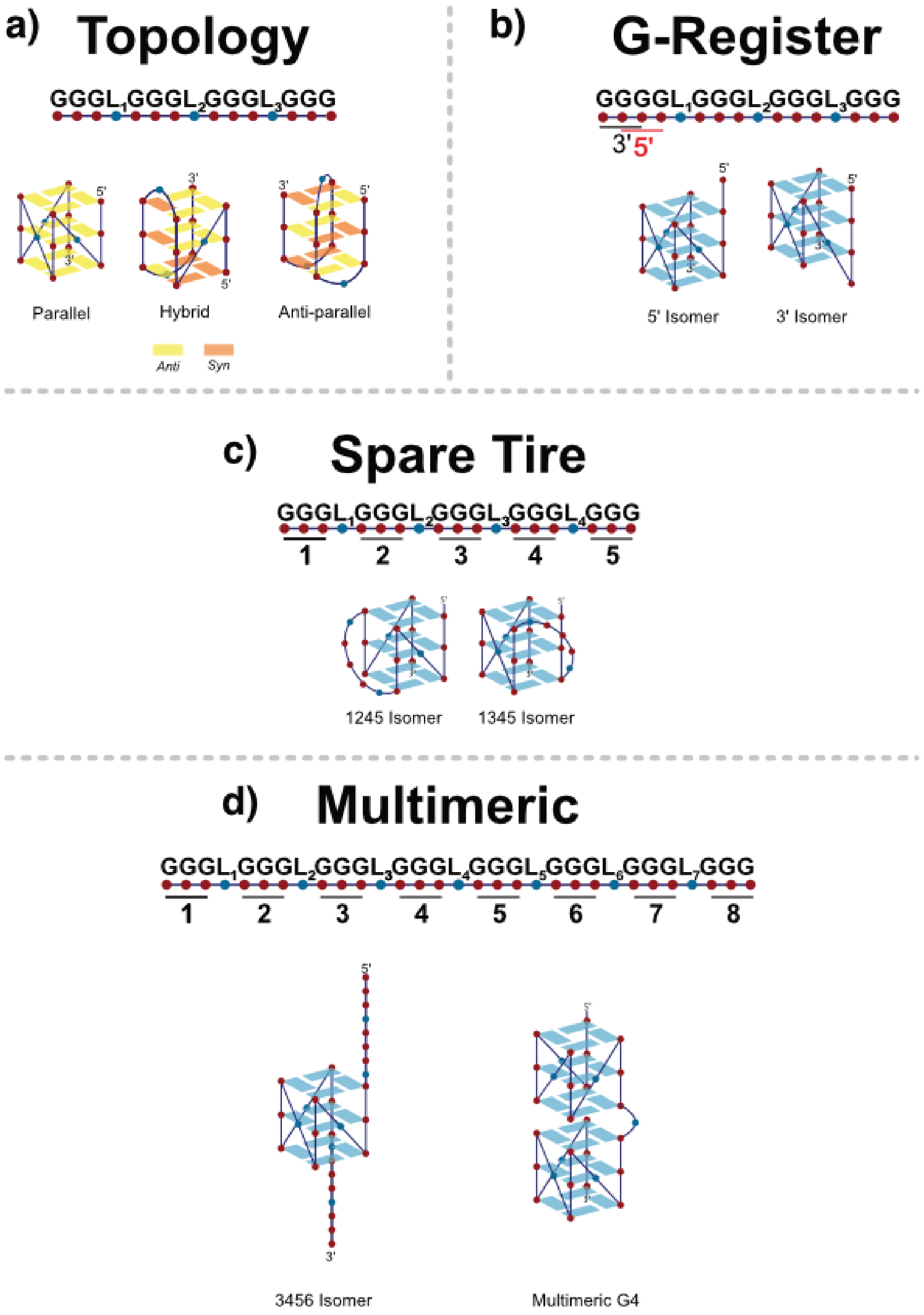


Figure 1.13: Possible G4 isomers. a) Topology isomers arising from different syn and anti patterns. b) G-register isomers resulting from tracts with more than 3 guanines. c) Spare tire isomers resulting from more than 4 G-Tracts. d) Folding frustration in multimeric quadruplexes.

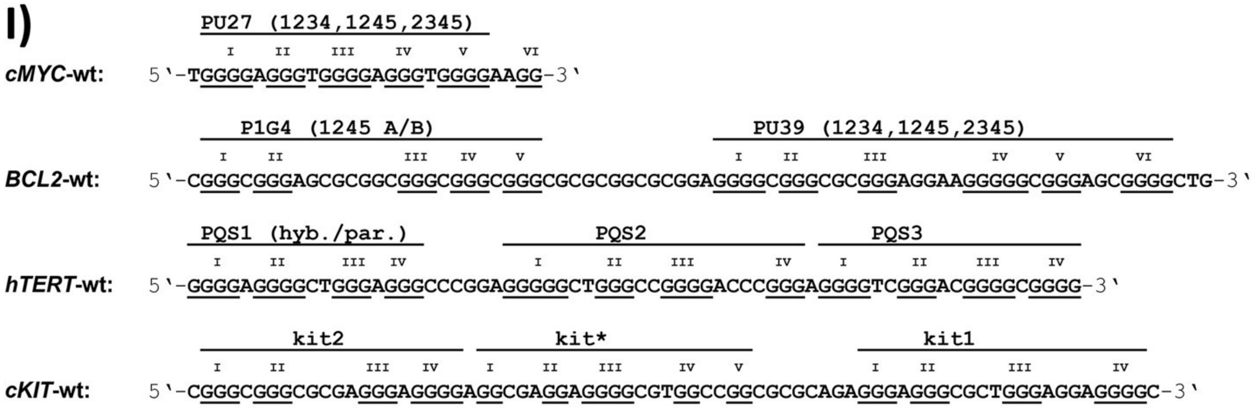

I)
PU27 (1234,1245,2345)
I II III IV V VI
cMYC-wt: 5'-TGGGGAGGGTGGGGAGGGTGGGGAAGG-3'
P1G4 (1245 A/B)
PU39 (1234,1245,2345)
I II III IV V I II III IV V VI
BCL2-wt: 5'-CGGGCGGGAGCGCGGCGGGCGGGCGGGCGCGGCGCGGAGGGGCGGGCGCGGGAGGAAGGGGGCGGGAGCGGGGCTG-3'
PQS1 (hyb./par.)
PQS2
PQS3
I II III IV I II III IV I II III IV
hTERT-wt: 5'-GGGGAGGGGCTGGGAGGGCCCGGAGGGGGCTGGGCCGGGGACCCGGGAGGGGTCGGGACGGGGCGGGG-3'
kit2
kit*
kit1
I II III IV I II III IV V I II III IV
cKIT-wt: 5'-CGGGCGGGCGCGAGGGAGGGGAGGCGAGGAGGGGCGTGGCCGGCGCGCAGAGGGAGGGCGCTGGGAGGAGGGGC-3'

Figure 1.14: Naturally occurring G4 sequences. Sequences of c-MYC, BCL2, hTERT, and cKIT quadruplexes. Validated G-tracts are underlined and well studied individual G4s are highlighted. Adapted from Grün et al. with permissions[139].

While the previous examples have all been model sequences, naturally occurring G4 sequences often have more guanines than can form a single G4 core (i.e. more than four tracts of three G's, or G-tracts with more than three G's), potentially leading to large degrees of structural polymorphism (*Figure 1.14*). Because of this, G4 folding is challenging to study with conventional techniques. For example, most spectroscopic methods are not able to differentiate between different intermediates and folds, and instead measure ensemble effects[139]. More sophisticated structural techniques such as NMR are unable to resolve the large number of peaks caused by this structural diversity, making assignment difficult[139]. Nevertheless, spectroscopy is one of the main biophysical techniques used to study G4 stability, structure, and kinetics. In these cases folding data from the ensemble often resembles a two-state, all or none mechanism[218] where the unfolded DNA strand proceeds through a single transition state towards the folded structure. This type of analysis provides important insight into the folding mechanism of quadruplexes. However, effects such as negative activation enthalpy[219], are often observed, likely due to a zippered folding mechanism which has been observed for other DNA structures such as duplexes[220], triplexes[221], and i-motifs[222]. Experimentalists can use mutations to study single isomers and use mathematical models to help probe the kinetic mechanisms, which provides a deep level of insight into G4s folding and dynamics[180, 210].

1.2.4 Higher order DNA assemblies in nanotechnology

Nucleic acids have a remarkable structural diversity which, along with their ease of synthesis and bioavailability, has given them prominence in the nanotechnology field[223]. The simple rules of Watson-Crick base pairing allow sequences to be designed with highly predictable behavior. Combinations of the different structures described in Section 1.2.2. have been used to create novel 2- and 3-dimensional assemblies which have applications in the biomedical and biotechnology fields[223]. The fabrication of these structures uses two main approaches: top down and bottom up[224]. Top-down fabrication begins with a large structure which is then shaped and reduced in size to the desired assembly, akin to folding a large piece of paper in the art of origami. Bottom-up fabrication uses small discrete units to build the larger desired structure, akin to the use of bricks to build houses.

1.2.4.1 Nucleic acid junctions

Nucleic acid junctions represent an important tool in designing higher order nucleic acid structures. They occur when multiple stands of nucleic acids cross over each other to create a branched junction and can be exploited to make 2- and 3-dimensional structures using simple duplex DNA. The most famous of these is the Holliday junction which is a naturally occurring DNA structure that is involved in the process of homologous genetic recombination and was first described by Robin Holliday in 1964 (Figure 1.15*a*)[225]. It is built from two pairs of near parallel helices which cross over, as shown in Figure 1.14. In the 1980's Ned Seeman theorized that other types of DNA junctions could be created, theoretically allowing for the creation of novel 2- and 3-D structures[226]. In practice, the Holliday junction was too flexible to form some more complicated structures, which led to development of double crossover junctions (DX) (*Figure 1.15b*)[227]. These junctions have two DNA strands coupled via two crossover events and exhibit much higher rigidity than the simple Holliday junction, thus serving as a basis for the creation of more complicated nanostructures described below.

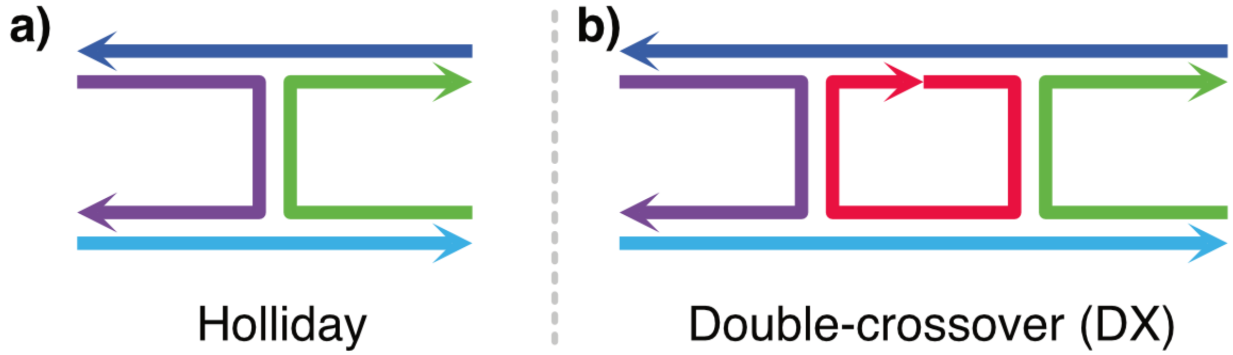


Figure 1.15: Cartoon representation of DNA junctions. Different DNA strands are shown in different colours, and arrows point on the 5'-3' direction. a) Holliday junction. b) Double-crossover (DX) junction.

1.2.4.2 DNA Origami

The concept of DNA origami revolves around the systematic folding of a long, single-stranded DNA molecule, often derived from a bacteriophage, into a predetermined structure[228]. This is a top-down assembly approach which is achieved by using numerous short "staple" oligonucleotides, to bind specific sections of the long strand, guiding its folding into the desired geometry via DX junctions (*Figure 1.16a*). Dr. Paul Rothemund, who introduced the term "DNA origami" in 2006, demonstrated this technique's potential by folding DNA into diverse two-dimensional shapes, such as stars, triangles, and even intricate designs like smiley faces (*Figure 1.16b*)[229]. Since 2006 DNA origami has evolved to more complex 3-D shapes and structures via intricate crossover patterns[230], for example allowing researchers to create hollow boxes which respond to external stimuli[231], polyhedral meshes[232], and DNA robots that can sort cargo[233]. Due to the predictable nature of DNAs interactions, a variety of software has been developed to predict both 2- and 3-dimensional structures, giving the sequences of all DNA staple strands required to form a specific structure[228].

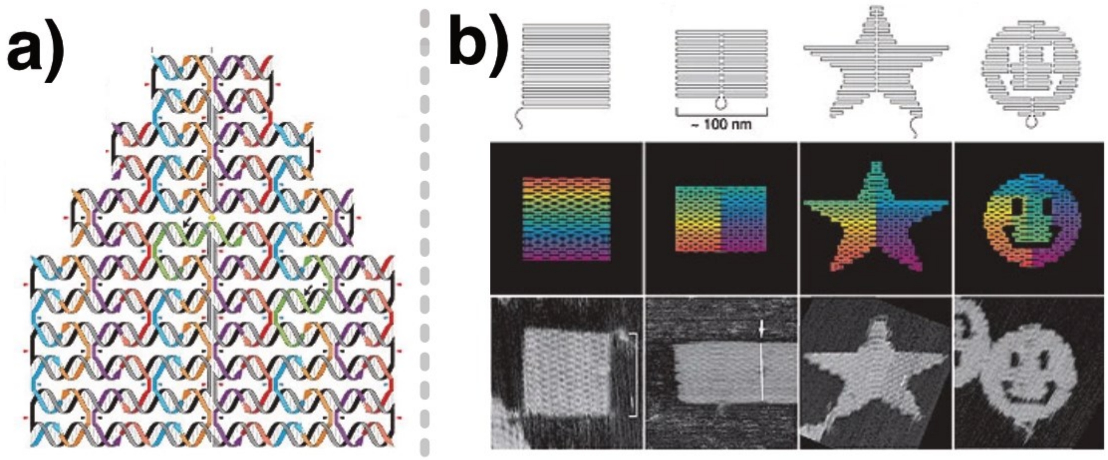


Figure 1.16: DNA origami structures. a) Two-dimensional representation of a DNA origami fold. b) Different two-dimensional shapes created from DNA origami. Adapted from Rothemund with permission[229].

1.2.4.3 Wireframe structures

DNA wireframe structures are generally constructed through a bottom-up methodology, distinct from the aforementioned DNA origami[234]. This approach involves the self-assembly of numerous short DNA sequences into larger complex structures (*Figure 1.17a*). A notable advantage of this method lies in its avoidance of the synthesis or purification processes of long DNA strands[235]. Large assemblies are formed through the predictable self-assembly of easily synthesized smaller structures. Furthermore, the absence of predetermined length facilitates the creation of longer assemblies. Wireframe structures require fewer unique strands, leading to a significant reduction in production costs. The construction of elongated repetitive structures typically involves the design of two complementary building blocks: "rungs" forming the base structure, often in the shape of a lower-order polygon, and "linking strands" connecting these rungs (*Figure 1.17b*)[236]. Linking strands can be either short nucleotides complementary to both ends of a rung, leading to polydisperse assembles, or they can be synthesized to yield a backbone of a predesigned length, resulting in monodisperse assemblies[237].

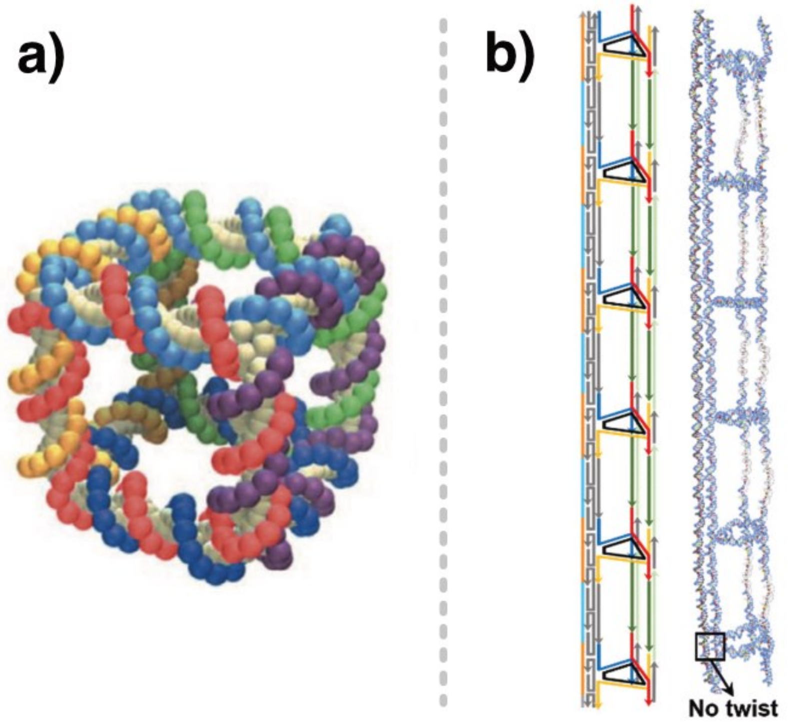


Figure 1.17: Wireframe DNA nanostructures. a)Polyhedral DNA structures, adapted from Seeman et al. with permission[223]. b) DNA nanotubes made from small rungs and connected with a long backbone, adapted from Saliba et al. with permission[236].

1.2.4.4 Supramolecular co-assembly

While the natural assembly of DNA using only four bases is sufficient for design of nanostructures, moving to structures with different functions can be challenging. One way to expand the DNA alphabet is to use non-natural nucleotides, however this can be synthetically challenging and expensive. Other approaches involve the use of small molecules that co-assemble with DNA[238, 239]. These small molecules can be incorporated without the need for modifying the DNA strand. One example of these molecules is cyanuric acid (CA). CA has previously been shown to form large supramolecular assemblies with the small molecule melamine, which bears a resemblance to adenosine[240]. Indeed, CA is complementary to two sides of adenosine, and early experiments showed that mixtures of polydeoxyadenosine (polyA) strands would assemble with cyanuric acid to form long hexameric fibres (*Figure 1.18a*)[239]. Importantly, when in the bound form only two of the three faces of CA are involved in binding, which allows the third face of the CA molecules to be functionalized (*Figure 1.18b*)[241]. These structures form impressively stable hydrogels which exhibit self-healing and pH responsiveness[242]. However, understanding their assembly can be quite challenging.

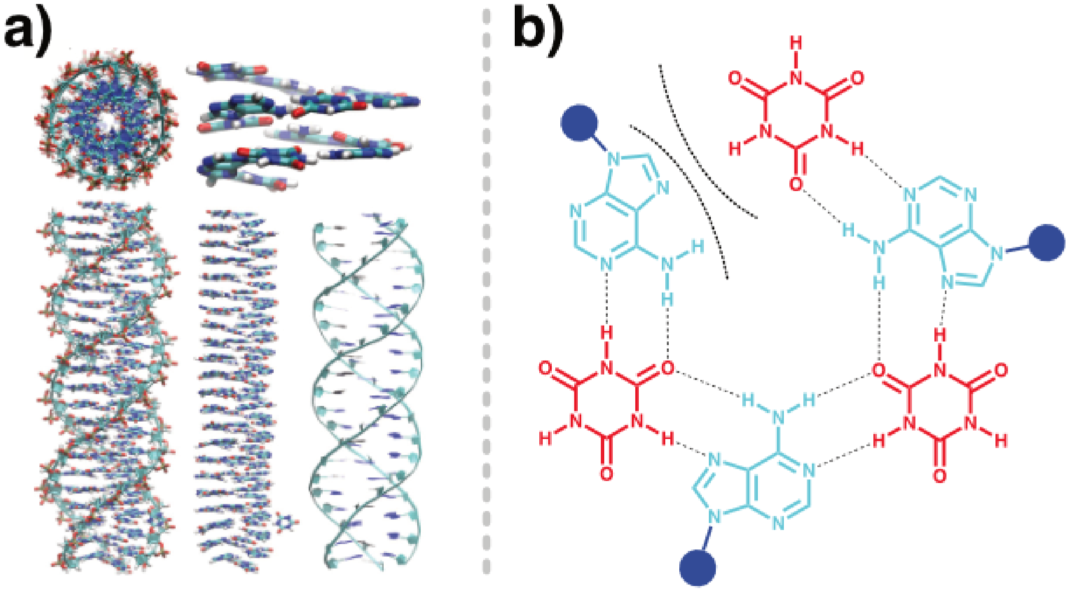


Figure 1.18: Structure of supramolecular fibres created from polyadenosine and cyanuric acid. a) Model of helicine structure, adapted from Alenaizan et al. with permission[243]. b) Hydrogen bonding pattern of polyA-CA fibres adapted from Hennecker et al. with permission[244].

These have complex assembly pathways[245], along with large energetic barriers, hindering their assembly kinetics[246]. *Harkness et al.* was able to show that these structures assemble in a nucleation-elongation model. This model was first developed by Fumio Oosawa and Michiki Kasai to describe the linear aggregation of macromolecules[247], and later expanded on by Robert Goldstein and Lubert Stryer (*Figure 1.19a*)[248]. In this model, the co-assembly of CA and polyA is modelled as monomers (*M*) coming together into elongated fibres. As discussed in Chapter 4 of this thesis, this is an oversimplification of the true mechanism, but still provides insight into how these fibres assemble. The nucleation-elongation model (*Figure 1.19*) has two distinct stages: nucleation and elongation. In the nucleation stage, assembly and disassembly of monomers (*M*) up to a certain size, referred to as the nucleus (M_s), is modelled with one set of association (k_{n+}) and dissociation constants (k_{n-}) to give the equilibrium constant ($K_n = k_{n+}/k_{n-}$)[249]. In the elongation stage, fibres larger than the nucleus are described with another set of constants (k_{e+}, k_{e-}, *and* $K_e = k_{e+}/k_{e-}$). Importantly, for cooperative assembly the stability of the fibres is greater than the stability of the pre-nucleated species ($K_e^{-1} < K_n^{-1}$), resulting in the growth of these fibres only when the free monomer concentration is greater than the stability of the fibres ($[M] > K_e^{-1}$)[249]. This monomer concentration is often reffered to as the critical monomer concentration ($[M]_c$), and can be used as a measure of the stability of these fibres ($[M]_c = K_e^{-1}$).

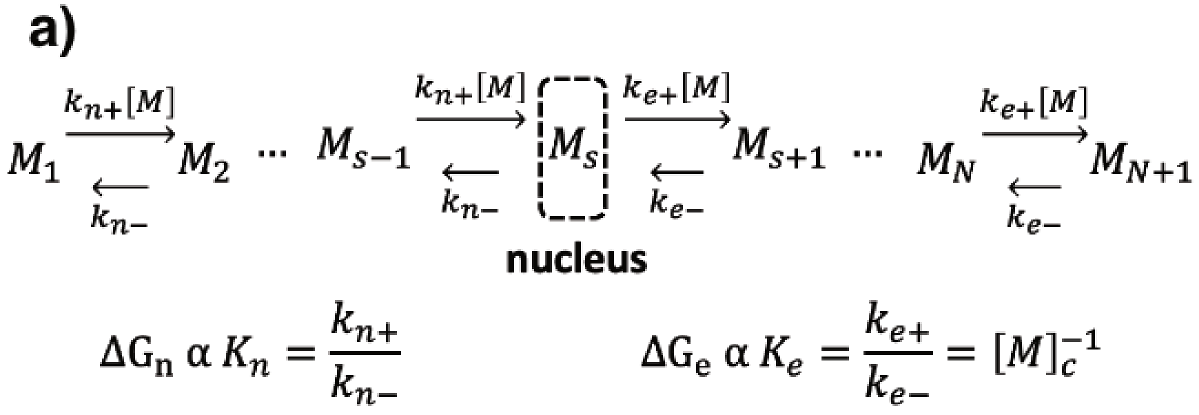

a)
$M_1 \underset{k_{n-}}{\overset{k_{n+}[M]}{\rightleftarrows}} M_2 \cdots M_{s-1} \underset{k_{n-}}{\overset{k_{n+}[M]}{\rightleftarrows}} M_s \underset{k_{e-}}{\overset{k_{e+}[M]}{\rightleftarrows}} M_{s+1} \cdots M_N \underset{k_{e-}}{\overset{k_{e+}[M]}{\rightleftarrows}} M_{N+1}$
nucleus
$\Delta G_n \; \alpha \; K_n = \frac{k_{n+}}{k_{n-}}$
$\Delta G_e \; \alpha \; K_e = \frac{k_{e+}}{k_{e-}} = [M]_c^{-1}$

Figure 1.19: Nucleated-elongation model. a) kinetic and thermodynamic representation of the full model.

1.2.5 Thermal analysis of nucleic acid structures

Measuring the thermodynamic stability and kinetics of nucleic acid structures gives important insight into their chemical interactions and the mechanisms underlying their formation. Even short nucleic acid sequences can have intricate folding (intramolecular) or assembly (intermolecular) pathways. A common strategy for measuring their stability is using thermal denaturation/renaturation experiments[218]. In this approach, nucleic acids are generally first held at a high temperature (>90 °C), driving disassembly/unfolding of the structure. The temperature is then lowered at a constant rate ($\frac{dT}{dt}$) which shifts the equilibrium towards the folded/assembled form. As this process occurs, the relative populations of the folded/assembled and unfolded/disassembled structures are monitored, typically either spectroscopically using techniques such as absorbance[210], fluorescence[250], circular dichroism[251], and nuclear magnetic resonance[252], or calorimetrically using differential scanning calorimetry[253]. The resulting profiles can be analyzed to extract equilibrium populations[210] as well as both thermodynamic and kinetic parameters[218]. In order to extract this kind of information, thermal denaturation/renaturation profiles must be fit to physical chemical models. Take for example, the intramolecular folding of a DNA strand follow the scheme

$$U \underset{k_U}{\overset{k_F}{\rightleftarrows}} F \qquad \text{(Scheme 1.1)}$$

where *U* and *F* are the unfolded and folded forms and k_F *and* k_U are the rate constants for folding and unfolding respectively. When under thermodynamic equilibrium, the system can be described in terms of its temperature dependent equilibrium constant using the van ‘t Hoff equation

$$K_{eq} = \frac{k_F}{k_U} = e^{-\frac{\Delta G}{R*T}} \qquad \text{(Equation 1.1)}$$

where K_{eq} is the equilibrium constant, ΔG is the free energy of folding, R is the ideal gas constant, and T is the temperature. The fraction of folded species at a given temperature can be calculated using

$$\theta_T = \frac{K_{eq}}{1+K_{eq}} \qquad \text{(Equation 1.2)}$$

Heating and cooling then produce sigmoidal transitions between the folded and unfolded species, with lower temperatures favouring folding. In the case of a system which is not at equilibrium, *Equation 1.2* will not adequately explain the observed behaviour of thermal traces. Instead, differential equations must be written to describe the time dependence of each species. For *Scheme 1.1* this leads to the following two equations

$$\frac{d[U]}{dt} = -k_F[U] + k_U[F] \qquad \text{(Equation 1.3)}$$

$$\frac{d[F]}{dt} = k_F[U] - k_U[F] \qquad \text{(Equation 1.4)}$$

The temperature dependences of both $\mathrm{k_F}$ and $\mathrm{k_U}$ are described by the Arrhenius equation

$$k = Ae^{-\frac{E_a}{R*T}} \qquad \textit{(Equation 1.5)}$$

Where *k* is a rate constant (*k_U* or *k_F*), *A* is the pre-exponential factor and *E_a* is the activation energy. These differential equations often have to be numerically integrated. However, for some simple mechanisms, analytical descriptions are possible. The integration of these equations provide values of $\theta = \frac{[F]}{[U]+[F]}$, which can be related back to thermal analysis traces.

This thesis makes use of temperature-controlled UV-spectroscopy to monitor the folding/assembly and unfolding/disassembly of nucleic acid structures. UV-spectroscopy represents a rapid and inexpensive way to monitor these processes since there are often measurable differences in the absorbance of light in the region of 250-290 nm between unstructured and structured DNA. It does not require the use of expensive fluorophores, and multiple samples can be measured in parallel giving it an advantage over single sample machines such as differential scanning calorimeters.

UV-spectroscopy relates absorbance to the concentration of DNA using Beers law

$$A = \varepsilon cl \qquad \textit{(Equation 1.6)}$$

Where A is the absorbance, ε is the extinction coefficient at the measured wavelength for the specific species, c is the species concentration, and l is the pathlength of the cuvette. A typical thermal trace contains three distinct regions (*Figure 1.20a*); 1) A high temperature region where DNA is completely unfolded/disassembled. 2) a transition region (generally sigmoidal but can be more complicated when more than two species are present), where populations of both unfolded/disassembled and folded/assembled species are present. 3) A low temperature region where the DNA is completely folded/assembled. Both the high temperature and low temperature regions may be sloped, this can be due to temperature dependent changes in ε caused by subtle rearrangements of the structure, degradation of the sample, spectrophotometer drift, or evaporation at high temperatures and condensation at low temperatures[218]. Absorbance traces are converted into fractions or concentrations of folded/assembled or unfolded/disassembled by fitting linear baselines to the high-temperature and low-temperature regions (*Figure 1.20b*).

The fraction of folded/assembled DNA is given by[218]

$$\theta_T = \frac{L0_T - A_T}{L0_T - L1_T} \qquad \textit{(Equation 1.7)}$$

Where θ_T is the fraction of folded DNA, A_T is the measured absorbance, $L0_T$ is the high-temperature baseline and $L1_T$ is the low-temperature baseline, all at temperature *T.* A typical thermal analysis will include both a renaturation trace (found from cooling the sample), and a denaturation trace (found from heating the sample). These two traces are required to determine if the system is at thermal equilibrium under the chosen temperature scanning rate (*Figure 1.20c*), or if the system is not under thermal equilibrium (*Figure 1.20d)*, resulting in thermal hysteresis and a distinct offset of the denaturation and renaturation profiles, with the denaturation profile is offset to higher temperatures, and the renaturation profile is offset to lower temperatures; the true equilibrium curve lies somewhere between the two traces (*Figure 1.20d)*. Physical models which describe how the different species in a system change with temperature can be fit to these traces to give a wealth of information on the folding/assembly and unfolding/disassembly thermodynamics and kinetics, as well as information on the mechanisms dictating their formation[218, 246].

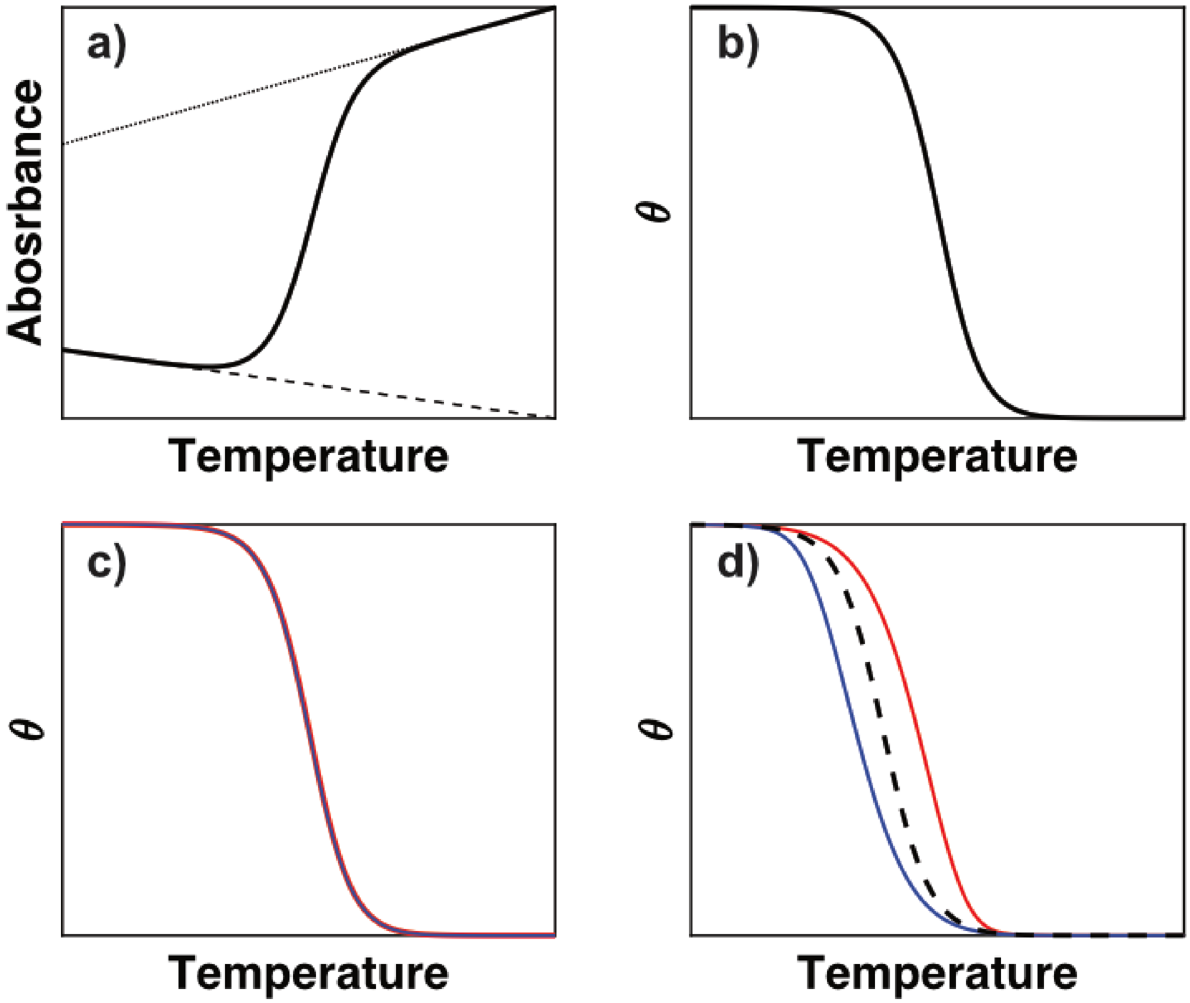


Figure 1.20: Thermal analysis by UV-spectroscopy. a) Raw absorbance traces, high temperature baselines is shown as the dotted black line, low temperature baseline is shown as the dashed black line. The experimental absorbance showing the transition between unfolded/disassembled and folded/assembled is shown as the solid black line. b) Baseline corrected absorbance profile to give a plot of fraction folded/assembled as a function of temperature. c) thermal analysis traces of a system at equilibrium d) Thermal analysis traces showing a system which is not at equilibrium, the true equilibrium is shown as the black dashed line. In panels c and d, denaturation profiles are shown in red and renaturation profiles are shown in blue.

1.3 Proteins

1.3.1 Amino acids

Proteins are essential biomacromolecules which play fundamental roles in biology[254]. They have functions including but not limited to, structural[255], signaling[256], and catalytic processes[257]. As with all other biomacromolecules, they are built from the combination of small building blocks, in this case amino acids, which dictate their overall structure and function. Amino acids are molecules which contain amino and carboxyl functional groups in addition to specific side chains bonded to a central α-carbon (*Figure 1.21a*). There are twenty side chains which are directly coded for in the human genome, all of which have specific shapes, charges, reactivity, hydrogen-bonding patters, and hydrophobicity.

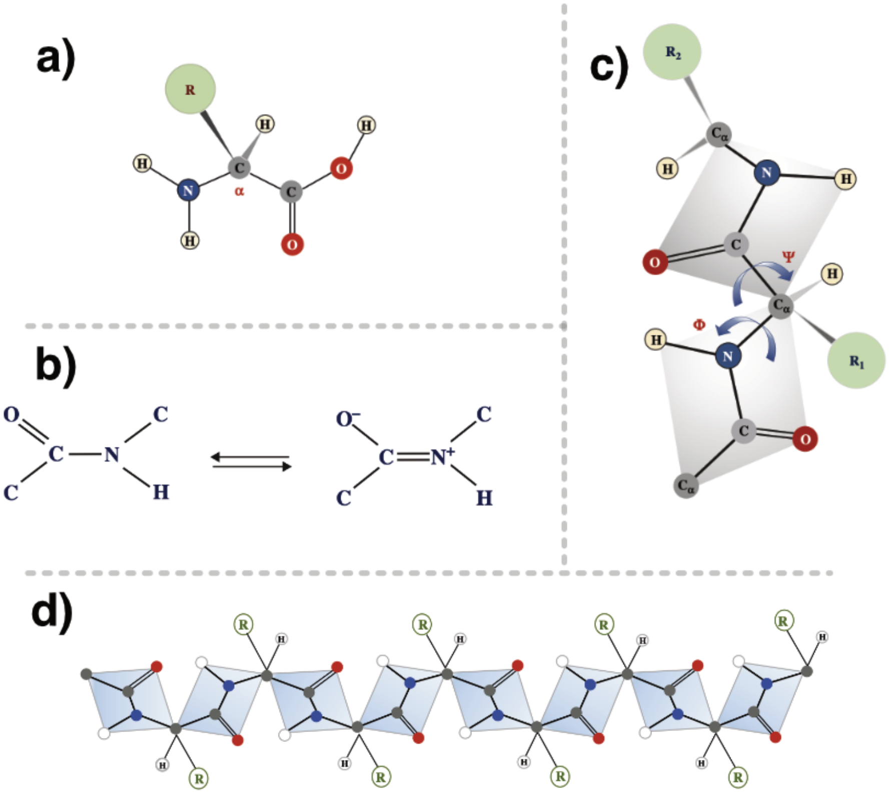


Figure 1.21: Structure of polypeptides. a) structure of a single amino acid, the R group refers to the side chain which differs between amino acids. b) The resonance structures between adjacent amino acids. c) Structure of two amino acids, showing the planes of each amino acid and the possible rotations of the ϕ and φ angles. d) The representation of a polypeptide chain. Adapted from Pal with permissions[254].

Amino acids are joined together by a covalent bond between the carboxyl group of one amino acid and the amino group of another. This leads to the loss of water from the single amino acid structures, and the formation of a C-N bond, commonly referred to as the peptide bond. Rotation around this bond is prevented by two resonance structures, which leads to the peptide group having a rigid, planar structure (*Figure 1.21b*). The other two bonds, between the C_α–C and N–C_α are single bonds. Thus, amino acids are free to rotate these bonds, leading to two defining angles: phi (ϕ) which describes the rotation around N–C_α and psi (φ) which describes the rotation around C_α–C (*Figure 1.21c*). Long chains of amino acids are referred to as polypeptides. The order of amino acids in a protein is referred to as its primary structure (*Figure 1.21d*). By convention, the sequence of amino acids is written from the amino terminus to the carboxyl terminus. When in a polypeptide chain, amino acids can be referred to as residues, and are generally described by a single letter which is unique to each side chain. Proteins are formed from one of more polypeptides and can contain hundreds or thousands of amino acids which then form different secondary, tertiary, and quaternary structures.

1.3.2 Secondary structure

Combinations of different ϕ and φ angles give rise to different secondary structures of the amide backbone. The α helix is one of these structures, which forms when the carboxyl group of the *nth* forms a hydrogen bond with the amino group of the $(n+4)^{th}$ residue (*Figure 1.22a*). This leads to a distinctly cylindrical structure where every carboxyl and amino group is hydrogen-bonded (not including residues at either the start or end). Adjacent residues are rotated by 100° and offset by 1.5 Å which leads to 3.6 residues per turn and a pitch of 5.4 Å. The orientation of each residue in the α helix leads to the dipole moments of individual residues being aligned, creating a macrodipole where the amino-terminal end of the helix is positive, and the carboxyl-terminal end is negative.

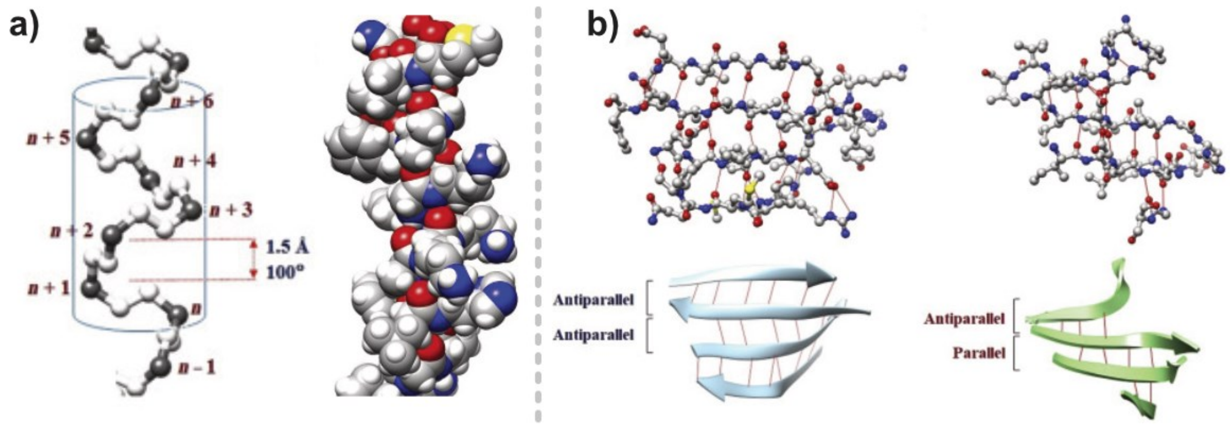


Figure 1.22: Secondary structures of polypeptide chains. a) The α helix. b) the β sheet. Adapted from Pal *with permissions*[254]*.*

The β sheet occurs when two or more polypeptides are arranged side by side with each other in either a parallel or anti-parallel fashion (*Figure 1.22b*). In contrast to the α helix, β sheets have nearly fully extended polypeptide chains. Hydrogen bonds form between the amino groups of one polypeptide strand and the carboxyl group of another, which join these strands. β sheets can be relatively flat, or adopt a slightly twisted shape. The α helix and β sheet make up the two most common secondary structures, however proteins can adopt other structural elements which include β turns, β hairpins, and Ω loops, along with other hydrogen bonding patterns that help give the protein its shape and function.

1.3.3 Tertiary and quaternary structure

The tertiary structure is the overall three-dimensional structure adopted by a single polypeptide chain. It includes the specific positions and interactions of each atom in a polypeptide, including those in the amide backbone, and those of the individual side chains. The tertiary structure is influenced by the different hydrogen bonding patterns between both the side chain and backbones, along with van der waals forces, hydrophobic effects, and electrostatic effects. Water molecules are often found on the surface of the structures and can sometimes be used as structural elements to bridge hydrogen bonding interactions. Polypeptides tend to form compact globular structures. While small polypeptide chains tend to be approximately spherical, larger polypeptides can fold into more than one globular cluster, which are referred to as domains (*Figure 1.23a*). Domains are able to fold independently, and often have specific functions in the protein.

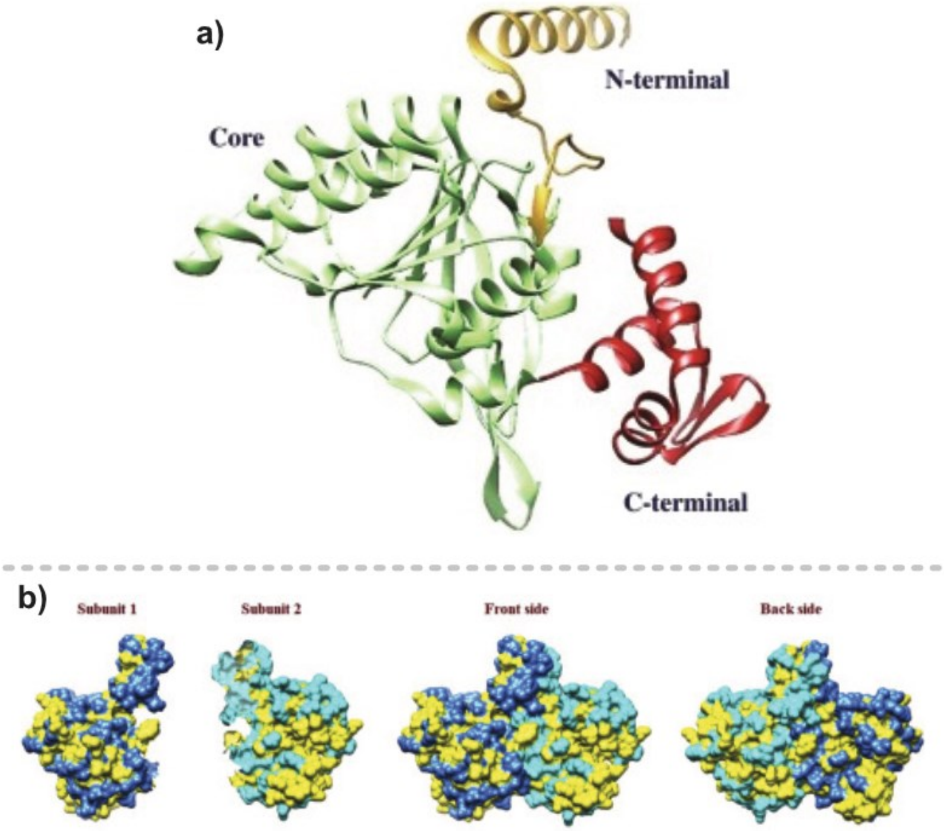


Figure 1.23: Tertiary and quaternary structures. a) the tertiary structure of RecA showing three distinct domains along with multiple β sheets and α helices. b) The quaternary structure of two subunits of undecaprenyl pyrophosphate synthetase from E. Coli. Adapted from Pal *with permissions*[254]*.*

Quaternary structures occur when proteins are made up of more than one polypeptide chain. These individual chains are called subunits, and have their own primary, secondary, and tertiary structures. Subunits are joined by complementary interactions, with hydrogen bonding, electrostatics and hydrophobic effects playing large roles in the stabilization of this interface (*Figure 1.23b*).

1.3.4 Enzymes

Enzymes are proteins that act as biological catalysts to accelerate chemical reactions in biological contexts. They have important roles in many cellular processes, including but not limited to, sensing, metabolism, transport, and regulation[258-261]. Their prevalence in these critical functions have made them attractive drug targets, as many diseases involve the modulation of enzyme activity to some extent[262]. Beyond their clinical relevance, they are highly desirable in industrial applications where they can lead to a reduction in process time, lowering the number of required reactions for a specific transformation, and decreasing the amount of waste produced[263, 264]. Furthermore, enzymes have been engineered to process substrates which have no natural enzyme, leading to the ability to generate specific chemical conversions[265]. They are able to accelerate reactions up to 10^{19}-fold[266], which they can accomplish in a variety of different ways. The simplest of these ways is by stabilizing the reaction transition state through hydrogen-bonding[267] or electrostatic interactions[268]. Besides transition state stabilization, enzymes can also lower the solvent reorganization energy[269] and their dynamics can aid the substrate progressing along the reaction pathway[270]. Finally, some enzymes have been shown to change the reaction mechanism completely[271].

1.3.4.1 Michaelis-Menten and Briggs-Halden kinetics

The simplest form of enzyme catalysis involves three distinct reversible steps, which are depicted in *Scheme 1.2*:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \underset{k_{-c}}{\overset{k_c}{\rightleftarrows}} EP \underset{k_{-r}}{\overset{k_r}{\rightleftarrows}} E + P \qquad \textit{(Scheme 1.2)}$$

1) The binding of the substrate (*S*) to the enzyme (*E*), to form the enzyme-substrate complex (*ES*). This reversible binding is described by the association (k_1) and dissociation (k_{-1}) rate constants. **2)** The conversion of substrate to product to produce the enzyme-product complex (*EP*). This step is described by the rate of chemical conversion (k_c) and the reverse rate of chemical conversion (k_{-c}). **3)** The release of product (*P*) from the enzyme-product complex (*EP*), which is described by the rate of product release (k_r) and binding (k_{-r}).

The two most common enzyme kinetic models, referred to as Briggs-Haldane[272] and Michaelis-Menten, simplify *Scheme 1.2* into a two-state process (*Scheme 1.3*):

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \xrightarrow{k_{cat}} E + P \qquad \textit{(Scheme 1.3)}$$

where both the chemical conversion of the enzyme-substrate (*ES*) to enzyme-product (*EP*) complex and the release of product (*P*) from the enzyme-product complex (*EP*) are assumed to be both fast and irreversible. (i.e. k_c >> k_{-c} and k_r >> k_{-r}). When these assumptions are made, k_c and k_r can be grouped into a single catalytic rate constant ($k_{cat} \ = \frac{k_r * k_c}{k_r + k_c}$), and the rate of product formation can be described as

$$\frac{d[P]}{dt} = v = k_{cat}[ES] \qquad \textit{(Equation 1.8)}$$

where the concentration of the enzyme-substrate complex (*ES*) is described by

$$\frac{d[ES]}{dt} = k_1[E][S] - k_{-1}[ES] - k_{cat}[ES] \qquad \textit{(Equation 1.9)}$$

In both the Michaelis-Menten and Briggs-Haldane kinetic models, the steady state approximation ($\frac{\mathrm{d[ES]}}{\mathrm{dt}} = 0$) is used to simplify *Equation 1.8* and *Equation 1.9* to

$$\frac{d[P]}{dt} = v = \frac{k_{cat}[E]_0[S]}{[S] + K_m} \qquad \textit{(Equation 1.10)}$$

where *[E]₀* is the total concentration of enzyme. In Briggs-Haldane kinetics, K_m is equal to $K_m = \frac{k_{-1} + k_{cat}}{k_1}$. In Michaelis-Menten kinetics, the free enzyme and substrate are assumed to be in rapid equilibrium, with the dissociation rate of the substrate being much larger than the catalytic rate ($k_{-1} \gg k_{cat}$). In this case K_m can be simplified to just be the dissociation constant of the substrate ($K_m = K_d = \frac{k_{-1}}{k_1}$). Importantly, both cases predict the same hyperbolic dependence of substrate concentration on the rate of catalysis (*Figure 1.24a*).

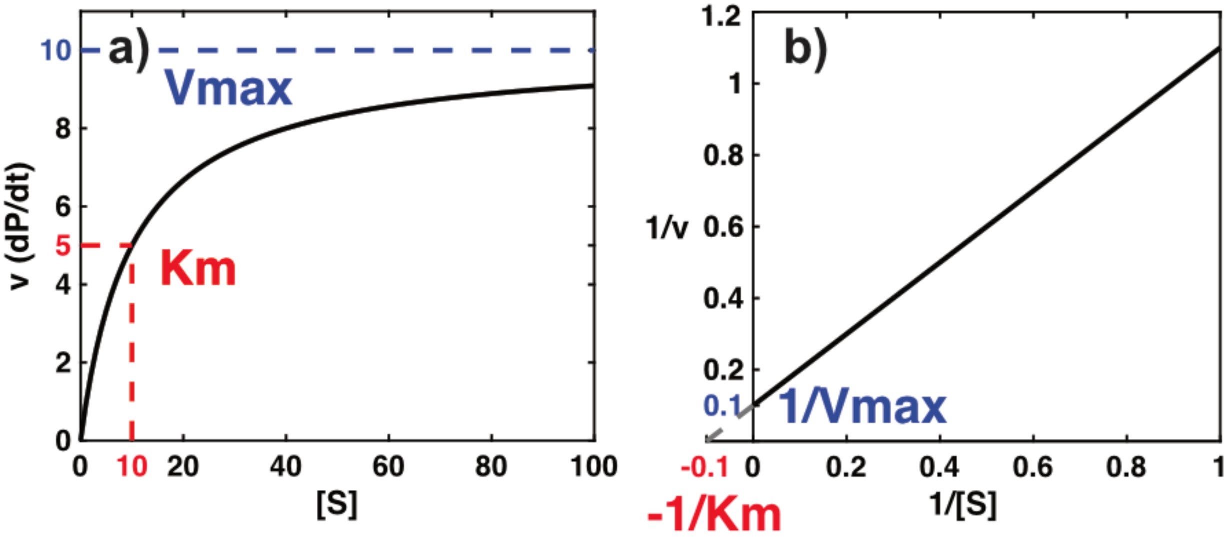


Figure 1.24: Graphs to visualize enzyme activity. a) The dependence given by Equation 1.10 is shown as the black solid line. The maximum rate of product formation (V_{max}) is shown as the blue dashed line. The K_m is given by the substrate concentration when v = ½ V_{max} and is shown as the red dashed line. b) The Lineweaver-Burk plot which linearizes panel a. Here the y-intercept is $1/V_{max}$ and the x-intercept is $-1/K_m$.

At high substrate concentrations (*[S]* >> K_m), *Equation 1.10* simplifies to

$$\frac{d[P]}{dt} = V_{max} = k_{cat}[E]_0 \qquad \textit{(Equation 1.11)}$$

Which is defined as the maximum velocity of the reaction and occurs when all the enzyme in solution is bound to substrate. Furthermore, *Equation 1.10* shows that the substrate concentration which produces a rate of catalysis equal to half of the V_{max} is equal to the K_m. In the rest of this thesis, *Equation 1.10* is referred to as Michaelis-Menten kinetics, and K_m is referred to as the Michaelis constant.

Finally, it is often beneficial to linearize *Equation 1.10* by taking the inverse of each side, which gives

$$\frac{1}{v} = \frac{1}{[S]} * \frac{K_m}{V_{max}} + \frac{1}{V_{max}} \qquad \textit{(Equation 1.12)}$$

Thus, a plot a $\frac{1}{\mathrm{v}}$ vs $\frac{1}{[\mathrm{S}]}$, commonly called a Lineweaver-Burk plot (*Figure 1.24b*), produces a line where the y-intercept is equal to $\frac{1}{\mathrm{V_{max}}}$ and the x-intercept is equal to $-\frac{1}{\mathrm{K_m}}$.

As discussed below, molecules which inhibit enzyme activity change one or both of these parameters and produce very distinct trends when plotted on a Lineweaver-Burk plot.

1.3.4.2 Enzymes in drug discovery

Much of modern drug-discovery is focused on modulating the activity of enzymes. This is often accomplished by developing small molecules which interact with specific enzymes that are dysregulated in a certain disease, restoring their activity back to normal levels[273], or by targeting enzymes present only in a specific bacteria[274] or virus[275]. This is development is typically a multi-step process which begins with identification of the enzyme of interest, followed by high-throughput screening assays to find molecules which inhibit their activity. These high-throughput screens are increasingly done in silico with the aid of molecular docking programs[276], but are also commonly done in vitro via fluorescence[277], mass spectrometry[278], among others. Molecules which are found to interact favorable with the enzyme of interest are referred to as "hits" and are then optimized into "leads" by introducing chemical modifications while characterising their potency and selectivity *(Figure 1.25)*. This is done by combining knowledge of their structure and potency to develop a deep understanding of their structure-activity relationships. Lead compounds then proceed to both pre-clinical and clinical trials, before they can be sold as drugs[279]. This development can be very costly, with only 1 in every 10 compounds that enter clinical trials becoming a drug[280], and new drugs having an average development cost of 2.6 billion[281].

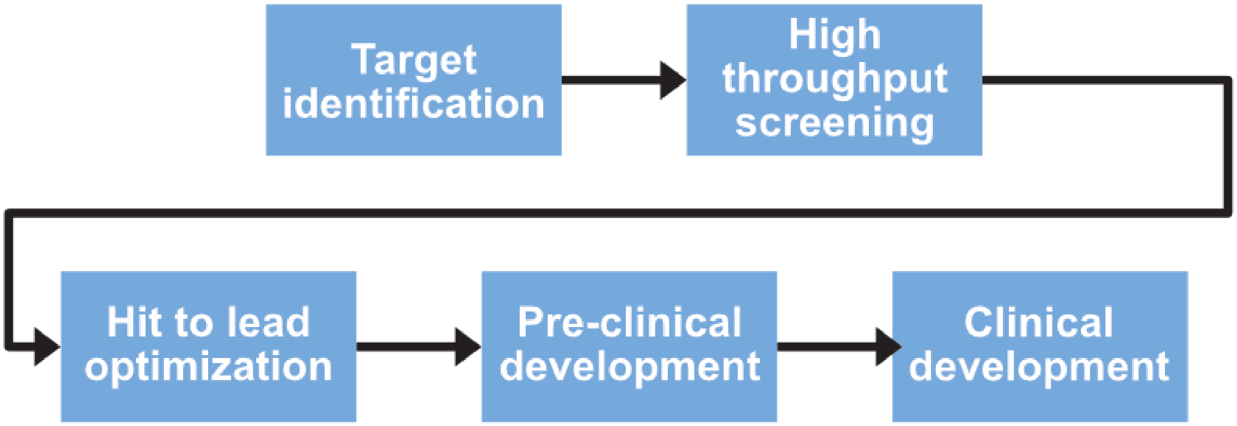


Figure 1.25: Drug development workflow from target identification to clinical development.

1.3.4.3 Enzyme inhibition

The early stages of drug development are typically focused on characterizing the potency of inhibitors *in vitro*. This potency is usually expressed in terms of the affinity between the inhibitor and the target enzyme, usually by extracting the dissociation constant (K_d), which is commonly referred to as the inhibition constant (K_i). Inhibitors can interact with either the free enzyme (E) (*Scheme 1.4*), the enzyme-substrate complex (ES) (*Scheme 1.5*), or both.

$$E + I \overset{K_{i(1)}}{\longleftrightarrow} EI \qquad \textit{(}\text{Scheme 1.4}\textit{)}$$

$$ES + I \overset{K_{i(2)}}{\longleftrightarrow} ESI \qquad \textit{(}\text{Scheme 1.5}\textit{)}$$

Where $K_{i(1)}$ is the dissociation constant for the inhibitor binding to the free enzyme

$$K_{i(1)} = \frac{[E][I]}{[EI]} \qquad \textit{(Equation 1.13)}$$

And $K_{i(2)}$ is the dissociation constant for the inhibitor binding to the enzyme-substrate complex

$$K_{i(2)} = \frac{[ES][I]}{[ESI]} \qquad \textit{(Equation 1.14)}$$

Equation 1.13 and *Equation 1.14* can be combined with *Equation 1.10* to give a full description of possible enzyme inhibition patterns to give

$$\frac{d[P]}{dt} = v = \frac{k_{cat}[E]_0[S]}{[S]\left(1+\frac{[I]}{K_{i(2)}}\right)+K_m\left(1+\frac{[I]}{K_{i(1)}}\right)} \qquad \textit{(Equation 1.15)}$$

This equation describes mixed inhibition, which occurs when the inhibitor binds to both the free enzyme, preventing substrate binding, and the enzyme-substrate complex, preventing conversion of the substrate to the product. This leads to changes in both Michaelis-Menten parameters (K_m and k_{cat}) (*Figure 1.26a*). Furthermore, the limits in this equation lead to the three main forms of inhibition:

$$\frac{1}{K_{i(2)}} \rightarrow 0, \quad \frac{d[P]}{dt} = v = \frac{k_{cat}[E]_0[S]}{[S]+K_m\left(1+\frac{[I]}{K_{i(1)}}\right)} \qquad \textit{(Equation 1.16)}$$

1) Competitive inhibition (*Equation 1.16*) occurs when the inhibitor binds to only the free enzyme and is able to block the binding of the substrate. Often times, these types of inhibitors are structural mimics of the substrate and bind into the active site of the enzyme[282]. In competitive inhibition, only the K_m of the enzyme is affected, leading to a Lineweaver-Burk plot where increasing inhibitor concentration leads to a decrease in the

x-intercept, but no effect on the y-intercept (*Figure 1.26b*). This lowers the enzyme activity when at lower substrate conditions, but still sees the same activity when *[S]* >> K_m.

$$\frac{1}{K_{i(1)}} \rightarrow 0, \quad \frac{d[P]}{dt} = v = \frac{k_{cat}[E]_0[S]}{[S]\left(1+\frac{[I]}{K_{i(2)}}\right)+K_m} \qquad \textit{(Equation 1.17)}$$

2) uncompetitive inhibition (*Equation 1.17*) occurs when the inhibitor binds to the enzyme-substrate complex and blocks the conversion of substrate to product. This type of inhibition is much less common than the other two[283], however there have still been several described in the literature[284-286]. Since the conversion of substrate to product in the enzyme-substrate complex is lowered, the k_{cat} changes where k_1 and k_{-1} are unaffected. This leads to both lower values of both Michaelis-Menten parameters (*Figure 1.26c*).

$$\frac{1}{K_{i(1)}} = \frac{1}{K_{i(2)}} = \frac{1}{K_i}, \quad \frac{d[P]}{dt} = v = \frac{k_{cat}[E]_0[S]}{([S]+K_m)*\left(1+\frac{[I]}{K_i}\right)} \qquad \textit{(Equation 1.18)}$$

3) Non-competitive inhibition (*Equation 1.18*) occurs when the affinities for both the free enzyme and the enzyme-substrate complex are equal. When this happens, only the k_{cat} will change, this leads to a Lineweaver-Burk plots of increasing inhibitor concentration having the same x-intercept, but different slopes and y-intercepts (*Figure 1.26d*).

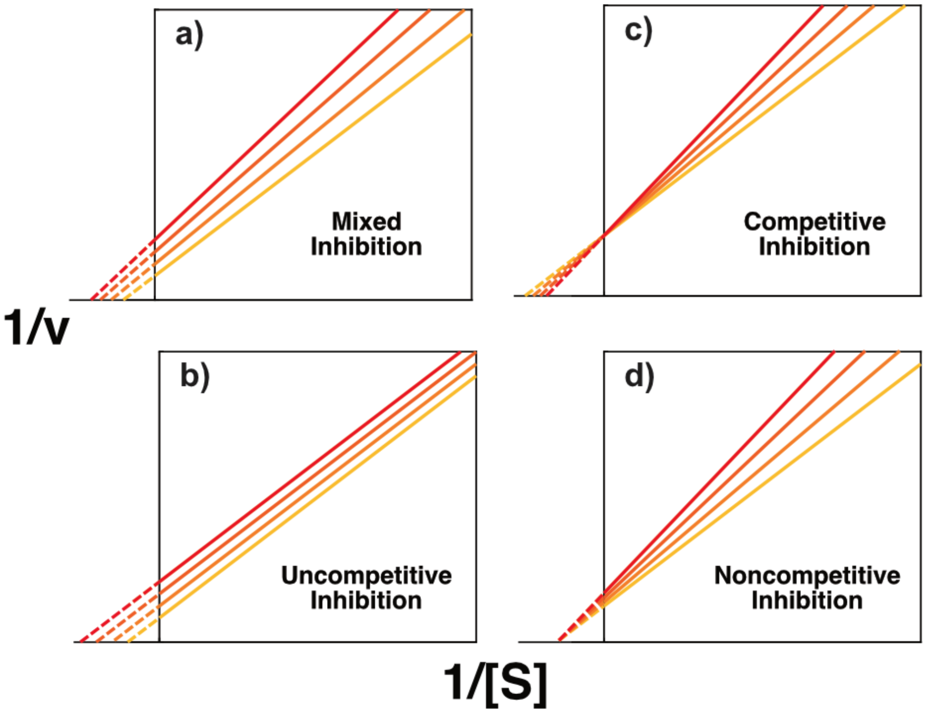


Figure 1.26: Types of inhibition patterns viewed on a Lineweaver-Burk plot. a) Mixed inhibition, b) competitive inhibition, c) uncompetitive inhibition, d) noncompetitive inhibition. Increasing inhibitor concentrations are shown from a orange to red gradient, with orange indicating low inhibitor concentration and red indicating high inhibitor concentration.

In a drug discovery pipeline, it is often too labour intensive to measure the full inhibition profiles for all inhibitors. In this case, the IC_{50} is generally measured as a way to determine the potency of inhibitors relative to each other. An IC_{50} is commonly defined as either the concentration of inhibitor required to lower enzyme activity by 50%, or the concentration of inhibitor required to reduce product formation at a given time by 50%[287]. In order to find the IC_{50}, measurements of enzyme rate or product formation are taken at different inhibitor concentrations and normalized to that of the enzyme and substrate alone. When the %activity or %of product formation is plotted vs the log of inhibitor concentration a sigmoidal curve is obtained. When the enzyme binds exactly one inhibitor, this graph can be fit to the following equation

$$\%activity = \frac{100}{1+\frac{[I]}{IC_{50}}} \qquad \textit{(Equation 1.19)}$$

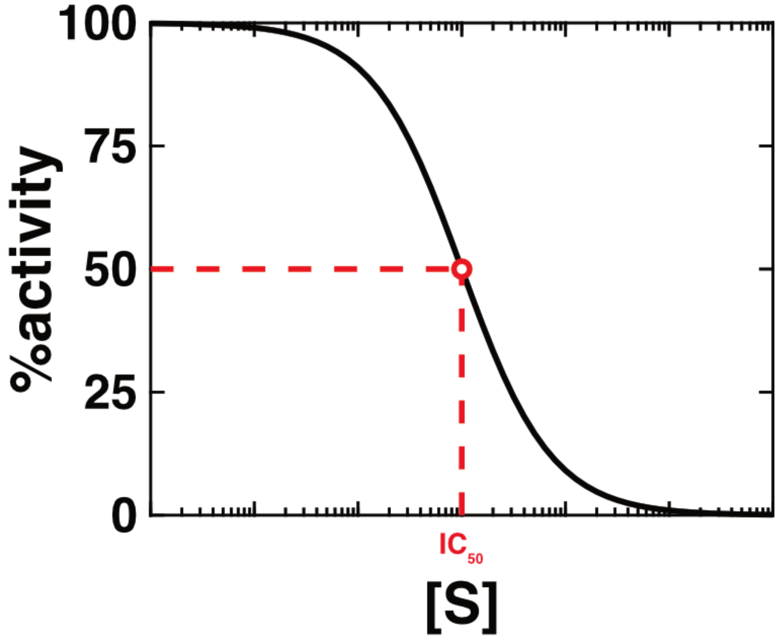


Figure 1.27: Determination of IC_{50}. Substrate concentration is plotted on a log axis.

More complicated forms of this equation can account for multiple inhibitors binding and different upper and lower asymptotes[288]. Furthermore, in the case of competitive inhibition, the K_i can be directly calculated from the IC_{50} as long as the K_m of the substrate is known by using the Cheng-Prusoff relationship[289]

$$K_i = \frac{IC_{50}}{1+\frac{[S]}{K_m}} \qquad \textit{(Equation 1.20)}$$

1.3.4.4 Covalent inhibition

While the aforementioned types of inhibition are by far the most common, inhibitors which form covalent bonds with the target of interest have become increasingly popular. These types of inhibitors were often not pursued due to concerns regarding their toxicity and off target effects[290]. However, many blockbuster drugs were found to be covalent long after their initial discovery[291]. These includes Clopidogrel[292], an anti-coagulant, and Aspirin[293], a common pain-killer. In fact, nearly a third of currently available drugs form covalent bonds with their targets[294]. Due to this discovery, design of covalent drugs has been increasing as they have several beneficial properties over their non-covalent counterparts.

Covalent inhibitors bind to their target enzyme in a two-step reaction following

$$E + I \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} EI \underset{k_{rev}}{\overset{k_{inact}}{\rightleftarrows}} E - I \qquad \textit{(Scheme 1.6)}$$

Where *EI* is a non-covalent complex, similarly to what is seen in typical non-covalent inhibition ($K_i = \frac{k_{off}}{k_{on}}$), and *E-I* is the covalently bound enzyme-inhibitor complex. Here, k_{inact} represents the rate of covalent bond formation and k_{rev} represents the cleavage of the covalent bond. It should be noted that this mechanism is not necessarily exclusive to covalent inhibition but can also explain two-step inhibitors which don't form a covalent bond, such as those who bind a target that then undergoes a slower structural rearrangement.

One advantage of covalent inhibitors over their noncovalent counterparts, is the strength of the covalent bonds, which leads to very slow k_{rev} and long drug-residence times[291]. When $k_{rev} << k_{inact}$ inhibitors effectively become irreversible, making them potent therapeutics. While this is a distinct advantage of covalent inhibitors, they are not without their disadvantages. Among these is that their more complex mechanism means that classical inhibitor analysis is unable to fully quantify their inhibition. Take for example a measurement of IC_{50}. In the case of a noncovalent inhibitor, the analysis is quite trivial. However, when the inhibitor forms a slow irreversible covalent bond, the IC_{50} value becomes time-dependent, with longer experimental times leading to lower IC_{50} values.

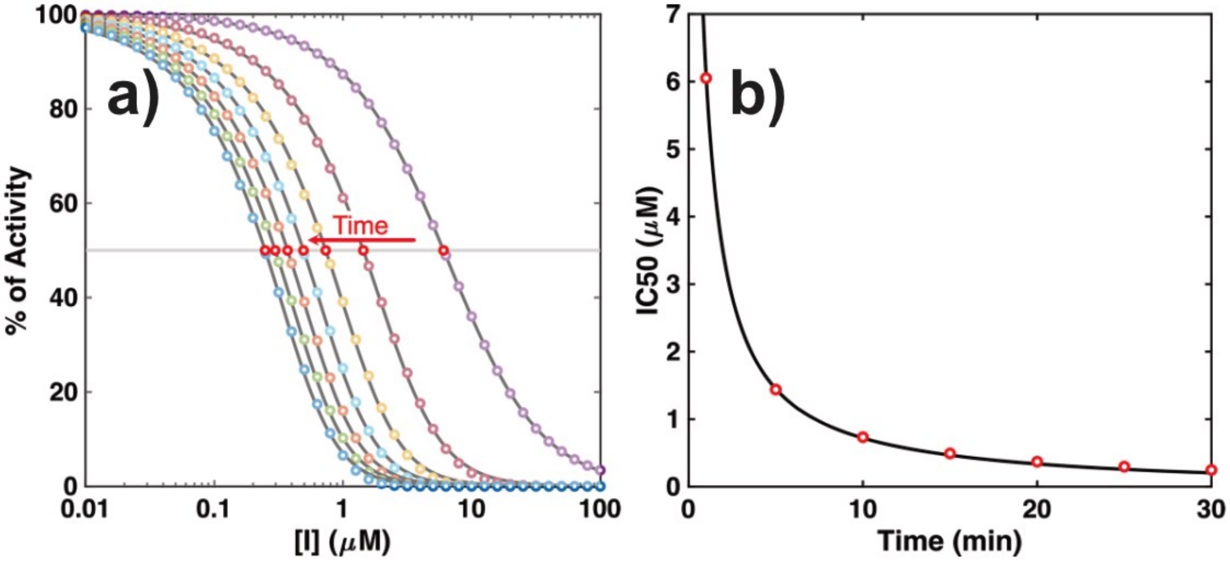


Figure 1.28: Time-dependent IC_{50} measurements of an irreversible covalent inhibitor. a) IC_{50} traces as a function of time. B) IC_{50} values vs time.

$$IC_{50}(t) = K_i * \left(1 + \frac{[S]}{K_m}\right) * \left(\frac{2 - 2e^{-\eta_{IC_{50}} * k_{inact} * t}}{\eta_{IC_{50}} * k_{inact} * t}\right) \qquad \text{(Equation 1.21)}$$

$$\eta_{IC_{50}} = \frac{IC_{50}(t)}{K_i * \left(1 + \frac{[S]}{K_m}\right) + IC_{50}(t)} \qquad \text{(Equation 1.22)}$$

In fact, when taken to the limit of t → infinity, all irreversible covalent inhibitors have the same $IC_{50} = \frac{[E]}{2}$ (assuming one active site per enzyme). Although taking the IC_{50} of inhibitors at a constant time can provide a good measurement of the potency of those molecules,[295] it does not provide any information on whether the inhibitor tightly binds the enzyme and then slowly forms a covalent bond (low K_i, low k_{inact}), or if the inhibitor binds weakly to the enzyme and reacts quickly (high K_i, fast k_{inact}). This an important distinction as an inhibitor which reacts too quickly can lead to off target effects. Furthermore, having information on both the K_i and k_{inact} aids in the optimization of initial hits to leads during a drug discovery pipeline.

A second common way to measure covalent inhibition is through inhibitor concentration-dependent progress curves analysis (IDPC), which requires measuring product formation by the enzyme at different concentrations of inhibitor. These experiments require a way to observe either product formation or substrate consumption as a function of time, which is often done spectroscopically using techniques such as fluorescence. In the example below, the observable comes from monitoring the increasing fluorescence caused by product formation. In its simplest form, it requires the reaction to be performed under conditions of constant product formation (i.e. linear product formation as a function of time in the no-inhibitor control), however corrections for non-linear behaviour can be added[296]. Two-step inhibitors following *Scheme 1.6* show decreasing activity with increasing inhibitor concentration, and distinctly non-linear product formation (*Figure 1.29a*). This non-linear behaviour is caused by the enzyme becoming irreversibly inhibited over time. In the case of two-step inhibition, the traces can be fit independently to the following equation

$$F_t = F_0 + \frac{v_i}{k_{obs}} * (1 - e^{-k_{obs} * t}) \qquad \textit{(Equation 1.23)}$$

Where F_t is the fluorescence at time t, F_0 is the background fluorescence, v_i is the initial slope of the plot, and k_{obs} is the observed rate of inactivation. This analysis provides a series of k_{obs} values which, when plotted vs inhibitor concentration, gives a hyperbolic curve (*Figure 1.29b*). This curve can fit to the following equation

$$k_{obs} = \frac{[I]}{[I] + K_i\left(1 + \frac{[S]}{K_m}\right)} k_{inact} \qquad \textit{(Equation 1.24)}$$

where K_m is the Michaelis-Menten constant, and *[S]* is the concentration of substrate during the measurement.

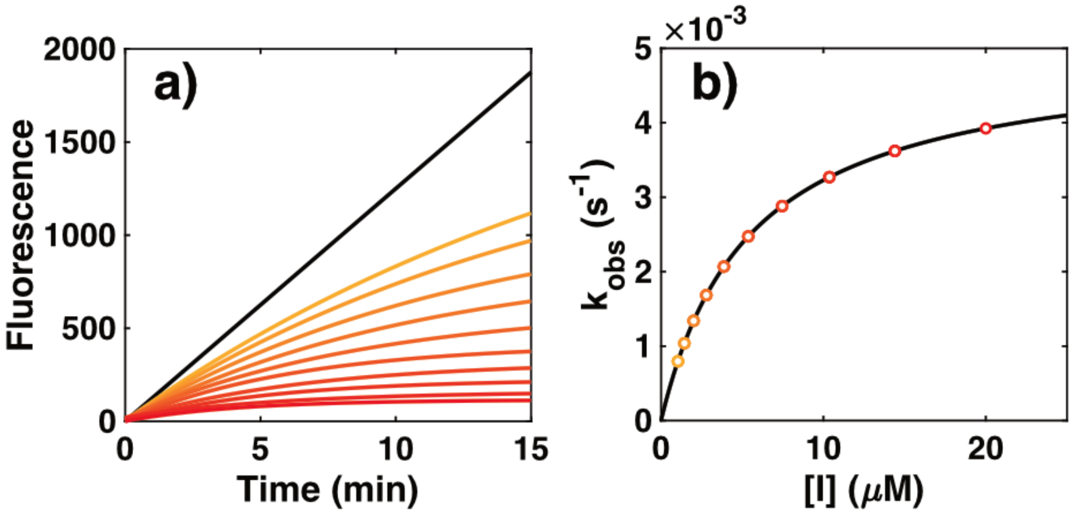


Figure 1.29: Inhibitor concentration-dependent progress curves analysis. a) two-step inhibition following Scheme 1.6, with K_i = 5 μM and k_{inact} = 5e-3 s^{-1}. b) hyperbolic curve of k_{obs} values fit from panel b. Orange to red curves and dots represent increasing inhibitor concentrations. Black lines in panel a/b represent no-inhibitor controls. The black line in panel b represents the curve generated by Equation 1.24 *and the corresponding parameters for panel a.*

1.3.4.5 Measuring enzyme inhibition using isothermal titration calorimetry

There are many ways to measure the velocity of an enzymatic reaction. These include techniques such as NMR spectroscopy[297], UV-Vis spectroscopy[298], Fluorescence spectroscopy[299], mass spectrometry[300], electrophoresis[301], chromatography[302], and isothermal titration calorimetry[303]. All of the aforementioned techniques, besides ITC, measure the concentration of either product or substrate as a function of time and relate this back to the enzyme velocity by taking a first derivative of the data. ITC on the other hand, can relate the velocity of the enzyme directly to the experimental signal measured by the ITC using the following equation

$$\Delta P = v * \Delta H_{react} * V \qquad \text{(Equation 1.25)}$$

Where v is the velocity of the enzyme, ΔH_{react} is the molar enthalpy of the reaction, V is the volume of the ITC's cell, and ΔP is the instantaneous power measured by the ITC instrument.

This is an important distinction, as it gives ITC a unique advantage over other methods, since extracting the Michaelis-Menten parameters requires measuring the velocity of the enzyme at different substrate concentrations. Furthermore, by measuring how these parameters change as a function of inhibitor concentration, key parameters such as the affinity of the inhibitor (K_i) and the mechanism of inhibition can be extracted. Measuring the velocity of the enzyme directly is not the only advantage ITC has over other techniques. It also does not require the use of spectroscopically active, or labelled substrates, and can be done even in the presence of a high concentration of UV-absorbing material such as bovine serum albumin. This makes ITC a near universal enzyme activity assay, as demonstrated in 2001 by Matthew J Todd and Javier Gomez, who were able to measure the Michaelis-Menten parameters for all six classes of enzymes[303].

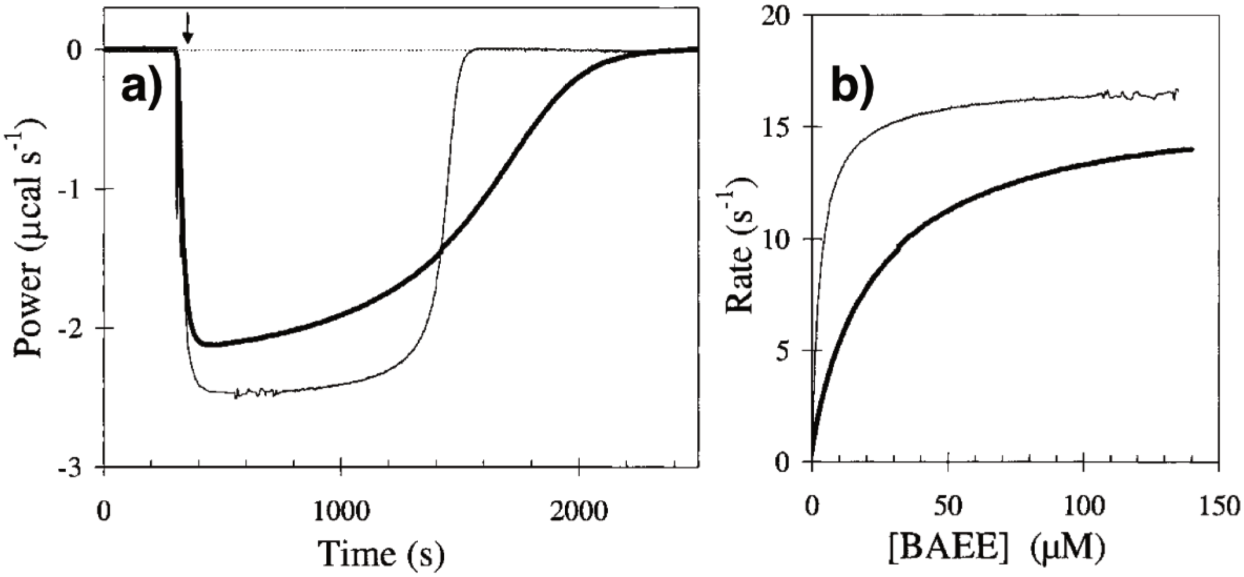


Figure 1.30: Measuring enzyme velocity with a single injection ITC experiment. a) Raw ITC trace. b) Conversion of panel a into a plot of enzyme turnover vs the substrate concentration using Equation 1.25 *and Equation* 1.26*.* The reaction is trypsin catalyzing the hydrolysis of BAEE with (thick line) and without (thin line) benzamidine. Adapted from *Todd and Gomez* with permissions[303].

A typical ITC enzyme activity assay involves the injection of a substrate, held in the syringe of the calorimeter, into the enzyme, held in the sample cell of the calorimeter. This initiates the reaction, and, as the injection is occurring (typically over the period of 1-

80 seconds), the differential power begins to increase (in the case of an endothermic reaction) or decrease (in the case of an exothermic reaction). After the injection is complete the velocity of the reaction will lower back to the baseline signal of the instrument (*Figure 1.30a*). The experimental trace can be converted into a measure of enzyme velocity as a function of substrate concentration (*Figure 1.30b*), by first converting the power measured by the ITC (ΔP) to enzyme velocity (v) using *Equation 1.25* and then converting the measured heat ($\sum \Delta P$) to substrate concentration using the following equation.

$$S_t = S_0 * \left(1 - \frac{\int_{t=0}^{t} \Delta P_t}{\int_{t=0}^{\infty} \Delta P_t}\right) \qquad \textit{(Equation 1.26)}$$

Where ΔP_t is the ITC signal at time t, and S_0 is the initial substrate concentration.

In order to extract the Michaelis-Menten parameters, the final concentration of substrate in the cell after the injection is complete should be sufficiently high (S>> Km). This high concentration is required as, at lower substrate concentrations *Equation 1.10* simplifies to

$$\frac{d[P]}{dt} = v = \frac{k_{cat}[E]_0[S]}{K_m} \qquad \textit{(Equation 1.27)}$$

Where only the ratio of the Michaelis-Menten parameters can accurately be extracted[304]. A single injection ITC experiment is able to measure the velocity of an enzyme at a range of different substrate concentrations, thus when this experiment is performed under ideal conditions, a single injection experiment is able to extract both the K_m and k_{cat} of an enzymatic reaction.

For the past several years, members of the Mittermaier lab have been expanding the toolkit of ITC methods available to experimentalists. For example, *Di Trani et al.* developed a way to fully characterize non-covalent inhibition in a single multi-injection experiment[305]. This experiment requires the sample cell to contain the enzyme of interest, and the syringe to contain a mixture of a inhibitor and substrate. Multiple injections of this mixture are made into the sample cell, and the Michaelis-Menten parameters are extracted from each injection individually. In each injection, the same amount of substrate is titrated into the sample cell and consumed fully. However, each injection also adds a certain amount of inhibitor. This inhibitor is not consumed at any point during the experiment, so each injection is measuring the velocity of the enzyme under both different substrate concentrations and different inhibitor concentrations. The resulting ITC trace is shown in *Figure 1.31a,* where sequential injections have progressively shorter and broader peaks, which shows that the velocity of the enzyme is reduced with each injection.

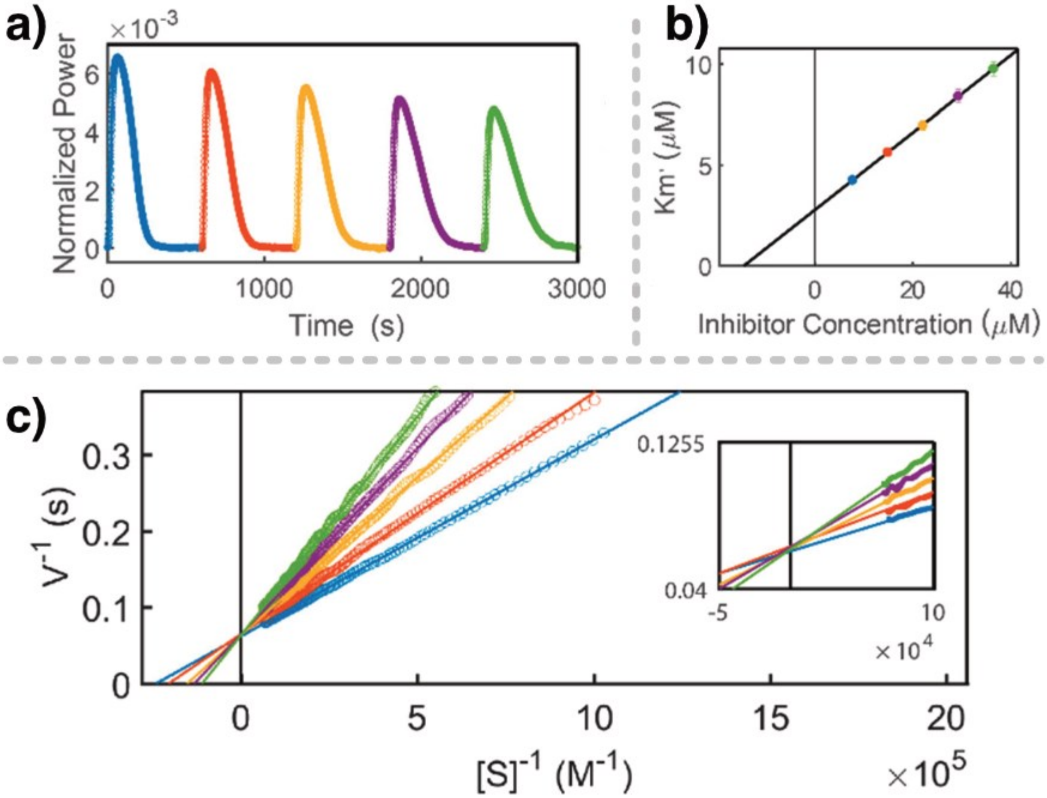


Figure 1.31: Multi-injection ITC experiment characterizing the inhibition of enzymatic reaction. a) Normalized ITC signal. b) Each injection from panel a converted into a Lineweaver-Burk plot, showing competitive inhibition. Adapted from *Di Trani et al.* with permissions[305].

Each of these peaks can then be converted into a Lineweaver-Burk plot (*Figure 1.31c)*. In this specific example, the trend shows that as the inhibitor concentration increases, the maximum velocity of the enzyme remains constant, but the Michaelis-Menten constant (*K*$_m$) increases after each injection. This is characteristic of competitive inhibition as shown in *Figure 1.26b*, where the apparent *K*$_m$ ($K_{m(app)}$) increases as a function of inhibitor concentration according to

$$K_{m(app)} = \mathrm{K_m}\left(1 + \frac{[\mathrm{I}]}{\mathrm{K_{i(1)}}}\right)$$

Thus, a plot of apparent K_m vs inhibitor concentration produces a straight line, with the y-intercept being the uninhibited K_m, and the x-intercept being the negative of the K_i.

In another example *Di Trani et al.* were able to quantitatively model the trace of an ITC signal using the instruments response function enabling them to measure rapid time-scale kinetics[306]. They used this fitting technique to develop two new multi-injection ITC experiments, the kinetics of inhibition and the kinetics of initiation[307]. In the kinetics of inhibition experiment, a mixture of enzyme and a large excess of substrate are held in the sample cell of the ITC. Thus, at the start of the ITC experiment the reaction is occurring at a constant rate. Upon injection of inhibitor into the sample cell, the rate of the reaction slows down gradually until coming to a new equilibrium. This gradual decrease in activity is caused by the inhibitor slowly binding to the enzyme. This binding can be described by

$$\frac{d[EI]}{dt} = k_{on}[E][I] - k_{off}[EI] \qquad \textit{(Equation 1.28)}$$

Where k_{on} and k_{off} are the on and off rate of the inhibitor respectively. The velocity of the enzyme will be affected by the loss of free enzyme ([E]) to the enzyme inhibitor complex (*EI*), following

$$\frac{d[P]}{dt} = v = \frac{k_{cat}[S]([E]_0 - [EI])}{[S] + K_m} \qquad \textit{(Equation 1.29)}$$

Thus, by fitting these equations to multiple injections of inhibitor, both the on and off rates can be extracted.

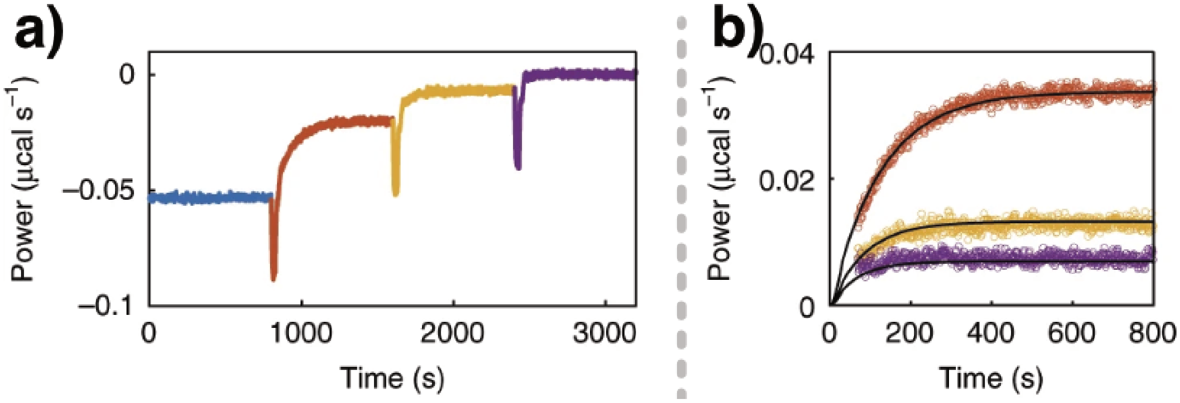


Figure 1.32: Kinetics of inhibition and initiation ITC experiments. ITC trace of kinetics of inhibition experiment. b) Overlay of each injection in panel a, fit to Equation 1.28 and Equation 1.29 is shown in black. Adapted from Di Trani et al. with permissions[307].

1.4 Thesis objectives

Biomacromolecules play a pivotal role in a diverse set of biological processes, and measuring their fundamental physical characteristics is key to understanding their biological relevance and developing novel therapeutics to modulate their function. These molecules fold and assemble through complex pathways which can have large energetic barriers. Conventional techniques are not always suitable for characterizing these processes. This thesis aims to increase the experimental methods available by combining common laboratory equipment with modern day computational power and mathematical modeling to develop techniques which address the need for the swift and cost-effective characterization of biomacromolecules. Chapter 2 delves into a global-fitting analysis tailored for non-equilibrium thermal denaturation experiments, applied specifically to unravel the folding dynamics of guanine quadruplexes (G4s). These four-stranded, non-canonical nucleic acid structures are implicated in cancer and exhibit multiple folding pathways, potentially influencing their biological function. Chapter 3 introduces the concept of guanine quadruplex containing regions (G4CRs), contiguous DNA stretches with the potential to form multiple stable G4 structures. GReg, a bioinformatic algorithm which is able to find and characterize G4CRs within a sequence of DNA is applied to human promoter sequences. In Chapter 4, an experimental approach involving cyclic heating and cooling ramps is presented to measure thermodynamic information on slowly assembling supramolecular structures. This technique, which can extract information previously unattainable with other methods, is employed to study the co-assembly of poly-adenosine strands and cyanuric acid into long supramolecular fibers, and is discussed in terms of small-molecule loading. Finally, Chapter 5 details a method for measuring the binding kinetics of covalent inhibitors using isothermal titration calorimetry. This approach offers faster and more robust characterization of these inhibitors, eliminating the need for modified substrates, as ITC directly measures the rate of enzymatic catalysis. Together, these innovative approaches constitute valuable additions to the researcher's toolkit, enabling rigorous characterizations of biomacromolecular folding, assembly, and function.

Chapter 2: Parallel reaction pathways accelerate folding of a guanine quadruplex

2.1 Preface

The work in the following chapter was a result of a collaborative effort with Robert Harkness. Rob was a previous member of the Mittermaier lab and spent a lot of time training me when I started. He helped drive a lot of the conclusions of this chapter and helped me optimize my experimental and fitting procedures to extract the necessary kinetic parameters from my thermal hysteresis traces. Rob had previously looked at how structural heterogeneity effected the thermodynamic stability of the c-MYC quadruplex[210], and I took the main experimentalist role in incorporating the techniques from his paper to address how this structural heterogeneity affected the kinetics of the c-MYC quadruplex. We are co-first authors of the manuscript from which this chapter was adapted. Rob validated our use of guanine to inosine mutations to create structural mimics of each G4 isomer using a combination of NMR and CD. Furthermore, he ran simulations to show how a three-state mechanism involving a slightly populated intermediate can still be analyzed as a two-state transition by incorporating sloped baselines. I have removed these sections from the chapter, as they represent a large body of work done which was not performed by me and have instead referenced his conclusions where appropriate. Our results were also orthogonally validated with isothermal NMR photo-caging experiments performed by our collaborators under the supervision of Prof. Harald Schwalbe. One again I have removed this section of the paper and referenced the results in the necessary sections. I measured and analyzed all of the thermal hysteresis data which drove the main conclusions of the chapter that the c-MYC quadruplex contains four distinct assembly pathways to different isomers, and that the presence of these pathways leads to a net folding acceleration of greater than 2.5-fold.

This chapter was adapted with permission from: Harkness, R. W.†, Hennecker, C.†, Grün, J. T., Blümler, A., Heckel, A., Schwalbe, H., & Mittermaier, A. K. (2021). Parallel reaction pathways accelerate folding of a guanine quadruplex. *Nucleic acids research*, 49(3), 1247-1262.

2.2 Abstract

G-quadruplexes (G4s) are four-stranded, guanine-rich nucleic acid structures that influence a variety of biological processes such as the transcription and translation of genes and DNA replication. In many cases, a single G4-forming nucleic acid sequence can adopt multiple different folded conformations that interconvert on biologically relevant timescales, entropically stabilizing the folded state. The coexistence of different folded conformations also suggests that there are multiple pathways leading from the unfolded to the folded state ensembles, potentially modulating their folding rate and biological activity. This chapter details the development of an experimental method for quantifying the contributions of individual pathways to the folding of conformationally heterogeneous G4s that is based on mutagenesis, thermal hysteresis kinetics experiments, and global analysis. We applied this method to the regulatory Pu22 G4 from the c-MYC oncogene promoter, which adopts at least four distinct folded isomers. We found that the presence of four parallel pathways leads to a 2.5-fold acceleration in folding; that is, the effective folding rate from the unfolded to folded ensembles is 2.5 times as large as the rate constant for the fastest individual pathway. Since many G4 sequences can adopt many more than four isomers, folding accelerations of more than an order of magnitude are possible via this mechanism.

2.3 Introduction

G-quadruplexes (G4) are four-stranded, helical, nucleic acid structures formed when guanine (G)-rich tracts in DNA or RNA come together to form G-tetrads, arrangements of four planar, Hoogsteen-hydrogen bonded Gs that are stacked to form the core G4 structure, with a cation (typically, Na^+ or K^+) coordinated between each pair of tetrads (*Figure 2.1A,B*)[63, 124]. G4s are typically associated with sequences following the pattern 5'-$G_{3\text{-}5}N_{1\text{-}7}G_{3\text{-}5}N_{1\text{-}7}G_{3\text{-}5}N_{1\text{-}7}G_{3\text{-}5}$-3'[59, 119], where three G-tetrads form the canonical core structure. However, divergent patterns are also sometimes observed that include G4s with bulges[308], vacant positions in their cores[309], and those formed with two or more than three G-tetrads[310, 311]. G4-forming sequences are found in genomes from all three domains of life[140, 312, 313] as well as in mitochondria[314], and viruses[315]. G4s are implicated in the regulation of many biological processes such as DNA replication[316], transcription[317], mRNA translation[318], alternative splicing[319], and post-translational protein processing[320]. Intriguingly, G4 sequences frequently adopt multiple folded conformations of similar free energy, leading to conformational ensembles of interconverting structures. For example, we have shown that many G4 sequences within the promoter regions of human genes contain G-tracts with non-identical numbers of Gs[321]. This type of sequence produces an ensemble of different strand-shifted conformations we term G-register (GR) isomers[59, 321], where each structure incorporates a unique subset of the available Gs in the G4 core (*Figure 2.1a*). Previously, we have investigated the influence of adopting multiple GR isomers on the thermal stability of the c-MYC Pu18, VEGFA, and PIM1 G4s[321]. Through mutational analysis, thermal melting-based global fitting, and computer simulations, we demonstrated that these G4s all populate an ensemble of GR isomers at equilibrium. The minimum number of GR isomers is given simply by the number of G-tract alignments afforded by their additional core dG residues. We found that for the PIM1 G4, GR isomers can also have different topologies such that G-tract shifting is accompanied by strand reversals, pointing to the existence of a complex conformational energy landscape. Remarkably, in all cases we found that these structural dynamics thermodynamically stabilize G4s by reducing the equilibrium entropy penalty for folding in a manner that correlates with the number of GR isomers populated by the sequence[321]. Since G4 stability is known to have a direct influence on gene expression levels[322, 323] and regulatory proteins can to bind to different GR isomers[324], these dynamics offer an additional level of control. In another recent NMR folding kinetics study, we explored the kinetic mechanism of GR exchange dynamics[216]. We demonstrated that the GR exchange mechanism can depend on the particular G4 sequence. For example, we found for the c-MYC Pu18 and hTERT promoter G4s that GR isomers exchange with one another largely via the unfolded state in a three-state manner; that is, each GR isomer transits through the unfolded state without populating other structural intermediates. For the c-MYC Pu-18, the exchange can also occur to some extent through direct G-tract sliding[216].

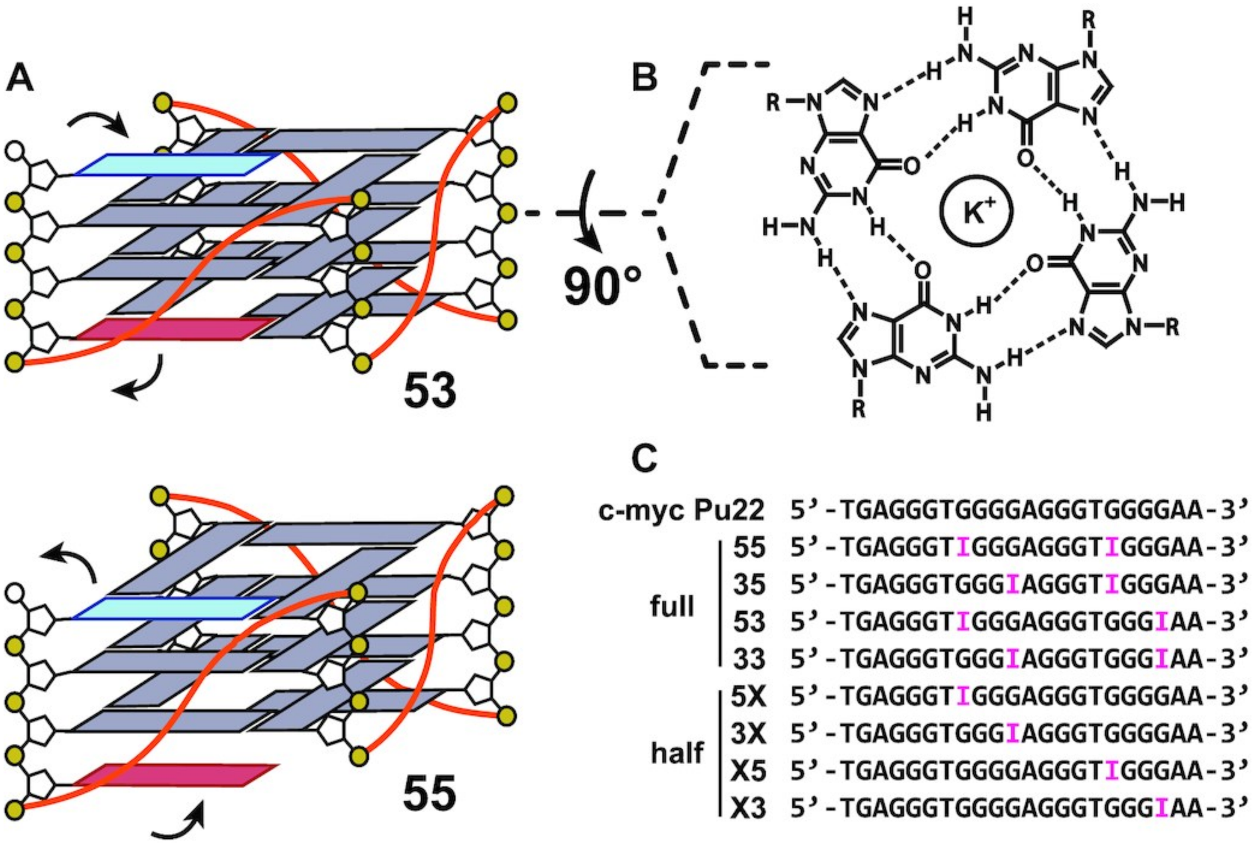

A
53
55
90°
B
K+
C
c-myc Pu22 5'-TGAGGGTGGGGAGGGTGGGGAA-3'
full
55 5'-TGAGGGTIGGGAGGGTIGGGAA-3'
35 5'-TGAGGGTGGGIAGGGTIGGGAA-3'
53 5'-TGAGGGTIGGGAGGGTGGGIAA-3'
33 5'-TGAGGGTGGGIAGGGTGGGIAA-3'
half
5X 5'-TGAGGGTIGGGAGGGTGGGGAA-3'
3X 5'-TGAGGGTGGGIAGGGTGGGGAA-3'
X5 5'-TGAGGGTGGGGAGGGTIGGGAA-3'
X3 5'-TGAGGGTGGGGAGGGTGGGIAA-3'

Figure 2.1: G4 structures and sequences were investigated in this study. (A) GR isomers with exchanging dG residues colored red and blue. (B) G-tetrad structure. (C) The WT c-MYC Pu22 and trapped mutant sequences were investigated in this study. dG to dI mutations to the WT sequence are shown in pink. Full corresponds to the fully trapped sequences that contain dual mutations, whereas half indicates the half-trapped sequences with a single mutation that are capable of undergoing two-state GR exchange. Numbers indicate the direction that the G-tract is locked by mutation; for example, the 33 sequence has both exchanging G-tracts locked in the 3' direction. The exchanging G-tract for the half-trapped sequences is denoted by X.

The roles of G4s in the cell have been rationalized in terms of thermodynamic equilibria between duplex and G4 forms of DNA [325]. Yet, many of the processes that G4s help to control are fundamentally nonequilibrium and folding kinetics could become highly influential[326, 327]. Several groups have studied G4 folding in a double-stranded DNA background where multiple folded G4 populations were observed in coexistence with the unfolded or duplex states[328, 329] pointing to a rich dynamic repertoire that could help regulate biological function. For example, during transcription and replication, helicases actively unwind the stable DNA duplex, briefly offering single-stranded regions unique opportunities to fold. mRNA secondary structures are unwound by the ribosome with each round of translation. The relative rates of G4 formation versus other processes such as polymerase and ribosomal engagement and movement and interactions with binding proteins could represent decisive factors in G4 function. For instance, RNA Pol II generates a transcription bubble of roughly 18-25 unwound base pairs[330]. The enzyme can spend several minutes at the site of initiation before achieving promoter escape[331] – it often pauses at promoter-proximal sites[332] – and the rates of polymerization can vary greatly even within the same gene, providing a highly variable and dynamic environment for single-stranded DNA to adopt the G4 fold. There is evidence that passage of RNA Pol II can promote G4 folding, blocking DNA replication and generating genome instability in yeast[317]. In another example, synthesis of the lagging strand during DNA replication involves generating thousands of bases of single stranded DNA while the polymerase synthesizes the complimentary strand through a backstitching mechanism[333]. The DNA, thus exposed, is protected through interactions with replication protein A (RPA), which binds 30 single-stranded nucleotide stretches with sub-nanomolar affinity[334, 335], preventing reannealing of the DNA and degradation by nucleases. A recent model of RPA function proposes that proteins bind individually to single-stranded DNA and then close the gaps between them by sliding to form a contiguous coating[334], potentially leading to brief and fluctuating opportunities for G4 formation. Finally, folding of mRNA into G4 structures has been reported to impede translation in both prokaryotes and *eukaryotes in vitro* and *in vivo*[327, 336], although some aspects remain controversial[337]. Secondary structures are absent from mRNAs immediately after they emerge from RNA Pol II or the ribosome. Therefore translation inhibition by G4s depends on the balance between their folding rates and the delays between transcription and subsequent rounds of translation[338]. An additional level of dynamic control can exist when the G4-forming region can adopt meta-stable alternative folds such as hairpin structures[327, 339] or partly-structured folding intermediates like three-stranded triplexes[340-342]. Thus, characterizing G4 folding mechanisms and elucidating rules for how the primary nucleotide sequence influences folding rates are key to understanding how G4s function in the cell.

G4 folding has been studied by a variety of approaches such as rapid mixing with cations[216, 343], thermal melting combined with singular-value decomposition[344], atomic force microscopy[345], or single-molecule Förster resonance energy transfer (smFRET)[346, 347]. In general, folding studies of G4s have focused on quantifying the properties and populations of partly structured intermediates along the folding trajectory, though some G4s have been found to fold in a two-state manner[321]. Here we have investigated the possibility that conformational heterogeneity of the folded-state ensemble may be an important factor in controlling how quickly G4s are able to fold, and one that until now has been largely overlooked. We previously investigated how conformational heterogeneity influences thermodynamics, and how different folded conformers interconvert[216, 321]. However, the relationship between conformational heterogeneity and folding rates has been largely unexplored and is important for understanding G4 function. In essence, this relationship is based on the idea that the transition from the unfolded state to each conformational isomer making up the folded state can be considered a separate folding pathway, particularly when the exchange between folded isomers is slow[321]. In principle, the existence of multiple parallel pathways accelerates biomolecular folding, since the apparent macroscopic folding rate constant is equal to the sum of the microscopic rate constants for the individual pathways[348]. However, to our knowledge, this effect has never been quantified for G4s. Here we investigate the relationship between pathway multiplicity and the folding kinetics of G4s for the regulatory c-MYC oncogene promoter Pu22 G4, which we have found to adopt a minimum of four distinct GR isomers in the folded state with conserved parallel topologies. Consequently, the formation of this ensemble can be described by folding through four parallel pathways. This G4 controls 80-90% of the expression of the potent c-MYC oncogene[349], with c-MYC estimated to be deregulated in >50% of human cancers[350]. Our approach combines mutagenesis, thermal hysteresis (TH) kinetics melting experiments[218, 246], and a novel global fitting procedure to dissect the kinetic contributions of individual pathways to the overall folding rate. We find that folding of the Pu22 G4 is accelerated by roughly a factor of 2.5 due to the existence of four parallel pathways, with a macroscopic folding rate about 2.5 times as large as that of the most rapid individual route. Orthogonal kinetic experiments performed using NMR to monitor the folding of G4s following laser photolysis of caging groups gave results in good agreement with the thermal hysteresis analysis[351]. Isothermal folding through multiple pathways leads to a transient conformational ensemble where the populations of folded isomers are controlled kinetically rather than thermodynamically; the equilibrium populations are then gradually reached via GR exchange. The principle of folding acceleration by parallel reaction pathways extends to more complex GR conformational ensembles that can have 20 or more distinct structural isomers, potentially giving an order of magnitude rate increases. Our TH-based approach is widely applicable to complex nucleic acid folding mechanisms and provides a simple and effective method for unraveling the contributions of parallel reaction pathways.

2.4 Results

2.4.1 Trapping individual GR isomers by systematic mutation

We studied the c-MYC Pu22 G4 sequence, 5'-TGAGGGTGGGGAGGGTGGGGAA-3', which is derived from the promoter region of the human c-MYC oncogene[352]. This sequence contains four dG residues in its 2nd and 4th G-tracts and three in the 1st and 3rd (*Figure 2.1C*). Thus there are two ways for both the 2nd and 4th tract to align with respect to the to the 1st and 3rd, giving rise to four GR isomers in total (*Figure 2.1a*)[321]. Exchange between c-MYC GR isomers takes place on the minutes timescale (~2×10^{-2} min^{-1}) at ambient temperatures[216]; therefore, the transition from the unfolded state to each of the folded isomers can be considered a distinct folding pathway. In order to study each pathway separately, we used site-directed mutagenesis to trap the 2nd and 4th G-tracts shifted in either the 5' or 3' direction with respect to the 3 G-tetrads of the G4 core. We substituted surplus dG residues in the 2nd and 4th G-tracts with deoxyinosine (dG>dI) which lacks an N2 amino group but is otherwise identical to guanine. The hypoxanthine base can therefore closely mimic the physical properties of surplus guanines in loop positions, but cannot form the full complement of H-bonds in a G-tetrad and is therefore disfavored within the G4 core[351]. In what follows, we refer to the c-MYC Pu22 as wild-type (WT), sequences containing two dG>dI substitutions as fully-trapped mutants, and those with a single dG>dI mutation as half-trapped mutants. We have previously shown that trapped mutants are excellent structural and thermodynamic mimics of the corresponding WT GR isomers[321, 351]. For instance, the Pu22 double mutant

5'-TGAGGGTIGGGAGGGTIGGGAA-3' mimics the WT GR isomer in which both the 2nd and 4th G-tracts are shifted in the 5' direction, and we refer to this as the 55 fully-trapped mutant. The Pu22 single mutant 5'-TGAGGGTGGGGAGGGTIGGGAA-3' mimics the two WT GR isomers in which the 4th G-tract is shifted in the 5' direction and the 2nd G-tract can shift in either 3' or 5' directions, and we refer to this as the X5 half-trapped mutant. We systematically mutated positions 8, 11, 17, and 20 in the Pu22 G4 to generate a library of 4 fully-trapped (55, 33, 35, 53) and 4 half-trapped (5X, 3X, X5, X3) mutants, in addition to the WT sequence.

2.4.2 Folding kinetics characterized by thermal hysteresis

We used TH measurements to characterize the folding kinetics and thermodynamic stabilities of the Pu22 WT G4 and trapped mutants. The TH approach is based on spectrophotometric detection of folding/unfolding as the temperature is varied, similarly to simple thermal melting experiments. In the case of TH, the temperature ramp rate is adjusted to be fast compared to the length of time required by the molecules to equilibrate such that populations of folded and unfolded states lag behind their equilibrium values. This causes the melting mid-point on the up-scan to occur at a higher temperature than the refolding mid-point on the down-scan. The gap between heating and cooling profiles increases with increasing scan rates and gives detailed information on the folding and unfolding rate constants of the system[246]. The transition temperatures themselves are related to the thermodynamic stability of the system.

We measured TH datasets for the Pu22 WT G4 and trapped-mutant sequences at 6 temperature scan rates ranging from -4 to 4 °C min^{-1} (*Figure 2.2*). The data were baseline (See Materials and methods, *Supplementary Figure 2.4*) and temperature corrected (*Supplementary Figure 2.3*) as described previously[246]. TH absorbance thermograms were collected at both 260 and 295 nm for the fully trapped mutants to test whether folding is two-state, since the presence of well-populated intermediates such as triplexes can lead to large discrepancies between folding/unfolding curves collected at these two wavelengths[341]. Data for the fully-trapped mutants collected at the two wavelengths overlaid closely (*Supplementary Figure 2.1*), suggesting that partly structured folding intermediates with very different spectroscopic properties are not appreciably populated and the fully-trapped mutant G4s fold in an effectively two-state manner. We further tested this assumption by comparing the kinetic parameters extracted from global fits to the TH datasets at both wavelengths, finding similar values in all cases (details given below, *Table 2.1*, Supplementary Table 2.14, and *Figure 2.3*), which further supports the idea that folding is effectively two-state for these mutants. Finally, CD spectra collected over a range of temperatures superimpose extremely closely to one another, suggesting that no well populated structural intermediates are formed at temperatures near the transition (*Figure 2.4*). This agrees with several other studies of the Pu22 G4. For instance, mechanical unfolding measurements obtained using magnetic tweezers were consistent with a two-state transition[353]. Similarly, a two-state kinetic model gave good agreement with TH measurements for dT-variants of the trapped mutants studied here (dG>dT instead of dG>dI substitutions)[354]. Finally, it has been shown that under some circumstances, Pu22 folding can involve well-populated folding intermediates[355], but as we discuss below, this does not preclude the rigorous use of two-state equations to describe folding in the current study. Specifically, Gray et al. found that when folding of the dT-variant of the 33 fully trapped mutant (dT-33) is initiated by the rapid addition of K^+ ions, antiparallel intermediates are formed early in the process, followed by slow conversion to the parallel topology ground state[355]. Similarly, mass spectrometric analyses of K^+-initiated G4 folding detected long-lived, incompletely K^+-coordinated G4 folding intermediates[56]. However, we have found that folding reactions starting from a K^+-free unfolded state (i.e. K^+-initiated) differ fundamentally and are slower by about two orders of magnitude, compared to reactions starting from a K^+-equilibrated unfolded state, such as in TH experiments[216]. The folding intermediates detected in K^+-initiated folding experiments likely represent kinetic traps that are not present when the unfolded state is allowed to equilibrate with K^+ ions, as they do in this study. Gray et al. also analyzed circular dichroism thermal unfolding traces of the dT-33 fully trapped mutant using singular value decomposition. They found that unfolding involved three-states: folded (F), unfolded (U) and an intermediate (I). The difference in enthalpy between F and I was only 12% of the total enthalpy of unfolding, and the reconstructed parallel topology CD spectra of the F and I states were very similar. Therefore, the I state appears to involve a modest rearrangement or slight partial unfolding of the F state.

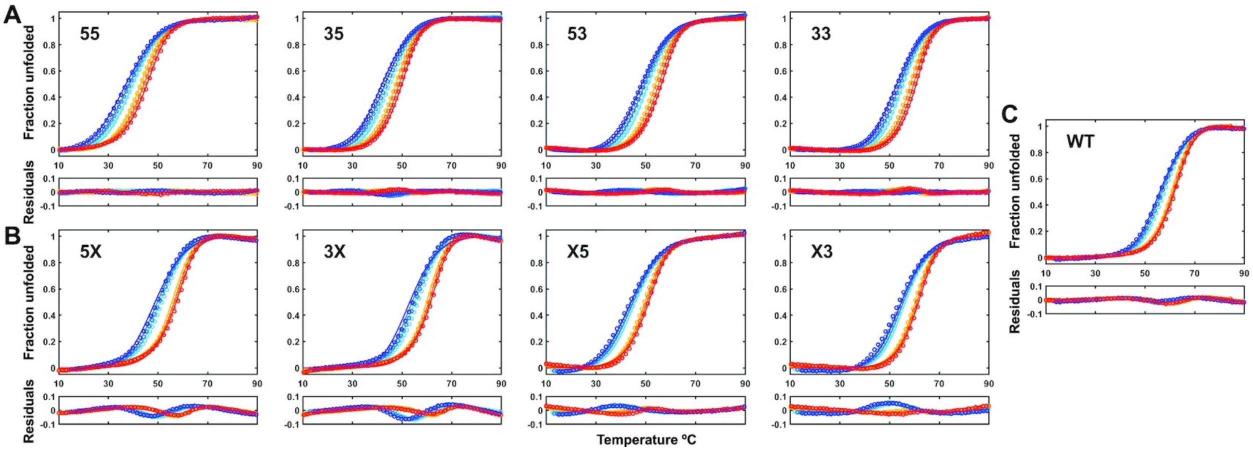


Figure 2.2: Global fits of a parallel pathways model to TH data for the WT and trapped-mutant G4s. TH datasets (295 nm) for the fully trapped (A), half-trapped (B) and WT (C) G4s. Fit residuals are shown in the subpanel below each dataset. Light to dark blue and orange to red indicate slowest to fastest annealing and melting scan rates, respectively. Experimental data are shown as colored circles, while optimized global fit data are colored lines.

GR isomer	E_F (kJ mol^{-1})	k_F (min^{-1})	E_U (kJ mol^{-1})	k_U (min^{-1})	T_m (°C)
55	-36.0 ± 0.7	(387 ± 6)×10^{-3}	121 ± 1	(162 ± 3)×10^{-3}	41.5
35	-47.9 ± 0.8	(572 ± 6)×10^{-3}	140.4 ± 0.8	(72.1 ± 0.9)×10^{-3}	46.1
53	-54.6 ± 0.6	(917 ± 8)×10^{-3}	144.5 ± 0.4	(23.2 ± 0.2)×10^{-3}	52.5
33	-54.4 ± 0.5	(1220 ± 9)×10^{-3}	164.3 ± 0.3	(7.2 ± 0.1)×10^{-3}	57.0

Table 2.1: Kinetic parameters extracted from global fits to the c-myc Pu22 and trapped mutants. TH datasets were collected at 295 nm according to Scheme 2.1c with k_{ex} = 0 min^{-1}. Rate constants are reported at T_0 = 37 °C. Equilibrium T_m values were extracted from two-state fits of slow scanning (0.2 °C min^{-1}) thermal melts monitored at 295 nm (Figure 2.2). Errors were calculated using a Monte-Carlo approach as described in the materials and methods[356].

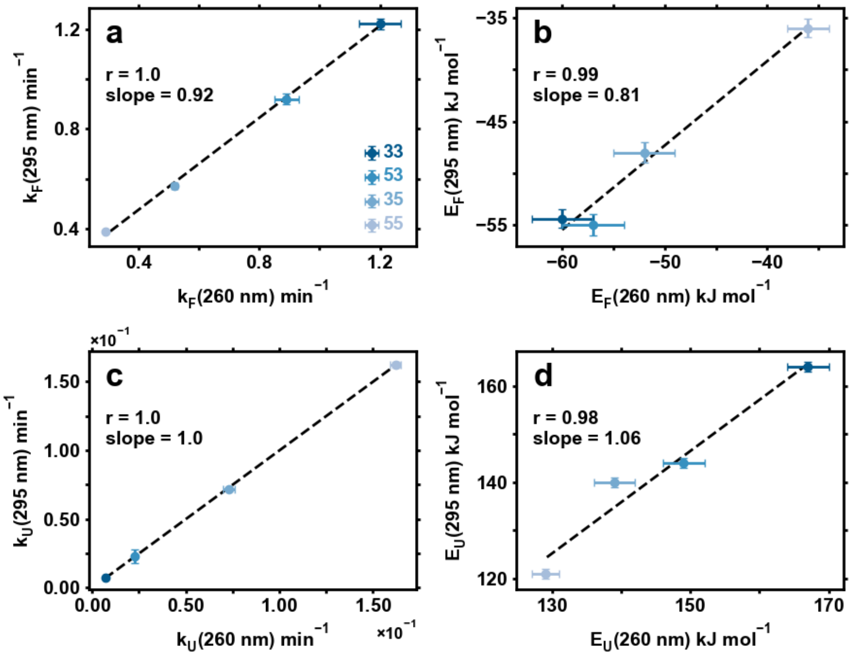


Figure 2.3: Correlations of optimized folding (a,b) and unfolding (c,d) kinetic parameters from global fits to the c-MYC Pu22 and trapped mutant TH datasets collected at 295 and 260 nm respectively. In (a,c), rate constants are reported at 37 °C.

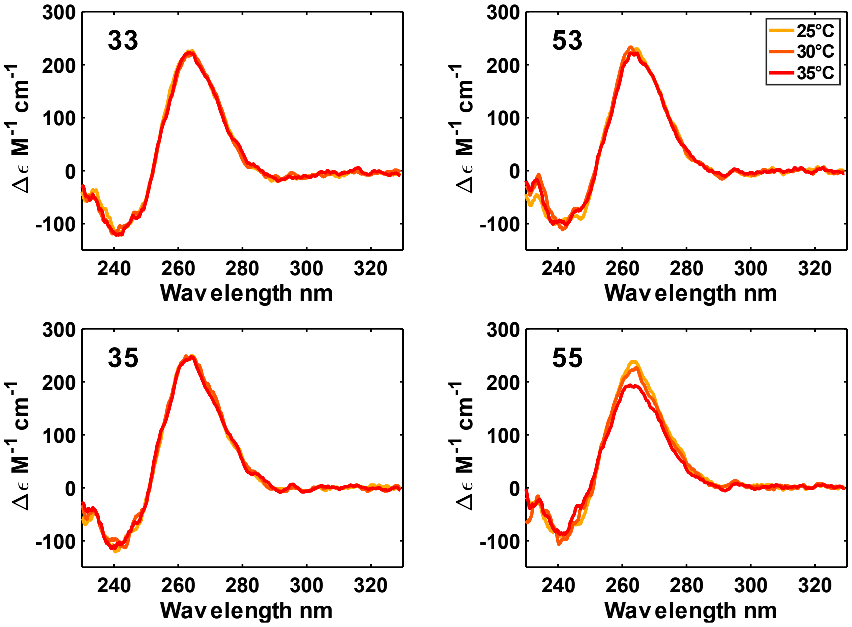


Figure 2.4: CD spectra for fully trapped mutants at three different temperatures (yellow - 25°C, orange - 30°C, red - 35°C).

Equilibrium melting temperatures (i.e. thermal stability) for the trapped mutants followed the order 55 < 35 < 53 < 33 (*Table 2.1*), which is consistent with equilibrium melting and differential scanning calorimetry experiments performed previously on fully trapped mutants of the slightly shorter c-MYC Pu18 G4[321]. In general, the melting transitions occurred at higher temperatures for the half-trapped mutants than for the fully-trapped ones, and the melting transitions for the WT occurred at the highest temperatures of all (*Table 2.3*). We have previously shown that the Pu18 WT G4 melts at a higher temperature than any of its fully-trapped mutants due to entropic stabilization of the folded state by GR exchange[321]. We expect the situation is similar here as well, with the half-trapped and WT molecules able to adopt 2 and 4 GR isomers, respectively, compared to the fully-trapped molecules which have only a single GR isomer available. Conformational entropy from GR exchange thus dictates that both the half-trapped and WT are more stable than the single most stable GR isomer within their conformational ensembles.

Visual inspection of the TH datasets indicates that all four potential pathways contribute to the effective folding rate of the Pu22 WT G4 folding. The temperature gap between the apparent melting points on the heating and cooling scans (ΔT_m) decreases as the rates of folding and unfolding increase (note that at the equilibrium T_m these rates are equal). A mutation that decreases the folding rate but does not affect the unfolding rate will lead to a lower apparent melting temperature, slower kinetics, and larger ΔT_m. Alternatively, a mutation that increases the unfolding rate will also lead to a lower apparent melting point, but with faster kinetics and a smaller ΔT_m (*Figure 2.5*). Here, the ΔT_m for the WT is smaller than those of the half-trapped mutants, which in turn are smaller than those of the fully trapped mutants (*Table 2.2*). This implies that the effective folding rate of the WT is greater than those of the half-trapped mutants, which are greater than those of the fully trapped mutants. Since each mutation reduces the number of folding pathways by a factor of 2, and in every case the apparent folding rate decreases, this implies that all four folding pathways makes non-negligible contributions to the overall folding rate of the WT.

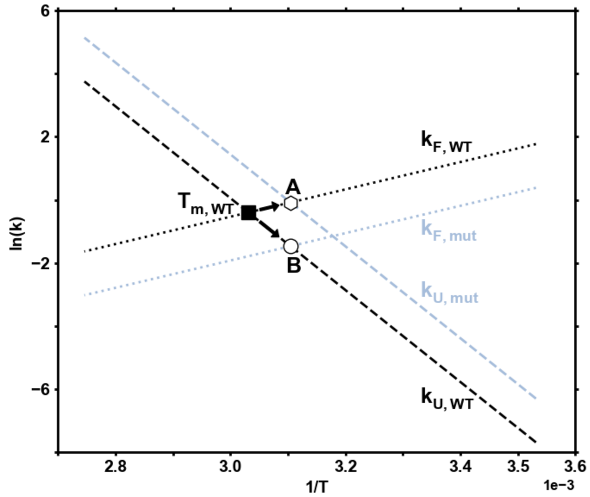


Figure 2.5: A hypothetical Arrhenius plot of G4 folding and unfolding rates. The activation energy for folding and unfolding are positive and negative, as observed experimentally. The intersection of the two lines is the equilibrium melting point (T_m) where k_U = k_F. A decrease in the folding rate (from black dotted to blue dotted line) or increase in the unfolding rate (from black dashed to blue dashed line) shifts the T_m to lower temperature. In the case of decreasing k_F, the resulting k_{ex} = k_U + k_F at the new T_m is smaller (point B) and more hysteresis is expected. In the case of increasing k_U, the resulting k_{ex} at the new T_m (point A) is larger and less hysteresis is expected.

G4	ΔT_m °C
c-MYC Pu22 WT	3.5 ± 0.2
3X	4.4 ± 0.2
5X	4.4 ± 0.2
X3	4.3 ± 0.2
X5	4.3 ± 0.2
33	4.8 ± 0.2
53	5.4 ± 0.2
35	5.0 ± 0.1
55	5.9 ± 0.1

Table 2.2: TH values (ΔT_m) as a model-free measure of parallel-pathway folding acceleration. TH values were calculated from the experimental datasets collected at 295 nm as the difference between the maxima of the first derivatives of the slowest heating and cooling scans for each G4, i.e. at the temperature where the G4 is 50% unfolded. A linear interpolation was used to estimate the 50% point between the nearest two flanking data points. Errors reported are the standard deviation of the values from three replicate TH datasets for each G4.

2.4.3 Globally fitting a parallel pathways folding model

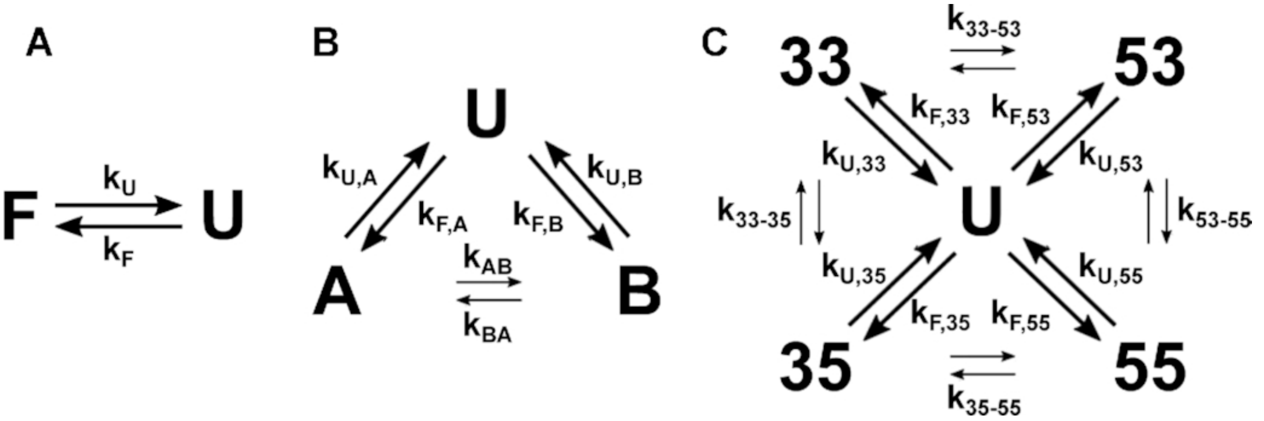
The qualitative analysis of ΔT_m values described above suggested that increasing numbers of folding pathways lead to increasing folding rates. In order to test this conclusion, we analyzed whether TH data for G4s with two and four possible pathways (half-trapped and WT) are quantitatively consistent with TH data for the fully trapped mutants, which approximate folding through each individual pathway. TH kinetic data have typically been analyzed assuming two-state folding behavior described by temperature-dependent folding k_F and unfolding k_U rate constants (*Scheme 2.1a*)[218]. Briefly, in this approach the spectroscopic thermal melting data are used to estimate the fraction of folded (θ_F) and unfolded molecules ($1 - \theta_F$) as a function of temperature, by applying appropriate folded and unfolded baselines. The shapes of the TH profiles are then given by the expressions:

$$\left(\frac{d\theta_F}{dT}\right)_{heating} = \left(k_F\left(1-(\theta_F)_{heating}\right) - k_U(\theta_F)_{heating}\right)\left(\frac{dt}{dT}\right)_{heating} \qquad \textit{(Equation 2.1)}$$

and

$$\left(\frac{d\theta_F}{dT}\right)_{cooling} = \left(k_F\left(1-(\theta_F)_{cooling}\right) - k_U(\theta_F)_{cooling}\right)\left(\frac{dt}{dT}\right)_{cooling}, \qquad \textit{(Equation 2.2)}$$

where the values, $(\theta_F)_{heating,cooling}$, and slopes, $(\frac{d\theta_F}{dT})_{heating,cooling}$, of the curves are obtained directly from the data and the inverses of the temperature ramp rates, $(\frac{dt}{dT})_{heating,cooling}$, are set by the user. At any given temperature, this gives a system of two equations and two unknowns (k_F, and k_U) which are obtained algebraically as a function of temperature. The temperature dependences of the rate constants can then be fitted to extract the activation energies from an Arrhenius plot. Examples of this type of classical analysis for the sequences studied here are shown in *Figure 2.6*.



Scheme 2.1: Folding mechanisms for the WT c-MYC Pu22 ensemble and trapped mutants.The fully trapped G4s adopt a single folded (F) conformation from the unfolded state (U). (B) The half-trapped G4s fold into two GR isomers (A and B) from the unfolded state, which can then slowly equilibrate by direct interconversion (indicated by small arrows)[216]*. (C) The WT c-MYC Pu22 ensemble primarily folds by directly adopting the four GR isomers from U in parallel, with slow GR exchange between isomers.*

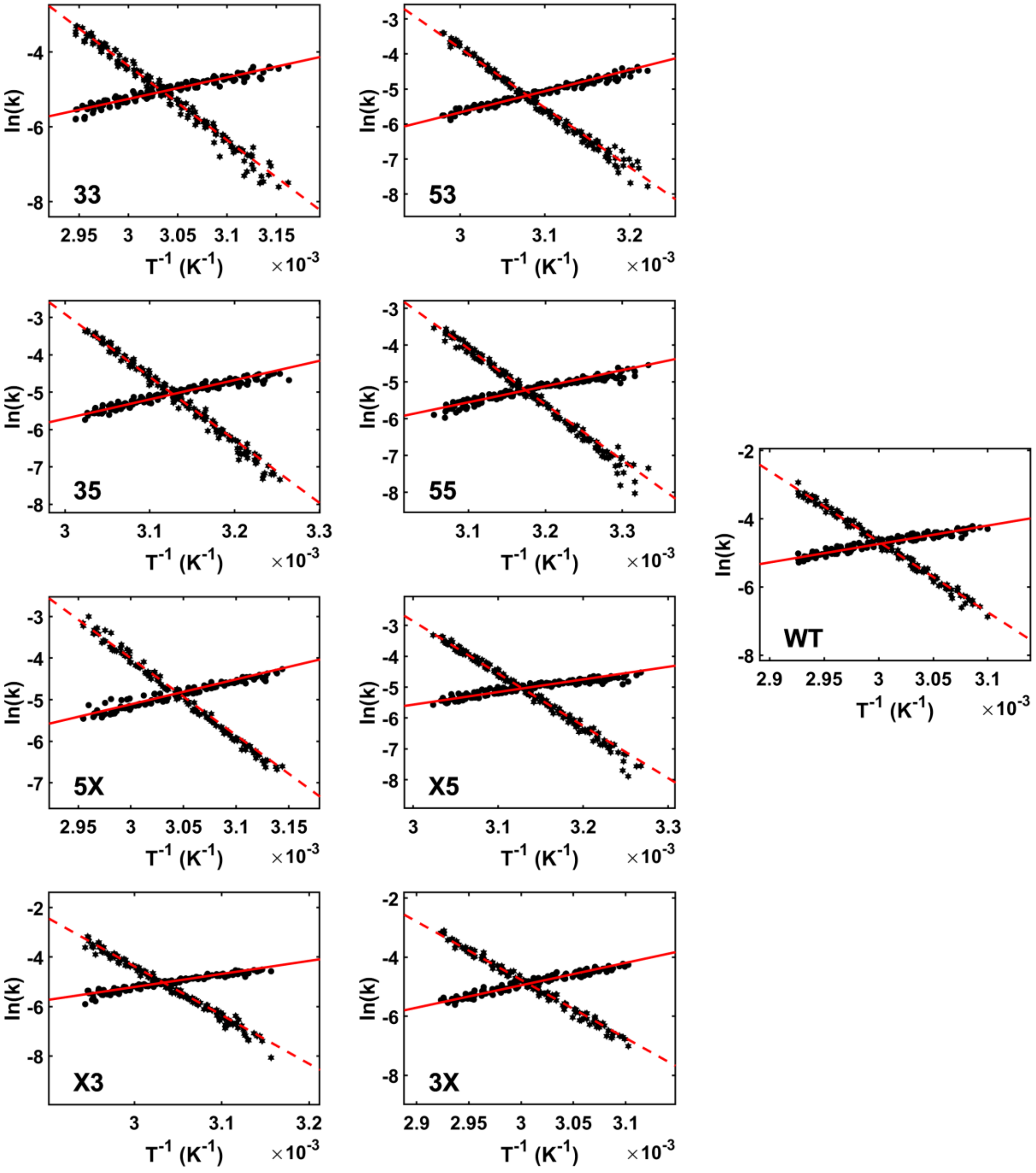


Figure 2.6: Arrhenius plots from classical two-state analysis of TH profiles. The folding (k_F) and unfolding (k_U) rate constants calculated from the experimental datasets are shown as circles and stars respectively, while the corresponding rate constants calculated from two-state fits to each dataset are shown as solid and dashed red lines respectively. Arrhenius plots calculated at all experimentally employed scan rates are overlaid. The equilibrium melting temperatures are at the intersections of the folding and unfolding lines.

The analysis of TH data for half-trapped and WT G4s quickly becomes more complicated, due to the existence of multiple GR isomers in the folded state ensemble, as visualized directly by NMR spectroscopy. For example, in the case of a half-trapped mutant whose folded state ensemble consists of the GR isomers A and B, the kinetics of the system is now described in terms of two folding rates, $k_{F,A}$ and $k_{F,B}$, two unfolding rates, $k_{U,A}$ and $k_{U,B}$, as well as the rates of GR exchange, k_{AB} and k_{BA}, for A→B and B→A, respectively (*Scheme 2.1B*). The fraction of folded molecules is the sum of fractions of A and B isomers, $\theta_F = \theta_A + \theta_B$, and the shape of the TH profile is given by

$$\frac{d\theta_F}{dT} = \left(\left(k_{F,A} + k_{F,B}\right)(1 - \theta_F) - k_{U,A}\theta_A - k_{U,B}\theta_B\right)\frac{dt}{dT}, \qquad \textit{(Equation 2.3)}$$

where the relative amounts of the A and B isomers obey the equations

$$\frac{d\theta_A}{dT} = \left(k_{F,A}(1 - \theta_F) + k_{BA}\theta_B - \left(k_{AB} + k_{U,A}\right)\theta_A\right)\frac{dt}{dT}, \qquad \textit{(Equation 2.4)}$$

$$\frac{d\theta_B}{dT} = \left(k_{F,B}(1 - \theta_F) + k_{AB}\theta_A - \left(k_{BA} + k_{U,B}\right)\theta_B\right)\frac{dt}{dT}, \qquad \textit{(Equation 2.5)}$$

and the principle of detailed balance [357] requires that

$$k_{BA} = k_{AB}\frac{k_{F,A}}{k_{U,A}} \cdot \frac{k_{U,B}}{k_{F,B}}. \qquad \textit{(Equation 2.6)}$$

In the case of the Pu22 WT G4, there are four folding rates, four unfolding rates, four GR exchange rates and 12 associated activation energies (*Scheme 2.1C*). Note that the half-trapped and WT datasets are very well fit by two-state folding models (*Figure 2.7*). The apparent folding rates extracted from these fits match the ΔT_m analysis above, with each half-trapped mutant generally folding faster than their two corresponding fully trapped mutants, and the WT generally folding faster than the half-trapped mutants (*Table 2.3*). However, this simple model does not capture the underlying equilibrium since the WT and half-trapped mutants contain multiple, slowly exchanging folded isomers. Clearly, data from the heating and cooling scans are insufficient to extract all of the half-trapped parameters algebraically, or even by nonlinear least squares fitting to a single dataset.

Although it is not possible to reliably extract all the relevant folding kinetic parameters for half-trapped and WT G4s from analyses of their individual TH datasets alone, we can still rigorously test whether their data are consistent with a parallel pathways folding mechanism by fitting the data for all sequence variants globally. We applied an extension of a global fitting approach we had developed previously to analyze the thermodynamics of GR exchange[210]. We simultaneously analyzed the data for the Pu22 G4 WT and all fully trapped and half-trapped mutants with the assumption that the kinetic parameters for any folding or unfolding transition in the WT or half-trapped mutant are equal to those of the corresponding fully trapped mutant, and that all rate constants obey an Arrhenius temperature dependence (see Materials and methods). This produces a global fit where kinetic parameters for 12 independent transitions are extracted from data for 9 different sequence variants, which compare favorably to the simple two-state case where kinetic parameters for two transitions (folding and unfolding) are extracted from data for one sequence variant. The global fit provides a quantitative test of the parallel folding pathways model. Large systematic deviations between calculated and experimental TH profiles in the global fit would indicate that either the half-trapped or WT G4s do not follow a parallel pathways mechanism as depicted in *Scheme 2.1C* and/or the fully-trapped G4 folding kinetics are not good measures of the individual pathways. However in actuality, we found that the parallel pathways model gave excellent simultaneous agreement with all nine datasets (*Figure 2.2*), providing validation for the model and justifying the use of fully-trapped mutants to mimic the individual GR isomers. The sum of squared residuals from the global fit is only ~3.8-fold greater than that obtained from fitting data for each sequence variant independently (*Figure 2.7*), even though the global fit contains 20 fewer adjustable parameters overall (16 versus 36 kinetic parameters in the global and independent fits, respectively) and therefore represents the simpler model. Furthermore, the global fit yields rate constants and activation enthalpies for all folding and unfolding transitions of the WT (listed in *Table 2.1*), providing some insight into how the WT Pu22 c-MYC G4 folds.

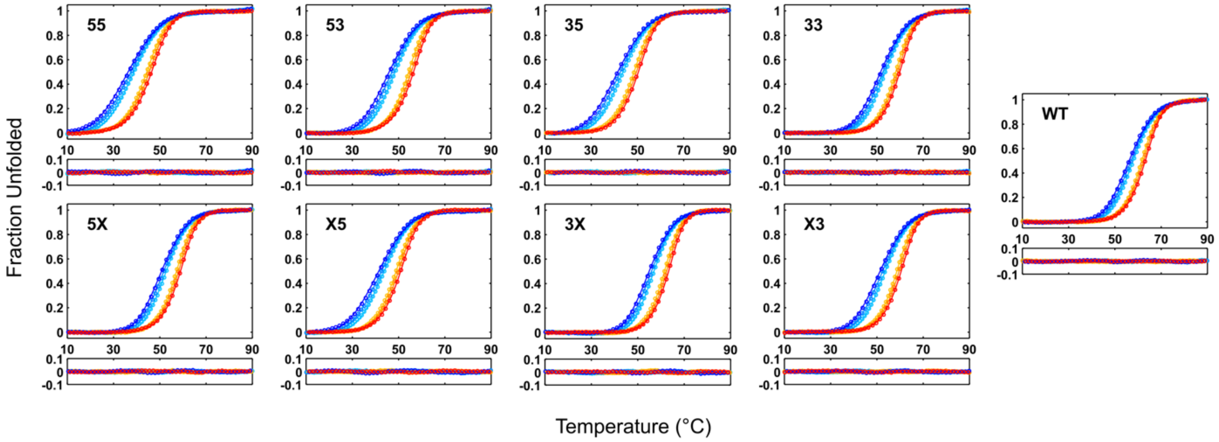


Figure 2.7: Independent two-state fits to TH data for the c-MYC Pu22 WT and trapped mutant G4s. Each TH profile was independently fit to a two-state kinetic model according to Scheme 2.1a in the main text (see Results and Materials and methods). Fit residuals are shown below each panel. The independent two-state fits here used a total of 36 kinetic parameters (4 each of the 9 independent fits), compared to the global fit (Figure 2.2) which only requires 16 kinetic parameters that are shared between all 9 datasets and is therefore the simpler kinetic model. The sum of the residual sum-of-squares (RSS) for the 9 independent fits is 2.17×10^{-5}, while the global fit RSS is 8.24×10^{-5}.

G4 sequence	**E_F kJ mol^{-1}**	**k_F min^{-1}**	**E_U kJ mol^{-1}**	**k_U min^{-1}**	**T_m °C**
WT	-45 ± 2	0.87 ± 0.01	172 ± 2	0.079 ± 0.002	59.9
55	-36 ± 1	0.227 ± 0.003	126 ± 1	1.09 ± 0.02	41.8
53	-50 ± 1	0.369 ± 0.002	141 ± 1	0.256 ± 0.002	51.7
35	-42 ± 5	0.32 ± 0.02	140 ± 5	0.67 ± 0.03	46.5
33	-47 ± 2	0.540 ± 0.006	165 ± 2	0.116 ± 0.002	56.4
5X	-49 ± 1	0.638 ± 0.004	154 ± 1	0.190 ± 0.002	55.3
X5	-34 ± 1	0.338 ± 0.004	141 ± 1	0.690 ± 0.007	46.3
3X	-63.1 ± 2	0.87 ± 0.01	164 ± 2	0.079 ± 0.002	59.4
X3	-43.9 ± 1	0.539 ± 0.005	164 ± 1	0.117 ± 0.002	56.5

Table 2.3: Kinetic parameters extracted from two-state fits reported at 50°C. Note that the reference temperature of 50°C was chosen because these fits are less well constrained than the global fit, so we have less confidence in the extrapolation to 37°C.

The rates of folding into each of the four GR isomers were similar at 37 °C, with only ~5-fold variation between the different folding pathways. In contrast, the unfolding rates differed by up to 15-fold between the most- and least-stable GR isomer. This suggests that the specific molecular interactions that give the GR isomers different stabilities are largely formed after the G4 folding transition state. All of the activation energies for folding were negative on the order of -35 to -55 kJ mol^{-1}, implying that enthalpically favorable interactions are made in the transition states for folding, as have been observed previously for other G4s[358]. The activation energies for unfolding were all large and positive (~120–170 kJ mol^{-1}) and followed an identical order to the stabilities of the isomers at 37 °C and their melting temperatures, i.e. 55 < 35 < 53 < 33. Unlike the folding and unfolding transitions, the rates of direct GR exchange (i.e. transitions from one GR isomer to another without unfolding) were not precisely defined by the data. However, we could place limits on how rapidly these transitions might occur. We performed global fits fixing all four direct GR exchange rates ($k_{ex} = k_{AB} + k_{BA}$) to values between 1×10^{-6} and 1×10^{5} min^{-1} while optimizing all other kinetic parameters (*Supplementary Figure 2.2*). We obtained optimal agreement with k_{ex} = 0.1 min^{-1}, i.e. on a similar timescale to the unfolding rate constants, and slightly worse fit qualities at k_{ex} below this value. Fixing k_{ex} to faster values resulted in substantially worse agreement with the data as evinced by larger residual sum of squared differences (RSS) between experimental and calculated values. This agrees well with a recent dynamic NMR study of GR exchange rates that found a Pu18 half-trapped mutant undergoes transitions between GR isomers on a timescale that is approximately 10-fold faster (k_{ex} ~2 × 10^{-2}min^{-1}) than global unfolding (k_U ~2 × 10^{-3} min^{-1}) at 25 °C[216]. Importantly, all k_F, k_U, E_F and E_U values were not at all sensitive to the particular value of k_{ex} (≤1 × 10^{3} min^{-1}), so all the values reported in *Table 2.1* are well-defined by the fits. We further tested the robustness of the extracted global fit folding rate constants by generating pairwise parameter correlation surfaces (See Materials and methods section, *Supplementary Figure 2.5*) that additionally show that the parameters are well-defined in the global fit. We note that this dataset was collected in 2 mM K^{+} to produce folding kinetics that was slow enough for TH and isothermal NMR folding measurements (see below). This is far below the biological K+ concentration of ~140 mM[359], and therefore we collected an additional TH dataset for the fully trapped mutants at 5 mM K+ to test whether these results held at higher K+ concentrations. We found a strong linear correlation between rate constants in 2 and 5 mM K+, with approximately 2- to 3-fold faster folding and slower unfolding, respectively, in 5 mM K+ at 37 °C (*Figure 2.8*). This suggests that the folding rates of the GR isomers in the Pu22 WT ensemble scale similarly with respect to K+ concentration, and the results of our model can be extrapolated to higher, more biologically relevant salt conditions. The extracted kinetic parameters from either the 2 or 5 mM K+ TH datasets reveal the extent to which parallel reaction pathways accelerate folding of the Pu22 WT G4. The effective rate constant for the transition from the unfolded state to the ensemble of folded conformations is given by the sum of folding rate constants for the four pathways leading to the four GR isomers,

$$k_{F,WT} = k_{F,33} + k_{F,35} + k_{F,53} + k_{F,55} \qquad \textit{(Equation 2.7)}$$

giving a value of $k_{F,WT}$ = 3.1 min^{-1} in 2 mM K+. For comparison, the rate constant for the fastest folding and most stable GR isomer (33) is $k_{F,33}$ = 1.2 min^{-1}, meaning that folding is accelerated by at least a factor of 2.5 due to the presence of multiple pathways. It is worth noting that compared to the average rate constant, folding acceleration is always equal to the number of pathways, four in this case. In other words,

$$\frac{k_{F,WT}}{\langle k_{F,33,35,53,55} \rangle} = 4 \qquad \textit{(Equation 2.8)}$$

where angled brackets indicate the mean value.

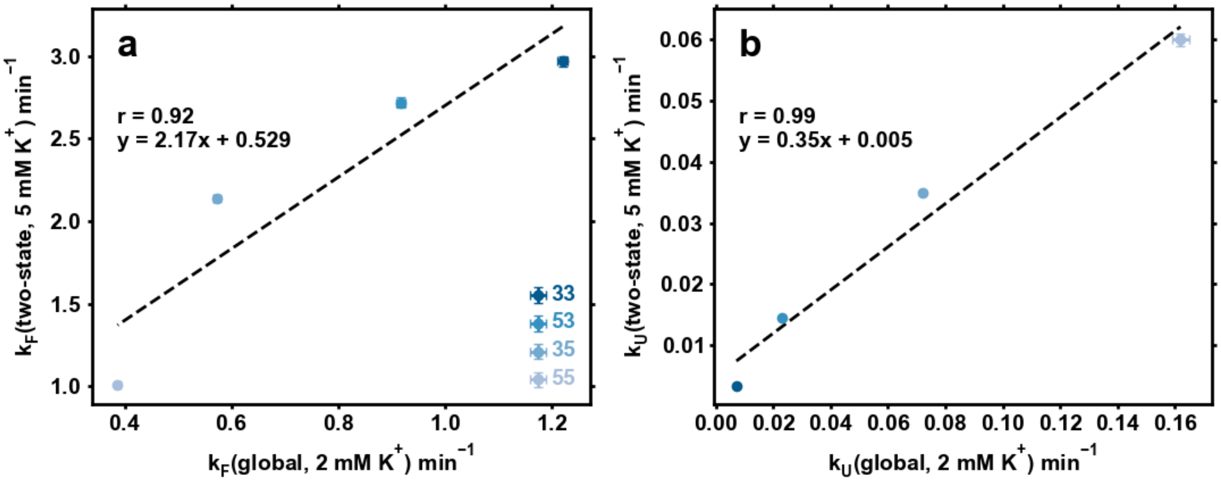


Figure 2.8: Correlation of rates for the fully-trapped c-MYC Pu22 mutants in the presence of 2 and 5 mM K$^+$ *at 37 °C. Folding (a) and unfolding (b) rates in 2 and 5 mM K*$^+$*. Individual two-state fits were performed on the fully-trapped G4 datasets in 5 mM K*$^+$ *and compared with results of the global fit performed on all data at 2 mM K*$^+$*.*

The different folding and unfolding rates for the four reaction pathways imply the existence of some interesting nonequilibrium behavior when an unfolded Pu22 WT chain is allowed to fold under ambient conditions. We performed a numerical simulation of the folding reaction using the measured kinetic constants (see Materials and methods) according to *Scheme 1C* with k_{ex} = 0 and the optimal value of 0.1 min^{-1}, and plotted the relative amounts of each GR isomer as a function of time (*Figure 2.9*). Interestingly, the populations of the 35 and 55 isomers built up relatively quickly to reach maxima near ~1 min (*Figure 2.9a*) and then decreased by up to a factor of 10 at longer time points (*Figure 2.9b*). At very long times, the relative populations are thermodynamically controlled and the folded ensemble consists of 77%, 18%, 4% and 1% of the 33, 53, 35 and 55 GR isomers, respectively, in agreement with our previous studies of c-MYC G4s[210]. In contrast, at the early stages of folding, populations are determined by the rate rather than the equilibrium constants of folding and the ensemble consists of approximately 40%, 30%, 18% and 12% of the 33, 53, 35 and 55 GR isomers, respectively, which explains the transient buildup of the less stable 35 and 55 forms. Similar results were obtained for simulations including direct GR isomer exchange at the optimal value of k_{ex} (*Figure 2.9*), with only slightly more rapid equilibration of the folded ensemble. Thus, the composition of the Pu22 G4 structural ensemble is different at short and long time-points following the initiation of folding. Furthermore, the effective rate constant for unfolding depends on the instantaneous populations of GR isomers according to

$$k_{U,WT} = k_{U,33}\theta_{33} + k_{U,53}\theta_{53} + k_{U,35}\theta_{35} + k_{U,55}\theta_{55}. \qquad \textit{(Equation 2.9)}$$

Initially, $k_{U,WT} \approx 0.043$ min^{-1} while at very long times, $k_{U,WT} = 0.014$ min^{-1}. Thus the average lifetime of a folded G4 (=1/$k_{U,WT}$) is about 20 minutes in the ensemble of conformations that forms initially, but increases to over an hour as the populations of more stable GR isomers are enriched with time.

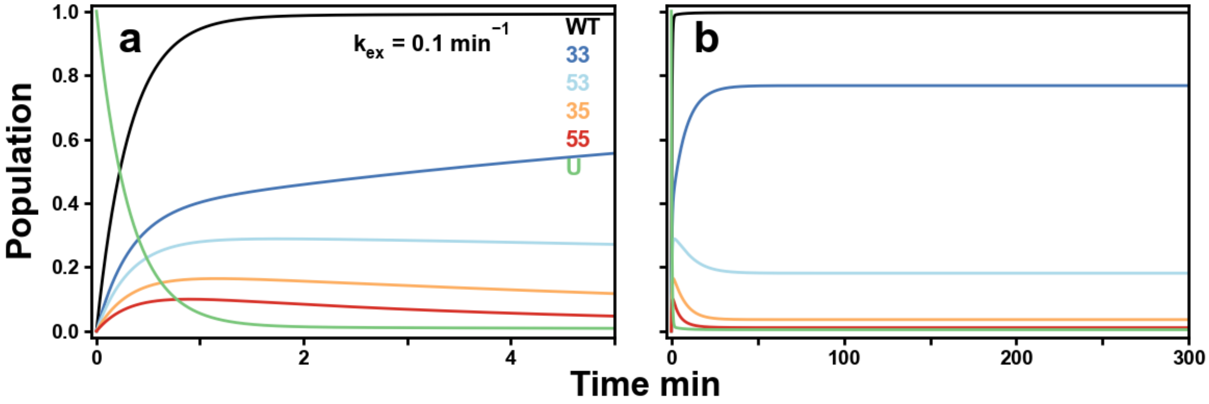


Figure 2.9: Isothermal folding simulations for the c-MYC Pu22 WT ensemble including direct GR isomer interconversion at the optimal rate constant k_{ex} = 0.1 min^{-1}. Simulations were performed according to Scheme 2.1c and the parameters from Table 1 in the main text (see Supplementary Methods). Short and long timescales are shown in (a) and (b).

2.5 Discussion

A key question in this research is the extent to which conformational heterogeneity in the folded state implies the existence of parallel pathways that accelerate effective folding rates. Specifically in this case, does the existence of multiple GR isomers in the Pu22 WT folded ensemble lead to more rapid adoption of a folded state, starting from an unfolded chain? If the commitment to a particular GR isomer occurs after the transition state for G4 folding, then the answer is likely no. The folding rate would depend only on the height of a single dominant barrier, regardless of the number of possible GR isomers existing on the far side, which would be separated from each other by smaller energy barriers. In contrast, if the commitment to a given GR isomer occurs early in the folding process then there is a separate transition state for each GR isomer. Each pathway corresponds to a separate folding pathway, whose rates sum to produce a net acceleration. There are several lines of evidence suggesting that the second scenario applies to the G4s studied here. First, the exchange between different GR isomers occurs quite slowly. Rates are on the same timescale and enthalpic barriers are similar to those for complete unfolding of the G4[216]. If commitment to a given G-register occurred after the transition state, one would expect GR isomers to be separated by energy barriers smaller than the one separating the folded and unfolded states, which is not what is observed. Second, model-free ΔT_{m} values and effective two-state folding rates extracted from our TH data indicated that fully-trapped G4s with a single GR isomer generally fold more slowly than half-trapped mutants with two GR isomers, which largely fold more slowly than the WT with four. This is consistent with the idea that increasing the number of possible GR isomers accelerates folding. Finally, the global analysis verifies that the folding rates of the Pu22 WT and half-trapped mutants can be quantitatively accounted for in terms of parallel folding pathways, where the rate of each pathway is given by the folding rate of the corresponding fully trapped mutant.

The existence of parallel folding pathways in protein energy landscapes has long been recognized, both theoretically and experimentally[360-365], and pathway multiplicity has been directly linked to the overall folding rate. For example, Aksel and Barrick studied consensus ankyrin repeat proteins (CARPs), where the number of distinct folding pathways is equal to the number of repeats[348]. They characterized the folding kinetics of CARPs of different lengths and observed that the folding rate constant increased in proportion to the number repeats present, directly demonstrating acceleration due to multiple folding pathways. The folding mechanisms of nucleic acids have been less intensively studied than those of proteins; however, there are examples of molecules with multiple folding reaction pathways similar to those observed here. For example, in a study of the vertebrate telomeric i-motif sequence, Lieblein *et al.* observed formation of the thermodynamically more stable 5′E intercalation topology through a rapidly folding and slowly unfolding 3′E topology intermediate[69]. Bessi *et al.* studied a telomeric G4 sequence, which initially formed two different topologies with similar rate constants[215].

The initial folding reaction was followed by a period of several days over which the population ratio shifted from about 1:1 to 4:1 for the two isomers. Similarly to what was seen here, the effective overall folding rate for the telomere G4 is about twice as large as for either of the individual isomers, and the effective unfolding rate decreases by about a factor of two as the populations slowly equilibrate.

One unique aspect of systems undergoing GR exchange is that the existence of parallel folding pathways can be inferred directly from the nucleotide sequence, unlike other nucleic acids where multiple pathways are only revealed by detailed biophysical analyses. In principle, an estimate of the number of GR isomers is given by[210]

$$R_T = \prod_{i=1}^{4}(n_i - 2) \qquad \textit{(Equation 2.10)}$$

where n_{i} is the number of dG residues in the i_{th} G-tract and a G4 core of three G-tetrads is assumed. When a G4 folds with only two G-tetrads[310], even more GR isomers become available to the sequence, since there would then be n - 1 possible alignments for a given G-tract with respect to the other G-tracts that form the core as opposed to n - 2 in *Equation 2.10*. We note that this equation does not account for additional conformational heterogeneity that may exist in addition to, or superposed on GR exchange. In principle, each GR isomer can have a different topology, or populate multiple alternate topologies, as we have previously shown for the PIM1 and hTERT promoter G4s[210, 216]. Some G4s form bulges where internal Gs are extruded from the core and the flanking Gs shift inward to fill these gaps[308]. As well, some DNA sequences include a fifth, or 'spare tire' G-tract[366], with G4s forming from different subsets of four of the five G-tracts[367, 368]. Additional conformations present in the folded ensemble would likely lead to additional parallel folding pathways that could further accelerate folding. If we make the very rough approximation that the folding rate of each pathway is comparable, then the total acceleration factor due to the existence of parallel pathways is simply equal to the number of pathways (similar to *Equation 2.8*). *Equation 2.10* gives a conservative estimate for the number of GR isomers that can be formed by a given DNA sequence and the associated acceleration factor for G4 folding. We previously analyzed a database of human gene promoter regions for potential G4-forming sequences that could undergo GR exchange, according to *Equation 2.10*. Of roughly 28 000 putative G4s, nearly 20%

could adopt 12 or more GR isomers and 5% could form >20[210]. Assuming that folding rates to the individual GR isomers are relatively similar (as they are for Pu22) many of these G4s could fold more than an order of magnitude faster than they would if they only adopted a single GR isomer.

There is an important distinction to be made between parallel folding pathways and misfolding. In the Pu22 G4, as well as the telomeric G4 and i-motif mentioned above, some or most of the folding pathways lead to conformational isomers that are not highly populated at equilibrium. These rarer states are formed transiently after folding is initiated, but essentially disappear at long time-points. This begs the question of whether their formation represents folding acceleration or rather kinetic trapping by transiently misfolded species. The best answer to this question lies in the details of biological function. In many cases, it is believed that G4s exert their biological effects simply by virtue of being folded. For instance, folded G4s are proposed to act as physical obstacles that impede the procession of DNA and RNA polymerases[316]. Alternatively, G4 folding in the nontemplate strand of DNA could reduce its ability to displace RNA:DNA hybrids formed after RNA synthesis, leading to stalling of the polymerase[317]. Similarly, G4 folding in both open reading frames and untranslated regions of mRNAs can reduce translation and influence ribosomal frameshifting and co-translational protein folding[327]. In all of these examples, it would be expected that folding to both thermodynamically favored and disfavored states would have similar biological consequences. Thus, reaction pathways to rare GR isomers can be considered to legitimately accelerate folding. In other cases, a greater degree of structural specificity might be expected. There are examples of transcription factors and other proteins that recognize the G4 fold[170]. Nucleolin, which binds G4s in the c-MYC and VEGF gene promoter regions[369], binds different c-MYC GR isomers with different affinities[324]. Similarly, it has been reported that different helicases are active on different G4 folds[370]. Nevertheless, the ability of G4s to influence transcription, translation and replication simply by folding suggests that, for the most part, a reaction pathway leading to any member of the folded conformational ensemble contributes to the activity of a functional G4.

The relationships between G4 sequence, folding kinetics and biological function are still in the very early stages of being uncovered. Many of the relevant processes are highly complex and orchestrated through the coordinated activity of diverse molecular machines. Furthermore, G4 stability is highly sensitive to the concentrations of K^+ and Na^+ ions, molecular crowding, and mechanical stress (e.g. helical over- or under-winding)[52, 371, 372] and folding rates are likely to be similarly affected[373]. This study was performed on short, homogeneous oligonucleotides, at a concentration of K^+ lower than that found intracellularly (2 mM versus ~140 mM), and in the absence of crowding agents, so the absolute rates of folding are different (and likely much slower) than those occurring *in vivo*. Nevertheless, the principle of folding acceleration by parallel reaction pathways applies universally, since the rates of all pathways likely scale similarly as solution conditions vary, as we have shown in *Figure 2.8*. The combination of mutagenesis, TH, and global fitting that we have developed provides a rapid and inexpensive of way of mapping folding landscapes with parallel pathways, and requires only that folding occurs slowly enough to produce thermal hysteresis and that individual folded isomers can be trapped with nucleotide substitutions. Parallel G4 folding pathways associated with GR isomerization are likely common in nature, can be easily identified by sequence analysis, and lend themselves to being characterized by this approach. These techniques thus provide an avenue toward a better understanding of the complex dynamics underlying G4 function.

2.6 Materials and methods

2.6.1 Sample Preparation

Oligonucleotide samples were purchased from the Yale Keck Oligonucleotide Synthesis facility (Yale University, USA) or were synthesized using a MerMade 12 oligonucleotide synthesizer with standard solid-phase phosphoramidite chemistry. Samples for TH measurements were subjected to cartridge purification and analyzed by LC-mass spectrometry for purity. DNA strands were dissolved in MilliQ water and concentrations were calculated using nearest neighbour extinction coefficients[374]. Prior to usage, the DNAs were HPLC purified and desalted. All experiments were performed in TH buffer: 10 mM lithium phosphate, pH 7.0, supplemented with 2 mM KCl. The buffer pH was titrated using 1 M LiOH to avoid the further addition of stabilizing Na^+ or K^+

cations. Our reaction conditions employed 2 mM K^+ ions, since higher concentrations (physiological is approximately 140 mM[359]) lead to much higher melting temperatures with undetectably small amounts of hysteresis and short or nonexistent unfolded baselines.

2.6.2 Circular dichroism spectroscopy

CD experiments were performed using a JASCO J-810 (JASCO, USA) spectropolarimeter with a cell path length of 0.1 cm. Spectra were recorded with 10 µM samples and at temperatures of at 25, 30, and 35 °C. The samples were scanned three times from 330 to 230 nm for signal averaging. The CD spectra were baseline corrected using a buffer blank.

2.6.3 Thermal hysteresis measurements

TH datasets were collected using a Cary Win-UV spectrophotometer (Agilent Technologies, USA) and cuvettes with a 1 cm path length. Absorbance profiles were measured as a function of temperature at 260 and 295 nm over the interval of 5–90 °C with 10 µM samples. Scan rates of ±2, ±3 and ±4 °C min^{-1} were chosen to induce TH between the heating and cooling scans. The samples were equilibrated at high and low temperatures for 5 min. A layer of mineral oil was placed on top of the sample solution in the cuvettes to mitigate evaporation. Measurements were repeated in triplicate yielding a total of 27 datasets (3 replicates for each of the 9 G4 sequences studied herein). We applied baseline and temperature corrections to all datasets as described previously[246] to account for deviations between the true solution temperature and the block temperature reported by the instrument. TH datasets were collected in both 1 and 0.1 cm cuvettes to assess the influence of heat transfer in the two cell volumes at the rapid scanning rates employed here (*Supplementary Figure 2.3*). We measured nearly identical TH datasets using either cuvette, implying that heat transfer throughout the cell volume is efficient in both cases, even at very fast scanning rates. TH measurements collected in 1 cm cuvettes represent the primary dataset used below.

2.6.4 Thermal hysteresis global fitting

TH datasets for the c-MYC Pu22 WT and trapped mutants were globally fit with a kinetic model that assumes the temperature dependences of the folding and unfolding rate constants obey the Arrhenius equation

$$k(T) = k_0 e^{\frac{E_a}{R}\left(\frac{1}{T_{ref}} - \frac{1}{T}\right)} \qquad \textit{(Equation 2.11)}$$

where k_0 is the rate constant at the reference temperature T_0, E_a is the activation energy, and R is the ideal gas constant. The fully- and half-trapped mutant profiles were fit simultaneously with the wild-type dataset using the following rate equations

Fully-trapped

$$\frac{d}{dT}[33] = \left(k_{F,33}[U] - k_{U,33}[33]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.12)}$$

$$\frac{d}{dT}[35] = \left(k_{F,35}[U] - k_{U,35}[35]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.13)}$$

$$\frac{d}{dT}[53] = \left(k_{F,53}[U] - k_{U,53}[53]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.14)}$$

$$\frac{d}{dT}[55] = \left(k_{F,55}[U] - k_{U,55}[55]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.15)}$$

Half-trapped

$$\frac{d}{dT}[X3] = \left(\left(k_{F,33} + k_{F,53}\right)[U] - k_{U,33}[33] - k_{U,53}[53]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.16)}$$

$$\frac{d}{dT}[X5] = \left(\left(k_{F,35} + k_{F,55}\right)[U] - k_{U,35}[35] - k_{U,55}[55]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.17)}$$

$$\frac{d}{dT}[3X] = \left(\left(k_{F,33} + k_{F,35}\right)[U] - k_{U,35}[35] - k_{U,33}[33]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.18)}$$

$$\frac{d}{dT}[X5] = \left(\left(k_{F,35} + k_{F,55}\right)[U] - k_{U,35}[35] - k_{U,55}[55]\right)\frac{dt}{dT} \qquad \textit{(Equation 2.19)}$$

Pu22 WT

$$\frac{d}{dT}[WT] = \begin{pmatrix} \left(k_{F,33} + k_{F,53} + k_{F,35} + k_{F,55}\right)[U] \\ -k_{U,33}[33] - k_{U,53}[53] - k_{U,35}[35] - k_{U,55}[55] \end{pmatrix}\frac{dt}{dT} \qquad \textit{(Equation 2.20)}$$

where dt/dT is the inverse experimental scan rate in min °C^{-1}. We also performed fits and simulations where the rate equations were modified to include direct GR isomer exchange terms (see main text Results) according to

$$\frac{d}{dT}[33] = \left(k_{F,33}[U] - k_{U,33}[33] - (k_{33-53} + k_{33-35})[33] + k_{53-33}[53] + k_{35-33}[35]\right) \qquad \textit{(Equation 2.21)}$$

The interconversion rates were set by invoking the principle of detailed balance

$$\frac{[33]}{[U]}\frac{[53]}{[33]}\frac{[U]}{[53]} = \frac{k_{F,33}}{k_{U,33}}\frac{k_{33-53}}{k_{53-33}}\frac{k_{U,53}}{k_{F,53}} = 1 \qquad \textit{(Equation 2.22)}$$

and defining

$$k_{ex} = k_{33-53} + k_{53-33} \qquad \textit{(Equation 2.23)}$$

from which it can be shown that

$$k_{33-53} = k_{ex}\frac{K_{F,53}}{K_{F,53}+K_{F,33}} \qquad \textit{(Equation 2.24)}$$

where $K_{F,53} = \frac{k_{F,53}}{k_{U,53}}$ and $K_{F,33} = \frac{k_{F,33}}{k_{U,33}}$ are the equilibrium constants for folding into the 33- and 53-shifted GR isomers. The reverse rate constant k_{53-33} is then determined by *Equation 2.24*). We have shown the case for the 33 trapped mutant as an example, though similar equations hold for the other trapped mutants by substitution of the appropriate rate and equilibrium constants.

The set of folding rate equations for the WT and trapped mutants was numerically integrated as a function of temperature using the ordinary differential equation (ODE) solvers in MATLAB. Absorbance TH profiles were calculated from the numerically integrated concentrations according to

$$A(T) = A_F(T)\theta_F(T) + A_U(T)\theta_U(T) \qquad \textit{(Equation 2.25)}$$

where $\theta_F(T)$ and $\theta_U(T)$ are the fractions of the folded and unfolded states respectively, and $A_F(T)$ and $A_U(T)$ are the linear folded and unfolded absorbance baselines. Fractions of the folded state were given by

Fully-trapped

$$\theta_{F,33}(T) = \frac{[33]}{C_T} \qquad \textit{(Equation 2.26)}$$

Half-trapped

$$\theta_{F,3X}(T) = \frac{[33]+[35]}{C_T} \qquad \textit{(Equation 2.27)}$$

c-MYC Pu22 WT

$$\theta_{F,WT}(T) = \frac{[33]+[35]+[53]+[55]}{C_T} \qquad \textit{(Equation 2.28)}$$

where C_T is the total experimental concentration and for brevity, we have shown only the calculations for the 33 and 3X trapped mutants. The fraction of the unfolded state was given by

$$\theta_U(T) = \frac{[U]}{C_T} = 1 - \theta_F(T). \qquad \textit{(Equation 2.29)}$$

The linear folded and unfolded absorbance baselines were calculated using

$$A_F(T) = m_F T + b_F \qquad \textit{(Equation 2.30)}$$

$$A_U(T) = m_U T + b_U \qquad \textit{(Equation 2.31)}$$

where m_F, m_U, b_F, b_U are the slopes and intercepts for the folded and unfolded baselines respectively (*Supplementary Figure 2.4*). All 27 TH datasets (triplicate measurements for each of the WT and 8 trapped mutant G4s) were fit by minimizing the residual sum-of-squared (RSS) differences between the experimental and modelled data

$$RSS = \sum_{n=1}^{27} \sum_k \left(A_{exp,n}(T_k) - A_{model,n}(T_k, \xi)\right)^2 \qquad \textit{(Equation 2.32)}$$

where T_k is the kth experimental temperature, A_{exp} and A_{model} are the experimental and fitted TH data respectively, and ξ is the set of folding and unfolding rate constants and activation energies for the four GR isomers. Optimized fraction unfolded profiles from the global fits are displayed in the main text.

2.6.5 Isothermal parallel pathway folding simulations

Isothermal folding simulations were performed by numerical integration of the time-dependent rate equations for the unfolded state, wild-type, and trapped mutants, (*Equation 2.11* to *Equation 2.29* and *Scheme 2.1c* with k_{ex} = 0 or 0.1 min $^{-1}$) with the optimized folding parameters from the TH global fitting analysis (Table 1, 295 nm) at the reference temperature of 37 °C. The simulations were initiated from an initial condition θ_U = 1 and all other populations set to 0.

2.6.6 Monte Carlo errors

Experimental errors for the TH data were calculated by simulating 500 experiments with random error equal to that of the experimental data for the 9 G4 sequences studied herein. The resulting synthetic datasets were globally fit in an identical fashion to the experimental data according to Scheme 2.1 in the main text and the methods, and the errors in Table 1 were reported as one standard deviation of the 500 extracted parameter sets.

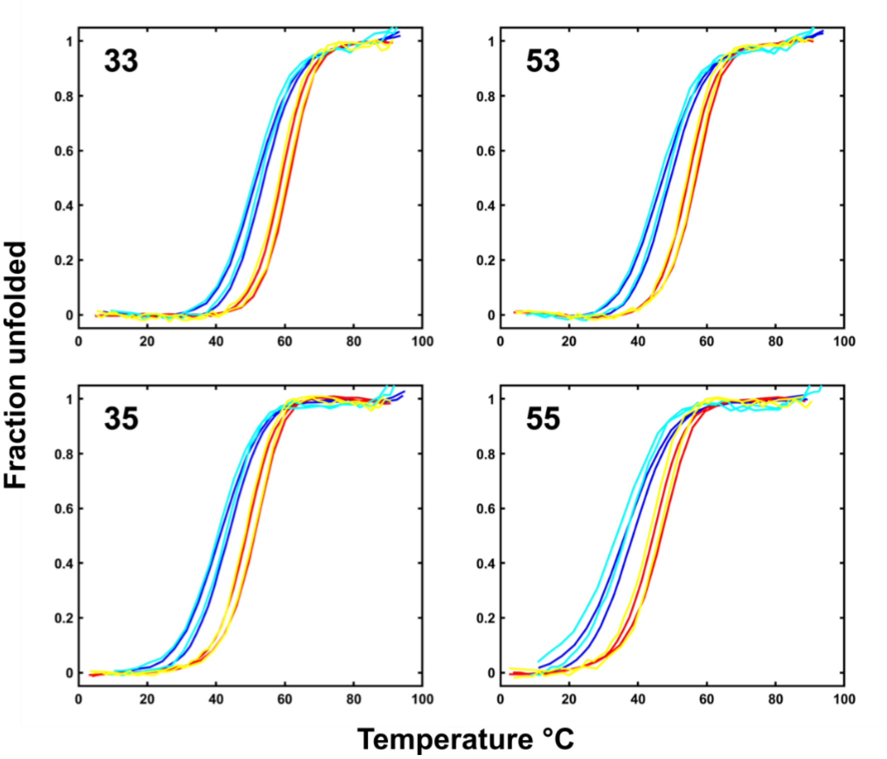
2.6.7 Residual sum of squares parameter correlation contour plots

Contour plots were created by performing a grid search for each pair of fitted on rates for each fit and calculating the RSS at each point. The confidence level (CL) at each RSS value was calculated as[375]

$$CL_{i,j} = F_{CDF}\left(\left(\frac{RSS_{i,j}}{RSS_{min}} - 1\right) * \left(\frac{Dof}{M}\right), M, DoF\right) \qquad \textit{(Equation 2.33)}$$

Where F_{CDF} is the F distribution cumulative density function and RSS_{min} is the residual sum of squares at the minimum found via global fitting. $RSS_{i,j}$ is the RSS at the point i,j for each of the i^{th},j^{th} pair of fitted parameters. DOF is the degrees of freedom of the fit (equal to the number of points – M), and M is the total number of fitted parameters. The resulting %CL contour plots are shown in *Figure 2.1*.

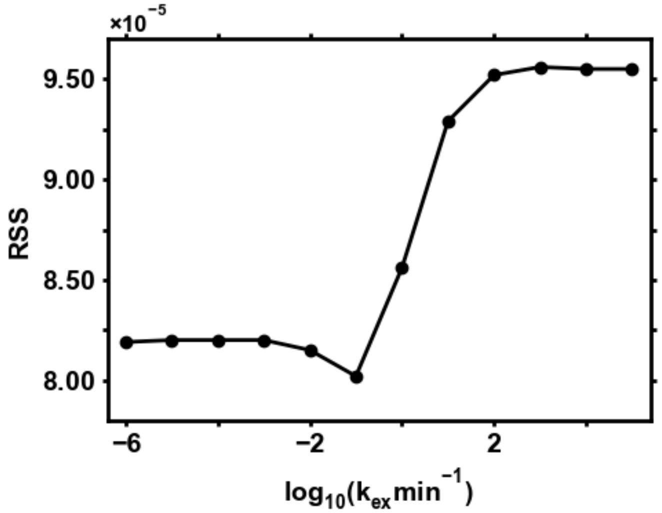
2.7 Supplementary Information



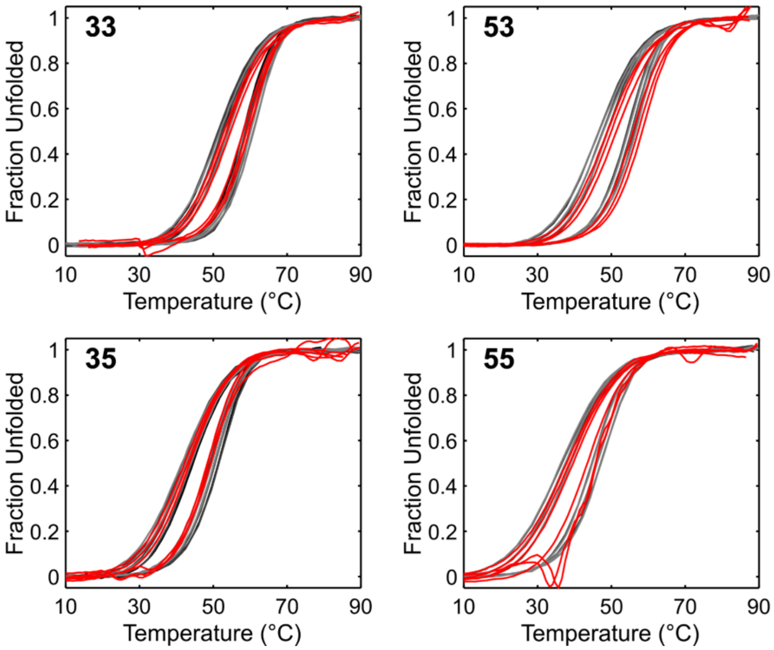
Supplementary Figure 2.1: Overlays of baseline-corrected experimental TH data for the fully-trapped mutants collected at 260 and 295 nm. The 295 nm dataset for each mutant is shown as dark blue (cooling) and red (heating) lines, while the corresponding 260 nm dataset is shown as light blue (cooling) and yellow lines (heating). Only the slowest and fastest heating and cooling scans are shown here for clarity (±2 and ±4 °C min^{-1} respectively).

GR isomer	E_F kJ mol^{-1}	k_F min^{-1}	E_U kJ mol^{-1}	k_U min^{-1}
55	-36 ± 2	$(290 \pm 4)\times10^{-3}$	129 ± 2	$(162 \pm 3)\times10^{-3}$
35	-52 ± 3	$(520 \pm 10)\times10^{-3}$	139 ± 3	$(73 \pm 3)\times10^{-3}$
53	-57 ± 3	$(890 \pm 40)\ \times10^{-3}$	149 ± 3	$(23 \pm 1)\times10^{-3}$
33	-60 ± 3	$(1200 \pm 70)\times10^{-3}$	167 ± 3	$(7 \pm 0.6)\times10^{-3}$

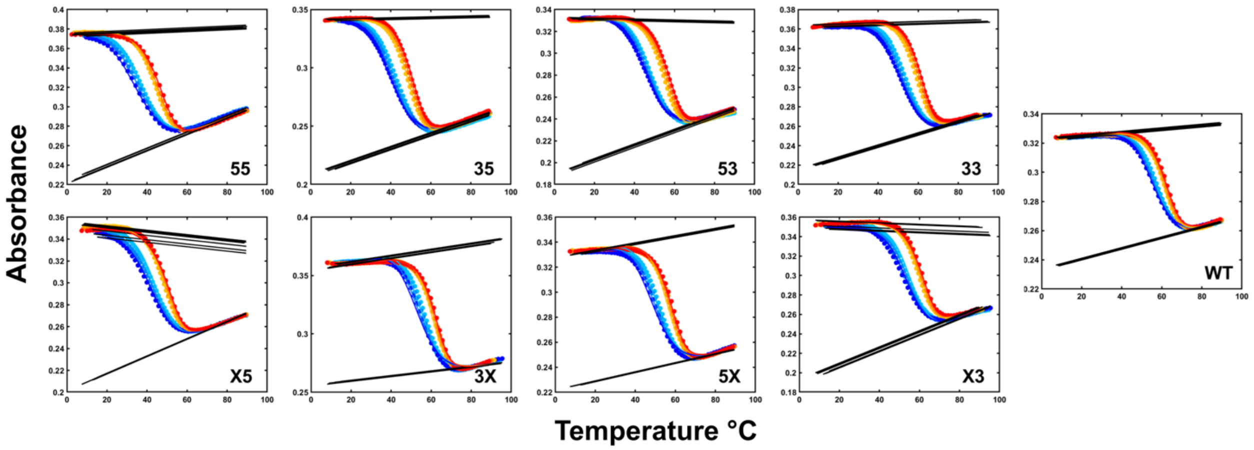
Supplementary Table 2.14: Parameters extracted from global fits of the c-MYC Pu22 WT and trapped mutant TH datasets at 260 nm. Fit to Scheme 2.1c in the main text with k_{ex} = 0 min^{-1}. Rate constants are reported at 37 °C. Errors were calculated according to the covariance matrix approach[376].



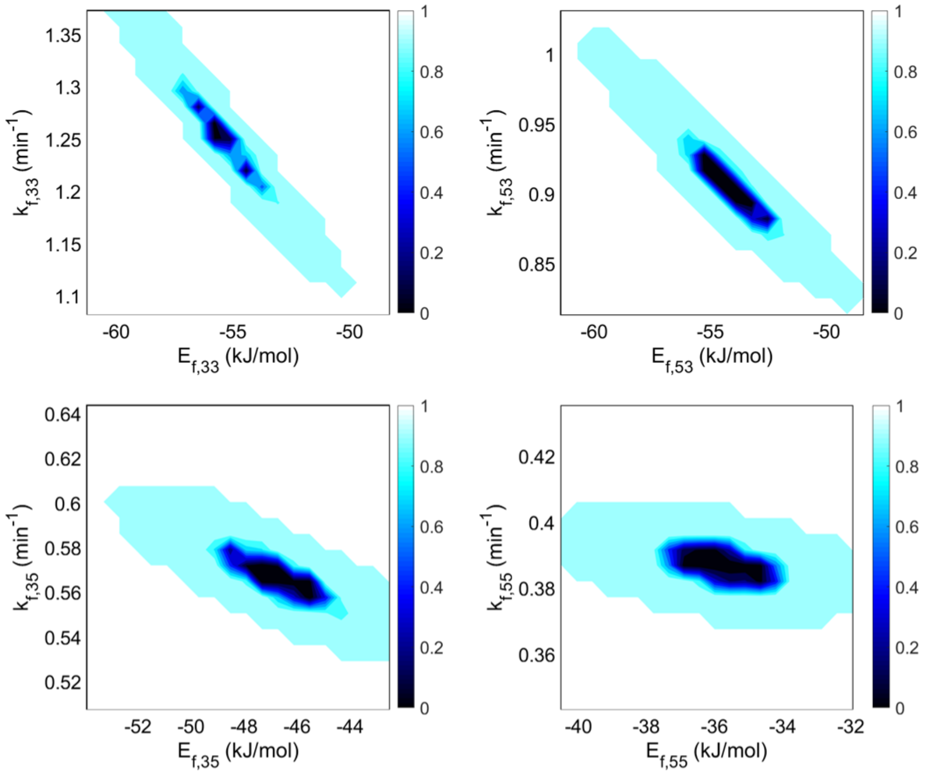
Supplementary Figure 2.2: TH global fit residual sum-of-squares (RSS, 295 nm) as a function of direct GR isomer interconversion rate k_{ex}.

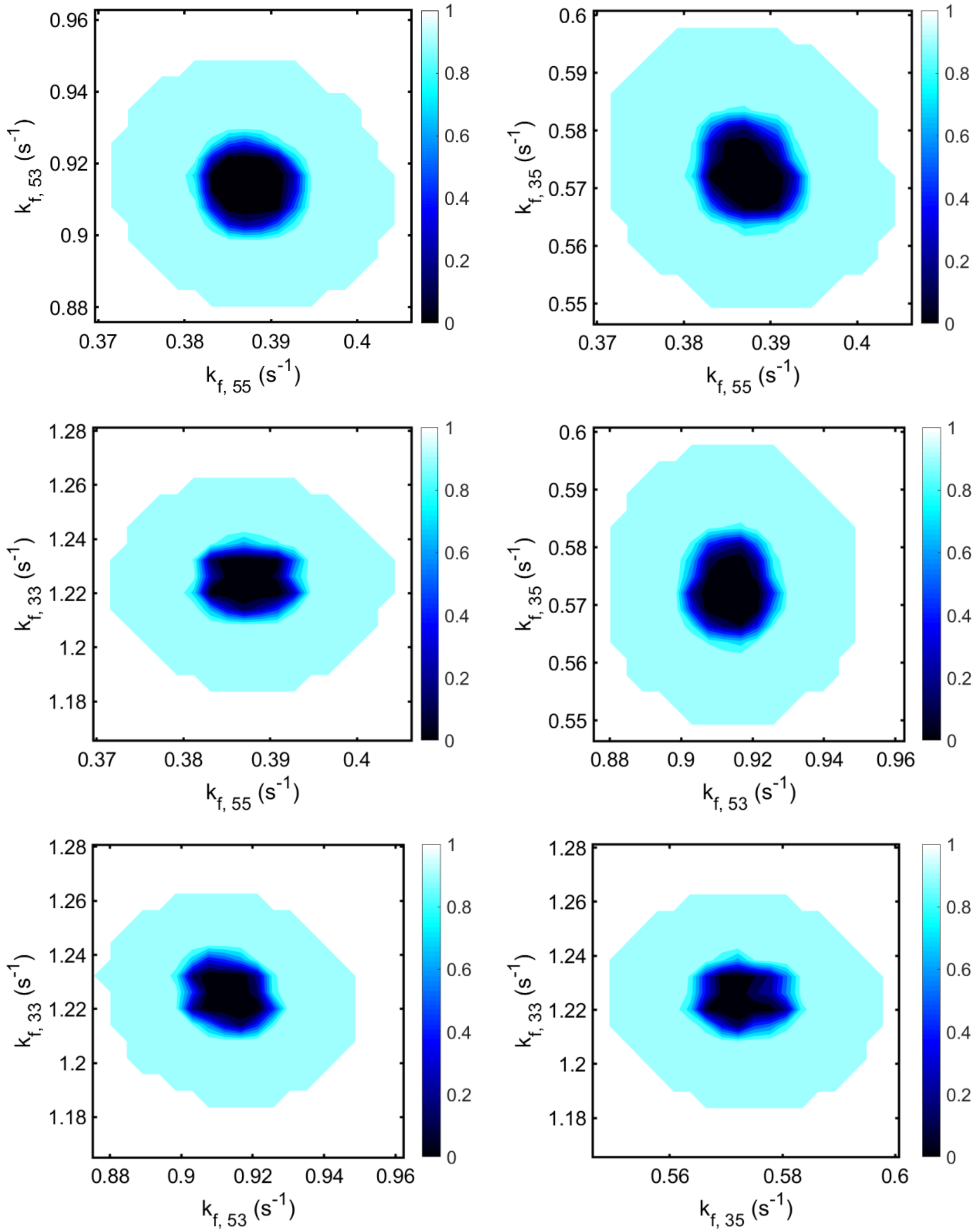


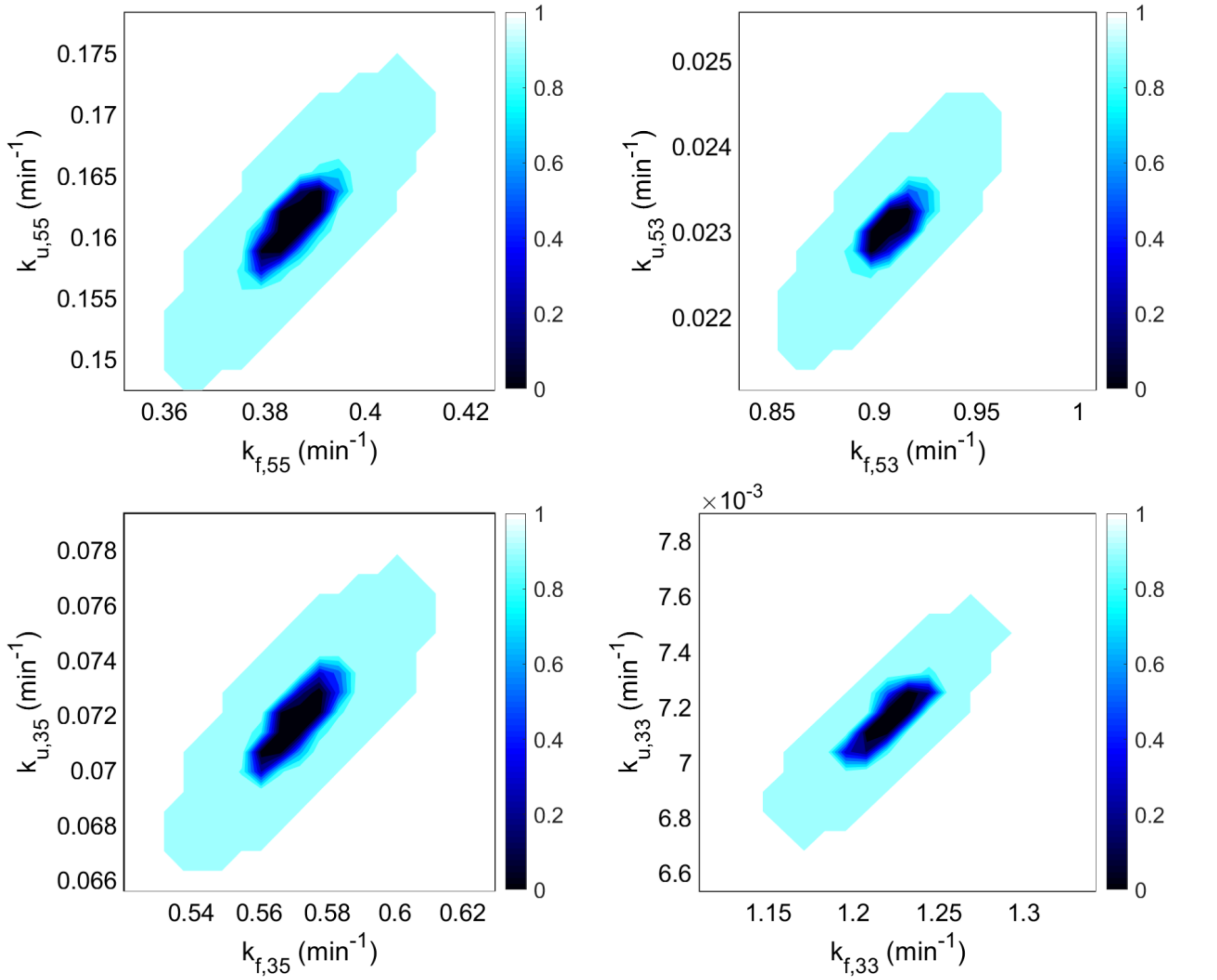
Supplementary Figure 2.3: Temperature correction of TH data sets. Overlays of TH data for the fully–trapped mutants collected in a 1 cm cuvette (triplicates shown in black, grey, and light grey respectively) and 0.1 cm cuvette (red). Data were temperature corrected as previously described[246].



Supplementary Figure 2.4: Baseline correction of TH datasets. Temperature-corrected c-MYC Pu22 WT and trapped mutant G4 TH data (295 nm) are shown with folded and unfolded baselines in black. Slow to fast heating and cooling scan rates are shown as orange to red and light blue to dark blue respectively. Datasets were converted to fraction unfolded profiles and displayed in Figure 2.2.




$k_{f,53}$ (s^{-1})
0.96
0.94
0.92
0.9
0.88
0.37
0.38
0.39
0.4
$k_{f,55}$ (s^{-1})
1
0.8
0.6
0.4
0.2
0
$k_{f,35}$ (s^{-1})
0.6
0.59
0.58
0.57
0.56
0.55
0.37
0.38
0.39
0.4
$k_{f,55}$ (s^{-1})
1
0.8
0.6
0.4
0.2
0
$k_{f,33}$ (s^{-1})
1.28
1.26
1.24
1.22
1.2
1.18
0.37
0.38
0.39
0.4
$k_{f,55}$ (s^{-1})
1
0.8
0.6
0.4
0.2
0
$k_{f,35}$ (s^{-1})
0.6
0.59
0.58
0.57
0.56
0.55
0.88
0.9
0.92
0.94
0.96
$k_{f,53}$ (s^{-1})
1
0.8
0.6
0.4
0.2
0
$k_{f,33}$ (s^{-1})
1.28
1.26
1.24
1.22
1.2
1.18
0.88
0.9
0.92
0.94
0.96
$k_{f,53}$ (s^{-1})
1
0.8
0.6
0.4
0.2
0
$k_{f,33}$ (s^{-1})
1.28
1.26
1.24
1.22
1.2
1.18
0.56
0.58
0.6
$k_{f,35}$ (s^{-1})
1
0.8
0.6
0.4
0.2
0



Supplementary Figure 2.5: Global fit parameter correlations and confidence intervals. Confidence level plots were generated according to the Supplementary Methods for the primary TH dataset in 2 mM K$^+$, 1 cm cuvette conditions. Colour bars indicate the normalized confidence level, with 95% being approximately at the cyan-white interface. Only correlation surfaces between the folding rate constants, folding rate constants and activation energies, and folding and unfolding rate constants are shown.

Chapter 3: Structural polymorphism of guanine quadruplex-containing regions in human promoters

3.1 Preface

The work in this chapter follows from the results of Chapter 2, where we found that the c-MYC G4 had parallel folding pathways which led to a net acceleration in its folding. This led us to ask the simple question: What was the prevalence of structural polymorphism of G4s in human promoters? To address this, I developed a new bioinformatic algorithm to find G4 motifs from a sequence of DNA and combine them into regions which overlapped with each other, called the GReg algorithm. There are many possible G4 motifs, and not all of them have the same stability. As described in this chapter, stability decreases with increasing loop length, and with increasing the number and size of bulges. In order to only search for highly stable G4 motifs, Lynn Yamout developed *Equation 3.3* which I then incorporated into the GReg algorithm. Lynn Yamout also wrote and analyzed all of the data present in Section 3.4.2, which details her development of *Equation 3.3*. In order to make the GReg algorithm easy to use and more widely available, it was adapted into a python-based webserver. This work was done by Chuyang (Amos) Zhang, with help from David Hiraki. The description of this webserver is found in Section 3.4.7. I, with guidance from Prof. Anthony Mittermaier, completed all of the rest of the work for this chapter.

This chapter was adapted with permission from: Hennecker, C., Yamout, L., Zhang, C., Zhao, C., Hiraki, D., Moitessier, N., & Mittermaier, A. (2022). Structural polymorphism of guanine quadruplex-containing regions in human promoters. *International Journal of Molecular Sciences*, 23(24), 16020.

3.2 Abstract

Intramolecular guanine quadruplexes (G4s) are non-canonical nucleic acid structures formed by four guanine (G)-rich tracts that assemble into a core of stacked planar tetrads. G4-forming DNA sequences are enriched in gene promoters and are implicated in the control of gene expression. Most G4-forming DNA contains more G residues than can simultaneously be incorporated into the core resulting in a variety of different possible G4 structures. While this kind of structural polymorphism is well recognized in the literature, there remain unanswered questions regarding possible connections between G4 polymorphism and biological function. Here we report a detailed bioinformatic survey of G4 polymorphism in human gene promoter regions. Our analysis is based on identifying G4-containing regions (G4CRs), which we define as stretches of DNA in which every residue can form part of a G4. We found that G4CRs with higher degrees of polymorphism are more tightly clustered near transcription sites and tend to contain G4s with shorter loops and bulges. Furthermore, we found that G4CRs with well-characterized biological function tended to be longer and more polymorphic than genome-wide averages. These results represent new evidence linking G4 polymorphism to biological function and provide new criteria for identifying biologically relevant G4-forming regions from genomic data.

3.3 Introduction

Intramolecular guanine quadruplexes (G4s) are four-stranded nucleic acid structures formed when four tracts of contiguous guanine residues, separated by three loops, come together in stacked planar tetrads stabilized by Hoogsteen hydrogen bonding and metal coordination (*Figure 3.1a*). Putative G4s are plentiful in the human genome, and are found in functional regions including origins of replication, introns, 5’ and 3’ untranslated regions, as well as in promoter regions, where they help to regulate gene expression[38, 135, 187, 377-379]. The stability of these structures is influenced by a variety of different factors such as the presence of different cations, pH, and molecular crowding[380]. Some of the best-characterized G4s have relatively simple structures consisting of four tracts of 3 Gs, with all 12 G residues engaged in the core structure[381, 382]. However, in general, G4-forming DNA sequences are polymorphic. They contain more Gs than can be simultaneously incorporated into a single structure, resulting in ensembles of different conformations with different subsets of Gs engaged in the core[383-386]. These polymorphisms can take many different forms (*Figure 3.1b*). For example, stable G4s can form from G-tracts that contain non-G residues which are bulged out from the core structure [308, 387, 388]. Alternatively, when the tracts contain different numbers of Gs, the strands can effectively slide with respect to one another, a type of motion we refer to as G-register exchange [210, 216, 351]. For example, the Pu18 sequence from the human *MYC* promoter (AGGGTGGGGAGGGTGGGG) has two tracts of 3 Gs and 2 tracts of 4 Gs and can form 4 different G-register isomers: aGGG**t**GGG**ga**GGG**t**GGGg, aGGG**tg**GGG**a**GGG**t**GGGg, aGGG**t**GGG**ga**GGG**tg**GGG, and aGGG**tg**GGG**a**GGG**tg**GGG, where guanine residues participating in the tetrad core are capitalized and loop residues are in bold type. Of these, the first isomer is the most thermodynamically stable [210]. As well, there are several examples of biologically relevant G4s that contain extra (>4), or "spare tire" G-tracts, such that stable G4s can form from different subsets of four of the tracts[389, 390]. For example, the Pu27 sequence from the human *MYC* promoter comprises the Pu18 sequence (above) and a 5th G-tract appended at the 5' end. Three alternative G4s have been reported for this region in which the tetrad core is formed from tracts 1234, 1245, and 2345 (the Pu18 sequence)[211, 389]. Similarly, G-rich minisatellite DNA can contain dozens of consecutive G-tracts[391, 392] and can form enormous numbers of different G4 folded states involving different subsets of G-tracts, potentially leading to a highly frustrated energy landscape[180]. Finally, many different chain topologies are possible, with the strands running parallel or anti-parallel to one-another[393] or adopting snap-back conformations[394]. G4s composed of G-tracts containing between two[395] and six[40] consecutive Gs have been reported. G4s can be stabilized by DNA hairpin formation in the loops and bulges[396, 397], additional hydrogen bonding of core guanines to loop residues[398], and stacking of adjacent G4s[399] leading to some highly non-canonical G4 structures[400].

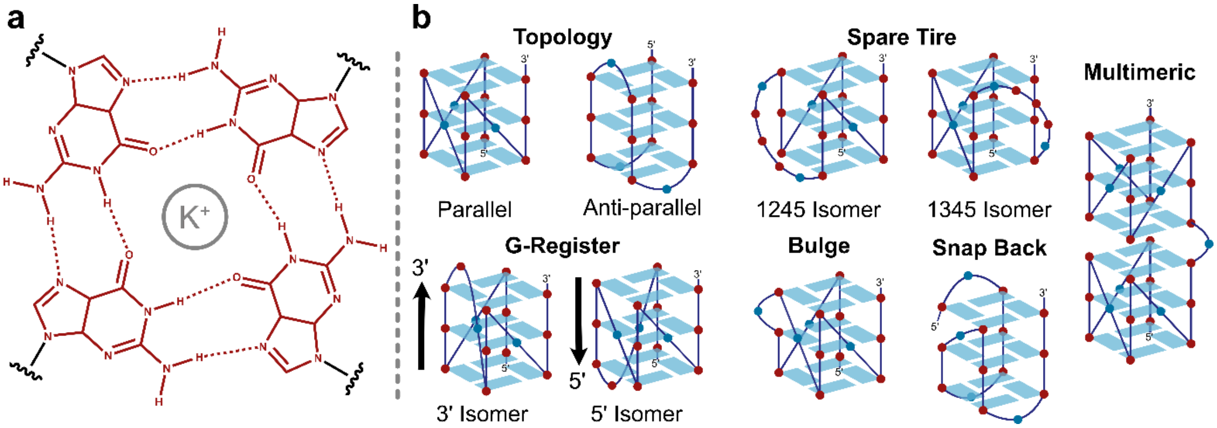


Figure 3.1: Guanine quadruplex structures. a) Hydrogen bonding pattern of a G-tetrad in the G4 core. b) Representative structural features of G4s: Topological isomers, Spare tire isomers, G-register isomers, Bulges, Snap back structures, and multimeric G4s.

There are several ways in which the polymorphism of G4s has been proposed to impact their biological function[121, 139, 401-403]. For example, we have found that the existence of multiple G register isomers can accelerate folding[351] and stabilize the folded state[210], with implications for G4 function. It has been shown that the presence of spare tire G-tracts can provide resilience to DNA damage[390]. G4-binding proteins such as nucleolin can differentiate between different folding isomers (G4s), while different DNA helicases have strong preferences for unwinding different G4 structures[404]. The fact that a single G-rich DNA sequence from a gene promoter region can fold into multiple different structures has been identified as hurdle to designing effective specific G4-targetting drugs[401]. Many of best-characterized G4s with validated biological functions are capable of undergoing both spare tire (>4 G-tacts) and G-register (G-tracts of different lengths) isomerism. For example, the well-studied Pu-27 sequence from the promoter of the oncogene *MYC* contains 2 tracts of GGG and 3 tracts of GGGG. When 8 or more G tracts are present, then in principle 2 or more adjacent G4s can form. For example, a stretch of 68 nucleotides from the telomerase (*TERT*) gene promoter contains 5 tracts of GGG and 7 tracts of GGGG and can fold into two[405] or three[406] consecutive G4s.

While seemingly common in naturally occurring G4s, these types of structural polymorphism have not yet been surveyed in a systematic manner, thus their prevalence and functional significance remain poorly understood. Recent reviews have split the bioinformatic approaches for predicting the locations of G4s into the categories of regular expression matching, scoring, sliding window scoring, machine learning, and specialized tools [134]. For example, the *quadparser* algorithm focuses on identifying short segments of DNA that are likely able to adopt a stable fold, by matching the sequence of interest to a consensus motif [119]. Alternatively, the *G4hunter* algorithm predicts the tendency of DNA to form G4s by evaluating G richness and G skewness (i.e. the density of tracts containing consecutive Gs) [133]. Newer approaches such as the *QPARSE* algorithm have identified the prevalence of multimeric G4s in the human genome, along with sequences that can form hairpin loops [407]. However, none of these methods quantify the extent of G4 structural polymorphism. Thus, the extent, prevalence, distribution, and functional significance of G4 polymorphism remains largely unknown.

We have set out to evaluate G4 structural polymorphism in promoter regions of the human genome. To begin, we recognized that the basic functional unit is not a single G4 structure, but rather a G-rich stretch of DNA that can adopt between one and a multitude of different G4 folds. We defined a G4 containing region (G4CR) as a contiguous stretch of DNA containing G-tracts, such that each G-tract can, in principle, form a stable G4 with the G-tracts on either side. In other words, G4CRs may be thought of as regions of DNA where every single residue has the potential to be included in at least one G4, in either a core, bulge, or loop position. We found that G4CRs can be anything from about 15 nucleotides (nt) ($(GGGN)_3GGG$) to several hundred nt in length. They can form between one and several thousand structural distinct G4s, with up to about 25 simultaneously adjacent. In parallel, we also calculated the structural multiplicity of individual G residues, which we defined as the number of structural distinct G4s that incorporate a particular G residue into the core. Since multiplicity is defined on a per-G basis, this provides a simple approach for mapping structural polymorphism onto a DNA sequence with single-residue precision. We found that multiplicity values can vary between about one and a thousand for different G residues within a single G4CR. Intriguingly, the degree of polymorphism within a G4CR appears related to both the location of the G4CR relative to the transcription start site, and also the lengths of loops and presence of bulges in the G4 structures it forms. Furthermore, a variety of well-characterized biologically relevant G4 containing regions from the *MYC*, *VEGFA*, *BCL2*, *KIT*, and *KRAS* gene promoters have greater than average degrees of polymorphism. This is new evidence that polymorphism itself may be related to the biological function of G4CRs and provides new criteria for identifying potentially important regulatory sites in DNA.

3.4 Results

3.4.1 Calculating G4 polymorphism.

Predicting the number of possible different G4 conformations available to a single stretch of G-rich DNA becomes intractable if every source of structural variation is considered. For example, we are not aware of any method for reliably predicting *a priori* which topology or topologies (parallel, antiparallel, hybrid) a particular DNA sequence will favour. Non-canonical inter-G4 stacking and additional base-pairing between core and loop and flanking residues are usually only identified by biophysical analyses and full three dimensional structural elucidation [408], although this is complicated somewhat by the fact that crystal packing can influence G4 structure [409]. Therefore, we have opted to simplify the problem by considering only a subset of the structures that G-rich DNA can adopt. Specifically, we calculated the number of ways to incorporate different sets of twelve G residues into a stable, three-tetrad, G4 core, while ignoring topology and non-canonical or higher-order interactions. We considered potential G4s with loop lengths of up to 7 residues [119] and bulges with lengths of up to 3[308]. For example, a hypothetical sequence $G_3TG_4TG_3TG_3TG$ could adopt 4 different structures: GGG**t**GGG**gt**GGG**t**GGG**tg**, GGG**tg**GGG**t**GGG**t**GGG**tg**, GGG**t**GGG**gt**GGG**tg**GG(**t**)G, and GGG**tg**GGG**t**GGG**tg**GG(**t**)G, where G's in the core are capitalized, looped residues are lowercase and bolded, and bulged residues are lowercased, bolded, and put in parenthesis. Bulged G's were not considered, since the non-bulged variant would always be more stable. We note that, while two-tetrad G4s exist, they are usually quite unstable in the absence of additional non-canonical interactions, which is beyond the scope of this analysis[395, 410]. G4s containing more than 3 tetrads have been reported[40, 61], although other studies suggest that strand-shifted conformations with 3-tetrad cores may be still be preferred, even when all G-tracts contain more than 3 G's[41]. Furthermore, while we have restricted our analysis to a maximum loop length of 7, structures with larger loops are present in the human genome as well as other G4 forming sequences which do not fit our simplified model[135]. Thus in the following analysis, it should be remembered that we have considered an important but partial subset of the possible G4 structures formed by any given DNA sequence. The true number of accessible conformations is likely somewhat larger.

3.4.2 Predicting stable G4 structures.

In order to realistically evaluate G4 polymorphism, it is important to account for the fact that some G4 structures are more likely to form than others. Previous studies have related experimental melting temperatures (T_m) of G4s to the lengths of the loops ($L_{1\text{-}3}$)[411] and number (N_b) and size (L_b) of bulges[308]. Unsurprisingly, T_m decreases with increasing total loop length ($L_1+L_2+L_3$) (*Figure 3.2a*). The dependence is reasonably well fit by a linear relationship, however the prediction breaks down for the most stable G4s with the shortest loops. For example, the calculated melting temperature of the extremely stable $(G_3T)_3G_3$ sequence (77.6°C) indicated by the red circle in *Figure 3.2a* is actually lower by 6.4°C than the measured melting temperatures of the longer-loop variants (84°C and 81.8°C, respectively), when the reverse should be true. Notably, when data for a single loop are examined in detail, T_m values show a curvilinear decrease with increasing L (*Figure 3.2b*), which is perhaps to be expected since theory suggests that the entropic cost of closing a loop during macromolecular folding varies as the logarithm of the length of the loop [412]. Plotting T_m as a function of log(L) produces far better linear relationships (*Figure 3.2e*). When data for all variants are plotted as a function of $\log(L_1)+\log(L_2)+\log(L_3)$ ($=\log(L_1L_2L_3)$), a linear relationship is obtained (*Figure 3.2d*) with an improved R^2 value compared to *Figure 3.2a*. The T_m predicted for $(G_3T)_3G_3$ (at the y-intercept) is 6°C higher than those determined experimentally for the longer-loop variants, as expected. The melting temperature can therefore be estimated for this dataset by the empirical equation

$$T_m(L_1, L_2, L_3) = a - b \cdot log\{L_1L_2L_3\} \qquad \textit{(Equation 3.1)}$$

where *a*=89.9°C and *b*=19.2°C.

Stability data have also been reported for G4s containing different numbers of bulges of different lengths[308]. Introduction of a bulge leads to a roughly 20°C reduction in T_m, and each additional residue in the bulge reduces the T_m by about another 10°C. We found that the data are well fit by the empirical relationship

$$T_m(N_b, L_b) = c - d \cdot N_b - f \cdot (L_b - N_b) \qquad \textit{(Equation 3.2)}$$

where N_b and L_b are the total number of bulges and the total number of bulged residues, respectively. For data obtained with 12 mM K^+, *c*= 89.9°C, *d*= 20.5°C, and *f*= 8.5°C (*Figure 3.2c*). For data obtained with 60 mM K^+, *c*= 97.8°C, *d*= 19.4°C, and *f*= 8.5°C (*Figure 3.2f*). Since the 12mM K^+ data and 60mM K^+ data produced similar values for *d* and *f* we took the mean of these two parameters giving *d* = *20*°C and *f* = *8.5*°C*.* We combined *Equation 3.1* and *Equation 3.2* to estimate the stabilities of all predicted G4s on the basis of their loop lengths, and number and length of bulges according to:

$$T_m^{est}(L_1, L_2, L_3, N_b, L_b) = a - b \cdot log\{L_1L_2L_3\} - d \cdot N_b - f \cdot (L_b - N_b) \qquad \textit{(Equation 3.3)}$$

It should be kept in mind that the T_m^{est} parameter is not a precise predictor of the melting temperature, given the scatter evident in *Figure 3.2a* and *Figure 3.2d*, and possible contributions of non-canonical and higher order interactions in naturally occurring G4s [400]. Rather, a high T_m^{est} value indicates that a putative G4 structure has short loops and a small or no bulge while a low T_m^{est} values is indicative of long loops and/or bulges, with the values tied to physically reasonable melting temperatures.

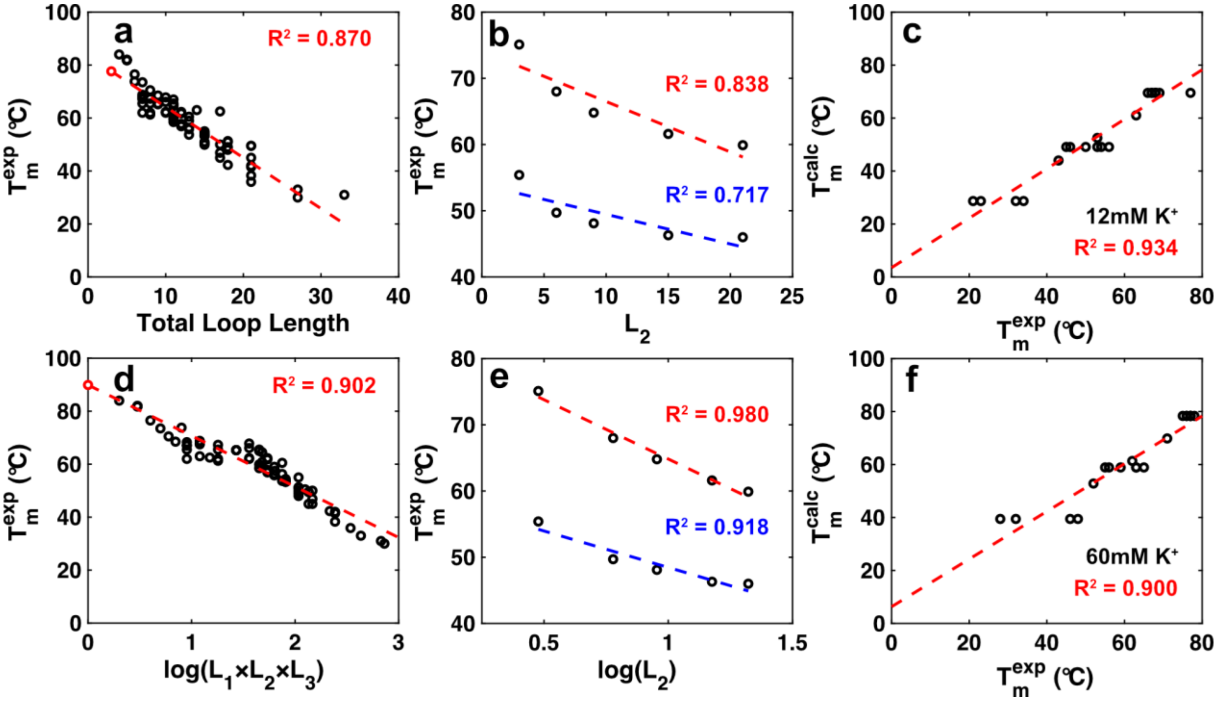


Figure 3.2: Effect of loop length on T_m. a) Dependence of experimental T_m on total loop length. b) Dependence of experimental Tm on the length of L_2 for L_1,L_3 = TTT (blue) and L_1,L_3 = T (red). c) Correlation between experimental and predicted T_m values for bulges of different lengths in 12mM potassium according to Equation 3.3. d) Dependence of experimental T_m on the logarithm of the product of the loop lengths. e) Dependence of experimental T_m on the logarithm of the product of the loop lengths for L_1,L_3 = TTT (blue) and L_1,L_3 = T (red). f) Correlation between experimental and predicted T_m values for bulges of different lengths in 60mM potassium according to Equation 3.3. Data in a,b,d,e taken from[411]. Data in c,f taken from[308].

3.4.3 Identifying G4 regions (G4CRs) in DNA sequences.

To locate G4CRs within a given stretch of DNA, we identified all sets of 12 core G residues within the sequence that can theoretically form a G4 with T_m^{est} ≥ 50°C according to *Equation 3.3*, together with the accompanying loops (limited to ≤7 nt) and bulges (limited to ≤3 nt). Regions encompassing overlapping sets of Gs were defined as G4CRs. *Figure 3.3* illustrates a hypothetical example in which the first G4CR ($G4CR_1$) contains 6 G-tracts and can form G4s with 8 distinct subsets of G residues. The second G4CR ($G4CR_2$) is separate from the first because no stable G4 includes both the 3’ G-tract of

G4CR1 and the 5’ G-tract of G4CR2. Note that for very G-rich sequences, the distinction between different G-tracts and between looped and bulged residues becomes somewhat blurred. For example, in G4CR2, the T is part of a loop in G41, G42, and G43 and a bulge in the third G-tract in G44 and G45. The G immediately following the T is part of the fourth G-tract in G41, part of a loop in G42 and G43, and part of the third G-tract in G44 and G45. Regardless, the locations and lengths of G4CRs can be assigned unambiguously.

Interestingly, the contributions of each G residue to the entire ensemble of G4s formed by the G4CR are quite different. For example, the Gs of the first tract in G4CR1 are only folded in 2 of the 8 possible structures. In contrast, those of the third G-tract are folded in all 8 of the possible structures. We refer to these relative contributions of each G as the folding multiplicity; values are listed for each G beneath the sequence in *Figure 3.3*. The values of multiplicity and the total number of G4s formed by a G4CR are related. Since each G4 is composed of exactly 12 G core residues in our simplified model, the total number of G4s is given by the sum of the multiplicities for all G residues in the G4CR, divided by twelve.

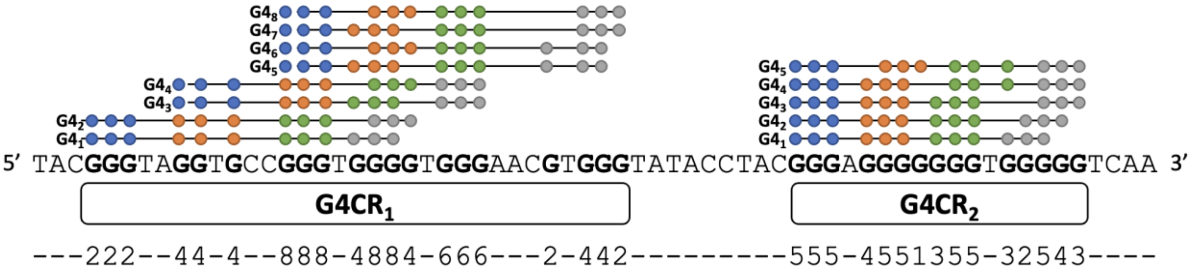


Figure 3.3: Hypothetical stretch of DNA containing two G4CRs. Filled circles above the sequence indicate the core G residues in each possible isomer. The multiplicity of each G residue is indicated at the bottom of the figure.

3.4.4 Characteristics of G4CRs in human promoters.

We analyzed portions of the human genome within -1999 and 2000 base-pairs of the transcription start sites (TSS) of 16528 genes using the first promoter listed in the eukaryotic promoter database[413], considering both coding and non-coding strands of DNA. A cumulative plot of the incidence of G4CRs is shown in *Figure 3.4*, where the curves give the number of genes with fewer than a given number of G4CRs. As found previously for individual G4s, most (85%) genes contain at least one G4CR within the examined region[133]. The median number of G4CRs per gene (i.e. the value at the 50th percentile) is 3, while about 1% of genes (154) have over 15 G4CRs. In some cases, these promoters contain G4CR-rich stretches that correspond to repetitive minisatellite DNA, as discussed below. We identified a total of 2847 G4CRs in these top 1% of genes. This is only about 5% of the total number of G4CRs we found in all genes (63,303), meaning that the statistics we report for G4CRs below are dominated by non-repetitive, non-satellite DNA.

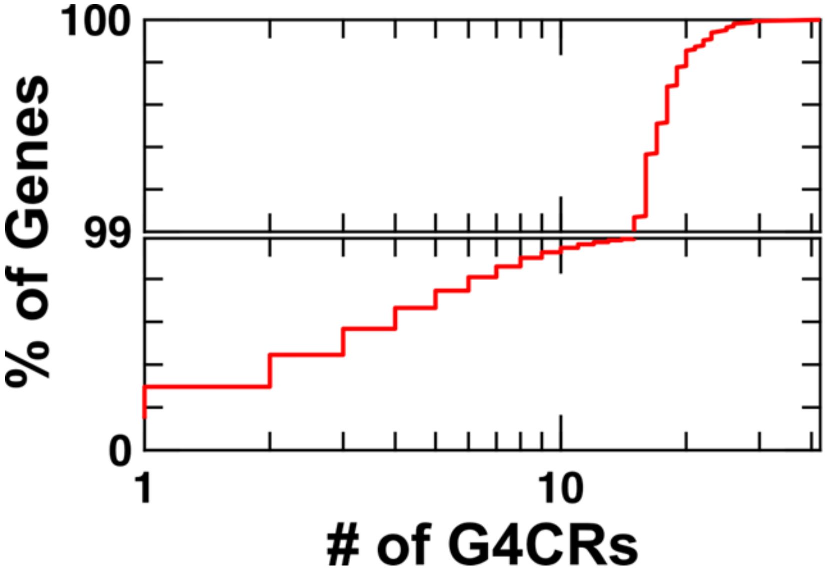


Figure 3.4: Cumulative plot of the number of G4CRs for human genes considering both coding and non-coding strands. The number of G4CRs is plotted on a logarithmic scale. The bottom panel represents the first 99% of genes whereas the top panel represents the top 1% of genes.

Figure 3.5a-d shows cumulative plots of the distributions of the lengths, G-content, total number of G4 isomers (N_{tot}), and total number of G4s that can form simultaneously in tandem (N_{tand}). Data for G4CRs from the coding and non-coding strands were pooled, since we did not observe any substantial differences between the two strands. The median length of a G4CR is about 25 nt, and 75% are shorter than about 32 nt (*Figure 3.5a*). However, a substantial number of G4CRs are much longer. Roughly 1%, or about 600 G4CRs are longer than 71 nt, which is a considerable number when one recalls that this refers to a continuous stretch of nucleotides where each nucleotide has the potential to form part of a G4 structure. *Figure 3.5e* shows cumulative plots of the estimated melting temperatures of the G4s predicted to be formed within the G4CRs. The median T_m^{est} is

60°C, with about 10% predicted to melt above 70°C. Identical T_m^{est} curves were obtained for G4CRs of different lengths (*Figure 3.5e*). The guanine content of most G4CRs is between about 50% and 75%, although a few G4CRs were identified with 100% G-content, i.e. stretches of poly-guanosine (*Figure 3.5b*). Interestingly, different T_m^{est} profiles were obtained for G4CRs with different G-content, such that G4s located in G4CRs with higher G-content tended to have higher estimated melting points (*Figure 3.5f*). For example, the fraction of G4s with T_m^{est} > 70°C was about 4-fold higher for G4s derived from G4CRs with >84% G than for those with ≤68% G. This implies that putative G4s from G4CRs with high G content have shorter loops and bulges compared to those from G4CRs with low G-content. The median number of different G4 isomers (N_{tot}) formed by a G4CR is 4, but there are many with far greater degrees of polymorphism (*Figure 3.5c*). 25% of G4CRs form greater than 14 different isomers, 1% form more than 179 G4 isomers, while 20 G4CRs can potentially form over 1000 distinct G4 structures each. Similarly to what was seen with G-content, T_m^{est} values are higher for G4s derived from G4CRs with larger N_{tot}, compared to those with lower N_{tot} (*Figure 3.5g*). In other words, G4s from G4CRs with greater degrees of polymorphism tend to have shorter loops and bulges. The overwhelming majority (>95%) of G4CRs can only form one G4 structure at a time (N_{tand}=1), as shown in Figure 4d. About 1% can form 3 or more tandem G4s and 0.1% (or about 100 G4CRs) can form 5 or more at one time. We did not observe any difference in the predicted stabilities of G4s derived from G4CRs with different N_{tand} values (*Figure 3.5h*).

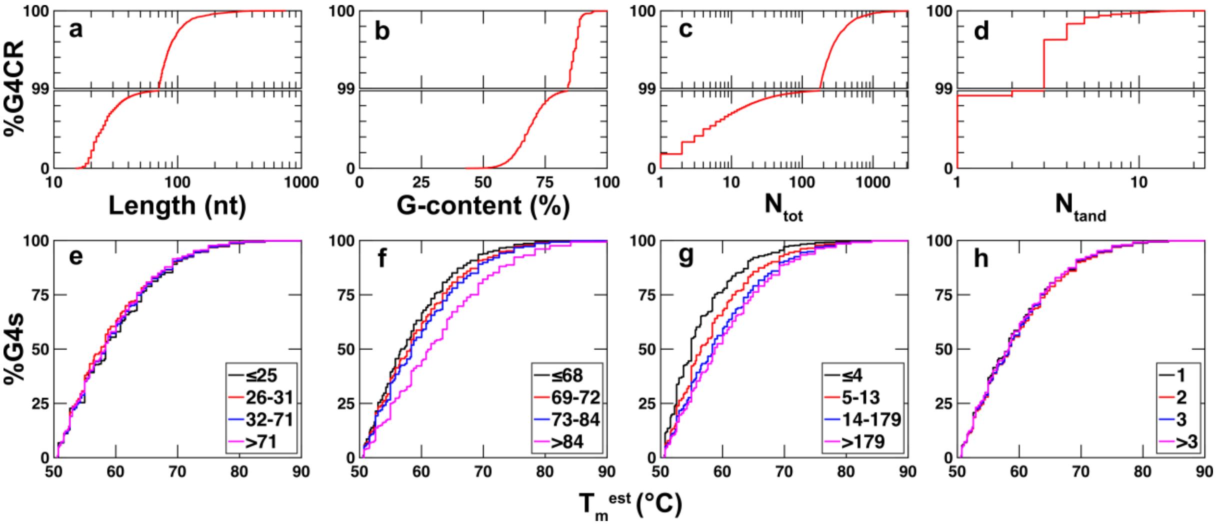


Figure 3.5: Characteristics of G4CRs. Cumulative plots of the a) Length, b) G-content, c) N_{tot}, d) N_{tand} of G4CRs found in the human promoters. Cumulative plots of the estimated melting temperatures for G4s derived from G4CRs binned by e) Length, f) percentage of residues that are G (G-content), g) Total number of G4 isomers (N_{tot}), h) maximum number of simultaneous tandem G4s (N_{tand}).

We next explored the correlation between the G4CR characteristics by constructing scatter plots for every pair of the four parameters discussed above (*Figure 3.6*). Surprisingly, apart from the strong and expected increase in number of tandem G4s with increasing G4CR length (*Figure 3.6b*), correlations among the other parameters were weak at best. For example, G4CRs with between about 25 and 60 nt in length can have anywhere from 2 to over 500 isomers (*Figure 3.6a*). Thus, the number of G4 isomers formed by a G4CR does not depend strongly on the length of the region. The very highest G contents were observed for the shortest G4RCs, implying that stretches of nearly 100% guanosine are limited to about 15-25 nt in length (*Figure 3.6c*). The lowest contents of G (≈40%) were observed for G4CRs about 40 nt in length. Conversely, the very longest G4CRs of several hundred bases (*Figure 3.6c*) and those with the largest number of tandem G4s (more than 10, *Figure 3.6f*) had intermediate G contents of 60-80%. These intermediate G content G4CRs formed anywhere between 1 and >1000 G4 isomers. Low G content (<50%) led to small N_{tot} values, likely due to a lack of G residues to form alternative structures. Very high G content (>90%) was associated with moderate N_{tot} values (<100), likely because these G4CRs tended to be short. The fact that the key characteristics of G4CRs are largely uncorrelated implies that G4CRs can be classified into multiple categories: short vs long, low vs high G content, and small vs large numbers of isomers (N_{tot}) (relative to the median values). Note that longer G4CRs are strongly associated with larger number of tandem G4s, so N_{tand} does not represent a separate parameter for the purposes of categorization. This rough division give 8 different classes of G4CR. It is an interesting question to which extent these different classes have divergent biophysical properties, functional relevance, or biological roles.

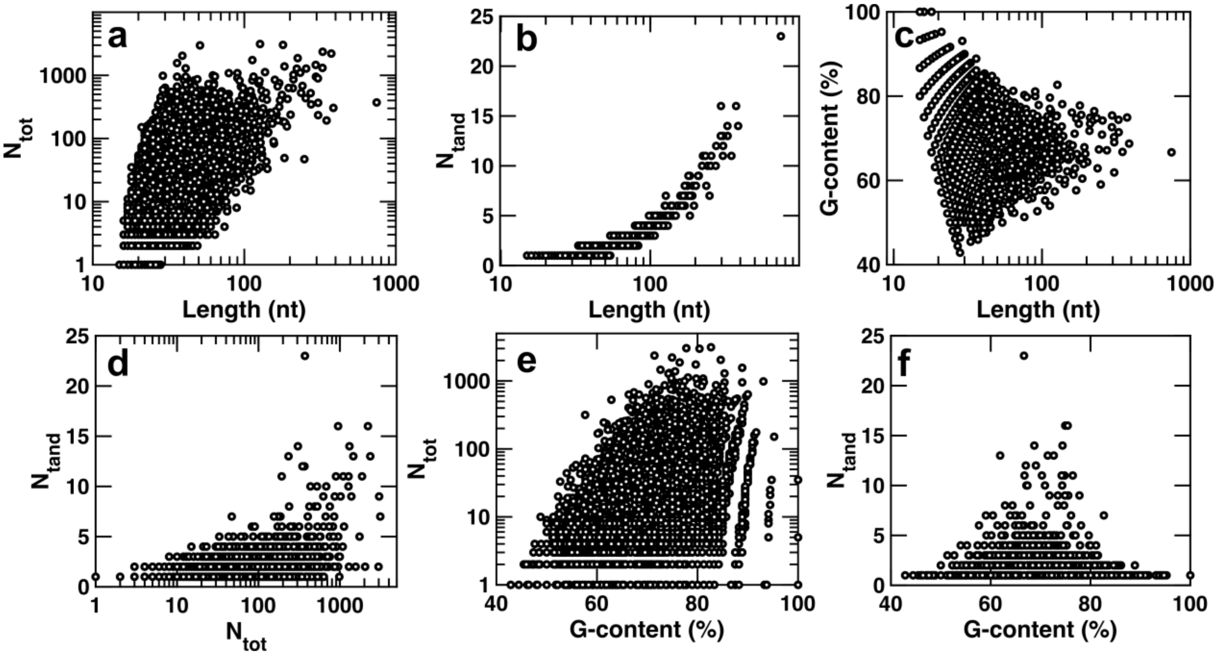


Figure 3.6: Scatter plots of each of the different features of G4CRs found in this study: Length, percentage of residues that are G (G-content), Total number of G4 isomers (N_{tot}), maximum number of simultaneous tandem G4s (N_{tand}).

3.4.5 Distribution of G4CRs in gene promoters.

It is well known that the distribution of putative G4 forming sequences within the human genome is distinctly non-random. G4 motifs are enriched adjacent to transcription start sites (TSSs) in gene promoters from humans[119, 133, 407] and a wide range of other eukaryotes[414], with asymmetric densities on coding vs non-coding strands, pointing to a general regulatory role for G4s[163]. Plots of G4 propensity calculated using the *quadparser*, *G4Hunter,* and *QPARSE* algorithms all produce characteristic sharp peaks about 100 nt immediately upstream from the TSS, equally sharp dips in the 100 nt immediately following the TSS, and broader peaks over the next roughly 500 nt[119, 133, 407]. The first peak and dip are evident on both coding and non-coding (template) strands, while the last peak is far more prevalent on the coding strand compared to the template strand. We set out to evaluate the extent to which the distribution of G4CRs matches these previous results and whether the distributions are correlated with the degree of polymorphism or other G4CR characteristics. We first calculated the fraction of genes in which a given position relative to the TSS lies within a G4CR bearing certain characteristics of length, G content, etc. for a region extending from -1999 (upstream) to +2000 (downstream) of the TSS of all 29598 promoters listed in the eukaryotic promoter database[413]. We then normalized the distributions to facilitate comparisons between G4CRs with differing characteristics. This was done by setting the sum of probabilities over all 4000 positions in the calculation window equal to 1, which accounts for the fact that the likelihoods of lying within longer and/or more common types of G4CRs (i.e. with characteristics close to the mode) are larger overall. *Figure 3.7* shows the normalized probability distributions for G4CRs for coding (*Figure 3.7a-d*) and non-coding (*Figure 3.7e-h*) strands. All of the plots closely mirror what was reported using the *quadparser*, *G4Hunter,* and *QPARSE* algorithms with sharp peaks upstream on both coding and non-coding strands and broad peaks downstream of the TSS, particularly on the coding strand. Interestingly, the sharpness of the distributions varies depending on the characteristics of the G4CR. For example, *Figure 3.7a*,e shows the normalized probabilities of lying within G4CRs of different lengths for the coding (*Figure 3.7a*) and non-coding (*Figure 3.7e*) strands. The pre-TSS (left) peaks are substantially sharper for longer G4CRs than for shorter ones, with highest peak for G4CRs with length>69 nt followed by those 31<length≤69 nt. As a result, the enrichment in G4CRs in the -200 to 0 relative to -1500 to -1300 regions (relative to TSS) is 11-fold for G4CRs longer than 69 nt compared with only 5-fold for G4CRs less than 25 nt. In other words, longer G4CRs are more tightly clustered immediately upstream from the TSS, compared to shorter ones. *Figure 3.7b,f* show similar data for G content. In this case, G4CRs with intermediate %G values (69-73% and 74-84%) exhibit the highest pre-TSS peaks in probability. In terms of polymorphism, G4CRs with the highest total number of G4 isomers (>168 and 14-168) show the sharpest clustering prior to the TSS, compared to G4CRs with fewer than 13 isomers, while G4CRs with 2 and 3 tandem G4CRS show a higher peak than those with more than 3 or just a single isomer. The greater clustering of long G4CRs with high numbers of G4 isomers that form 2 or more contiguous G4s points to a relationship between these characteristics and G4 biological function.

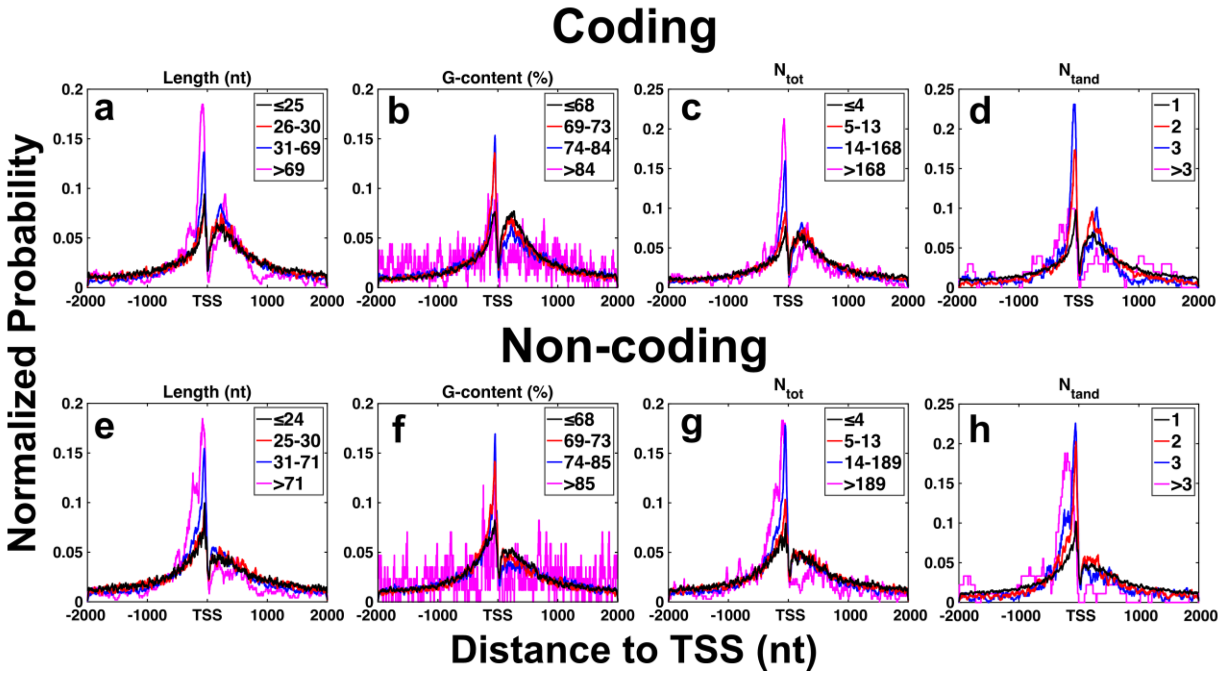


Figure 3.7: Distribution of G4CRs in gene promoters. Likelihood that a residue lies within a G4CR possessing certain characteristics, plotted as a function of the residue's position relative to the transcription start site (TSS), for human gene promoters. Colours represent the bottom 50% of G4CRs (black), 50-75% of G4CR, 76-99%, and top 1% of G4CRs. Specific values are given in the legend of each panel. Panels a-d are for the coding strand (Length, G-content, N_{tot}, and N_{tand} respectively). Panels e-h are for the non-coding strand (Length, G-content, N_{tot}, and N_{tand} respectively).

We then extended this analysis by considering promoter regions from a variety of organisms including the vertebrates *M.* mulatta (monkey), *M.* musculus (mouse), *R.* norvegicus (rat), *C.* familiaris (dog), *G.* gallus (chicken), and *D.* rerio (fish), the invertebrates *D.* melanogaster (fly), *A.* mellifera (bee), and *C.* elegans (worm), the plants *A.* thaliana (thale cress) and *Z.* mays (corn), the yeasts *S.* cerevisiae and *S.* pombe, and the parasitic protozoan *P.* falciparum (see Supplementary figures). Many of these organisms have been previously analyzed using the G4hunter algorithm and classified as containing a high density (human, monkey, mouse, rat, dog, and chicken), an intermediate density (fly, bee, and fish), a low density (yeasts, worm, and thalecress), or very low density (protozoan) of G4-forming sequences. We obtained similar results, with an average of 3.6 G4CRs identified per promoter in the high G4 density group, 0.19 G4CRs per promoter in the intermediate G4 density group, 0.08 G4CRs per promoter in the low G4 density group, and 0.005 G4CRs per promoter in the protozoan (*Table 3.1* and *Table 3.2*). We found that corn (which was not examined by G4hunter) has a relatively high abundance of G4CRs (0.96 per promoter). This is intriguingly similar to rice, which was included in the high G4 density group along with mammals and chicken, in the G4hunter study. We next examined distributions of G4CRs relative to the TSS for the species listed above. For all of the high G4-density vertebrate species, we observed clustering of G4CRs similar to that in humans (*Figure 3.11*). On both coding and non-coding strands, distributions show a large peak just upstream of the TSS, followed by a sharp dip at the TSS, and a second peak just downstream. In monkeys, the longest and most polymorphic G4CRs clustered more tightly near the TSS than shorter and less polymorphic ones, as we saw for humans. For other high G4-density organisms (mouse, rat, dog, and chicken) this trend was still present, albeit to a slightly lesser extent. For these species, the shortest and least polymorphic 50% of G4CRs were less clustered near TSS than longer and more polymorphic ones. But unlike humans and monkeys, the most tightly clustered G4CRs were not necessarily the top 1% of G4CRs in terms of length or polymorphism. For species with less G4-rich genomes, essentially no clustering of G4CRs was observed near the TSS at all, similarly to what we observed with randomly shuffled human promoter sequences (*Figure 3.16*). Thus, in animal genomes with a high propensity to form G4s, G4CRs are highly enriched near the TSS, and longer and more polymorphic G4CRs are more enriched than shorter and less polymorphic ones. Interestingly in corn, the distribution of G4CRs was very different than that seen in animals. Clustering was much more pronounced on the non-coding strand compared to the coding one, and the distribution had a single peak, with no sharp dip at the TSS (*Figure 3.13*).

Coding Strand						
Species	# Promoters	# G4CRs	#G4 motifs	Length	%guanosine	N_{tot}
H. sapiens	29598	61012	921566	25	65	4
M. mulatta	9575	17619	241709	24	65	4
M. musculus	25111	35503	857790	24	67	5
R. norvegicus	12601	15768	256013	24	66	4
C. familiaris	7545	20590	572821	25	67	6
G. gallus	6127	15340	375948	25	66	6
D. melanogaster	16972	2339	25600	22	63	3
A. mellifera	6493	464	6223	22	68	5
D. rerio	10728	800	37917	22	67	4
C. elegans	7120	290	16698	22	74	4
A. thaliana	10728	800	37917	22	67	4
Z. mays	17081	6165	188183	22	65	3
S. cerevisiae	5117	63	490	21	63	2
S. pombe	4802	24	79	21	64	2
P. falciparum	5597	13	46	20	76	2

Table 3.1: Statistics on analysis of the coding strand of eukaryote promoters. Length, %guanosine, and N_{tot} are reported as their median values.

Non-coding Strand						
Species	**# Promoters**	**# G4CRs**	**#G4 motifs**	**Length**	**%guanosine**	**N_{tot}**
H. sapiens	29598	50524	833421	24	65	4
M. mulatta	9575	14818	211580	24	65	4
M. musculus	25111	30783	862230	24	67	6
R. norvegicus	12601	13324	238733	23	67	5
C. familiaris	7545	18475	645158	26	67	6
G. gallus	6127	11463	322532	25	67	5
D. melanogaster	16972	2726	26668	22	63	3
A. mellifera	6493	373	6851	22	70	6
D. rerio	10728	790	49703	22	68	4
C. elegans	7120	616	33230	22	74	5
A. thaliana	10728	790	49703	22	68	4
Z. mays	17081	10195	360194	23	65	4
S. cerevisiae	5117	79	584	21	64	2
S. pombe	4802	56	312	21	63	2
P. falciparum	5597	14	84	18	67	1

Table 3.2: Statistics on analysis of the non-coding strand of eukaryote promoters. Length, %guanosine, and N_{tot} are reported as their median values.

3.4.6 Analysis of G4CRs with validated biological activity.

There are several oncogene promoters where ample experimental evidence exists to show that G4 formation is correlated with gene expression. It is therefore of interest to examine in some detail the G4CRs from these genes and compare their characteristics to those of G4CRs, in general. A useful tool in this regard is the calculation of multiplicities, i.e. the number of distinct G4 structures that include a particular G residue in the core, as these provide a measure of polymorphism at the single nucleotide level. As shown in *Figure 3.8*, about 20% of Gs in G4CRs have a multiplicity of 1, which matches our observation that 18% of G4CRs form only a single G4 structure. The median multiplicity of G residues in G4CRs is 4, while about 1% of G's have multiplicities of over 100.

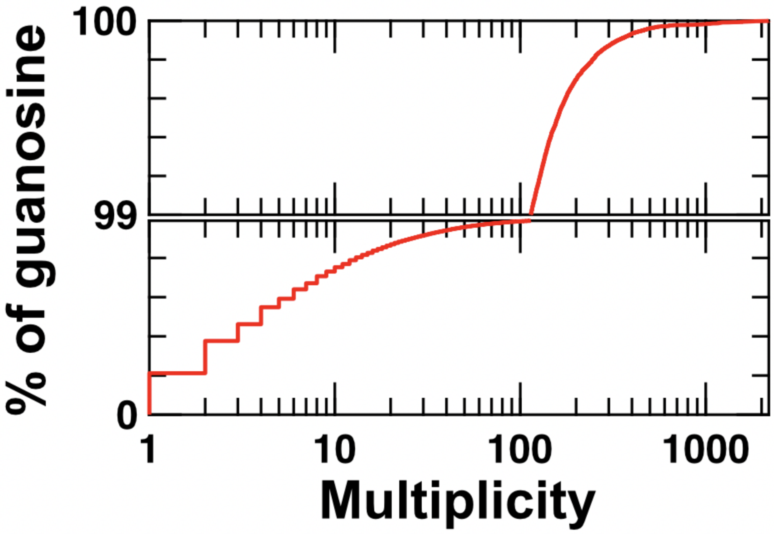


Figure 3.8: Cumulative plot of multiplicity for guanines participating in at least one G4 structure in a G4CR. Multiplicity is plotted on a logarithmic scale. The bottom panel represents the first 99% of multiplicities whereas the top panel represents the top 1% of multiplicities.

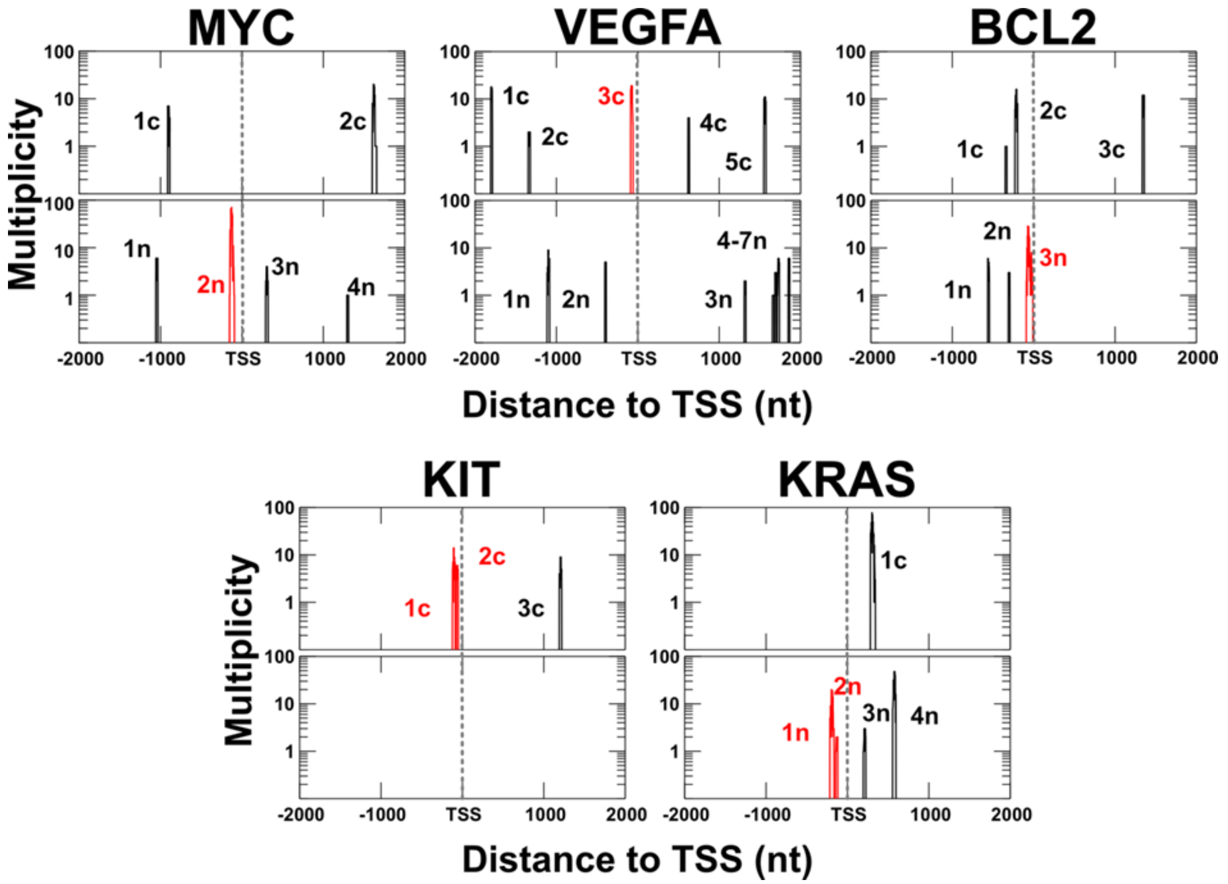


Figure 3.9: Multiplicities of G residues located within promoter regions of biologically relevant genes discussed in this paper plotted as a function of position relative to the transcription start site (TSS). The coding strand is shown as the top panel and all G4CRs are labeled with a lowercase c. The non-coding strand is the bottom panel and all G4CRs are labeled with a lowercase n. G4CRs were numbered from left to right, and multiplicity is plotted on a logarithmic scale. The G4CR corresponding to the most discussed in the literature is shown in red for each gene.

MYC								
G4CR	**Length (nt)**	**G-Content (%)**	**N_{tot}**	**N_{tand}**	**Distance to TSS**	**T_m (min)**	**T_m (median)**	**T_m (max)**
1c	27	74	7	1	-887	50.9	54.5	60.7
2c	54	69	21	1	1654	50.6	53.0	60.7
1n	25	64	6	1	-1034	56.3	61.2	64.5
2n	59	64	85	2	-94	50.7	60.3	84.1
3n	31	58	4	1	321	50.7	61.4	69.2
4n	21	62	1	1	1309	50.7	50.7	50.7

Table 3.3: Statistics on all the G4CRs found in the human MYC promoter. G4CRs found on the coding strand are indicated by "c" and G4CRs found on the non-coding strand are indicated by "n".

We first examined the promoter region of the *MYC* proto-oncogene, which is overexpressed in more than 50% of cancers [350]. The region -142 to -115 upstream of the TSS on the non-coding strand of DNA has been shown to fold into a parallel G4 structure[121, 386, 389]. Disruption of the G4 by mutation was shown to increase the expression of a reporter gene under the control of the *MYC* promoter by about 3-fold. Conversely, addition of a G4-stabilizing ligand decreases *MYC* gene expression, only when the promoter contains the G4 element[121]. Our analysis identified 6 G4CRs in the *MYC* promoter, 2 on the coding strand and 4 on the non-coding strand (*Figure 3.9* and *Table 3.3*). The previously studied G4 is contained in G4CR n2, which is the longest of the 6 (at 59 nt), contains the largest number of G4 isomers (at 85) and highest multiplicity values (at 70), and has the G4s with the highest values of T_m^{est} (at 84°C). The G4 isomer whose structure was solved by NMR spectroscopy and is commonly referred to as the "biologically relevant" conformation is predicted to have the highest melting temperature of the 85 isomers (tied with 3 others). Interestingly, the G4CR encompasses a longer region than is typically studied and allows a maximum of two G4s to form simultaneously. A recent report investigated this longer region experimentally, concluding that two G4s can, in fact, fold in tandem [402].

VEGFA								
G4CR	Length (nt)	G-Content (%)	N_{tot}	N_{tand}	Distance to TSS	T_m (min)	T_m (median)	T_m (max)
1c	23	74	20	1	-1786	50.7	57.4	80.7
2c	26	65	2	1	-1322	54.5	56.6	58.7
3c	30	70	19	1	-59	52.2	66.8	78.4
4c	19	79	4	1	634	55.0	59.5	64.1
5c	28	71	11	1	1575	50.6	53.0	58.7
1n	31	71	9	1	-1085	52.2	55.0	64.1
2n	19	74	5	1	-388	55.0	58.4	72.6
3n	22	64	2	1	1328	50.7	59.3	67.9
4n	22	64	1	1	1675	53.7	537	53.7
5n	20	70	3	1	1703	51.6	52..6	55.0
6n	27	67	6	1	1739	55.0	74.1	80.7
7n	19	79	6	1	1866	52.6	56.7	60.7

Table 3.4: Statistics on all the G4CRs found in the human VEGFA promoter. G4CRs found on the coding strand are indicated by "c" and G4CRs found on the non-coding strand are indicated by "n".

We next examined *VEGFA*, which is overexpressed and promotes tumour survival, growth, and metastasis in a range of human cancers[415, 416]. A region 50-85 nt upstream of the TSS forms a parallel G4[417, 418]. It is essential to *VEGF* expression[419], recruiting the transcription factor Sp1, which binds tightly to both duplex and G4 conformations[420]. Conversely, G4-binding ligands suppress *VEGF* expression[421]. We identified 12 G4CRs in the *VEGFA* promoter (*Figure 3.9* and *Table 3.4*). The one corresponding to the functional region (c3) is the second longest (30, as opposed to 31 nt for n1), has the second most G4 isomers (19 as opposed to 20 for c1), and has the third-highest maximum T_m^{est} (78, as opposed to 81 for c1 and n6). Notably, these other regions (c1, n1, and n6) are all more than 1 kb distant from the TSS. Of the 19 isomers we identified for the G4CR c3, the one we predicted to be the most stable is the one observed experimentally.

BCL2								
G4CR	**Length (nt)**	**G-Content (%)**	**N_{tot}**	**N_{tand}**	**Distance to TSS**	**T_m (min)**	**T_m (median)**	**T_m (max)**
1c	21	67	1	1	-329	55.0	55.0	55.0
2c	33	70	16	1	-198	50.6	54.6	58.4
3c	25	72	12	1	1359	55.8	60.6	64.9
1n	21	76	6	1	-544	50.7	55.3	56.5
2n	20	70	3	1	-292	51.6	52.6	55.0
3n	80	69	40	3	-11	50.7	57.5	69.9

Table 3.5: Statistics on all the G4CRs found in the human BCL2 promoter. G4CRs found on the coding strand are indicated by “c” and G4CRs found on the non-coding strand are indicated by “n”.

For *BCL2*, whose overexpression is linked to a large variety of cancers[422], deletion of a G4-forming region immediately before the TSS increases promoter activity[423], and when this region is placed upstream from a reported gene, G4-disruptive mutations increase expression while G4-binding ligands reduce it[397]. We found 6 G4CRs in the *BCL2* promoter region (*Figure 3.9* and *Table 3.5*). G4CR 3n, which corresponds to the previously identified region, is by far the longest (at 80 nt), has the by far largest number of G4 isomers (at 40), and has the G4 with the highest T_m^{est}. However, the dominant conformations determined in solution have long (>10 nt) hairpin loops, which are not captured by our algorithm[397, 422, 424].

KIT								
G4CR	**Length (nt)**	**G-Content (%)**	**N_{tot}**	**N_{tand}**	**Distance to TSS**	**T_m (min)**	**T_m (median)**	**T_m (max)**
1c	33	73	14	1	-90	50.7	52.6	76.5
2c	22	73	6	1	-51	50.7	56.7	66.8
3c	31	65	9	1	1220	52.2	55.6	64.5

Table 3.6: Statistics on all the G4CRs found in the human KIT promoter. G4CRs found on the coding strand are indicated by "c" and there are no G4CRs found on the non-coding strand of the KIT promoter.

The proto-oncogene *KIT*, which is associated with a large number of human cancers[425], has previously been found to contain 3 adjacent G4s, containing 3, 2, and 3 G-tetrads, respectively[390, 426]. Reporter gene assays have shown that disruptive mutations of the first G4 elevates gene expression, while in the second two, expression is suppressed, likely due to reduced recruitment of the transcription factor Sp1[390]. Furthermore, G4 stabilizing ligands reduce *KIT* expression in carcinoma cell lines[427]. We identified only 3 G4CRs in the *KIT* promoter region (*Figure 3.9* and *Table 3.6*). G4CR c1 encompasses the first G4 and part of the second, which with only 2-tetrads is not selected by our algorithm. This G4CR has a length (33), N_{tot} (14), and maximum T_m^{est} (76 °C) on par with the other functionally validated G4CRs examined here. The experimentally determined structure of the first G4 matches the most stable isomer identified for G4CR c1[428]. The G4CR c2 corresponds exactly to the third, previously studied G4, and its G4 isomer predicted to be the most stable by our algorithm matches the structure determined experimentally[426]. The three G4s of the *KIT* promoter are believed to be stabilized by higher order stacking interactions that are not captured by our algorithm[139, 390].

KRAS								
G4CR	**Length (nt)**	**G-Content (%)**	**N_{tot}**	**N_{tand}**	**Distance to TSS**	**T_m (min)**	**T_m (median)**	**T_m (max)**
1c	62	69	103	2	343	50.6	56.3	75.0
1n	52	65	25	2	-165	50.7	56.3	64.9
2n	26	58	2	1	-120	51.9	55.1	58.4
3n	31	71	3	1	225	53.0	63.4	66.8
4n	41	69	48	1	595	50.7	55.8	65.8

Table 3.7: Statistics on all the G4CRs found in the human KRAS promoter. G4CRs found on the coding strand are indicated by "c" and G4CRs found on the non-coding strand are indicated by "n".

The promoter of *KRAS* contains two G4-forming regions, termed the near-G4 and mid-G4[429, 430]. (A far-G4 region exists, but does not, in fact, form a stable G4 structure). The near-G4 region has been shown to recruit nuclear factors affecting gene expression[157]. In addition, binding of G4-ligands to the *KRAS* promoter decreases expression of a reporter gene, an effect that is primarily mediated by the mid-G4 region, rather than the near-G4[431]. Interestingly, the G4CR n1, which corresponds to the mid-G4 is longer (52 vs 26 nt), more polymorphic (25 vs 2 isomers), and has a higher maximum T_m^{est} (65 vs 58 °C), compared to G4CR n2, which corresponds to the near-G4 (*Figure 3.9* and *Table 3.7*). Considering all 62,791 G4CRs we identified in all gene promoter regions, the median length of a G4CR is 25, and median N_{tot} value is 4. Thus, among these biologically validated promoter G4s, there appears to be an over-representation of longer and more polymorphic G4CRs that contain G4s with shorter loops and fewer bulges. Applying this lesson to *KRAS*, the longest (62 nt), most polymorphic (103 isomers), and most stable (max T_m^{est} 75 °C) G4CR we identified is located about 300 nt downstream from the TSS on the coding strand (G4CR c1), and would therefore be transcribed into mRNA. There is already some evidence for mRNA G4 regulation of *KRAS* translation, involving several two-tetrad structures[432], but to our knowledge the G4CR c1 has not yet been investigated. We would conclude that characteristics of this G4CR would identify it as being a particularly interesting candidate for future study.

3.4.7 The GReg webserver

We have made our **G**4-Containing **Reg**ion (GReg) algorithm available as a webserver on our labs webpage (https://www.mcgill.ca/mittermaierlab/greg-webserver). This website allows users to enter an unlimited number of DNA sequences up to 32k nt in length. It provides both graphical and text output describing multiplicities (similar to *Figure 3.9*), lengths, positions, G-content, N_{tot}, and N_{tand} of all G4CRs present in the inputted sequences, as well as detailed listings of every putative G4 formed in each G4CR and its T_m^{est}.

3.5 Discussion

Polymorphism has long been recognized as an intrinsic property of G4s. In fact, some of the earliest algorithms (as well as later ones) used to enumerate G4s in genomic data explicitly recognized and eliminated multiple structures involving the same region of Gs to avoid overcounting[119, 133, 407]. Similarly, mutations are routinely used to trap single conformations and eliminate unwanted dynamics in structural studies[408]. However, these approaches discard a great deal of information on the nature, prevalence, and potential roles of G4 polymorphism. To some extent, the discussion thus far has been guided by our tendency to refer to the G4 structure as the biologically active unit. In cases with low polymorphism, it makes intuitive sense to assign a single conformation as the “biologically relevant” one, and account for simple dynamics in terms of “G-register exchange”[210] or “spare tire” motions[211]. However, nature provides many examples that do not fit neatly into this kind of single-structure, single function paradigm. An extreme example of this is minisatellite DNA, also known as variable number tandem repeats (VNTRs). These are repeating tandem units of more than two nucleotides that can be repeated several hundred times[433, 434]. The number of repeats differs from individual to individual and VNTRs are useful as genetic markers[433, 435]. When the repeating units are G-rich, VNTRs can form G4s. In fact, several characterized G4 structures are derived from short regions of VNTR DNA[391, 392]. In their entirety, VNTRs can produce a rich diversity of structures, in part because G-tracts repeated more than about 20 times in tandem can lead to a frustrated energy landscape with an enormous number of different folded G4 forms [180]. Furthermore, G4 folding has been linked to the genetic instability of VNTRs[436-438] while the number of repeats in VNTRs located in gene promoter regions is related to the expression levels of the corresponding proteins[439], echoing the influence of G4 folding on gene expression. In examples such as these, it makes little sense to ascribe the biological activity to a single folded G4. Instead, the concept of the G4CR, as defined in this study, provides an alternative definition of the biologically relevant unit that can be rigorously defined even when the structural folding landscape becomes exceedingly complex. In fact, our algorithm picked out several G4CRs that have previously been identified as VNTRs [440-442]. Notably, there is no clear division between short G4CRs that form a unique G4 and those that form thousands. The G4CRs we identified present continuums of length and polymorphism spanning several orders of magnitude. The concept of the G4CR provides a common framework for understanding the role of genomic G4s in all the different contexts in which they appear.

Our systematic survey quantitatively confirms the picture that has been emerging from studies of dozens of G4s with multiple isomers; polymorphism is a ubiquitous feature of G4 folding. In fact, only a minority of G4CRs (≈20%) contain a solitary G4 structure (Figure 4c). Even relatively short G4CRs that form only one G4 structure at a time can adopt up to hundreds of different folds sequestering different subsets of G residues in the core (Figure 5d). Furthermore, there are hints that higher degrees of polymorphism are related to biological function. We found that enrichment immediately upstream of the TSS is greatest for G4CRs that are longer and contain greater numbers of G4 isomers (Figure 6a,c). As well, some of the best studied G4CRs with validated activities in controlling gene expression are longer and more polymorphic than the median, as discussed above. This result opens up new possibilities for uncovering biological relevant sequences. More research is needed to clarify the relationships between G4CR characteristics and activity, however it already seems that longer, more polymorphic G4CRs may be good candidates for deeper study. As well, our bioinformatic search has uncovered a sizeable number of interesting regions that defy the paradigm of a single biologically relevant isomer. These have hundreds to thousands of different isomers, leading to a situation where polymorphism itself may play more of a role in governing the biophysics and activity of these DNA regions than the three-dimensional structure of any single isomer. Such highly polymorphic sequences are found in promoters of genes implicated in different forms of cancer and are listed in *Table 3.8*. Furthermore, some of the sequences we identify are unusually G rich (up to 83%) and unusually long (up to 369 nt). The complementary strands therefore contain regions that are equally C rich and equally long. Previous work has shown that i-motifs, four stranded structures formed by C-rich DNA are generally unstable at physiological pH [70], but that longer and more C rich sequences fold more readily [443]. I-motifs have putative roles in controlling gene expression, similarly to G4s[70]. Thus these exceptionally C-rich regions in gene promoters are highly interesting in their own right.

Gene	**Strand**	**Length (nt)**	**G-Content (%)**	**N_{tot}**	**N_{tand}**	**Distance to TSS**	**Diseases**
CAPN12	Non-coding	329	71	2363	13	-3	Pancreatic cancer[444]
USF2	Non-coding	90	80	907	4	-1	Small cell lung cancer,[445] breast cancer[446]
TTLL12	Coding	328	74	1311	14	3	Ovarian cancer[447]
ANO7	Coding	306	70	895	11	-48	Prostate cancer[448, 449]
RAE1	Coding	127	83	3134	7	-290	Lymphoid and epithelial tumors[450]

Table 3.8: A subset of G4CRs with high N_{tot}, N_{tand}, G-content, and length that are found close to the transcription start site for genes which are dysregulated in a number of different cancers.

It is useful to consider exactly how polymorphism may impact G4 function. We note that the isomers predicted by our algorithm are likely not equally populated. In most cases, the ensemble of structures will be dominated by the one or several most stable isomers, while many of the less stable ones may only be present at levels of a fraction of a percent at equilibrium. Nevertheless, the existence of many less stable isomers can still be functionally relevant. For example, we previously characterized G-register exchange among 4 structural isomers in a portion of the main G4CR from the *MYC* promoter. We found that even though a single isomer accounted for as much as 80% of the folded ensemble, the presence of 3 additional weakly populated isomers doubled the apparent folding equilibrium constant and increased the effective melting temperature by 3.4 °C due to entropic stabilization of the folded state[210]. Furthermore, the existence of the minor states increases the apparent folding rate by a factor of 2.5[351], since the folding of each of the four isomers represents a separate and parallel pathway to the folded state. This is particularly relevant to situations where G4 function relies more on kinetics (rapid folding) than on thermodynamics (high stability)[139]. While these effects are modest in the case of *MYC*, there are many G4CRs with orders of magnitude more isomers. Determining how the effects scale with the number of isomers is an interesting avenue for future research, as the impact on highly polymorphic G4CRs could be substantial. The existence of multiple weakly populated isomers is also relevant to the resilience of DNA to oxidative damage. Oxo-guanine residues located in a G4 core are not always accessible to base excision repair enzymes[451]. However, it has been shown that a fifth G-tract, if present, can replace the damaged G-tract, hence the “spare tire” terminology. This extrudes the oxo-guanine into a long loop where it becomes a substrate of the repair machinery [366]. Some of the polymorphism we have calculated using our algorithm falls explicitly into the category of spare tire dynamics with one complete G-tract replacing another. Other types of isomerization events might represent cryptic spare tire motions as well. In fact, the transition between any two isomers identified by our algorithm extrudes at least one G residue into a loop or flanking region, by definition. For example, the hypothetical oxidised G4 **gg**GG[oG]**t**GGG**t**GGG**t**GGG could isomerize to place the damaged G residue in a loop (**g**GGG**[og]t**GGG**t**GGG**t**GGG), without directly replacing one G-tract with another. While the relationship of G4 structure and accessibility to DNA repair enzymes is not yet fully understood [452], access to larger numbers of alternative structural isomers can be considered potentially protective against oxidative damage. As well, our analysis has shown that both higher G content and larger numbers of isomers are statistically correlated with the presence of more stable G4s (i.e. those with shorter loops and bulges, *Figure 3.5b,c*) We have examined this relationship in slightly more detail, comparing the distribution of T_m^{est} values as simultaneous functions of both %G and N_{tot}. We find that G4CRs with lower G content and fewer isomers than the medians (blue in *Figure 3.10a,b*) have lower T_m^{est} values than those with lower G content and more isomers (yellow). Similarly, G4CRs with higher G content and fewer isomers (orange) have lower T_m^{est} values than those with higher G content and more isomers (purple). Thus higher levels of polymorphism are directly correlated with G4s having shorter loops and bulges. We speculate that G-rich, highly polymorphic sequences are more evolutionarily accessible than unique, highly stable sequences such as $(G_3T)_3G_3$, implying that the evolutionary path towards stable G4s creates a large number of additional G4 isomers. Overall, both the functional (entropic stabilization, parallel folding pathways) and collateral (more stable individual G4 isomers) explanations are mutually consistent and both may be at play in explaining the ubiquity of G4 polymorphism.

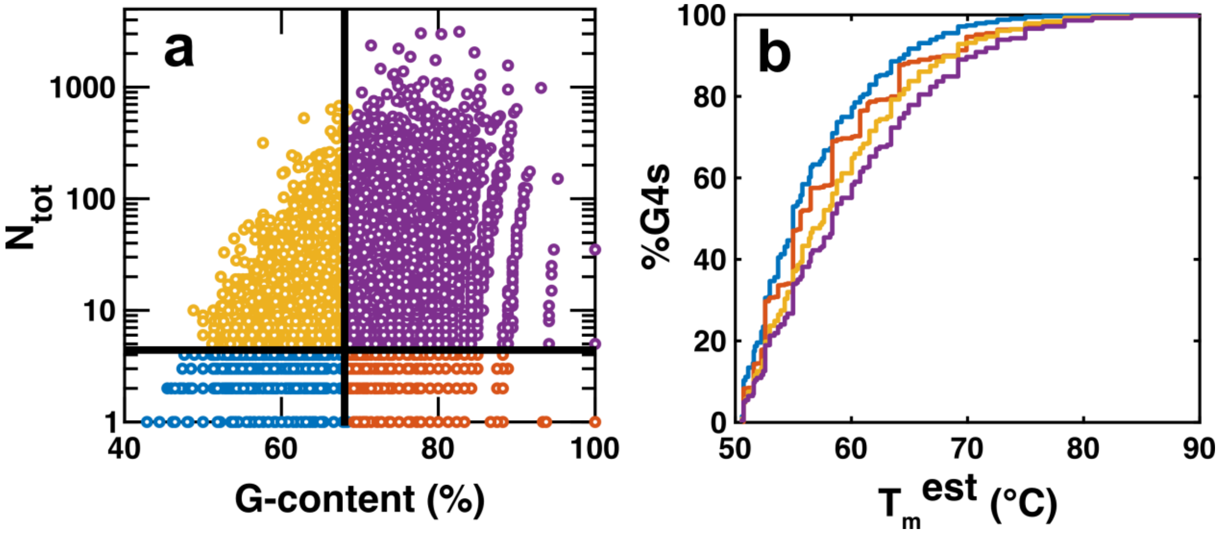


Figure 3.10: G-content and total number of isomers. a) Scatter plot of N_{tot} and G-content. N_{tot} is plotted on a logarithmic scale, whereas G-content is plotted on a linear scale. The graph is sectioned into four sections; G4CRs which contain less than the median value of N_{tot} and G-content (blue), G4CRs which contain more than the median value of N_{tot} but less than the median value of G-content (yellow), G4CRs which contain more than the median value of N_{tot} and of G-content (purple), G4CRs which contain less than the median value of N_{tot} but more than the median value of G-content (orange), b) Dependence of estimated melting temperatures for G4s contained in each of the different sections of a. Colours represented in b are the same as those in a.

It must be noted that our algorithm is quite conservative in the identification of G4 isomers. For example, it does not consider stabilizing interactions outside the G4 core, such as stacking of bases[398], or hairpin formation in loops and bulges[396, 397], or higher order G4/G4 interactions[399]. This inevitably led us to discard to G4 structures as unstable that might form readily in reality. We ignored G4s with only two tetrads[395], although these can be stabilized considerably by non-canonical interactions, and we discarded structures with loops longer than 7 nt, although much longer loops are sometimes observed in stable G4s[135, 411]. We also disallowed bulges containing G residues, since the isomer with the G in the core would be expected to be far more stable, even though bulged Gs are possible in principle and could be relevant in cryptic spare tire dynamics, as discussed above. The flip side is that all the isomers we predict are likely to be stable. Their presence may be obscured by more energetically favourable isomers that make up the bulk of the ensemble. However if all non-core G residues were replaced with mimics such as inosine[210, 351], it is highly probable that the 12 G residues predicted to be in the core would fold into a G4 with three tetrads. Thus, the very large numbers of isomers we calculate for many of the G4CRs likely underestimate rather than exaggerate the true number.

Additionally, given the prevalence of G4s in telomeres[453] and the promoter regions of oncogenes, and their ability to affect expression levels, targeting specific G4s with small molecule drugs has been an attractive new avenue towards developing cancer therapeutics[163]. However, the fact that most G4CRs can adopt multiple structures raises some fundamental questions about how this is best achieved. Selectivity is already an issue in targeting particular G4s, since unlike proteins, G4s possess similar cores and differ primarily in the identity of the loop residues. Even if a drug achieves specificity for a particular G4 with a unique three-dimensional structure, what is the likelihood that the target adopts alternative isomers and escapes binding? Conversely, to what extent might one of the many isomers of an off-target G4 bind the drug? Interfaces between tandem G4s have been proposed as more specific drug targets[401, 402]. However even here, polymorphism introduces complications. There are far fewer ways to fold a G4CR with the maximum number of tandem G4s than there are with a smaller number of G4s[180]. Therefore, most of the possible isomers formed by a G4CR lack the targeted interface, although this is potentially compensated by stabilizing higher-order interactions, which would increase the relative populations of the tandem structures. Thus, analyzing drug targets in terms of polymorphic G4CRs rather than as unique G4 structures seems a surer way to account for the complexity of this challenge.

Ultimately, the bioinformatic tools we have developed here and made publicly available on the GReg webserver will facilitate a better understanding of how G4 polymorphism intersects with biological function and evolution. They simplify the identification of longer and more polymorphic G4CRs, which we have found are positively associated with biological activity. They make it easy to identify the boundaries of G4 containing regions, thereby ensuring the full sequence can be analyzed. For example, our algorithm identified a longer region in the *MYC* promoter than had been typically studied; this longer region was just recently found to have structural relevance[402]. By mapping out the polymorphic landscape of G4CRs, our tools will make it easier to predict and interpret the effects of genetic variation on G4 formation, both between species and between individuals. This same information may also help us to better direct G4 ligands specifically to one stretch of G4-containing DNA over another. Finally, this bioinformatic analysis has underlined some fundamental unanswered questions regarding G4 containing regions. Why do some putative G4s occur in isolation and form a single unique structure, while others occur in the context of contiguous G4-forming regions hundreds of nucleotides long with potentially many thousands of alternate folded states? How do the biophysical properties of these various types of G4-containing regions differ and how does this impact biological function? We hope that the bioinformatic tools reported here will serve as a steppingstone towards answering these questions and developing a deeper biophysical understanding of G4s and their activity.

3.6 Materials and methods

3.6.1 The GReg algorithm

3.6.1.1 Generating possible G4 sequences (get_GQs)

Initially, a series of matrices ($G4_{matrix}$) were generated that encoded, in their rows, every possible G4 configuration, where core G positions were indicated by values of 1, loop positions were assigned 0, and bulged positions were assigned a value greater than 12 (13 was used here). The number of columns of a particular G matrix was given by the total number of loop and bulged residues plus 12 (core positions). We used a maximum loop length of 7 and bulge length of 3. Only one bulge was considered, since two or more bulges reduced the T_m^{est} below the selected threshold of 50°C. Thus 21 different G matrices were generated, containing between 15 columns (for three loops of 1 and no bulge) to 36 columns (for three loops of 7 and a bulge of 3). The rows were generated combinatorically, combining every possible length of the three loops with every possible bulge position and length. This gave a total of 8,575 different sequences: $7^3 \times 8 \times 3 + 7^3$ for (three loops of 1-7 residues)×(two possible bulge locations per G-tract)×(bulge lengths of one to three) + (the possible G4 sequences with no bulge). Next, the T_m^{est} corresponding to each row of each $G4_{matrix}$ was calculated using *Equation 3*, and only rows with $T_m^{est} \geq$ 50 °C were kept for further analysis resulting in 699 different G4 motifs. Thus each row of each $G4_{matrix}$ corresponded to a pattern of core, loop, and bulged residues that would produce a G4 structure with an estimated melting temperature greater than 50°C. For example, a hypothetical G4_{matrix} with 18 columns is shown below. The first row corresponds to a G4 with no bulges and loop lengths of [1,3,2]. The second row corresponds to a G4 with loop lengths of [1,1,2] and a bulge of two residues between the second and third G in the second G-tract. The third row corresponds to a G4 with no bulges and loop lengths of [1,4,1].

$$G4_{matrix} = \begin{pmatrix} 1\;1\;1\;0\;1\;1\;1\;0\;0\;0\;1\;1\;1\;0\;0\;1\;1\;1 \\ 1\;1\;1\;0\;1\;1\;(13)\;(13)\;1\;0\;1\;1\;1\;0\;0\;1\;1\;1 \\ 1\;1\;1\;0\;1\;1\;1\;0\;0\;0\;0\;1\;1\;1\;0\;1\;1\;1 \\ \vdots \end{pmatrix}$$

3.6.1.2 Converting original DNA sequence (convert_Seq)

Next, the DNA sequence of interest was converted into a column vector wherein all G residues were indicated by 1s and all other residues were indicated by 0s. Note that this encoding is different from the one used in the $G4_{matrix}$, above. The sequence was then split into possible G4CRs (pG4CRs), by removing all stretches of 0s longer than the maximum loop length (7), and retaining the intervening regions. Any pG4CR with a sum less than 12 were discarded, as these contain fewer than 12 G residues. For example, the following DNA sequence:

$$T\ C\ T\ A\ C\ A\ A\ A\ \boldsymbol{G\ G\ G}\ T\ \boldsymbol{G\ G\ G}\ A\ \boldsymbol{G}\ T\ \boldsymbol{G\ G\ G\ G}\ T\ \boldsymbol{G\ G\ G}\ T\ A\ T\ C\ T\ C\ A\ T\ \boldsymbol{G\ G}\ A\ \boldsymbol{G}\ C\ T\ C\ T\ T\ A\ C\ A$$

would be split into 2 pG4CRs:

$$pG4CR_1 = [1\;1\;1\;0\;1\;1\;1\;0\;1\;0\;1\;1\;1\;1\;0\;1\;1\;1]^T$$
$$pG4CR_2 = [1\;1\;0\;1]^T$$

where the second pG4CR would be immediately discarded.

3.6.1.3 Analyzing possible G4CRs (analyze_Seq)

The number of possible G4 structures in each pG4CR was evaluated by matrix multiplication with each $G4_{matrix}$ in turn. A sliding window with the same number of elements as the number of columns in the $G4_{matrix}$ was extracted from the pG4CR and multiplied by the $G4_{matrix}$, producing a column vector ($GReg_{vector}$) with the same number of rows as the $G4_{matrix}$. The value of each element of the $GReg_{vector}$ indicates whether or not the window contains G residues at all core positions indicated by the corresponding row in the G4$_{matrix}$. A value of 12 indicates that Gs are present at all core positions. A value less than 12 indicates that some core positions do not contain G residues in the window. A value greater than 12 indicates that a G is present in a bulged position, as illustrated below.

$$G4_{matrix} \cdot pG4CR = GReg_{vector}$$

$$\begin{pmatrix} 1\,1\,1\,0\,1\,1\,1\,0\,0\,0\,1\,1\,1\,0\,0\,1\,1\,1 \\ 1\,1\,1\,0\,1\,1\,(13)\,(13)\,1\,0\,1\,1\,1\,0\,0\,1\,1\,1 \\ 1\,1\,1\,0\,0\,0\,1\,1\,1\,0\,0\,1\,1\,1\,0\,1\,1\,1 \\ \vdots \end{pmatrix} \cdot [1\,1\,1\,0\,1\,1\,1\,0\,1\,0\,1\,1\,1\,1\,0\,1\,1\,1]^T = \begin{pmatrix} 12 \\ 25 \\ 11 \\ \vdots \end{pmatrix}$$

In the hypothetical example above (in which the length of the pG4CR happens to be equal to the width of the G4$_{matrix}$), only the [1,3,2] loop isomer (top row) is counted as a G4 structure in the pG4CR. The pG4CR contains G residues at bulge positions for the [1,1,2] loop isomer (second row) resulting in a value of 25. One core position in the [3,2,1] loop variant (third row) did not correspond to a G in the pG4CR, leading to a GReg$_{vector}$ element of 11. The analysis was repeated with the sliding window incremented across the entire pG4CR and for each G4$_{matrix}$. The total number of G4 isomers formed by a pG4CR was calculated as the total number of GReg$_{vector}$ elements that are equal to 12, summing over all positions of the sliding window and all G4$_{matrix}$s. pG4CRs that did not produce a single GReg$_{vector}$ element equal to 12 were discarded. pG4CRs which produce multiple GReg$_{vector}$ elements equal to 12, but which had sections of non-overlapping G4 motifs were split into separate G4CRs. The multiplicity of each guanine in the pG4CR was calculated by aligning each selected row of each G4$_{matrix}$ with the original sequence and summing over the columns.

3.6.2 Analysis of human promoters

The locations of all human promoters were downloaded from the Eukaryotic Promoter Database.[413] The promoter sequences were extracted from the GRCh38 build of the human genome with a window of -1999 to 2000bp surrounding the transcription start site. For the analysis of G4CRs in the human genome, 16528 unique promoters were analyzed using the first promoter labelled on the eukaryotic promoter database. Both coding and non-coding strands were analyzed together. Redundant promoters were not analyzed to avoid over counting G4CRs which appeared multiple times. When analyzing the positional dependence of G4CRs all 29598 promoters were analyzed, and coding and non-coding strands were analyzed separately.

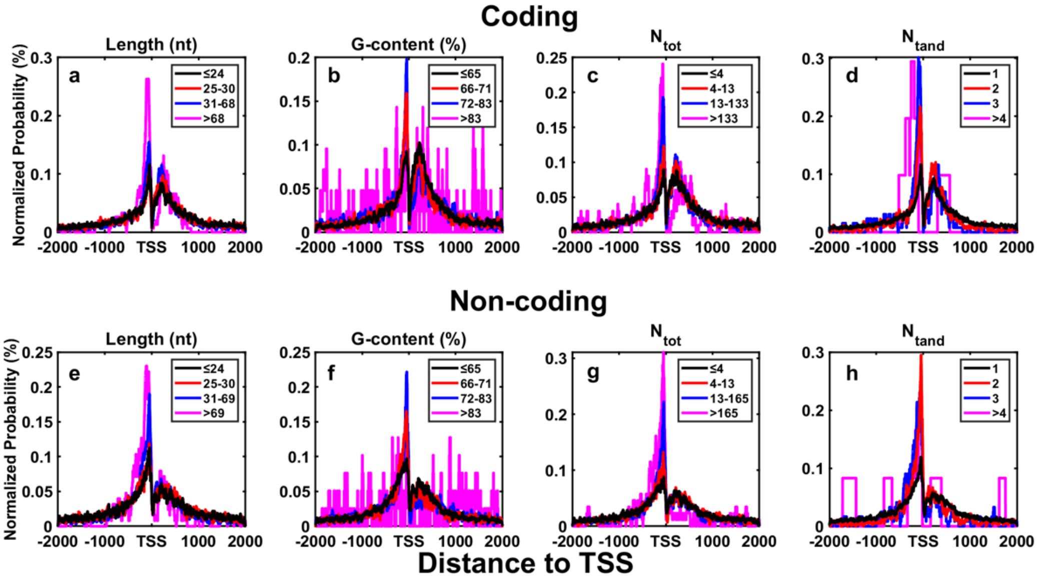
3.6.3 Scripting

The GReg algorithm and genome-wide searches were performed using in house MATLAB scripts, using MATLAB 2021a. Commented examples of each script for the GReg algorithm, an intuitive GUI for the GReg algorithm, and the python code used on the GReg webserver can be found at https://github.com/Christopher-Hennecker/GReg.

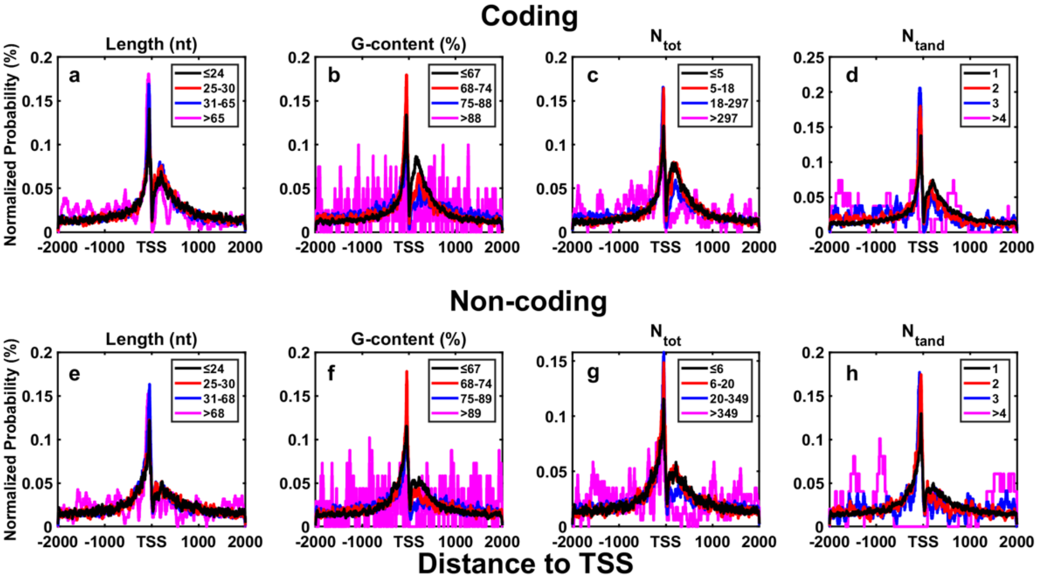
3.7 Supplementary Figures

Animals

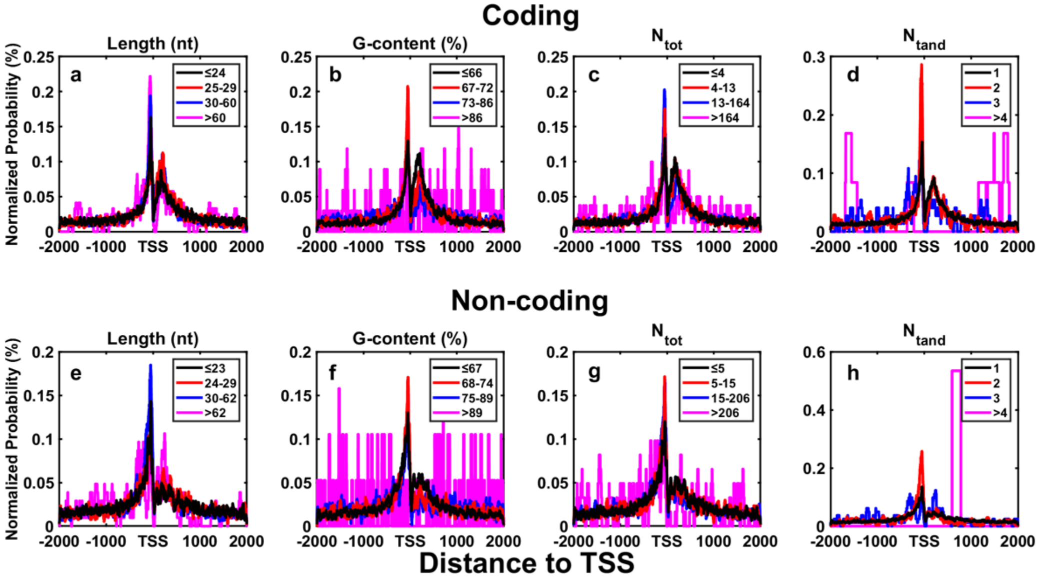
M. mulatta



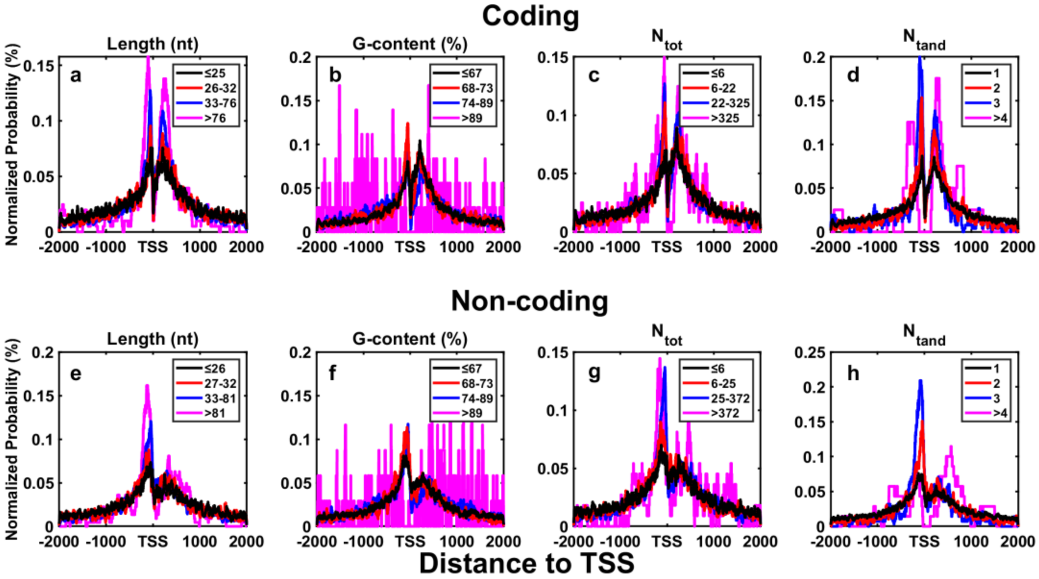
M. musculus



R. norvegicus



C. familiaris



G. gallus

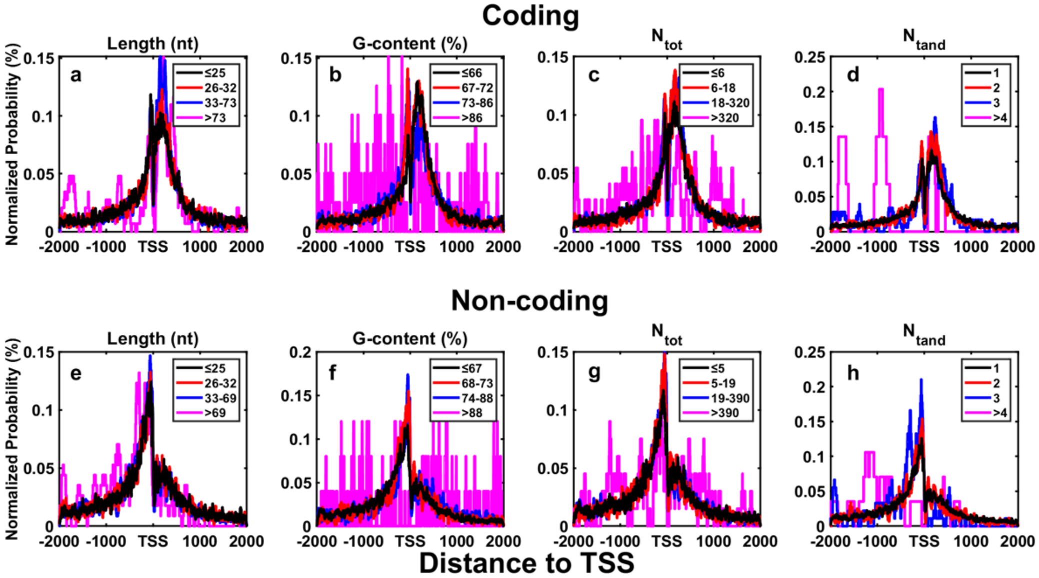
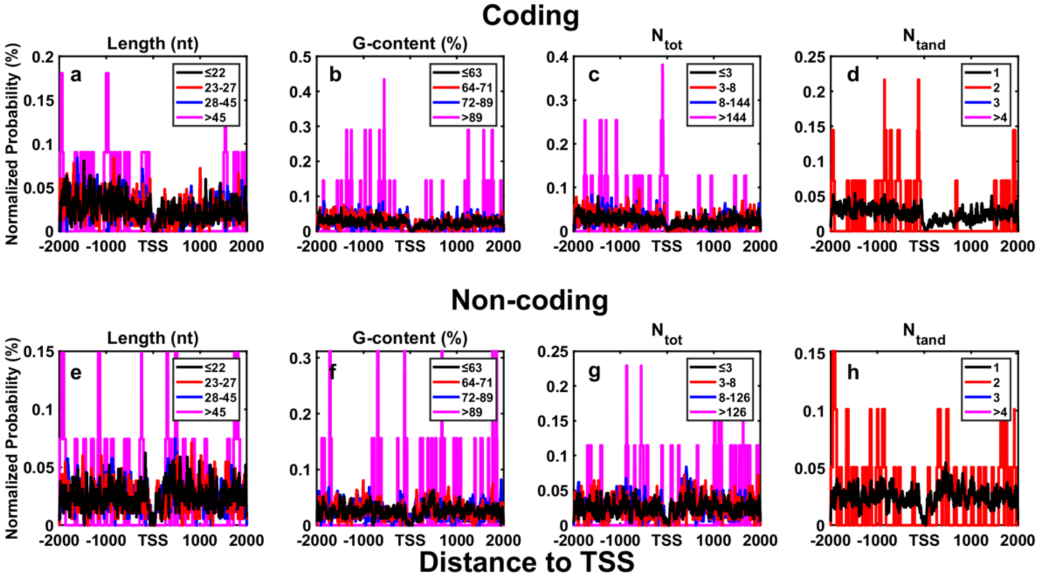


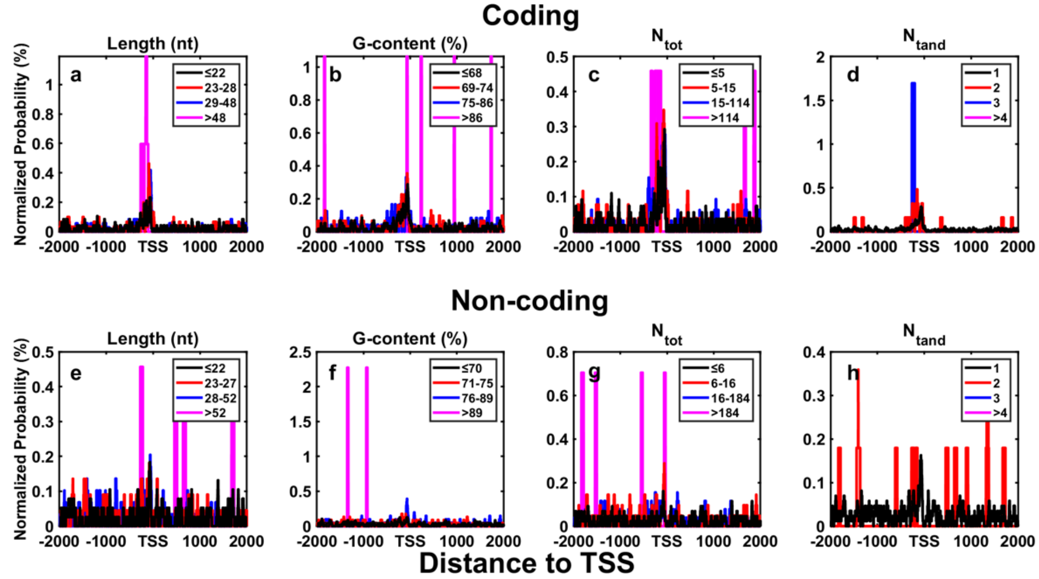
Figure 3.11: Distribution of G4CRs in animal promoters. Likelihood that a residue lies within a G4CR possessing certain characteristics is plotted as a function of the residue's position relative to the transcription start site (TSS). Colours and letters match those used in Figure 3.7.

Invertebrates

D. melanogaster



A. mellifera



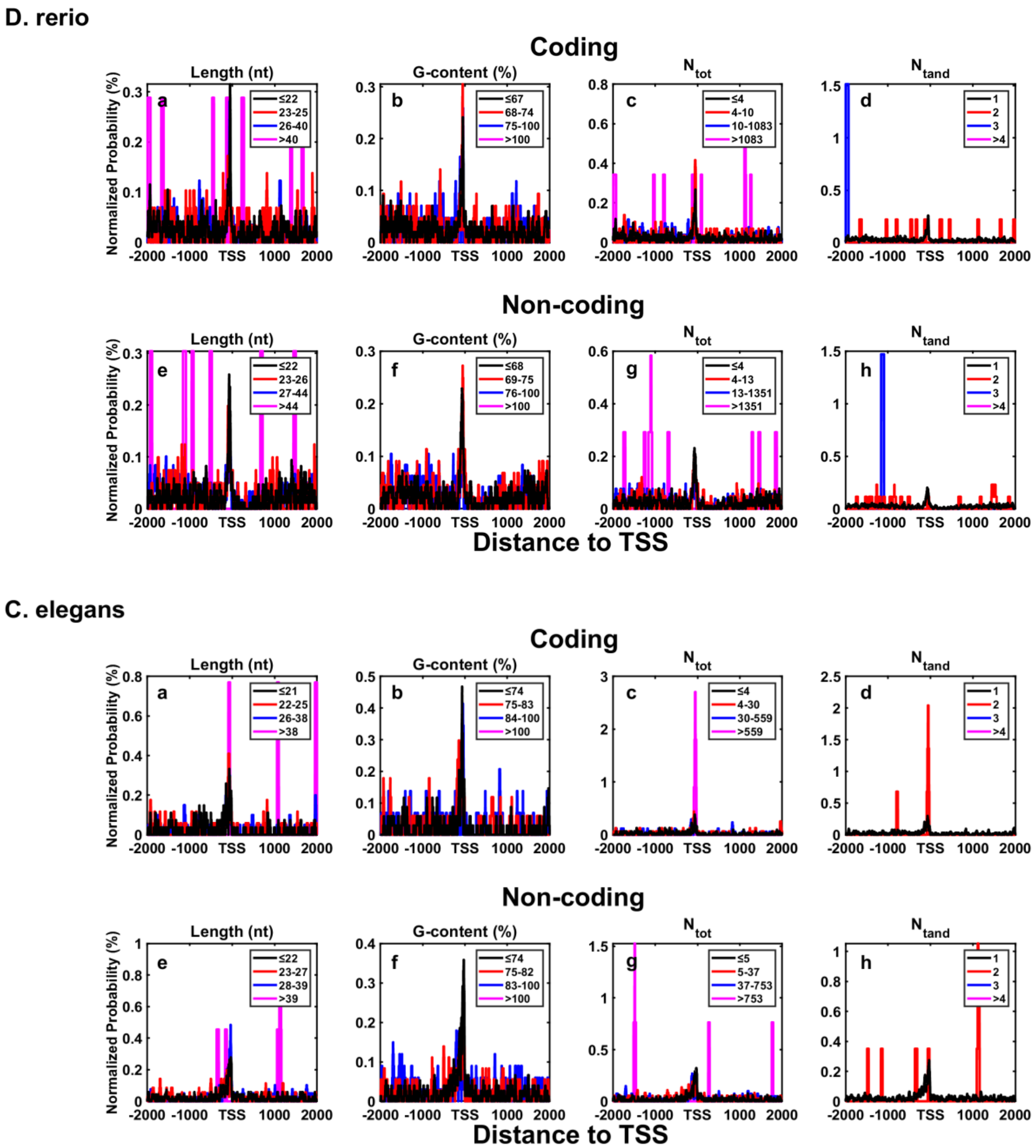


Figure 3.12: Distribution of G4CRs in invertebrate promoters. Likelihood that a residue lies within a G4CR possessing certain characteristics is plotted as a function of the residue's position relative to the transcription start site (TSS). Colours and letters match those used in Figure 3.7.

Plants

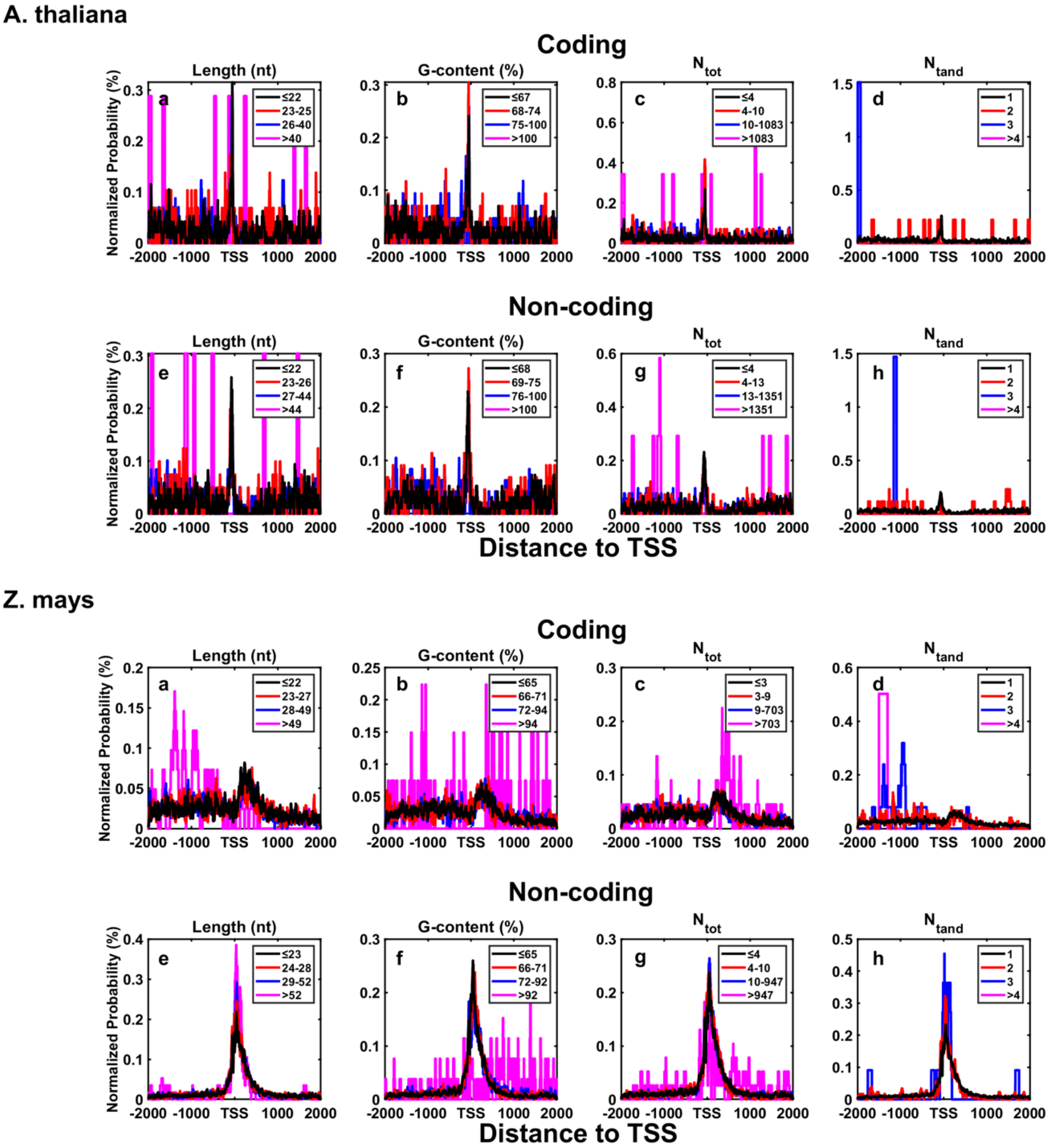


Figure 3.13: Distribution of G4CRs in plant promoters. Colours and letters match those used in Figure 3.7. Likelihood that a residue lies within a G4CR possessing certain characteristics is plotted as a function of the residue's position relative to the transcription start site (TSS).

Fungi

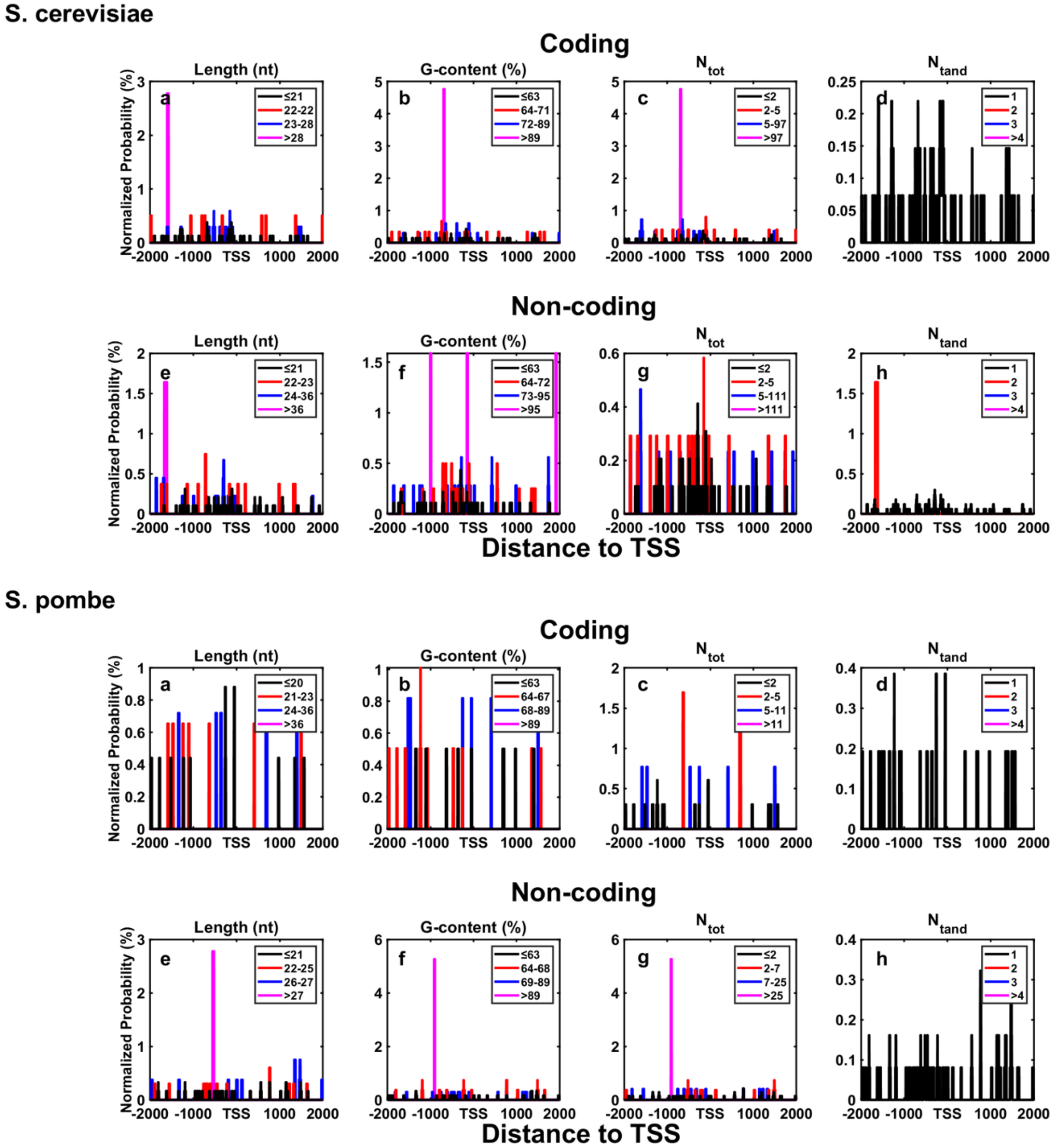


Figure 3.14: Distribution of G4CRs in fungil promoters. Likelihood that a residue lies within a G4CR possessing certain characteristics is plotted as a function of the residue's position relative to the transcription start site (TSS). Colours and letters match those used in Figure 3.7.

Protozoa

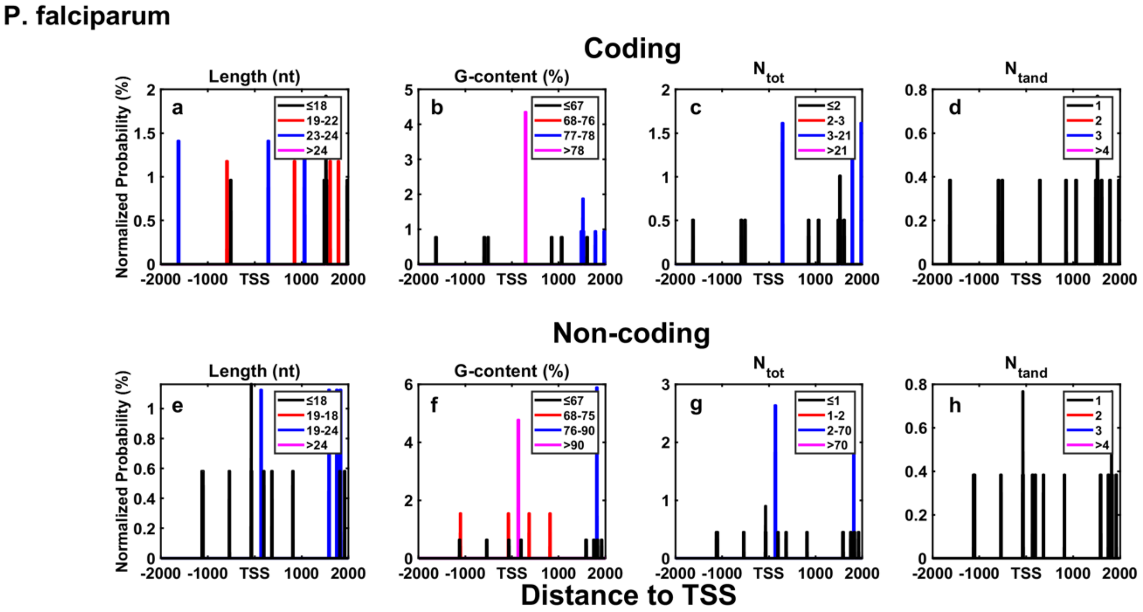


Figure 3.15: Distribution of G4CRs in protozoa promoters. Likelihood that a residue lies within a G4CR possessing certain characteristics is plotted as a function of the residue's position relative to the transcription start site (TSS). Colours and letters match those used in Figure 3.7.

Shuffled

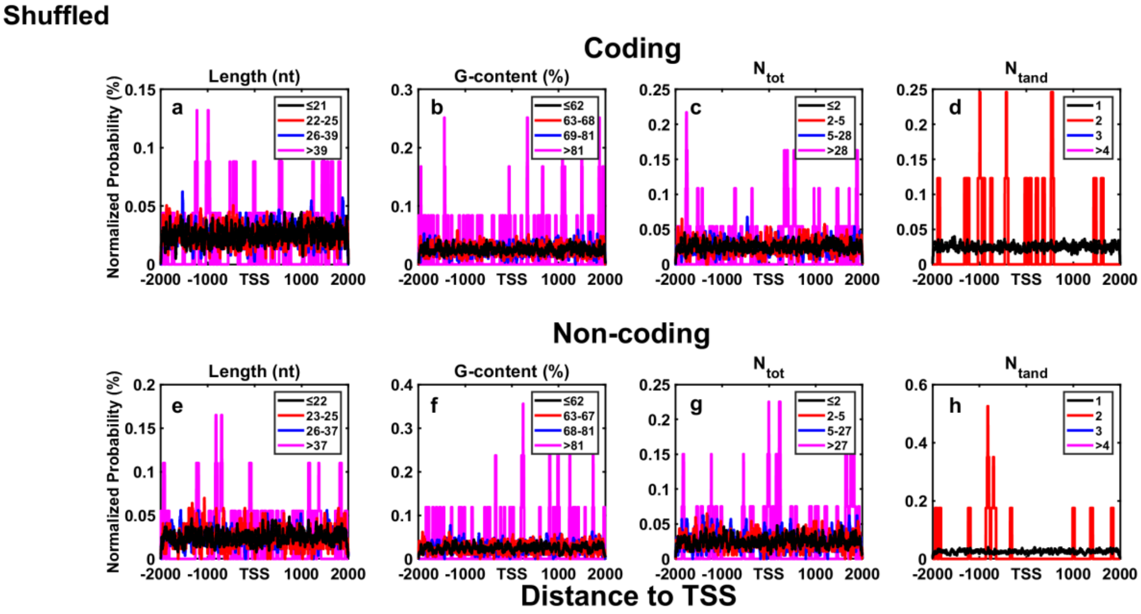


Figure 3.16: Distribution of G4CRs in shuffled DNA sequences. Likelihood that a residue lies within a G4CR possessing certain characteristics is plotted as a function of the residue's position relative to the transcription start site (TSS). Colours and letters match those used in Figure 3.7.

Chapter 4: Using transient equilibria (TREQ) to measure the thermodynamics of slowly assembling supramolecular systems.

4.1 Preface

The work in this chapter presents the first method to measure the thermodynamics of slowly assembling supramolecular systems reliably and robustly. It starts by using simulations to show how common methods fail to capture the thermodynamics of these systems, and how ignoring the effect of thermal hysteresis leads to completely incorrect characterization. It then introduces TREQ and discusses the theory behind how the method works and provides a guide for collecting and analyzing TREQ data. Finally, we use TREQ to study the copolymerization of polyadenosine DNA strands and the small molecule cyanuric acid (CA). We are able to understand the small-molecule loading efficiency of these fibres by measuring their stability as a function of CA concentration. We made the surprising discovery that about 33% of the CA binding sites in these fibres were unoccupied, and then developed a multivalent binding model to explain this behavior. In this chapter, Christophe Lachance-Brais performed nearly all of the experimental work, with the exception of the data found in *Figure 4.6*, *Figure 4.13*, and *Supplementary Figure 4.1*, which were data I collected. All of the code required to produce the simulations present in this chapter was developed by me, and I worked with Prof. Anthony Mittermaier to develop the mathematical models required to explain our results.

This chapter was adapted with permission from: Hennecker, C. D., Lachance-Brais, C., Sleiman, H., & Mittermaier, A. (2022). Using transient equilibria (TREQ) to measure the thermodynamics of slowly assembling supramolecular systems. *Science Advances*, 8(14), eabm8455.

4.2 Abstract

Supramolecular chemistry involves the non-covalent assembly of monomers into materials with unique properties and wide-ranging applications. Thermal analysis is a key analytical tool in this field, as it provides quantitative thermodynamic information on both the structural stability and nature of the underlying molecular interactions. However there exist many supramolecular systems whose kinetics are so slow that the thermodynamic methods currently applied are unreliable or fail completely. We have developed a simple and rapid spectroscopic method for extracting accurate thermodynamic parameters from these systems. It is based on repeatedly raising and lowering the temperature during assembly and identifying the points of transient equilibrium as they are passed on the up- and down-scans. In a proof-of-principle application to the co-assembly of polydeoxyadenosine containing 15 adenosines (polyA) and cyanuric acid (CA), we found that roughly 30% of the CA binding sites on the polyA chains were unoccupied, with implications for high-valence systems.

4.3 Introduction

Supramolecular chemistry is emerging as a rich source of diverse materials with novel and valuable properties. Potential applications range from drug-delivery and tissue regeneration to optical sensors and organic electronics[454]. This approach involves the non-covalent self-assembly of tens to thousands of monomeric units into larger structures with emergent physical properties that derive from both the structures of the individual components and their interactions and arrangement with respect to one another[455]. Reversible assembly has some distinct advantages compared to traditional covalent synthesis. The dynamic nature of supramolecular interactions allows bonds to break and reform leading to materials with self-healing properties. Furthermore, many supramolecular systems have the ability to generate multiple morphologies and sets of physical properties from a single set of building blocks with only small modifications of the assembly conditions[456]. Nevertheless, there are unique challenges associated with this approach. Chief among these is characterizing the products of a non-covalent assembly reaction. Much of the excitement surrounding supramolecular chemistry comes from the fact that there remains much to be understood regarding the relationships between the chemical structures of the monomeric units, the supramolecular architectures, and the emerging physical properties, and there is wide possibility for new and unexpected discoveries. However, this implies that the nature of supramolecular products is difficult to predict, and that rigorous structural and thermodynamic analyses are critical to advancing the field.

A variety of tools have been used to elucidate the structures produced by assembly, including atomic force, electron, and super-resolution microscopies, and solid-state NMR spectroscopy[457-459]. The stabilities of the assemblies are most commonly measured by thermal analysis. Most supramolecular structures dissociate when they are heated and reassemble when the monomer mixtures are cooled. This process can be quantified either by calorimetry[460] or by spectroscopically-detected melting and annealing[461, 462]. Detailed analyses of melting curves yield the enthalpies, ΔH, entropies ΔS, and free energies, ΔG, of assembly and shed light on the forces holding the supramolecular structures together[218]. This information is essential for determining structure/function relationships and the rational design and improvement of self-assembling systems[463, 464]. However, there exists a large class of supramolecular systems with extremely slow kinetics that only assemble or disassemble at useful rates when they are pushed far from equilibrium, i.e. under very highly stabilizing or destabilizing conditions. Common examples include amyloid fibrils, viral capsids, and a variety of self-assembling non-biological small molecules[246, 351, 463-477]. Interest in these kinds of slowly assembling supramolecular systems has grown in recent years, since they allow the size distributions of the resulting fibres to be tightly controlled[474, 476-478]. Current thermodynamic analyses rely on systems reaching equilibrium before the measurement is taken. In principle, this precludes thermodynamic analyses of slowly assembling systems, since equilibrium is not reached on practical timescales. Nevertheless, it is common practice in the supramolecular field to interpret non-equilibrium thermal data using equations derived for equilibrium systems, despite warnings in the literature that this is invalid[218]. Our mathematical simulations (see below) indicate this can lead to errors in reported thermodynamic parameters of >100% and equilibrium constants that differ from their true values by orders of magnitude. Thus a lack of reliable thermodynamic information for slowly-assembling systems is an impediment to the advancement of the supramolecular chemistry field.

We have developed a new experimental approach that can be performed using a standard temperature-controlled spectrophotometer and exploits transient equilibria (TREQ) to provide rigorous thermodynamic data on slowly assembling systems. Rather than waiting for the system to equilibrate (which can take days or weeks), the temperature is repeatedly raised and lowered, driving cyclic, non-equilibrium disassembly and assembly. We find that the system briefly passes through an instant of equilibrium on each up-scan and down-scan at which the rates of assembly and disassembly are equal. The temperatures and concentration values at which these moments of equilibrium occur can be clearly identified from the spectroscopic trace, allowing the full thermodynamic melting curve to be mapped in just a few hours.

As an example, we applied TREQ experiments to better understand the recently discovered co-assembly of polydeoxyadenosine (polyA) and the small molecule cyanuric acid (CA) into fibres whose biocompatibility and low cost make them promising candidates for tissue engineering and drug delivery[239]. A cross-section of the proposed structure (*Figure 4.1*) shows the adenosine of three different DNA strands hydrogen bonding to CA molecules in a continuous supramolecular helicene[241, 243]. We note that the ideal helicene structure has a 1:1 ratio of dA residues and CA molecules. We recently characterized the kinetics of polyA-CA fibre assembly using non-equilibrium melting methods[246]. Equilibration of the fibres near the melting point could take up to a month of constant instrument use. Using TREQ experiments, we determined the ΔG, ΔH, and ΔS values for adding a polyA chain to the end of a growing fibre in a single 10-hour experiment. By repeating these measurements at different concentrations of CA, we determined the minimum polyA:CA ratio necessary for assembly and made the surprising discovery that about 30% of the available CA binding sites are unfilled under our conditions. These results have implications for the future development of asymmetric systems involving components of very different valences, such as polyA and CA, and demonstrate the potential of the TREQ approach for learning about slowly assembling systems.

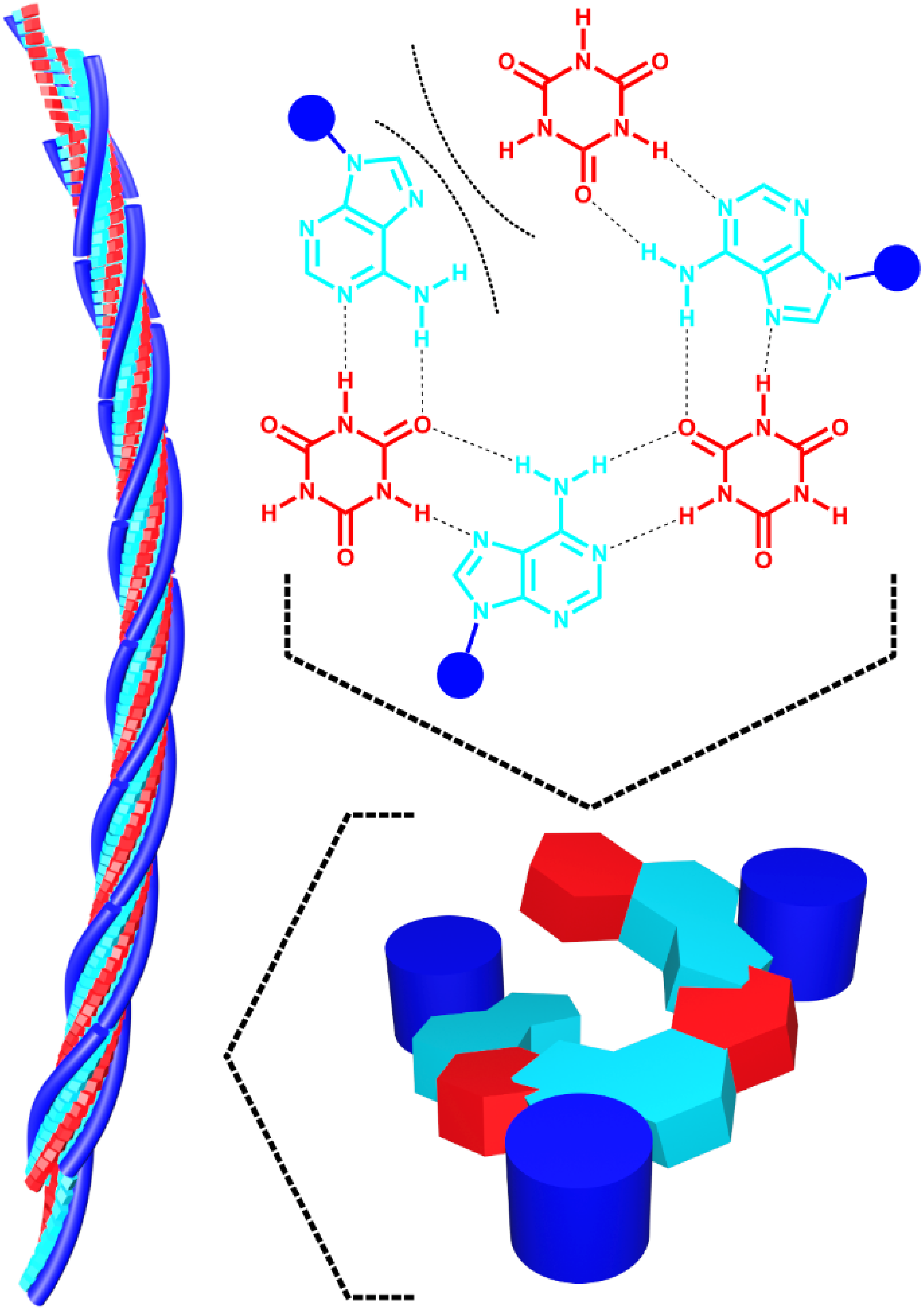


Figure 4.1: Putative supramolecular structure of polyA-CA fibres. Supramolecular fibres formed from the co-assembly of poly-adenosine strands and cyanuric acid (left). A cross section of a single hexameric helicene (right).

4.4 Results

4.4.1 Theory

Fibre assembly can be described by kinetic schemes such as the Goldstein-Stryer (GS) cooperative kinetic model:[239, 246, 248]

$$M_1 \underset{k_{n-}}{\overset{k_{n+}[M]}{\rightleftarrows}} M_2 \;\cdots\; M_{s-1} \underset{k_{n-}}{\overset{k_{n+}[M]}{\rightleftarrows}} \underset{\textbf{nucleus}}{\boxed{M_s}} \underset{k_{e-}}{\overset{k_{e+}[M]}{\rightleftarrows}} M_{s+1} \;\cdots\; M_N \underset{k_{e-}}{\overset{k_{e+}[M]}{\rightleftarrows}} M_{N+1}$$

Scheme 4.1: The Goldstein-Stryer cooperative kinetic model.

where M_N is a fibre containing *N* monomers. Association and dissociation of monomers from short oligomers less than the critical nucleus size, *s*, are described by the nucleation rate constants k_{n+} and k_{n-} respectively, while oligomers larger than *s* are described with the elongation rate constants k_{e+} and $k_{e–}$. An assembly parameter of great importance is the critical monomer concentration, $[M]_c$, at which the net rate of assembly or disassembly is zero, thus at this monomer concentration the system is at equilibrium. For rapidly-equilibrating systems, $[M]_c$ versus T curves can be measured directly by traditional melting or reannealing experiments and analyzed to obtain the enthalpies, entropies, and equilibrium dissociation constants for a monomer adding to the end of a fibre (ΔH_e, ΔS_e, and K_e, respectively) as well as the corresponding parameters for fibre nucleation[249]. For cooperative assembly, where nucleation is far less favourable than elongation, $[M]_c \approx K_e$ and a simplified analysis is commonly used; the maximum temperature at which fibres barely begin to form is identified as the elongation temperature, T_e, this temperature can be found by either fitting the elongation process or can be approximated from the assembly curve,[479, 480] while $[M]_c$ is equated to the total monomer concentration, c_T. The experiment is repeated several times at different c_T values (*Figure 4.2a*), where increasing c_T leads to an increase in T_e. A van ‘t Hoff plot of $\ln(c_T)$ vs $1/T_e$ is then used to extract values of ΔH_e and ΔS_e.

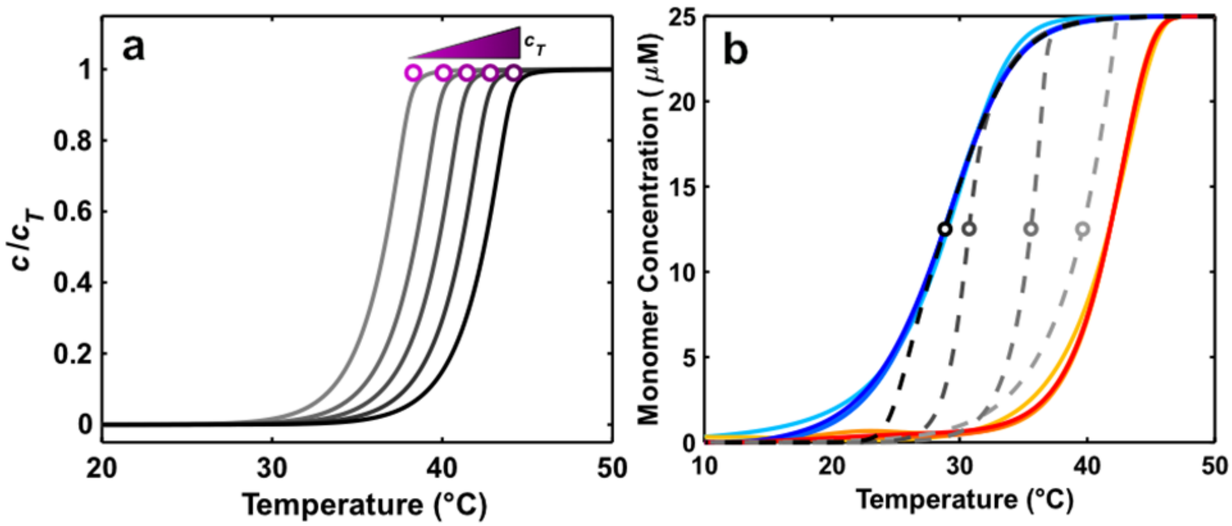


Figure 4.2: Traditional kinetic and thermodynamic analyses of supramolecular assembly. a) Simulated assembly curves for different total concentrations of monomer (c_T), increasing concentrations are shown as a gradient from grey to black, T_e values are shown as points using a purple gradient. b) Fibre assembly/disassembly simulated using the Goldstein-Stryer model (Scheme 4.1) and kinetic parameters that give similar melting and annealing curves (solid lines) with drastically different equilibrium curves (dashed lines). Heating curves are shown in red/orange and cooling curves are shown in blue/cyan. The offset between heating and cooling data is due to thermal hysteresis (TH). Simulation parameters are listed in Supplementary Table 4.1.

The situation is far more complicated for slowly assembling systems, such as polyA-CA fibres studied here. In these cases, the rate at which the system relaxes to equilibrium is far slower than available temperature scan rates, thus both folding (cooling) and unfolding (heating) occur out of equilibrium. The populations effectively lag behind the changing temperature such that the cooling and heating scans are offset, in a phenomenon known as thermal hysteresis (TH). Data for the up-scan lie to the right of the equilibrium $[M]_c$ vs T curve, and data for the down-scan lie to the left, as illustrated in *Figure 4.2b.* The resulting TH loops are rich in kinetic information, but are unsuitable for thermodynamic analyses, since the shape and location of the equilibrium curve is ill-defined, apart from the fact that it must lie somewhere between the heating and cooling scans[218, 246]. To illustrate, fibres obeying the GS assembly model can have very different thermodynamic parameters and equilibrium curves and yet produce nearly superimposable thermal hysteresis data (*Figure 4.2b*).

Nevertheless, data for systems exhibiting pronounced thermal hysteresis have frequently been analyzed as if they were obtained at equilibrium. Heating curves are typically used together with the concentration-dependent T_e approach described above[464, 473-475], although sometimes cooling scans have been employed instead[463, 470-472]. Interestingly, in their seminal 2003 review, Mergny and Lacroix point out that "*analysis of the concentration dependency of the denaturation profile only is seriously flawed*" and urge "*great caution about conclusions reached solely by analysis of the heating curves, a recurrent theme in the literature*", when thermal hysteresis is present[218]. To gain a clearer picture of the magnitude of the problem, we simulated TH data using GS parameters matching our polyA-CA system at different values of c_T and analyzed the resulting concentration dependent T_e values. Using heating scans, the extracted value of ΔH_e was 2.6-fold too large, using cooling scans it was 2-fold too small, and K_e values were incorrect by two to seven orders of magnitude (*Figure 4.3* and *Table 4.1*). In some studies[476, 477], different temperature scan rates produce superimposable heating data and it has been argued this validates their use in the concentration dependent T_e analysis. To test this hypothesis, we slightly modified our GS kinetic parameters to reproduce this effect and repeated the calculations. The resulting ΔH_e value was still about 1.8-fold too large (*Figure 4.3* and *Table 4.2*). Thus, commonly used thermal melting and reannealing experiments do not provide reliable thermodynamic data for slowly assembling systems. Notably our TREQ method reproduces the thermodynamic parameters in these simulations with a high degree of accuracy (*Figure 4.3, Figure 4.4* and *Table 4.1, Table 4.2*).

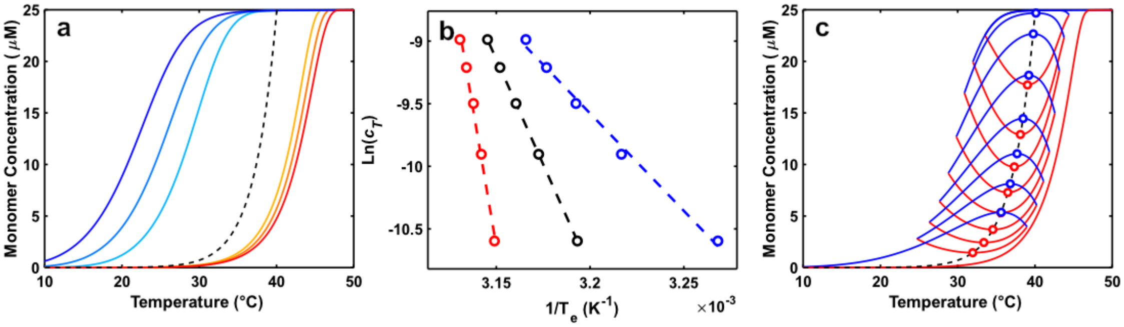


Figure 4.3: Thermodynamic analysis of simulated data. a) Simulated TH traces at 0.5, 1, and 2°C/min scan rates, showing hysteresis in both the heating and cooling traces. Cooling traces are shown as a cyan-blue gradient, heating traces are shown as an orange-red gradient, the true equilibrium trace is shown as the black dashed line. b) T_e analysis of the heating (red), cooling (blue) and equilibrium (black) curves at 25, 50, 75, 100, 125 uM total monomer concentration. The heating and cooling curves ran at 0.5°C/min. c) Simulated TREQ Analysis performed at 0.5°C/min, cooling traces are shown in blue and heating traces are shown in red. Extrema from each trace are shown as dots, and the true equilibrium is shown as a dashed black line.

Activation Energies		**Rate constants**		**Thermodynamic Constants**		**T_e Analysis Cooling**	**T_e Analysis Heating**	**TREQ Analysis**
E_{n+}	-14	k_{n+}	2.6 x 10^6	ΔG_e	-16	-12	-28	-16
E_{n-}	11	k_{n-}	5.9 x 10^3	ΔH_e	67	31	175	66
E_{e+}	-9	k_{e+}	1.7 x 10^6	ΔS_e	193	80	532	191
E_{e-}	58	k_{e-}	2.0 x 10^{-1}	K_e	1.2e-7	6.2e-6	3.5e-13	1.1e-7

Table 4.1: Kinetic and thermodynamic parameters used to simulate the data in Figure 4.3. TREQ values were found by fitting the elongation region of the transition to the model developed by Meijer and coworkers.[479, 480] E_{n+}, E_{n-}, E_{e+}, E_{e-}, ΔG_e and ΔH_e values are reported in kcal mol^{-1}, k_{n+} and k_{e+} are reported in M^{-1} min^{-1}, k_{n-}, and k_{e-} are reported in min^{-1}, ΔS_e is reported in cal mol^{-1} K^{-1}, K_e values are reported in M. Rate constants, equilibrium constants, and ΔG_e are reported at a reference temperature of 25˚C.

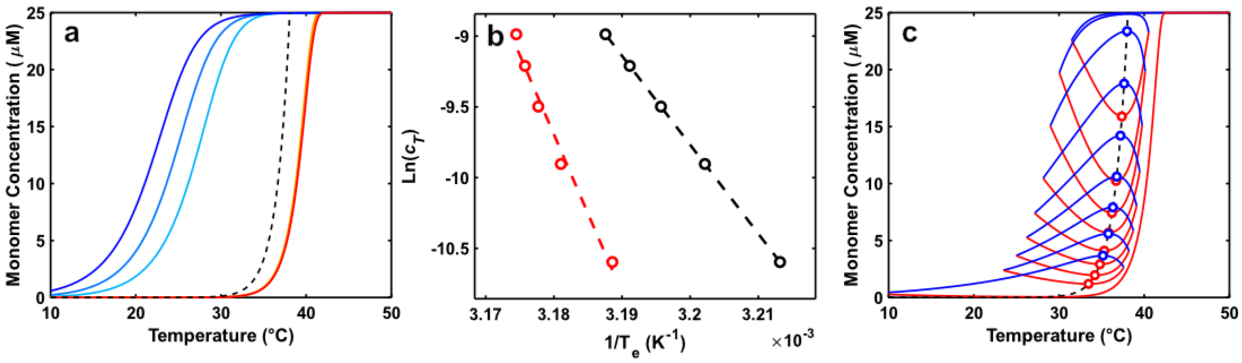


Figure 4.4: Thermodynamic analysis of simulated data with no hysteresis observed during heating. a) Simulated TH traces at 0.5, 1, and 2°C/min scan rates, showing hysteresis in only the cooling traces. Cooling traces are shown as a cyan-blue gradient, heating traces are shown as an orange-red gradient, the true equilibrium trace is shown as the black dashed line. b) T_e analysis of the heating (red), and equilibrium (black) curves at 25, 50, 75, 100, 125 uM total monomer concentration. The heating and cooling curves performed at 0.5°C/min. c) Simulated TREQ Analysis ran at 0.5°C/min, cooling traces are shown in blue and heating traces are shown in red. Extrema from each trace are shown as dots, and the true equilibrium is shown as a dashed black line.

Activation Energies		**Rate constants**		**Thermodynamic Constants**		**T_e Analysis Heating**	**TREQ Analysis**
E_{n+}	5	k_{n+}	1.5 x 10^5	ΔG_e	-19	-29	-19
E_{n-}	53	k_{n-}	1.5 x 10^2	ΔH_e	125	222	125
E_{e+}	-2	k_{e+}	9.8 x 10^5	ΔS_e	380	688	380
E_{e-}	123	k_{e-}	3.5 x 10^{-3}	K_e	3.6e-9	2.1e-13	3.6e-9

Table 4.2: Kinetic and thermodynamic parameters used to simulate the data in Figure 4.4. TREQ values were found by fitting the elongation region of the transition to the model developed by Meijer and coworkers.[479, 480] E_{n+}, E_{n-}, E_{e+}, E_{e-}, ΔG_e and ΔH_e values are reported in kcal mol^{-1}, k_{n+} and k_{e+} are reported in M^{-1} min^{-1}, k_{n-}, and k_{e-} are reported in min^{-1}, ΔS_e is reported in cal mol^{-1} K^{-1}, K_e values are reported in M. Rate constants, equilibrium constants, and ΔG_e are reported at a reference temperature of 25˚C.

Recent work from the Yamaguchi lab[469] has explored how the spectra of slowly equilibrating, self-assembling systems respond to repeated heating and cooling cycles[481]. Depending on the starting and ending temperatures and ramp rates, a rich diversity of shapes (thermal hysteresis loops) have been observed, providing qualitative information on the underlying assembly reactions. However, to date there has not been a straightforward way to extract quantitative thermodynamic information from these data.

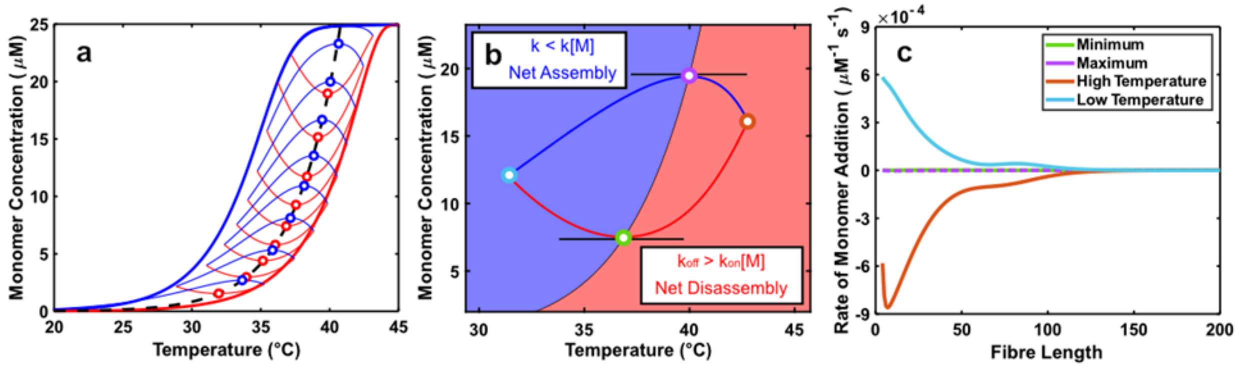


Figure 4.5: Analysis of a simulated TREQ experiment. a) Kinetic simulations of a typical hysteresis experiment (bold lines) and TREQ experiment (narrow lines). Cooling traces are shown in blue, heating traces are shown in red. The experiment begins by cooling from 45°C to 36°C, this is followed by the first up-scan (36°C to 44°C), a second down-scan (44°C to 35°C), a second up-scan (35°C to 43°C), a third down-scan (43°C to 34°C), etc. The equilibrium profile is shown as the dashed black line, with the extrema of each TREQ cycle shown as points. b) An isolated TREQ cycle: assembly occurs only in the blue shaded region; disassembly only occurs in the red shaded region. The interface of these two regions represents a system at equilibrium. Coloured points represent the position of calculated monomer flux in panel c. c) Calculated monomer flux of fibres for points shown in panel b, the horizontal extrema of the TREQ cycle have 100-fold less flux then either the high or low temperature values.

Our new TREQ approach uniquely fills this gap. In order to illustrate the fundamental principles, we performed kinetic simulations using the GS assembly model (*Scheme 4.1*) and parameters for polyA-CA fibres *(Figure 4.5a,* see Materials and methods*)*. The dashed black line indicates the equilibrium $[M]_C$ versus T curve, while the simulated heating and cooling scans give the left- and right-shifted blue and red curves, respectively. Thus the location of the true $[M]_c$ equilibrium curve is obscured by the thermal hysteresis.

The TREQ method is based on our discovery that repeatedly raising and lowering the temperature such that it repeatedly traverses the equilibrium curve reveals the precise locations of the hidden equilibria. Simulating TREQ data for polyA-CA assembly gives a series of concave-up and concave-down arcs on the heating and cooling scans, respectively (narrow red and blue curves *Figure 4.5a*). Strikingly, the $[M]_c$ values (black line) pass directly through the extrema (concentration maxima and minima) of the cooling and heating arcs. Thus experimentally determined extrema can be interpreted as a set of $[M]_c(T)$ values. The physical process underlying this behaviour can be understood as follows: for cooperatively assembled fibres, such as polyA-CA, equilibrium is reached when the rate of monomer addition to the end of a fibre ($k_{e+}[M]_c$) is exactly equal to the rate of monomer dissociation from the end of a fibre (k_{e-}), such that the net rate of fibre growth is zero (thus $[M]_c \approx K_e$)[249]. When $[M_1] < [M]_c$ there is net dissociation and $[M_1]$ increases with time, corresponding to the red region below the $[M]_c$ curve in *Figure 3b*. When $[M_1] > [M]_c$ there is net association and $[M_1]$ decreases with time, corresponding to the blue region above the $[M]_c$ curve. Every cooling scan starts in the red region with net dissociation (increasing $[M_1]$) and ends in the blue region with net association (decreasing $[M_1]$). As the temperature crosses the boundary where $[M_1]=[M]_c$, net fibre growth is zero, the arc is exactly horizontal, and the maximum is reached. Conversely, every heating scan starts in the blue region with decreasing $[M_1]$ and ends in the red region with increasing $[M_1]$. As the temperature crosses the $[M_1]=[M]_c$ boundary, the free monomer concentration is at a minimum. To validate this interpretation, we calculated the net rate of monomer addition to each length of fibre in the simulation. At the lower and upper limiting scan temperatures (orange and cyan), the rates of monomer addition and release are at least 100-fold greater than at the horizontal extrema of the heating and cooling arcs (green and purple) *(Figure 4.5c)*.

It must be noted that under certain conditions, polyA-CA co-assembly can deviate from the GS mechanism depicted in *Scheme 4.1*. For example, when polyA chains are mixed with CA at room temperature, fibres grow by a mixture of monomer addition (as described by the GS model) and coagulation (fibres joining end-to-end)[245]. The coagulation process introduces structural defects that can be backfilled with additional monomers. In contrast, when free monomers are gradually added to the system over a period of about an hour (through a process of proton dissipation), fibres grow almost exclusively by monomer addition and defects are rare[245]. Since fibre growth during a TREQ experiment occurs slowly as well, we would expect defects to also be rare in our experiments. In addition, polyA-CA chains are observed to form cable-like structures when formed under proton dissipation conditions[245]. We note samples subjected to TREQ heating and cooling cycles do not show evidence of cable formation by atomic force microscopy[244]. Nevertheless, it is worthwhile to discuss the potential effects of such higher order structures on the TREQ experiment. Cables and other forms of self-association may sequester fibre ends, possibly blocking monomer association and dissociation. However the termini of the cables are frayed into many individual polyA-CA fibres, where the processes of monomer association and dissociation can be safely assumed to be identical to those in isolated polyA-CA fibres[245]. The total rates of monomer uptake and release are both directly proportional to the number of exposed fibre ends[239, 246, 248]. Thus self-association would be expected to alter both rates by the same factor. In contrast, the value of $[M]_c$ and the thermodynamics of adding a monomer to a growing fibre do not depend on the number of exposed fibre ends. In the TREQ experiment, the shapes of the heating and cooling arcs depend on the kinetics of polymerization and depolymerization. Slower kinetics due to higher order structures that sequester fibre ends might be expected to produce flatter arcs. However, the locations of the extrema of the arcs are restricted to lying along the $[M]_c(T)$ curve, which is independent of the number of free ends. Thus, the TREQ experiment is expected to report the thermodynamics of forming individual fibres, but does not provide insight into whether or not fibres self-associate or the energetics of such processes.

4.4.2 Experimental validation of the TREQ experiment

It is not possible to experimentally confirm that TREQ data follow equilibrium values using the co-assembly of polyA and CA as a model system, since the process is so slow that the equilibrium curve is inaccessible to all other experimental techniques that could be used for cross-validation. We therefore turned to a much simpler system, the intramolecular folding of a DNA guanine quadruplex (G4) to experimentally test our approach. G4s are four-stranded, non-canonical nucleic acid structures composed of four tracts of consecutive guanine residues that form stacked, planar, guanine tetrads held together by Hoogsteen hydrogen bonds and coordination to monovalent cations[210]. Their folding reactions are effectively 2-state under many conditions,[210] and the timescale of folding can be tuned over several orders magnitude simply by adjusting the salt concentration. Heating and cooling scans collected for an intramolecular G4 (see Materials and methods) with a temperature ramp rate of 1 K min^{-1} are offset by about 6 degrees (*Figure 4.6a*), mimicking the TH observed for polyA-CA, albeit to a lesser extent. In contrast, data for the G4 obtained with a 0.1 K min^{-1} ramp rate are offset by only 0.5 degrees, meaning that they are close to equilibrium during both melting and refolding processes. This small amount of hysteresis, together with the simple folding mechanism, makes it possible to calculate the true equilibrium folding curve with a high level of confidence by performing a simple Arrhenius analysis (*Supplementary Figure 4.1*)[218]. We then performed TREQ analysis on the G4 sample with ±1 K min^{-1} ramp rates, by repeatedly raising and lowering the temperature over a window of roughly 5°C that shifted from (42.3-45.7) to (26.3-33.7) °C in 8 cycles while we monitored the spectroscopic absorbance at 295 nm. The high and low temperature absorbance regions were fitted to linear baselines and assigned 0% and 100% folded, respectively, giving the converted data shown in *Figure 4.6b*. Notably, the experimental equilibrium curve calculated from Arrhenius analysis of the TH experiments passes nearly exactly through the extrema of the TREQ heating and cooling arcs. The Arrhenius analysis of the 0.1 K min^{-1} ramp rates gave ΔH=148 ± 2 kJ mol^{-1} and ΔS=479 ± 6 J mol^{-1} K^{-1} (*Supplementary Figure 4.1*) and van ‘t Hoff analysis of the extrema of the TREQ experiment (described below) gave ΔH=146 ± 3 kJ mol^{-1} and ΔS=470 ± 10 J mol^{-1} K^{-1} (*Figure 4.6c*). Thus, the TREQ experiment closely reproduced the results of a traditional equilibrium melting measurement, in a special case where both measurements could be made on the same system.

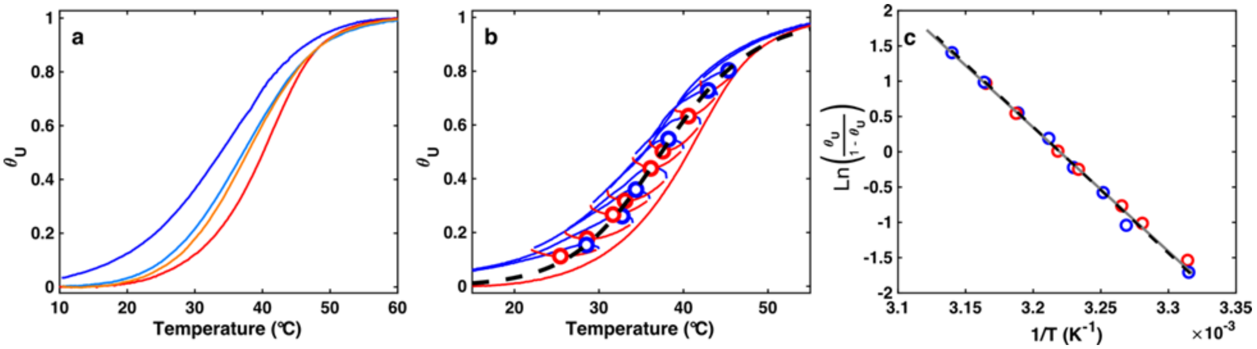


Figure 4.6: Experimental validation of the TREQ method using an intramolecular guanine quadruplex. a) Thermal hysteresis traces of intramolecular G4 folding. Heating and cooling scans at 1 K min^{-1} are shown in red and blue respectively. Heating and cooling scans at 0.1 K min^{-1} are shown in orange and light blue. b) TREQ data for intramolecular G4 folding, experimental traces obtained at a scan rate of 1 K min^{-1} are shown in blue for cooling and red for heating. Picked extrema are shown as circles, and the equilibrium curve found from analyzing 0.1 K min^{-1} hysteresis traces is shown as the black dashed line. c) Van 't Hoff analysis of experimental TREQ points, the line of best fit is shown as the grey solid line and the equilibrium curve found from analyzing 0.1 K min^{-1} hysteresis traces is shown as the black dashed line.

4.4.3 Guide for acquisition of TREQ data

The design of a TREQ experiment involves selecting an appropriate scan rate and choosing a series of temperature set points that define the cooling and heating scans, for example (cooling $T_1 \rightarrow T_2$), (heating $T_2 \rightarrow T_3$), (cooling $T_3 \rightarrow T_4$), (heating $T_4 \rightarrow T_5$), etc. Slower scan rates lead to better-defined TREQ maxima and minima but longer experiments. We would suggest scan rates on the order of 0.2 to 1 K min^{-1}. Slower assembly/disassembly kinetics require slower scan rates, although the success of the experiment is not particularly sensitive to the choice of scan rate. The selection of temperature set points is more critical as they will determine whether or not the system will pass through transient equilibria on the cooling and heating scans and generate a series of minima and maxima in the spectrophotometric data. We have developed a simple method for selecting the temperature set points that reliably produces high quality TREQ data. In the first step, a full cooling scan from maximum (T_{max}≈95°C) to minimum (T_{min}≈5°C) temperature is performed at the chosen scan rate, followed by a heating scan from T_{min} to T_{max}. The cooling scan will lie to substantially lower temperatures than the heating scan, due to thermal hysteresis, and the desired equilibrium curve lies somewhere between two. The temperatures at which the cooling scan generates 10, 20, 30, 40, etc. percent assembly are identified, yielding T_{C10}, T_{C20}, T_{C30}, T_{C40}, etc. respectively. The same analysis is performed for the heating scan, giving T_{H10}, T_{H20}, T_{H30}, T_{H40}, etc. We find that the oscillating cooling sequence $T_{max} \rightarrow T_{C10} \rightarrow T_{H10} \rightarrow T_{C20} \rightarrow T_{H20} \rightarrow \ldots \rightarrow T_{C90} \rightarrow T_{H90} \rightarrow T_{min} \rightarrow T_{max}$ reliably gives good quality TREQ data. The final heating scan is performed in order to obtain an adequate low temperature baseline. In principle, the oscillating heating sequence: $T_{min} \rightarrow T_{H90} \rightarrow T_{C90} \rightarrow T_{H80} \rightarrow T_{C80} \rightarrow \ldots \rightarrow T_{H10} \rightarrow T_{C10} \rightarrow T_{max} \rightarrow T_{min}$ also produces similar TREQ data, however we prefer to begin each experiment with a fully thermally denatured sample for the sake of reproducibility (i.e. the first sequence). Note that we chose 10% increments in assembly because our spectrophotometer software allows up to 20 scans to programmed in advance.

4.4.4 Guide for processing TREQ data

The first step in analyzing TREQ data is converting the raw spectroscopic output into fractions of folding or assembly, or equivalently, concentrations of unfolded or unassembled monomers. Thermal melting and annealing data typically have linear regions at temperatures below and above the transition.[218] The linear data points are identified by eye and fitted by linear regression to obtain the slopes, m_L and m_U, and y-intercepts b_L and b_U of the lower and upper linear regions, respectively. The fraction folded or assembled, θ(T), is then calculated as

$$\theta(T) = \frac{(m_L T + b_L) - S_{exp}(T)}{(m_L T + b_L) - (m_U T + b_U)} \qquad \textit{(Equation 4.1)}$$

where $S_{exp}(T)$ is the experimental absorbance (or fluorescence, or circular dichroism) measurement. The temperature-dependent monomer (M_1) concentration is calculated from θ(T) according to

$$[M_1](T) = \big(1 - \theta(T)\big)[M]_{tot} \qquad \textit{(Equation 4.2)}$$

where $[M]_{tot}$ is the total concentration of monomers in all assembled forms.

The reliability of the TREQ experiments depends on accurately pinpointing the extrema of the scans, i.e. choosing the values of T_{ext}, $[M]_{ext}$, where $[M]_{ext}$ is the maximum or minimum value of [M] on each cooling or heating scan and T_{ext} is the temperature at which this is reached. Simply picking the maximum- or minimum-valued datapoint of each arc is inaccurate. Due to instrument noise, roughly half of the measured points lie above the true curve and the other half lie below. The point with the largest (or smallest) value will therefore almost certainly over (or under) estimate the true $[M]_{ext}$ value. Furthermore, the convex and concave cooling and heating arcs are fairly broad, meaning that the temperature at which the single largest- or smallest-valued point occurs is strongly influenced by the stochastic nature of the experimental noise and will almost certainly differ from T_{ext}. We have developed two different approaches at two different levels of computational difficulty for accurately identifying the extrema. The first is simply to calculate a sliding window average, selecting the extreme value of the average as $[M]_{ext}$ and the centre of the window as T_{ext}. The second is to smooth the experimental data by fitting a curve to the data points and identifying the extremum of the fitted curve as T_{ext}, $[M]_{ext}$. We prefer to use an empirical polynomial function for smoothing rather than a mechanistic (eg GS model) calculation, since we wish to apply the TREQ approach even to systems where the precise kinetic mechanism is unknown.

To test the accuracy of these methods, we generated synthetic noisy TREQ data based on the GS fibre assembly model (see below) for which the true maxima and minima were known and compared these values with the results of the sliding window and polynomial smoothing calculations. The simulated TREQ data are shown in *Figure 4.7a* with dashed lines indicating the true (error-free) curves and black circles indicating synthetic data, sampled at 0.3°C intervals with 1% random noise. We found that a rolling average of 9 to 12 data points gave extrema close the true values. A van ‘t Hoff plot of the true data (dashed lines) and sliding average of ten points (purple circles) shows good agreement (*Figure 4.7b*). We repeated the calculation 1,000 times with a resampled selection of the data points, and the resulting standard deviations of the extrema are shown as error bars[482]. Next, we fit the upper halves of cooling curves and lower halves of heating curves to polynomial functions of different orders. We found qualitatively that polynomials of orders of about 5 to 15 delivered the best performance. Polynomials of lower orders were not able to faithfully reproduce the overall shape of the data and higher orders began to overly mimic the simulated noise. Extrema taken from fitted 5th-order polynomials (green circles) align with the true values even more closely than the sliding averages (purple circles). Bootstrapped uncertainties in the extrema were low; error bars are smaller than the symbols used in *Figure 4.7b*. We conclude that the polynomial smoothing approach provides a more accurate extraction of the extrema, however the performance of the sliding average approach is satisfactory and is simpler to apply if using standard spreadsheet software to analyze data. We have used polynomial smoothing throughout.

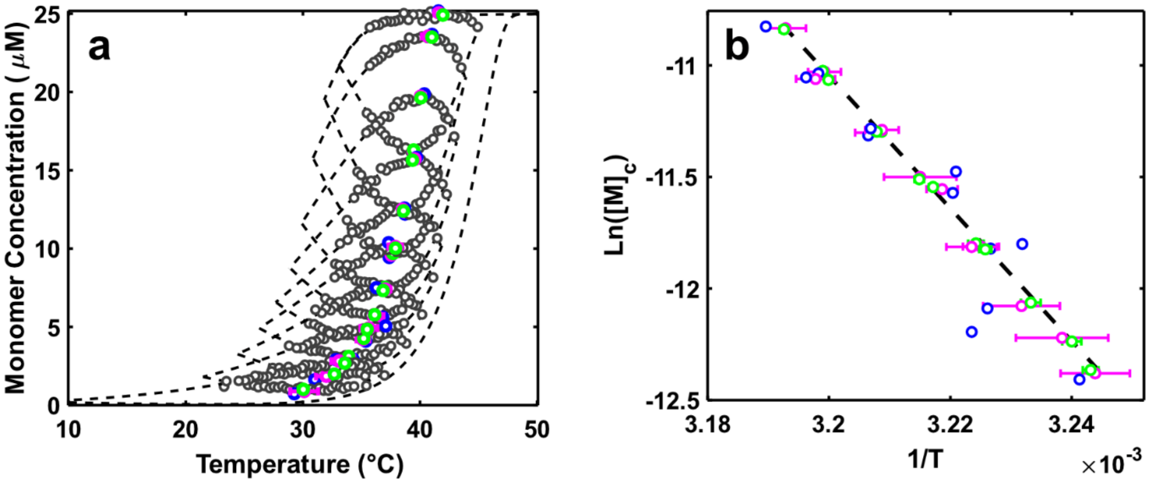


Figure 4.7: Analysis of simulated TREQ data with random noise. a) Dashed lines represent a simulated TREQ data, grey circles represent the top or bottom half of each trace with added random noise. b) van 't Hoff analysis of TREQ data, the dashed black line represents the true equilibrium curve. In both panels green circles represent extrema which were picked using a 5th order polynomial, magenta circles represent extrema which were picked using an averaging window of 10 data points and blue points represent extrema which were picked from raw data (i.e. max and min datapoints). Error bars in both monomer concentration and temperature are shown in both panels but are often smaller than the symbols.

4.4.5 Analysis of experimental TREQ data for polyA-CA coassembly

We performed a TREQ experiment on a mixture of CA and polyA chains (*Figure 4.8a*). The lower and upper absorbance regions were fitted to linear baselines and assigned 100% and 0% folded, i.e. $[M_1]$ = 0 and 25 μM, respectively. The fraction of folded monomers at a given temperature was taken as the difference between the measured absorbance and the lower baseline, divided by the difference between the upper and lower baselines (*Equation 4.1* and *Equation 4.2*), as is typically done in spectroscopic analyses of supramolecular assembly[218, 246, 463, 464, 471, 473, 476, 477]. The converted data are shown in *Figure 4.8b*, with blue and red indicating cooling and heating, respectively, and open circles placed at the extrema. These experimental arcs have a remarkable similarity to the calculations shown in *Figure 4.5a*. The y- and x-values of the extrema correspond directly to critical monomer concentration, $[M]_c$, and temperature pairs. As discussed above, $[M]_c$ values are equivalent to the equilibrium dissociation constant, K_e, for adding a polyA to the end of an elongating fibre, for this system. A van ‘t Hoff plot of $\ln([M]_c)=\ln(K_e)$ vs 1/T is linear with a slope of $-\Delta H_e/R$ and y-intercept of $\Delta S_e/R$ *(Figure 4.8c*), giving $\Delta H_e=$ 100 ± 2 kcal mol^{-1} and ΔS_e = 335 ± 7 cal mol^{-1} K^{-1}. Notably, although the values of ΔH_e and ΔS_e determined by TREQ differ from those obtained by kinetic fits to TH data by factors of 1.6 (*Supplementary Table* 4.2), repeating the TH analysis with ΔH_e and ΔS_e fixed to the TREQ-derived values gives good agreement with experimental data (*Figure 4.9*), illustrating the insensitivity of the kinetic fits to these thermodynamic parameters. In general, we would strongly recommend that, even if assembly kinetics are the main interest, the combination of TREQ and thermal hysteresis experiments provide more robust solutions than thermal hysteresis alone, as TREQ resolves ambiguity in the fitted rate constants and ratios thereof.

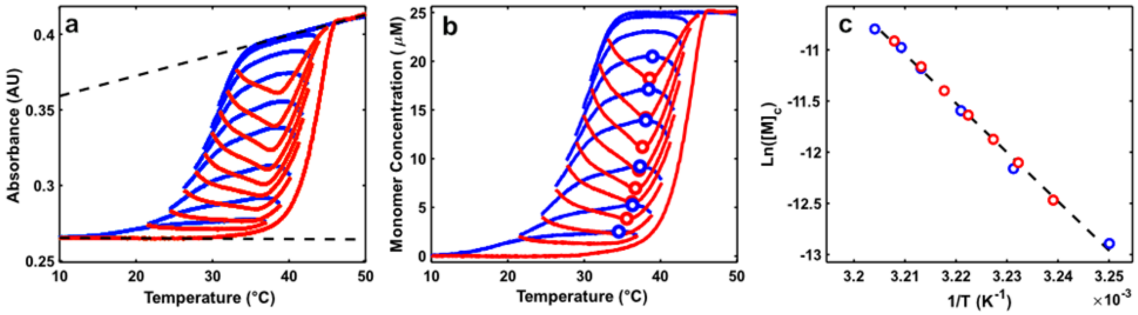


Figure 4.8: Analysis of TREQ data for polyA-CA co-assembly. a) Raw absorbance data for a 15mer polyA-CA coassembly with 25 μM dA$_{15}$ and 15 mM CA at pH 4.5, blue lines represent cooling traces and red lines represent heating traces. Unfolded (top black line) and folded (bottom black line) baselines are also shown. b) TREQ data processed according to Equations S1 and S2 with extrema of each cycle shown as points. c) Van 't Hoff analysis of experimental TREQ points, line of best fit is shown as the black dashed line.

Furthermore, the thermodynamic parameters provide a basis for comparing polyA-CA fibres to other nucleic acid structures. For example, polyA/polyT (dA$_{15}$dT$_{15}$) duplex dissociation is predicted to have approximately ΔH = 108 kcal mol^{-1} and ΔS = 335 cal mol^{-1} K^{-1} under similar solution conditions to those used here,[483] It is intriguingly similar to the values we measured for polyA-CA assembly (100 kcal mol^{-1} and 335 cal mol^{-1} K^{-1}). At first glance, we would have expected polyA-CA fibres to show much higher enthalpies and entropies than dAdT duplexes, since there are three strands rather than two, each dA forms twice as many hydrogen bonds and immobilizes a CA molecule in the putative polyA-CA structure (*Figure 4.1*). However partial vacancy of CA binding sites may help to reconcile these observations, as elaborated below.

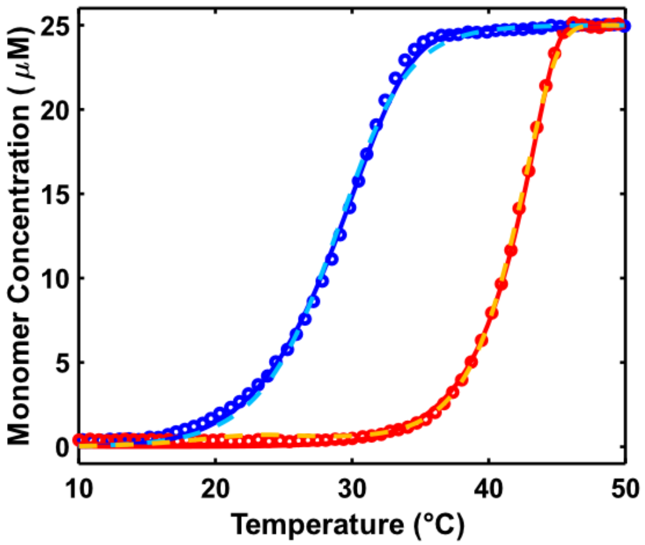


Figure 4.9: Fits of experimental TH curves for polyA-CA coassembly. With ΔH_e extracted from TH data alone (blue and red solid lines) and ΔH_e constrained to be equal to the value extracted from TREQ measurements (dashed cyan and orange lines). Simulation parameters are listed in Supplementary Table 4.2.

One of the great advantages of quantitative thermodynamic data is that much can be learned about the system of interest through careful analyses of how energetic parameters vary with changing conditions. For instance, the presumptive structure of polyA-CA fibres shows that one molecule of CA is present for every deoxyadenosine residue in each polyA chain. In other words, when one of the dA_{15} polyA chains binds the end of an elongating fibre, it should be accompanied by 15 CA molecules. While equilibrium dialysis experiments are consistent with this structure,[239] they have relatively low precision and the stoichiometry is very difficult to measure with accuracy. This property is of great interest since a CA:polyA stoichiometry of less than 15 would reveal the existence of defects, which could potentially be targeted with other small molecules. Thermodynamic data can help to resolve this issue, since the apparent dissociation constant, K_e, for a polyA chain binding to the end of the fibre should vary with CA concentration in a predictable way. For instance, if a polyA chain always brings with it c molecules of CA, i.e.

$$M_n + M_1 + cCA \overset{K_{eq}}{\longleftrightarrow} M_{n+1} \qquad \textit{(Equation 4.3)}$$

(following the nomenclature of *Scheme 4.1*), then the full equilibrium dissociation constant for the process is given by

$$K^{\circ}_{(T)} = \frac{[M_n][M_1][CA]^c}{[M_{n+1}]} \qquad \textit{(Equation 4.4)}$$

This is something of an over-simplification, as elaborated below, but for now it serves to illustrate the dependence of *K_e* on [CA]. For polyA-CA fibres, CA is always in great excess so that its concentration is effectively constant for any set of assembly conditions. The apparent polyA dissociation constant K_e is related to the full equilibrium constant according to

$$K_e = \frac{[M_n][M_1]}{[M_{n+1}]}\bigg|_{[CA]} = K^{\circ}_{(T)}[CA]^{-c} \qquad \textit{(Equation 4.5)}$$

with the temperature dependence of the standard equilibrium constant ($K^{\circ}_{(T)}$) described by

$$K^{\circ}_{(T)} = exp\left(-\frac{(\Delta H_{(T)} - T\Delta S_{(T)})}{RT}\right) \qquad \textit{(Equation 4.6)}$$

Therefore, measuring K_e at a series of different CA concentrations should produce offset van ‘t Hoff plots where the vertical distance between each line follows the stoichiometry of CA. To proceed, we noted that stabilization of polyA-CA fibres at high [CA] is largely entropic in nature, since it is primarily driven by differences in the entropy of dilution when dissociation of a polyA chain concomitantly releases *c* molecules of CA into solution.

We repeated the TREQ experiment at four CA concentrations between 7.5 and 15 mM (*Figure 4.10*). Van ‘t Hoff plots of the resulting K_e values are shown in *Figure 4.11.* Fitting *Equation 4.4* to this data set allows us to directly obtain the stoichiometry of CA. To account for the possibility of a temperature dependent enthalpy value we extracted global values of ΔH_e and ΔC_p. The heat capacity change of binding, ΔC_p, accounts for any temperature-dependent differences in the slopes of the different experiments according to:

$$\Delta H_e(T) = \Delta H_e(T_0) + \Delta C_p(T - T_0) \qquad \text{(Equation 4.7)}$$

$$\Delta S_e(T) = \Delta S_e(T_0) + \Delta C_p \ln\left(\frac{T}{T_0}\right) \qquad \textit{(Equation 4.8)}$$

The extracted ΔC_p = -0.6 ± 0.3 kcal mol^{-1} K^{-1} indicates that the enthalpy of adding a polyA chain to a growing fibre has only a slight temperature dependence. This is perhaps unsurprising, since ΔC_p values associated with nucleic acid folding are largely sequence dependent and have been observed to vary from slightly negative to positive values[484]. The global fit was in good agreement with experimental data points (*Figure 4.11* and *Table 4.3*). Surprisingly, the extracted stoichiometry coefficient, c = 10.4 ± 0.6, implies that 30% of possible CA binding sites are unoccupied in polyA-CA fibres under these conditions.

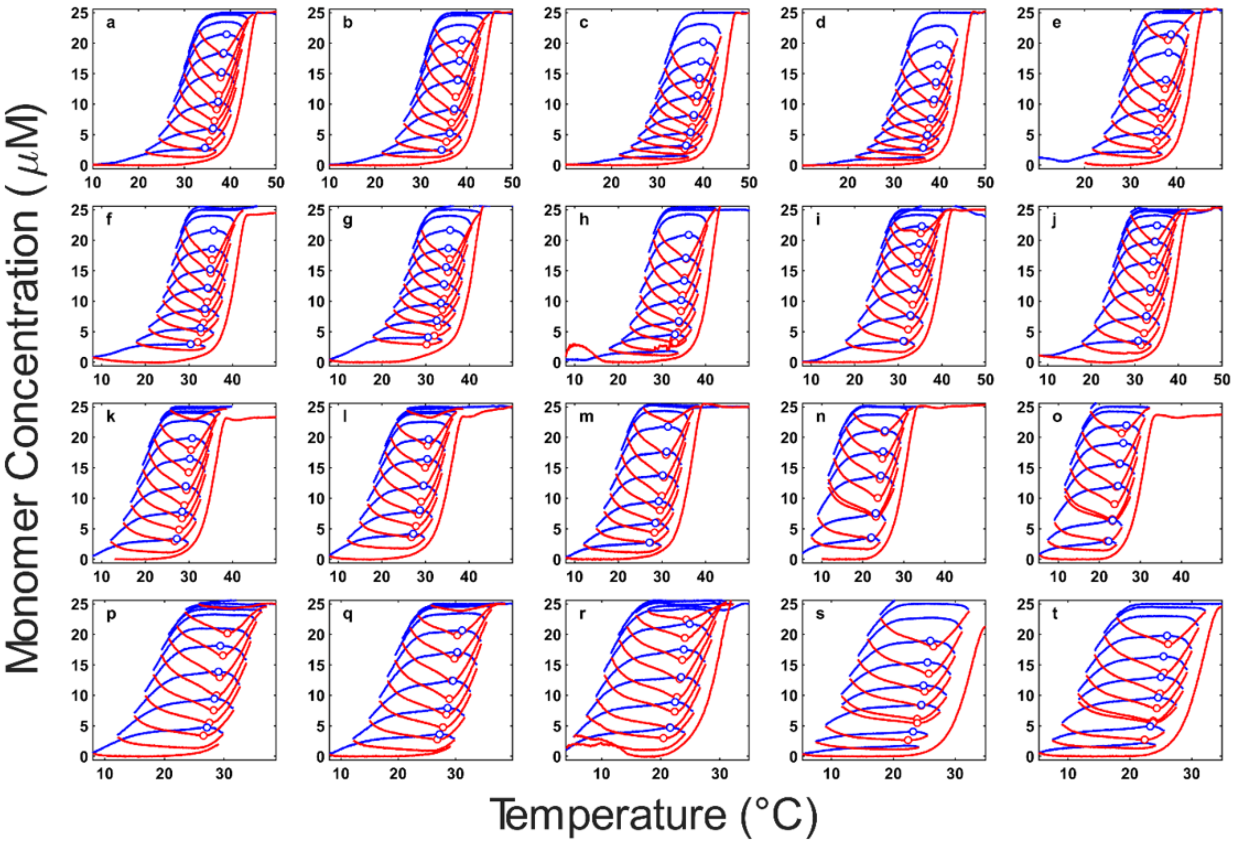


Figure 4.10: TREQ experiments of polyA-CA assembly. At 15mM (a-d), 12.5mM (e-h), 10mM (i-l), and 7.5mM (m-p) cyanuric acid. Cooling traces are indicated in blue, heating traces are indicated in red. Extrema are shown as circles.

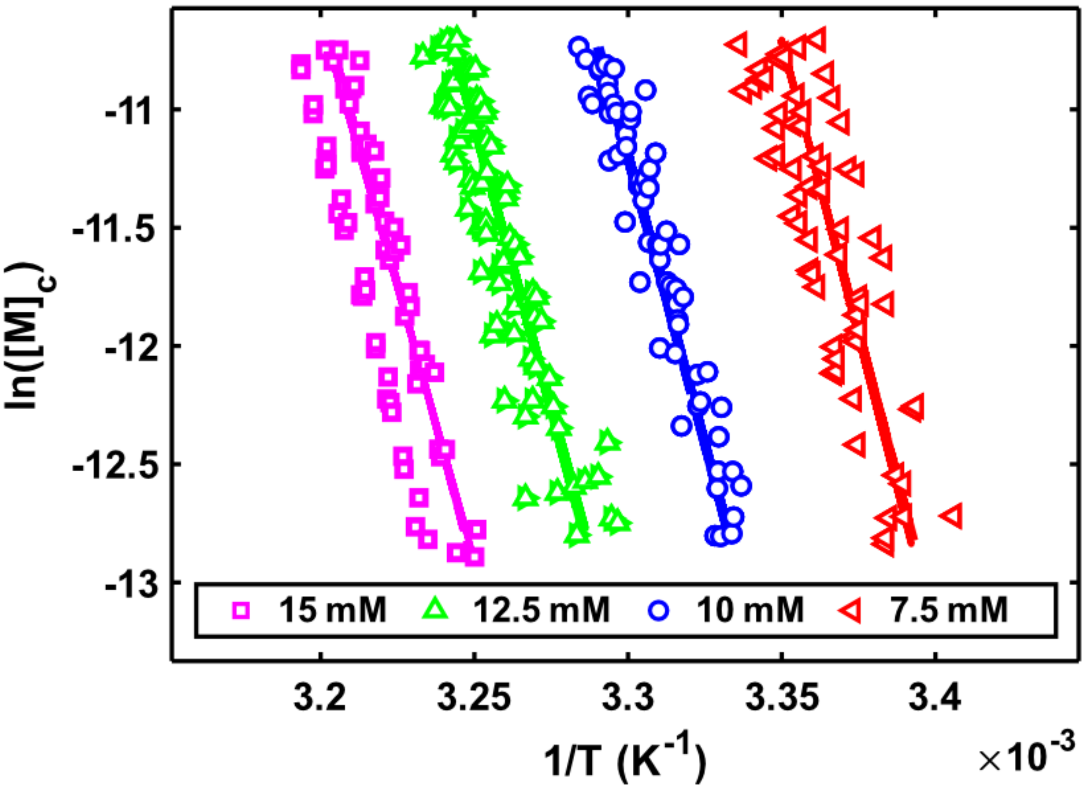


Figure 4.11: van ‘t Hoff plot of TREQ data obtained at different CA concentrations. Coloured symbols represent experimental data from TREQ traces, solid-coloured lines represent a global fit of Equation 4.5 and dashed coloured lines represent a global fit of Equation 9. Solid and dashed lines are virtually superimposed on each other. Experimental errors are smaller than the size of the symbols.

ΔH_e	99.5 ± 0.2 kcal mol^{-1}
ΔS_e	213.2 ± 0.5 cal mol^{-1} K^{-1}
ΔCp_e	-0.69 ± 0.02 kcal mol^{-1} K^{-1}
c_e	*10.15 ± 0.08*

Table 4.3: Thermodynamic parameters from a Van ‘t Hoff fit of the constant stoichiometry model in Figure 4.10. ΔS_e and ΔS_e are reported at a reference temperature of 25˚C.

4.4.6 Master equations for high-valence assembly

The thermodynamics of multivalent supramolecular assembly can be summarized in terms of two main trends: the "principle of maximum occupancy" which refers to the tendency of systems to evolve toward the most stable state with full occupancy of binding sites, and the "entropy factor" which favours the state of the system with the largest number of product species[485]. For most of the supramolecular systems studied to date, the valency (number of binding sites per monomer) is relatively small (<6), the principle of maximum occupancy dominates, and the all sites are generally filled in the assembled materials[486 487]. However, for high-valence monomers, such as the polyA chains studied here, the entropy factor strongly opposes the principle of maximum occupancy and more complex behaviour emerges. For example, each dA_{15} chain creates an additional 15 potential CA binding sites, on average, as it adds to the end of growing fibre; one site must be created for each additional dA residue to achieve the theoretical 1:1 dA:CA stoichiometry. The number of ways to fill *c* of the 15 binding sites with *c* molecules of CA is given by the binomial coefficient[488]

$$N_c = \frac{15!}{c!(15-c)!} \qquad \textit{(Equation 4.9)}$$

While there is only *N*=1 way completely fill all 15 binding sites (*c*=15), there exists a total of *N*=32,766 distinct ways fill the sites with 1≤*c*≤14 molecules of CA. A simplified model of this energy diagram is seen in Figure 4.12*b*, where partially filled states are higher in energy but are more numerous. Therefore, even though a polyA chain with 15 bound CA molecules may represent the single lowest energy configuration, there exists such an enormous number of partly-filled configurations that these dominate, with a broad distribution of CA uptake and just 10 of the 15 sites being filled on average as seen in *Figure 7c*.

This explanation implies that polyA chains can bring a variable number of CA molecules with them when they attach to the end of a growing fibre, which is inconsistent with *Equation 4*, where the stoichiometry is fixed. To resolve this inconsistency, we developed a simple combinatorial model to describe polyA-CA fibre elongation. There is a free energy penalty for bringing an unbound polyA chain in close proximity to the end of a fibre, $\Delta G_{polyA}=\Delta H_{polyA} - T\Delta S_{polyA}$. This is compensated by energetically favourable binding of CA molecules to the newly-created 15 binding sites. All CA molecules are assumed to bind with equal free energy $\Delta G_{CA}=\Delta H_{CA}-T\Delta S_{CA}$. The total free energy change for a polyA chain binding along with a specific configuration of *c* CA molecules is $\Delta G_{polyA} + c\Delta G_{CA}$. Overall, the apparent equilibrium dissociation constant for polyA chain binding is given by[489]

$$(K_e)^{-1} = K_{polyA}(1 + K_{CA}[CA])^{15} \qquad \textit{(Equation 4.10)}$$

where $K_{polyA}=\exp(-\Delta G_{polyA}/RT)$ and $K_{CA}= \exp(-\Delta G_{CA}/RT)$. The average number of CA molecules can be calculated using the following equation

$$\langle c \rangle = 15\frac{K_{CA}[CA]}{1+K_{CA}[CA]} \qquad \textit{(Equation 4.11)}$$

and the fraction of bound states with a given number of CA molecules can be calculated by

$$\theta_c = \left(\frac{15!}{c!(15-c)!}\right)\frac{K_{CA}[CA]^c}{(1+K_{CA}[CA])^{15}} \qquad \textit{(Equation 4.12)}$$

We fit *Equation 4.10* to the TREQ data, obtaining excellent agreement, and extracting ΔH_{polyA}, ΔS_{polyA}, ΔH_{CA}, and ΔS_{CA} *(Figure 4.11* and *Table 4.4).* These parameters allowed us to calculate the fractions of polyA chains with different numbers of CA molecules bound at different temperatures and [CA], providing a highly detailed description of assembly *(Figure 4.12c)*. Under highly stabilizing conditions of high [CA] and low temperature, the Equations predict that almost all binding sites are filled, in agreement with previous dialysis experiments[239]. Importantly, *Equation 4.10* and *Equation 4.11* explain why we observe 10 bound CA, and not more or less, even though experiments were performed at different [CA]. All experiments used 25 μM polyA, which means that we only detected K_e values between about 3 μM and 22 μM in all cases. This implies that the K_{CA}[CA] values are nearly identical in all experiments (since K_{polyA} does not change much with temperature) From *Equation 4.11*, this implies that <c> is very similar in all experiments, ranging from 10 to 11, and in excellent agreement with the simple fit described in the previous section.

High valence supramolecular systems have many useful properties that are only just beginning to be explored, such as the ability to self-heal, responsiveness to stimuli, and simple, inexpensive chemical derivatization. Examples include small molecule-directed nucleic acid assembly (CA + polyadenosine or polydeoxyadeonsine[239, 246]; melamine + polythymine[238]) and non-covalent polymer crosslinking via multiple metal chelation[486, 490] or host/guest interactions[491, 492]. *Equation 4.10* and *Equation 4.11* can serve as starting points for quantitatively describing assembly in such systems, where simple probabilistic considerations ensure that some of the binding sites will remain vacant under many conditions. Furthermore, we find that TREQ-derived data are sufficient to extract the relevant thermodynamic parameters robustly, providing a new avenue for gaining insight into these complex materials.

ΔH_{polyA}	-0.5 ± 0.5 kcal mol^{-1}
ΔS_{polyA}	8 ± 2 cal mol^{-1} K^{-1}
ΔH_{CA}	9.46 ± 0.05 kcal mol^{-1}
ΔS_{CA}	20.7 ± 0.2 cal mol^{-1} K^{-1}

Table 4.4: Thermodynamic parameters from Van ‘t Hoff fit of the independent sites model in Figure 4.11. The relatively large errors in ΔH_{polyA} and ΔS_{polyA} are caused by a correlation in the two parameters.

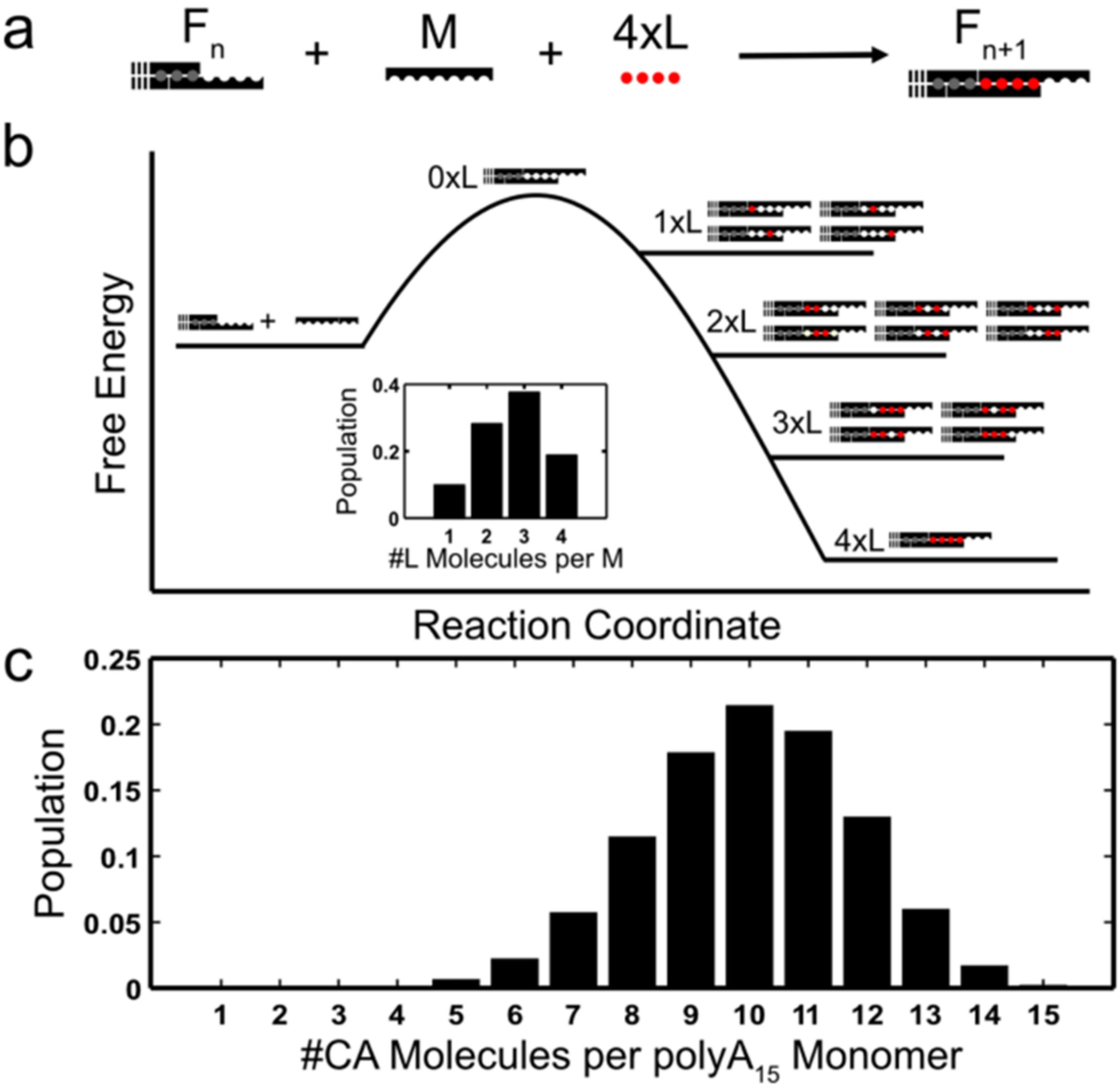


Figure 4.12: Mechanisms of high valence assembly. a) A simple constant stoichiometry assembly model where the end of a growing fibre (F_n) assembles with one monomer M and 4 ligand molecules L to create a fibre of length n+1 (F_{n+1}). b) A free energy diagram of a variable stoichiometry assembly model where the end of a growing fibre (F_n) can assembly with a monomer M and any number of ligand molecules L up to a maximum of 4, the insert represents the populations of each stoichiometry. c) The populations of each stoichiometry for the self assembly of polyA-CA fibres at 25˚C with a concentration of 12.5mM CA.

4.4.7 Generality of the Method

Our aim for the TREQ method is that it can be used as a general tool to determine the thermodynamic parameters of supramolecular assembly when standard thermal melting and annealing experiments are unsuitable for thermodynamic analysis. Towards this end, we have also tested the method on a tetrameric intermolecular guanine quadruplex (G4) in aqueous buffer, and zinc-porphyrin self-assembly in mixture of methylcyclohexane and chloroform. In both cases, we obtained series of concave-up and concave-down arcs, similar to those of the polyA-CA fibres (*Figure 4.13*). In parallel, we used computer simulations to model the TREQ experiment for different types of self assembling systems and observed two patterns of behaviour: either all the extrema aligned with the equilibrium curve or the maxima for the cooling curves and minima for the heating curves were offset from one another (*Figure 4.13*). This provides a useful guide for interpreting TREQ data on new systems of interest: when the extrema align, they can be used to trace out the equilibrium curve (as for polyA-CA fibres and the intermolecular G4). When they are offset, they cannot be directly equated to equilibrium temperature/concentration pairs, although the data are still information-rich. Furthermore, when the extrema are offset, the system can be assumed to have violated one or both of two criteria outlined below. To proceed we make the following definitions: We will use species to refer to any set of assemblies that are kinetically and spectroscopically indistinguishable, and which may or may not be structurally identical. For instance, all GS fibres larger than the nucleus grow or shrink at the same rate and they can be collectively considered a single species, even though they comprise individual fibres of different lengths. The spectroscopic TREQ measurements report the concentration of just one species. This is referred to as the probed species, while all others are referred to as unprobed. Fast and slow chemical kinetics are defined relative to the temperature scan rate. The two TREQ criteria are 1) the effective rates at which the probed species interconverts with all other significantly populated species must be slow and 2) the effective rates at which all significantly populated unprobed species interconvert with each other must be fast.

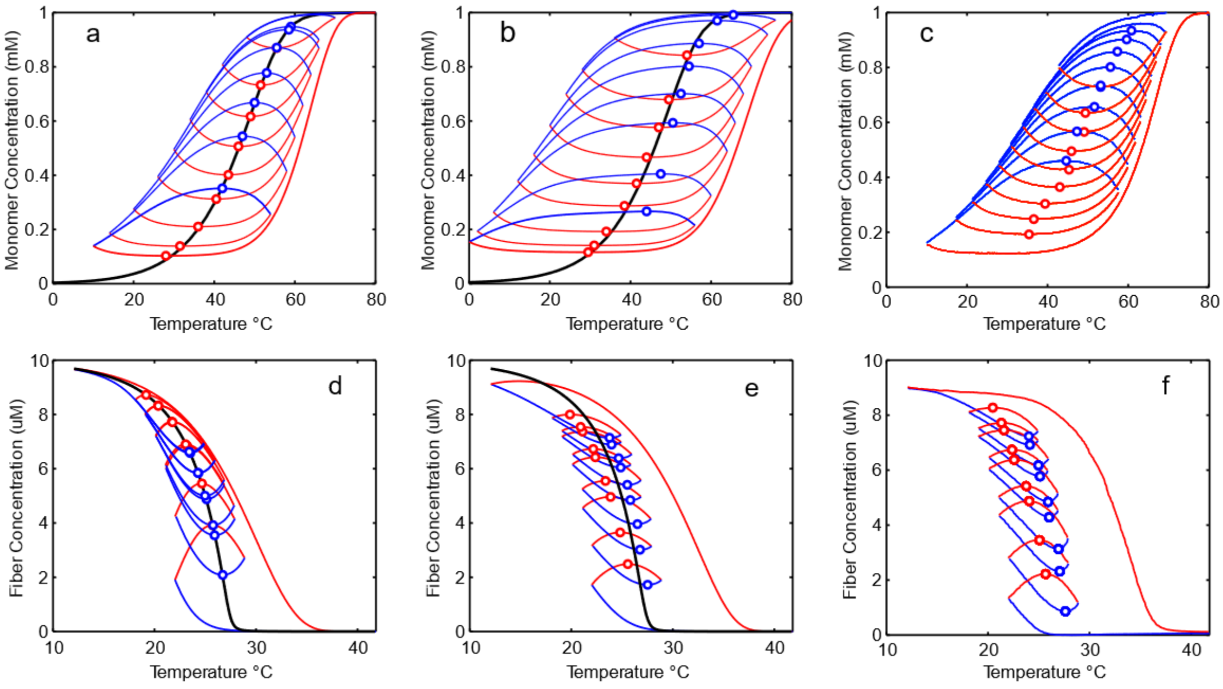


Figure 4.13: Simulated and experimental TREQ traces for different systems. Top row) TREQ traces for sequential tetramolecular GQ assembly. a) Kinetic traces which have minimal kinetic intermediates. b) Kinetics which allow for build up of dimer intermediates. c) Experimental TREQ data of GQ assembly showing that there are no kinetic intermediates. Bottom row) TREQ traces for a zinc porphyrin system which has parallel pathways (one Isodesmic, one cooperative). d) A system with fast Isodesmic aggregation kinetics and slow cooperative aggregation kinetics. e) A system with slow Isodesmic aggregation kinetics and slow cooperative aggregation kinetics. f) Experimental TREQ data of the zinc porphyrin system showing a system which has slow Isodesmic aggregation kinetics and slow cooperative aggregation kinetics. Kinetic parameters for each simulation can be found in Supplementary Table 4.3.

For the polyA-CA fibres, there are only two significantly populated species: monomers (probed) and fibres larger than the nucleus (unprobed). Computer simulations of TREQ data show that the extrema align with the equilibrium curve. However, we have also investigated assembly pathways that differ from the standard GS model. For instance, we previously studied the assembly of tetrameric guanine quadruplexes using thermal hysteresis[246]. The kinetics of assembly are consistent with a monomer ↔ dimer ↔ trimer ↔ tetramer pathway where only monomers (probed) and tetramers (unprobed) are significantly populated. Simulated TREQ data show that extrema closely follow the equilibrium curve (*Figure 4.13a*), in good agreement with experimental data where the extrema align (*Figure 4.13c*). In contrast, if we consider the situation where dimers (unprobed) are also well populated and in fast exchange with monomers, criterion 1 is violated since the probed species exchanges rapidly with a well populated unprobed species. Simulated TREQ maxima and minima are now offset in this scenario (*Figure 4.13b*). Finally, we studied a system that undergoes a parallel assembly mechanism. Tetra-amidated porphyrin molecules can assemble into either chiral fibres or achiral aggregates. As the temperature is reduced, the monomers first assemble into achiral aggregates that slowly convert to chiral fibres at low temperatures.[493] In this case, there are three well-populated species: chiral fibres (probed), achiral aggregates (unprobed), and monomers (unprobed). We performed a simulation in which achiral aggregates and monomers interconvert rapidly. This does not violate either criterion and the computed TREQ extrema align with the equilibrium curve (*Figure 4.13d*). We then performed a simulation in which achiral fibres and monomers interconvert slowly, in violation of criterion 2, and the computed maxima and minima are offset from the equilibrium curve (*Figure 4.13e*). Notably, this simulation closely matches experimental TREQ data for this system (*Figure 4.13f*) which shows the same pattern of offset extrema. Therefore, these data strongly suggest that interconversion between achiral fibres and monomers occurs slowly under these conditions and provide experimental validation for using offset extrema to identify situations that lie outside the scope of TREQ.

Fortunately, many slowly assembling supramolecular structures are amenable to the TREQ approach and, in these cases, it provides thermodynamic information that is not readily available from other sources. For example, a polyA-CA $[M]_c$ dataset similar to the one reported here would require a scan rate of <0.001 K min^{-1} in traditional melting measurements, leading to experiments on the impractically long timescale of a month. Our study demonstrates how the ready availability of high-quality thermodynamic dynamic data can lead to new insights, such as the prevalence of unfilled CA binding sites in polyA-CA fibres, and provides an opportunity to test theoretical developments, such as our master equation for high-valence assembly. These advances would not have been realistically possible for polyA-CA structures using previously existing methods.

A large number of slowly assembling supramolecular systems have been described in the literature, with only a subset referenced in this study[246, 351, 463-477]. This field is expected to expand in the coming years, since slow, nonequilibrium, nucleated assembly is a living polymerization process. The advantages of living polymers in supramolecular chemistry are an area of active research, with benefits already evident in the level of control they give over fibre length and monodispersity[474, 476-478]. Notably, thermodynamic information for slowly assembling systems is either completely lacking or determined using methods that we and others[218] have shown to be unreliable for such systems. We believe that the TREQ method presented here is a big step towards filling this gap in our knowledge. It can be applied to a wide variety of systems using common benchtop laboratory equipment and measurement times are on the order of 10 hours. The experiments are straightforward to set up and a typical analysis (eg van ‘t Hoff plot), can be performed entirely using standard spreadsheet software. We believe that the TREQ method will prove generally useful to the supramolecular chemistry community.

4.5 Materials and methods

4.5.1 Materials

4.5.1.1 Intramolecular G4

A 22mer mutant of the c-MYC G4 (TGAGGGTIGGGAGGGTGGGIAA) was synthesized using a MerMade-12 Oligonucleotide Synthesizer with standard solid-phase phosphoramidite chemistry.[351] The G4 Samples were cartridge purification and analyzed by LC-mass spectrometry for purity. DNA strands were dissolved in MilliQ water and concentrations were calculated using nearest neighbor extinction coefficients. buffer: 10 mM lithium phosphate, pH 7.0, supplemented with 250 uM KCl. The buffer pH was titrated using 1 M LiOH to avoid the further addition of stabilizing Na+ or K+ cations.

4.5.1.2 polyA-CA

Cyanuric acid (CA), tris(hydroxymethyl)aminomethane (Tris), magnesium chloride hexahydrate ($MgCl_2$•6 H2O), 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.

d(A_{15}) oligonucleotides were synthesized on a Mermade-12 synthesizer, purified by denaturing polyacrylamide gel electrophoresis (PAGE 20%, 1xTBE running buffer, 8 M urea) and desalted with Gel-Pak desalting columns from Glen Research. Purity of the strand was confirmed by HRMS (Calculated mass: 4635.18; Observed mass: 4634.28).

Stock solutions of 20 mM CA were prepared by dissolution in 100 mL of Milli-Q water in a volumetric flask and adjusted with acetic acid to pH 4.5. To properly dissolve and degas the solutions, they were heated at 65 °C and sonicated, then cooled down to room temperature before being used.

Samples of 100 µL of dA_{15} (25 µM) and CA (7.5, 10.0, 12.5 and 15.0 mM) in pH 4.5 $Mg(OAc)_2$ buffer (7.6 mM) were made in quadruplicates. A thin layer (~30 µL) of silicon oil was applied on top to prevent evaporation during experiments.

4.5.2 Instrumentation

4.5.2.1 Intramolecular G4

UV-Vis absorbance studies were performed using a 10 mm quartz cuvette with a 3mm aperture and monitored at 295 nm on an Agilent Cary 3500 Series UV-Vis Spectrophotometer equipped with a Peltier temperature controller and in-cell thermal probe. A thermal hysteresis scan was performed from 60-10 °C at 1 K min^{-1} and 0.1 K min^{-1} with an equilibration time of 30 min at both high and low temperatures. TREQ experiments were ran at 1 K min^{-1} with temperature windows chosen from the TH scans.

The maximum number of scans on the Cary 3500 is 10 so two TREQ experiments were ran and combined to create *Figure 4.6a and c*.

4.5.2.2 polyA-CA

UV-Vis absorbance-based quantification of d(A_{15}) was performed on a Nanodrop Lite spectrophotometer from Thermo Scientific. DNA purification by PAGE was carried out on a 20 x 20 cm vertical acrylamide Hoefer 600 electrophoresis unit.

UV-Vis absorbance studies were performed using a 1.0 mm quartz cuvette and monitored at 260 nm on an Agilent Cary 300 Series UV-Vis Spectrophotometer equipped with a Peltier temperature controller and water recirculator. A variable temperature range which started from 50-40 °C and went down to 10-4 °C was scanned at a rate of 0.5 °C/min and with an equilibration time of 30 min at the maximum and minimum temperatures. Argon gas and drierite were used to dry the chamber at temperatures below 10 °C.

4.5.3 Thermodynamic analysis

TREQ data for polyA-CA fibres (critical polyA monomer concentrations, $[M]_c$, as a function of temperature, obtained at different CA concentrations) were fitted using two different physical models. In both cases, $[M]_c$ values were equated to the equilibrium dissociation constant for adding a monomer to the end of a growing fibre (K_e). The first model invoked constant CA:polyA stoichiometry (*Equation 4.3* to *Equation 4.5*) and was essentially an extension of a classical van 't Hoff ln(K_e) vs 1/T analysis in which heat capacity changes and [CA] dependence are taken into account. The second model explicitly took into account the statistical effects of partially filling multiple binding sites (Master equations for high valence systems, *Equation 4.10*). In both cases, for each value of [CA], a $[M]_c$(T) dataset was calculated in a temperature range from 10-50˚C with a resolution of 0.01˚C. Each model's parameters were optimized using total least squares regression, which accounts for errors in both x- and y- dimensions. Fits were optimized by finding thermodynamic parameters to minimize the target function

$$RSS = \sum_{j=1}^{N} \sqrt{\left(\frac{\Delta_{\ln([M]_{c(j)})}}{\sigma_{\ln\,([M]_{c(j)})}}\right)^2 + \left(\frac{\Delta_{\frac{1}{T_{(j)}}}}{\sigma_{\frac{1}{T_{(j)}}}}\right)^2}$$

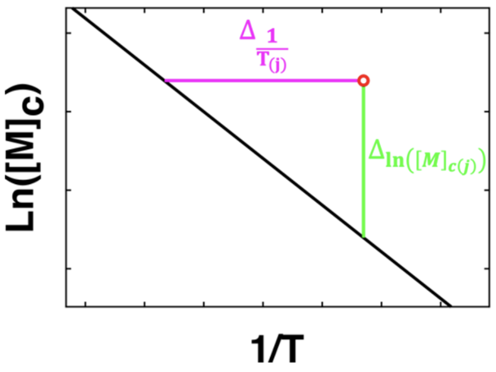


Figure 4.14: Minimization function for total-least squares regression and visualization of the horizontal (magenta) and vertical (green) distances of an experimental data point (red) to the simulated line (black).

Where $\Delta_{\ln([M]_{c(j)})}$ is the vertical distance of the jth experimental data point to the point on the simulated curve which minimized the horizontal distance. $\Delta_{\frac{1}{T_{(j)}}}$ is the horizontal distance of the jth experimental point to the point on the simulated curve which minimized the vertical distance. $\sigma_{\ln\,([M]_{c(j)})}$ and $\sigma_{\frac{1}{T}(j)}$are the experimental errors in the vertical and horizontal dimensions respectively. Errors for fitted parameters were calculated using a bootstrapping approach,[482] in which each bootstrap sample was obtained by random resampling of the original data. For example, if the original dataset contained N points, each bootstrap sample was constructed by randomly selecting N of these data points, such that points may be selected more than once or not at all. 500 bootstrap samples were constructed and fitted using the thermodynamic models described above. The errors in the extracted parameters were taken as the standard deviations of the 500 sets of parameters obtained for all bootstrap samples.

4.5.4 TREQ Simulations

TREQ experiments were simulated using the kinetic models described below. In all simulations the rate constants were assumed to have an Arrhenius temperature dependence following the equation

$$k(T) = k_0 e^{\frac{E_a}{R}\left(\frac{1}{T_{ref}} - \frac{1}{T}\right)} \qquad \textit{(Equation 4.13)}$$

Each set of differential equations were numerically integrated as a function of temperature using MATLABs built in ODE solver ode15s. Temperature windows were chosen from experimental data windows, and kinetic parameters can be found in *Supplementary Table* 4.1*, Table 4.1, Table 4.2, Supplementary Table* 4.2*, and Supplementary Table* 4.3*.*

4.5.4.1 TGGGG Assembly

Assembly of TGGGG strands into a guanine quadruplex was modelled as a sequential addition of monomers (M) into dimers (D), trimers (Tr) and tetramers (Q) using the following rate equations

$$\frac{d}{dt}[M] = 2k_{-1}[D] - 2k_1[M]^2 - k_2[M][D] + k_{-2}[Tr] - k_3[M][Tr] + k_{-3}[Q] \textit{(Equation 4.14)}$$

$$\frac{d}{dt}[D] = k_1[M]^2 - k_{-1}[D] + k_{-2}[Tr] - k_2[M][D] \qquad \textit{(Equation 4.15)}$$

$$\frac{d}{dt}[Tr] = k_2[M][D] - k_{-2}[Tr] + k_{-3}[Q] - k_3[M][Tr] \qquad \textit{(Equation 4.16)}$$

$$\frac{d}{dt}[Q] = k_3[M][Tr] - k_{-3}[Q] \qquad \textit{(Equation 4.17)}$$

4.5.4.2 polyA Assembly

polyA fibre formation was modelled following the Goldstein-Stryer model for cooperative self-assembly as described previously[248]. This model assumes reversible, cooperative stepwise addition of monomers (*M*) to nuclei (M_s), which then elongate to form fibres (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 fibres up to size *N* are explicitly described. A sparse Jacobian matrix was created to define the species which are related, this allowed for simulations of large fibre sizes (N = 1000). The Goldstein-Stryer model is described by the following rate equations

Monomer

$$\frac{d}{dt}[M] = -k_{n+}[M]\left(2[M] + \sum_{i=2}^{s-1}[M_i]\right) - k_{e+}[M]\left(\sum_{i=s}^{N-1}[M_i]\right)$$

$$+k_{n-}(2 * [M_2] + \textstyle\sum_{i=3}^{s}[M_i]) + k_{e-}\sum_{i=s+1}^{N}[M_i] \qquad \textit{(Equation 4.18)}$$

Pre-nucleus oligomers

$$\frac{d}{dt}[M_i] = \; k_{n+}[M]([M_{i-1}] - [M]) + k_{n-}([M_{i+1}] - [M_i]) \qquad \textit{(Equation 4.19)}$$

Nucleus

$$\frac{d}{dt}[M_s] = \; k_{n+}[M][M_{s-1}] - k_{e+}[M][M_s] + k_{e-}[M_{s+1}] + k_{n-}[M_s] \qquad \textit{(Equation 4.20)}$$

Post-nucleus fibres

$$\frac{d}{dt}[M_i] = \; k_{e+}[M]([M_{i-1}] - [M]) + k_{e-}([M_{i+1}] - [M_i]) \qquad \textit{(Equation 4.21)}$$

Fibre length N

$$\frac{d}{dt}[M_i] = \; k_{e+}[M][M_{N-1}] - k_{e-}[M_N] \qquad \textit{(Equation 4.22)}$$

4.5.4.3 Porphyrin Assembly

Zinc porphyrin assembly was modelled as a system with two distinct parallel pathways, where one pathway assembles via the Goldstein-Stryer model of assembly with pre-nucleated and post-nucleated rate constants of k_{n+}/k_{n-} and k_{e+}/k_{e-} up to a maximum length of N and a nucleus size s and one pathway forms Isodesmic aggregates which assembly with the rate constants k_{i+}/k_{i-} with a maximum length L. The parallel pathways model is described by the following rate equations.

Monomer

$$\frac{d}{dt}[M] = \; -k_{n+}[M]\left(2[M] + \sum_{i=2}^{s-1}[M_i]\right) - k_{e+}[M]\left(\sum_{i=s}^{N-1}[M_i]\right) + k_{n-}\left(2 * [M_2] + \sum_{i=3}^{s}[M_i]\right)$$

$$+k_{e-}\textstyle\sum_{i=s+1}^{N}[M_i] - k_{i+}[M](2[M] + \sum_{i=2}^{L-1}[I_i]) + k_{i-} * (2 * [I_2] + \sum_{i=3}^{L}[I_i]) \textit{(Equation 4.23)}$$

Pre-nucleus oligomers

$$\frac{d}{dt}[M_i] = \; k_{n+}[M]([M_{i-1}] - [M]) + k_{n-}([M_{i+1}] - [M_i]) \qquad \textit{(Equation 4.24)}$$

Nucleus

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

(Equation 4.25)

Post-nucleus fibres

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

(Equation 4.26)

Fibre length N

$$\frac{d}{dt}[M_N] = k_{e+}[M][M_{N-1}] - k_{e-}[M_N]$$

(Equation 4.27)

Isodesmic aggregates

$$\frac{d}{dt}[I_i] = k_{i+}[M]([I_{i-1}] - [M]) + k_{i-}([I_{i+1}] - [I_i])$$

(Equation 4.28)

Isodesmic aggregate length L

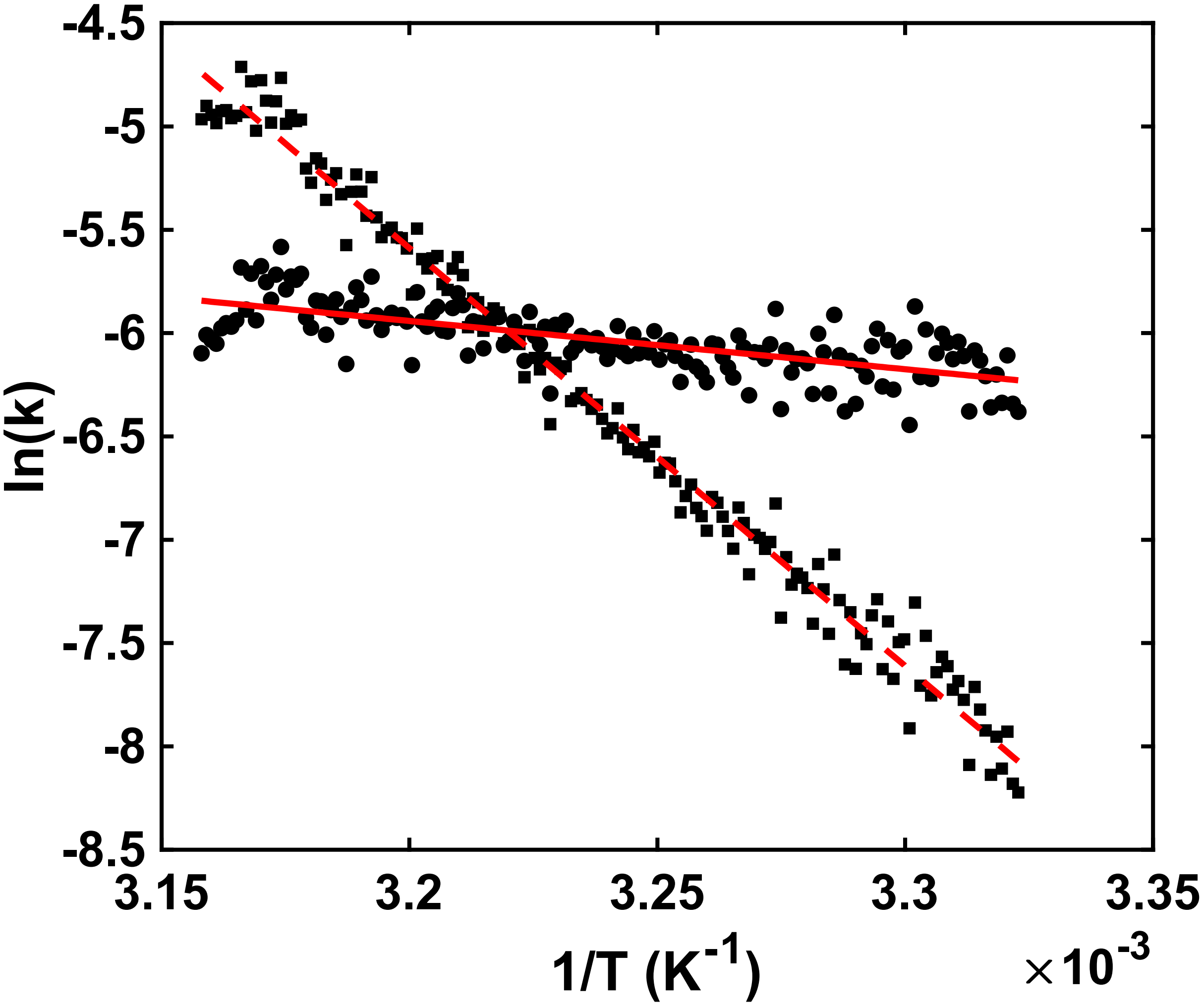
$$\frac{d}{dt}[I_L] = k_{i+}[M][I_{L-1}] - k_{i-}[I_L]$$

(Equation 4.29)

4.6 Supplementary information

$\Delta H_e = 50$: $\Delta S_e = 138$				**$\Delta H_e = 100$: $\Delta S_e = 301$**			
Activation Energies		**Rate constants**		**Activation energies**		**Rate constants**	
E_{n+}	-1.0	k_{n+}	2.9×10^5	E_{n+}	0	k_{n+}	9.2×10^5
E_{n-}	53	k_{n-}	54	E_{n-}	74	k_{n-}	8.6
E_{e+}	7.5	k_{e+}	3.1×10^5	E_{e+}	-30	k_{e+}	2.1×10^4
E_{e-}	58.5	k_{e-}	7.4×10^{-2}	E_{e-}	70	k_{e-}	8.1×10^{-4}
$\Delta H_e = 150$: $\Delta S_e = 472$				**$\Delta H_e = 200$: $\Delta S_e = 644$**			
Activation energies		**Rate constants**		**Activation energies**		**Rate constants**	
E_{n+}	0	k_{n+}	3.4×10^6	E_{n+}	0	k_{n+}	8.5×10^6
E_{n-}	81	k_{n-}	23	E_{n-}	78	k_{n-}	60
E_{e+}	-82	k_{e+}	5.9×10^3	E_{e+}	-129	k_{e+}	2.7×10^2
E_{e-}	68	k_{e-}	6.8×10^{-4}	E_{e-}	71	k_{e-}	5.1×10^{-4}

Supplementary Table 4.1: Kinetic parameters for each TH simulation in Figure 4.2a. Activation energies are given in kcal mol^{-1} rate constants are in M^{-1} min^{-1} and min^{-1} for forward and reverse steps respectively and reported at a reference temperature of 25˚C. ΔH_e values are given in kcal mol^{-1} and ΔS_e values are given in cal mol^{-1} K^{-1}.



Supplementary Figure 4.1: Arrhenius plots from classical two-state analysis of intramolecular G4 TH profiles taken at 0.1 K min^{-1}. The analysis was applied to the portion of the curve between 0.15 < θ_U < 0.75. The folding (k_F) and unfolding (k_U) rate constants calculated from the experimental datasets are shown as circles and squares respectively, while the corresponding line of best fit to each dataset are shown as solid and dashed red lines respectively. The equilibrium melting temperatures are at the intersections of the folding and unfolding lines, and the equilibrium profile can be calculated from the ratio of k_F and k_U[218].

Unconstrained Parameters **$\Delta H_e = 62$: $\Delta S_e = 197$**				**Constrained Parameters** **$\Delta H_e = 100$: $\Delta S_e = 335$**			
Activation energies		**Rate constants**		**Activation energies**		**Rate constants**	
E_{n+}	-24	k_{n+}	7.6×10^{6}	E_{n+}	-13	k_{n+}	5.2×10^{6}
E_{n-}	-5	k_{n-}	8.7×10^{3}	E_{n-}	62	k_{n-}	46
E_{e+}	-9	k_{e+}	6.9×10^{5}	E_{e+}	-31	k_{e+}	1.7×10^{4}
E_{e-}	53	k_{e-}	1.2×10^{-1}	E_{e-}	69	k_{e-}	7.5×10^{-4}
RSS	2.6×10^{-4}	T_{ref}	25	*RSS*	3.7×10^{-4}	T_{ref}	25

Supplementary Table 4.2: Kinetic parameters for each TH fit in Figure 4.9. Activation energies are given in kcal mol^{-1} rate constants are in M^{-1} min^{-1} and min^{-1} for forward and reverse steps respectively. ΔH_e values are given in kcal mol^{-1}, ΔS_e values are given in cal mol^{-1} K^{-1} and ΔCp values are given in kcal mol^{-1} K^{-1} in the constrained fit, a CA concentration of 15mM was used.

Panel A				Panel B			
Activation energies		Rate constants		Activation energies		Rate constants	
E_{1}	-5.4	k_{1}	3.0 x 10^{2}	E_{1}	-5.4	k_{1}	3.0 x 10^{2}
E_{-1}	14.4	k_{-1}	5.0 x 10^{3}	E_{-1}	14.4	k_{-1}	5.0 x 10^{3}
E_{2}	-4.0	k_{2}	1.6 x 10^{5}	E_{2}	-4.0	k_{2}	2.0 x 10^{4}
E_{-2}	15.9	k_{-2}	3.1 x 10^{-1}	E_{-2}	15.9	k_{-2}	3.9 x 10^{-2}
E_{3}	-3.8	k_{3}	8.2 x 10^{2}	E_{3}	-3.8	k_{3}	8.2 x 10^{2}
E_{-3}	37.4	k_{-3}	8.4 x 10^{-3}	E_{-3}	37.4	k_{-3}	8.4 x 10^{-3}
		T_{ref}	45			T_{ref}	45
Panel D				Panel E			
Activation energies		Rate constants		Activation energies		Rate constants	
E_{n+}	-12.0	k_{n+}	6.0 x 10^{7}	E_{n+}	-12.0	k_{n+}	6.0 x 10^{7}
E_{n-}	6.0	k_{n-}	1.7 x 10^{4}	E_{n-}	6.0	k_{n-}	1.7 x 10^{4}
E_{e+}	-12.0	k_{e+}	6.0 x 10^{7}	E_{e+}	-12.0	k_{e+}	6.0 x 10^{7}
E_{e-}	12.4	k_{e-}	1.7	E_{e-}	12.4	k_{e-}	1.7
E_{i+}	-12.0	k_{i+}	3.0 x 10^{6}	E_{i+}	-12.0	k_{i+}	3.0 x 10^{4}
E_{i-}	1.0	k_{i-}	9.8	E_{i-}	1.0	k_{i-}	9.8 x 10^{-2}
		T_{ref}	25			T_{ref}	25

Supplementary Table 4.3: Kinetic parameters for each TREQ simulation in Figure 4.13. Activation energies are given in kcal mol^{-1} rate constants are in M^{-1} min^{-1} and min^{-1} for forward and reverse steps respectively. Reference temperatures are in °C.

Chapter 5: Mechanistic Characterization of Covalent Inhibition by Isothermal Titration Calorimetry Kinetic Competition (ITC-KC)

5.1 Preface

The chapter details the development of a new method to measure both the affinity (K_i) and reactivity (k_{inact}) of two-step irreversible covalent inhibitors. This method is based around isothermal titration calorimetry (ITC), which is a technique that measures the heat released from chemical reactions. We use ITC to measure the velocity of an enzyme as a function of time while the enzyme is being inactivated by the inhibitors. We show how we are able to differentiate between simple inactive compounds, one-step reversible inhibitors, one-step irreversible inhibitors, and two-step irreversible inhibitors by performing two injections of enzyme into a mixture of substrate and inhibitor. We use simulations with random noise to determine the range of K_i and k_{inact} values that this method can accurately measure and compare it to conventional inhibitor concentration-dependent progress curve (IDPC) analysis and time-dependent IC_{50} ($TDIC_{50}$) analysis. Furthermore, we identified a systematic error in $TDIC_{50}$ analysis, and provide a new way of fitting this data to remove the error. We use this new technique to study 19 inhibitors, 10 of which have different covalent warheads, and 10 of which have different scaffolds. We discuss how the different warheads and scaffolds change both the affinity and reactivity of the molecules. The experimental ITC data from this chapter were all collected by Felipe Venegas. The IDPC analysis and synthesis of the compounds was done by Guanyu (Chris) Wang and Julia Stille. I performed all of the analysis and simulations of the ITC experiments, along with the comparisons to both IDPC and $TDIC_{50}$. Finally, Prof. Anthony Mittermaier and I wrote the manuscript, and I have adapted it for this thesis.

5.2 Abstract

Covalent inhibitors are increasingly sought after in drug discovery efforts due to their potential for high potency and specificity. Unlike non-covalent inhibitors which usually bind in a one-step mechanism, covalent inhibition typically requires at least two steps: formation of a non-covalent intermediate complex, followed by formation of a covalent bond, which locks the complex together. Rational optimization of covalent inhibitors requires quantitative information on both these steps, namely K_i, the dissociation affinity constant for the non-covalent complex and k_{inact}, the first order rate constant for covalent bond formation. Current methods for measuring these parameters are technically demanding, time consuming, and are not well suited for routine insertion into drug discovery pipelines. We have developed a new approach for measuring K_i and k_{inact} using isothermal titration calorimetry kinetic competition (ITC-KC) that overcomes many of these challenges. The technique measures the heat released by catalysis, making it a nearly universal approach, since virtually all enzymatic reactions produce a measurable signal. Furthermore, by measuring heat flow, our new ITC-KC method circumvents the weaknesses of current methods as it can measure enzyme activity directly and not through product formation or substrate depletion. We applied ITC-KC to a library of 19 potential inhibitors of 3CL^{pro} from SARS-CoV-2, obtaining results consistent with traditional inhibitor concentration dependent progress curve analyses. The ITC-derived K_i and k_{inact} parameters shed light on the complex interplay between the warhead and scaffold portions of covalent inhibitors and the affinity of the non-covalent intermediate complex and rate of covalent bond formation.

5.3 Introduction

Covalent inhibitors, which form covalent chemical bonds with their enzyme targets, have been historically disfavoured in drug development campaigns due to concerns over off-target effects. However, it is becoming increasingly recognized that covalent inhibitors can offer superior specificity, affinity, and residence times, compared to non-covalent drugs[291, 494]. In fact, several common medications have been found to be covalent drugs years after their initial discovery, including aspirin and penicillin[495]. There is now an increasing interest in discovering new covalent drugs for a wide variety of diseases[495].

However, there are currently barriers to developing these molecules. A crucial early step in the drug development process is the experimental characterization of potential hits[291]. This information is required to understand how changes in the chemical structure of a molecule alter its potency, and ultimately guides the optimization of hits into lead compounds and drug candidates. The approaches typically used to characterize structure-activity relationships in non-covalent inhibitors do not transfer well to covalent ones, due to fundamental differences in their binding mechanisms.

Most non-covalent inhibitors bind to their targets in a one step reaction,

$$E + I \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} EI \qquad \textit{(Scheme 5.1)}$$

where k_{on} and k_{off} are the kinetic rate constants for association and dissociation, and *E*, *I*, and *EI* are the enzyme, inhibitor, and inhibited complex, respectively. The experimental metric of potency is the equilibrium dissociation constant of the enzyme/inhibitor complex (K_i=k_{off}/k_{on}), where lower values indicate tighter binding. The value of K_i is usually obtained by measuring enzyme activity as a function of inhibitor concentration. The concentration of inhibitor required to reduce activity by 50% is referred to as the IC_{50}, which can be converted to a K_i value using the Cheng-Prusoff equation[289]. In contrast, covalent inhibitors bind their targets in at least two steps according to,

$$E + I \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} EI \underset{k_{rev}}{\overset{k_{inact}}{\rightleftarrows}} E - I \qquad \textit{(Scheme 5.2)}$$

where they first establish a non-covalent intermediate complex with the target (*EI*) and subsequently form a covalent chemical bond with it (*E–I*). In this case, the dissociation constant K_i=k_{off}/k_{on} refers to the affinity of the noncovalent intermediate while the first order rate constants k_{inact} and k_{rev} describe covalent bond formation and breakage, respectively. These molecules often react slowly, on the timescale of minutes or hours. Thus, potency can depend as much on the rate of formation as it does on the stability of the *E–I* state, since tight binding is meaningless if the covalently bound complex does not form on a therapeutically relevant time scale. For irreversible covalent inhibitors, the stability of the *E–I* state is effectively infinite, so the key property that distinguishes a good irreversible inhibitor from a poor one is the rate of formation of the *E–I* complex.

Optimizing covalent drugs for rapid *E–I* formation is complicated by the fact that there are distinctly different ways in which altering an inhibitor structure can affect binding kinetics. For example, a certain change that accelerates the formation of the covalently bound species (*E-I*) may do so by stabilizing *EI* (decreasing K_i), by increasing the rate of covalent bond formation (increasing k_{inact}), or by any changes in both parameters, possibly opposing ones, that result in a net increase in the overall rate of inhibition. Thus to establish meaningful structure activity relationships, we must be able to measure both K_i and k_{inact} for covalent drug candidates. Currently, most studies employ one of two approaches to obtain this information[296]: The rate of *E–I* formation can be determined as a function of inhibitor concentration, *[I]*, and the dependence analyzed by a linearized Kitz-Wilson plot or non-linear least squares fitting to give K_i and k_{inact}[496]. Alternatively, the dependence of IC_{50} values on incubation time can be fitted to yield these parameters[497]. However, these are time-consuming, labour-intensive, and technically challenging experiments. They involve running multiple enzyme assays (>10) for each inhibitor. As elaborated below, the data points most critical for accurate parameter estimation tend to be the least well defined. This weakness becomes increasingly acute for potent inhibitors with rapid k_{inact} values. Furthermore, commonly applied analysis techniques lead to systematic errors in the parameters extracted using the IC_{50} approach. Finally, only a subset of enzymes have readily available continuous (real-time) assays. Otherwise, one must use discontinuous assays in which ancillary techniques such as chromatography[302], electrophoresis[301], or mass spectrometry[498] are employed to quantify substrates and products. This adds considerable time and expense to the analysis and limits the number and accuracy of data points that can be collected. Consequently, mechanistic characterization is not routinely applied to covalent inhibitors, limiting their advancement[499].

We have developed a new approach for measuring K_i and k_{inact} values of covalent inhibitors that addresses many of the shortcomings of current methods. It is based on isothermal titration calorimetry (ITC), which was originally developed to measure the thermodynamics of host-guest interactions, but is increasingly used to characterize enzyme kinetics[303, 500]. ITC detects in real time the heat that is absorbed or released when a titrant in a syringe is injected into an analyte in the sample cell. In our assay, enzyme in the syringe is added to a sample cell containing both substrate and inhibitor, which compete kinetically to produce either catalysis or inhibition, respectively. Thus, we refer to this as a kinetic competition, or ITC-KC experiment. Since virtually all chemical reactions absorb or release heat, ITC is an essentially universal enzyme assay[303]. It can be used with natural substrates under near-physiological conditions, even with spectroscopically opaque solutions, and does not require downstream separation of substrates and products[501]. ITC detects heat flow even while injections (typically 1-80 seconds) are taking place, giving it many of the benefits of stopped-flow and rapid mixing devices[306]. Furthermore, ITC directly measures the instantaneous rate of catalysis. This contrasts with virtually all other enzyme assays in which substrate or product concentrations are measured, and reaction rates are calculated indirectly from time-dependent changes in concentration. This difference leads to substantial advantages in quantifying changes in enzyme activity while inhibitors are in the process of binding[307].

As proof of principle, we applied our ITC-KC method to a library of 19 potential inhibitors of the 3C-Like protease (3CL^{pro}) from SARS-CoV-2 we synthesized as part of our ongoing efforts to develop new COVID-19 and pan-coronavirus therapeutics[502]. 3CL^{pro} cleaves the viral polyprotein in a critical step of the replication cycle, and is the target of the clinically-approved drugs Paxlovid (Pfizer)[503] and Xocova (Shionogi)[504]. The ITC-KC experiment characterized each inhibitor in about an hour with just two injections of enzyme. Fortunately, the technique is suitable for automation, and using an auto PEAQ-ITC instrument (Malvern Panalytical), we could perform an initial analysis of the entire panel in under 24 hours. Results from the ITC-KC assay compared favourably with those of traditional fluorescence-based enzyme assays, providing validation for the method. Counter to our expectations, the data showed that both the reactive warhead and the scaffold portions of the molecules contribute to both the stability of the non-covalent intermediate (K_i) and the rate of covalent bond formation (k_{inact}), shedding new light on structure activity relationships for these important molecules.

5.4 Results

5.4.1 The isothermal titration calorimetry kinetic competition (ITC-KC) experiment

In the ITC-KC experiment, the syringe of the calorimeter contains a solution of enzyme, while the sample cell contains either the substrate alone or a mixture of substrate and inhibitor. Two short (5 seconds) injections of enzyme are made, spaced roughly 30 minutes apart, and the differential heat flow to the sample cell (*dQ/dt*) is measured as a function of time. Simulated traces are shown in *Figure 5.1* to illustrate the types of data that are expected for different inhibition strengths and mechanisms, based on the Michaelis-Menten parameters (*Supplementary Figure 5.1 and Supplementary Table 5.1*) we extracted for the 3CL^{pro} system. A simulated ITC trace for the negative control with no inhibitor is shown in *Figure 5.1a*. The deflection of the ITC signal following the first injection at *t*=10 minutes reports the heat flow due to the enzymatic reaction. Endothermic reactions produce positive deflections (as for 3CL^{pro} and simulated here), while exothermic reactions produce negative ones. The magnitude of the deflection is directly proportional to the velocity of the reaction according to[303]

$$\frac{dQ}{dt}(t) = -v(t) \times \Delta H_r \times V_{cell} \qquad \textit{(Equation 5.1)}$$

where *v(t)* is the reaction velocity, V_{cell} is the volume of the cell (200 μL in this case), and ΔH_r is the molar enthalpy of the reaction, which can be determined by dividing the total amount of heat absorbed (area of the peak) by the amount of substrate initially in the cell.

Thus, ITC traces provide quantitative, real-time readouts of enzymatic activity. The simulated heat signal decays in *Figure 5.1a*, returning to baseline after about 10 minutes, as all the substrate is consumed. The second injection of enzyme at about 45 minutes therefore produces no heat signal, since there is no substrate remaining at that time. The shape of the first peak provides enough information to determine the Michaelis Menten enzymatic parameters, K_m and k_{cat}, for the enzyme; the displacement (*dQ/dt*) gives the instantaneous velocity, *v(t),* while the fraction of substrate remaining at any time, *t*, is given by the area of the peak to the right of *t* divided by the area of the entire peak[305]. In practice, we fit the peak shapes using the ordinary differential equation solver routines in MATLAB, which also take into account the finite response time of the calorimeter[306].

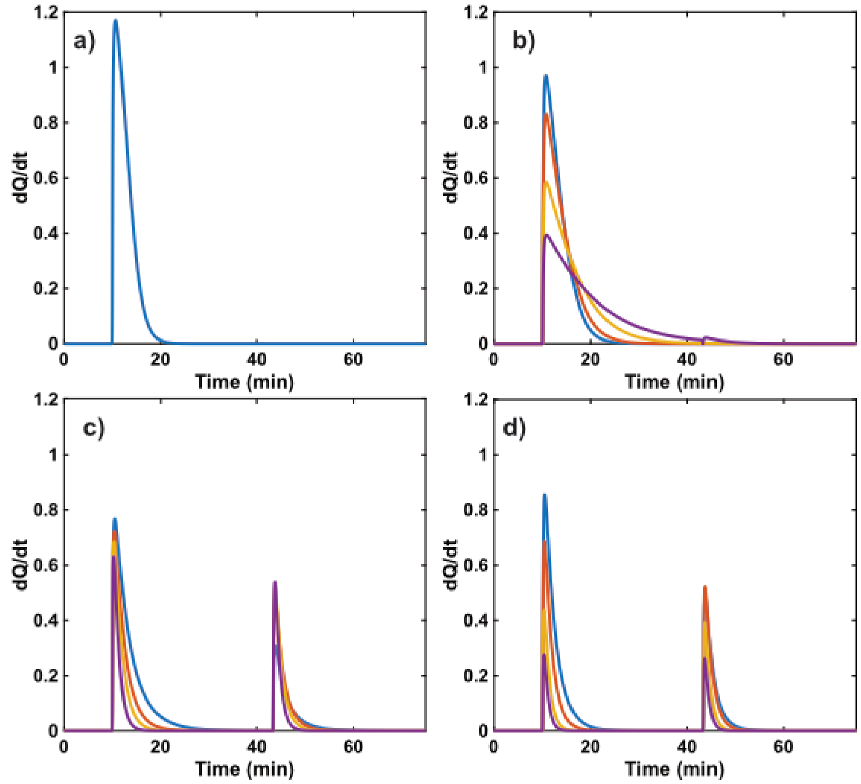


Figure 5.1: Simulated Isothermal Titration Calorimetry Kinetic Competition (ITC-KC) traces. a) No-inhibitor control. b) One-step rapid equilibrium with K_i = 100, 50, 20, and 10uM. c) Two-step rapid pre-equilibrium with K_i = 50uM and k_{inact} = 1e-2, 2e-2, 3e-2, and 5e-2 s^{-1}. d) Two-step rapid pre-equilibrium with K_i = 100, 50, 20, and 10uM and k_{inact} = 3e-2 s^{-1}. The concentration of inhibitor was set at 50uM for all simulations. Colours represent the different simulations listed in order: the first simulation is blue, the second is red, the third is orange, and the fourth is purple.

ITC-KC data obtained when the enzyme is injected into substrate/inhibitor mixtures fall into several different categories. ***I***) Inactive compounds give traces identical to that of the negative control (*Figure 5.1a*). ***II***) Compounds that rapidly and reversibly bind to the enzyme lead to shorter, elongated ITC peaks. The tighter the binding, the lower the initial velocity and the longer it takes the enzyme to complete the reaction, leading to broader peaks (*Figure 5.1b*). Nevertheless, once the signal returns to the baseline, all the substrate has been consumed, therefore the second injection of enzyme produces no peak. An exception occurs when the second injection is made before the first peak has returned to the baseline and some substrate remains (purple curve of *Figure 5.1b*). In this case, the second injection does produce a small peak. This can be avoided by increasing the spacing between the injections, increasing the amount of enzyme (to accelerate the reaction) or reducing the amount of inhibitor in the cell. These adjustments are not strictly necessary, as the analysis can be applied in either case. However, they are recommended, to ensure that all of the substrate is truly consumed in the first peak.

Conversely, compounds that completely inactivate the enzyme give peaks with much smaller areas, corresponding to less substrate cleavage than in *I* or *II*. In these cases, the rapid return of the ITC signal to the baseline is largely due to inhibition of the enzyme rather than exhaustion of the substrate. Consequently, the second injection produces a similar peak to the first, since there is substrate remaining in the sample cell (*Figure 5.1c/d*). This gives a simple visual test for the nature of inhibition; the presence of a second peak after the first returns to baseline clearly indicates that the inhibitor fully inactivates the enzyme on the seconds or minutes timescale. ***III*)** When EI is weakly populated (K_i>>[I]), the inhibitor does not instantaneously inactivate the enzyme when they are first mixed. Instead, the enzyme becomes progressively inhibited over time, following second order kinetics with a rate constant given by k_{inact}/K_i[296]. Thus the initial height of the ITC peak is identical to the negative control and returns to baseline in a manner that depends on the value of k_{inact}/K_i. (*Figure 5.1c*). ***IV*)** When E and EI are in rapid equilibrium and EI is substantially populated, the enzyme is immediately partly inhibited, leading to an ITC peak that is initially lower than the negative control by an amount that depends on the value K_i. The enzyme then becomes progressively more inhibited, and the ITC signal returns to baseline at a rate that depends on the value of k_{inact} (*Figure 5.1d*). Thus, for type II data, only the value of K_i can be extracted, for type III data, only the value of k_{inact}/K_i can be determined, and for type IV data, the values of both K_i and k_{inact} can be determined.

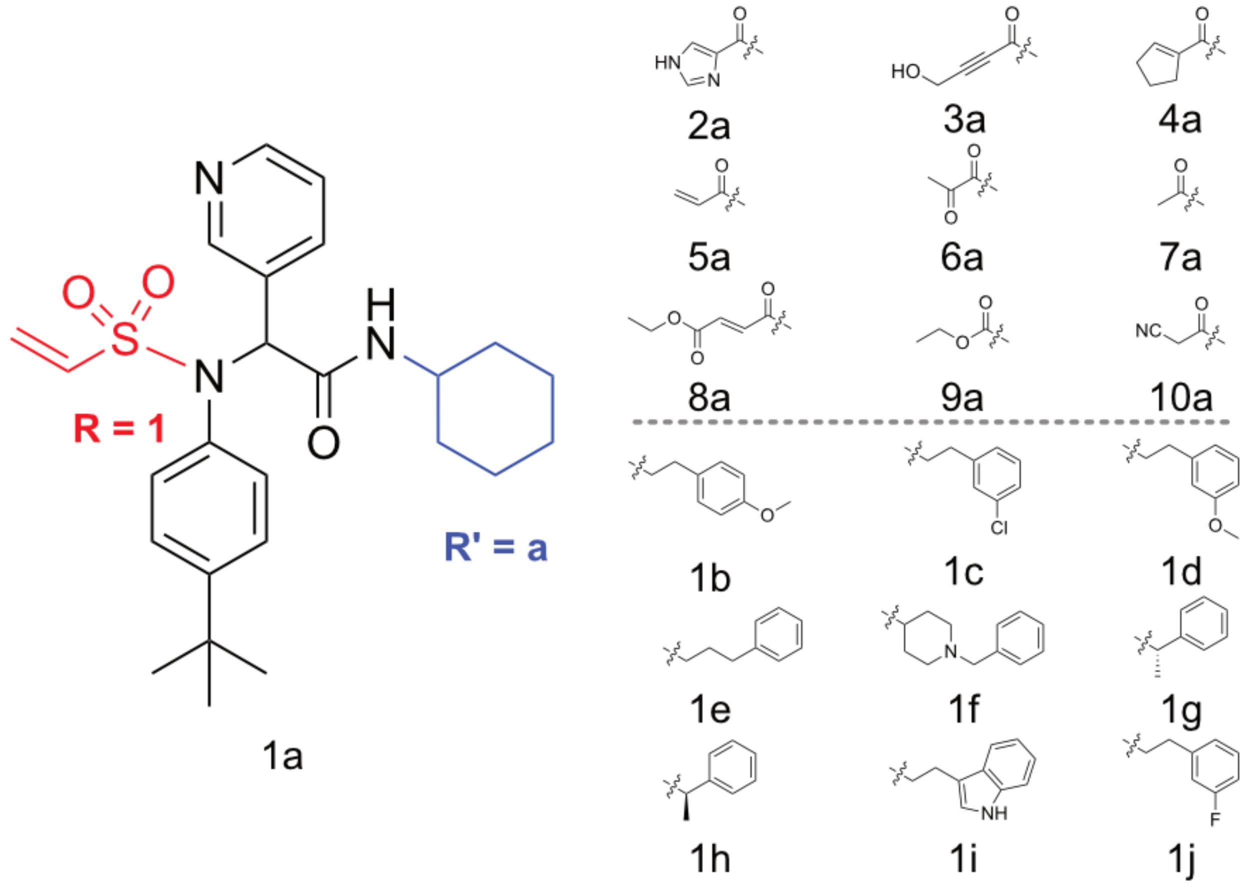


Figure 5.2: Chemical structures of the inhibitor library. 1a-10a contain the same chemical scaffold with different covalent warheads. 1a-1j have the same covalent warhead with different chemical scaffolds.

5.4.2 Mechanistic characterization of 3CL^{pro} inhibitors

To test our ITC-KC method, we selected a subset of compounds we had previously reported as potential 3CL^{pro} inhibitors[502]. These 19 molecules all share the same peptidomimetic core, shown in black in *Figure 5.2*. The first 10 molecules contain a cyclohexyl group at the *R′* position (blue). 9 of these bear different covalent warheads at the *R* position (red). Compound **2a** is the well-studied non-covalent inhibitor X77[505] with an imidazole at this location. The next nine molecules all contain a vinyl sulfonamide warhead with various replacements of the cyclohexyl group on the scaffold. In our nomenclature, the warheads at *R* are labelled **1** through **10** and the substitutions at *R′* are labelled **a** through **j**. This sampling of chemical space provides us with an opportunity to separately evaluate the influence of the reactive warhead and the scaffold on the inhibition mechanism.

We applied ITC-KC to the library of compounds described above (*Figure 5.2*) in an overnight experiment. No-inhibitor controls with 3CL^{pro} in the syringe and only substrate (a short peptide containing a native SARS-CoV-2 polyprotein cleavage site) in the sample cell were included in the series. The resulting ITC traces (*Supplementary Figure 5.1*) were fit to a Michaelis-Menten enzyme kinetic model, yielding the parameters (K_m= 290 ± 60 μM, k_{cat} = 2.2 ± 0.3 s^{-1}) which are close to the ranges of values previously reported for this enzyme[506-510]. For the inhibitors, experiments were initially performed with [I] = 50 μM. A second overnight experiment was then performed with [I] of either 10 or 100 μM; those compounds showing little inhibition with [I] = 50 μM were remeasured at a higher concentration while those giving substantial inhibition with [I] = 50 μM were remeasured at a lower concentration.

The ITC-KC data at two inhibitor concentrations were analyzed simultaneously according to *Scheme 5.2*, yielding values of K_i, 1/K_i, k_{inact}, k_{inact}/K_i, and the corresponding experimental uncertainties for each compound (see Materials and methods, *Figure 5.3, and Table 5.1*). The covalent inhibition type (I, II, III, or IV) was then assigned based on the relative values of the parameters and their uncertainties. A parameter was considered ill-defined and not significantly different from zero when its uncertainty was larger than its value. As summarized in *Table 5.1*, compounds were assigned as type I (inactive) when 1/K_i, k_{inact}, and k_{inact}/K_i were all ill-defined. Note that 1/K_i was used in this test, since its value is zero for compounds that do not interact, in contrast to K_i, whose value is infinite for non-binders. Compounds were assigned as type II (rapid equilibrium only) when only k_{inact} and k_{inact}/K_i were ill-defined, and 1/K_i was well-defined. Type III (slow, complete inhibition with negligible EI formation) was assigned when 1/K_i and k_{inact} were ill-defined and k_{inact}/K_i was well-defined. Type IV (slow, complete inhibition with substantial EI formation) was assigned when all the parameters were well-defined. As listed in Table 1, the library contained two type I compounds (with no reported parameters), two type II compounds (with reported K_i values), two type III compounds (with reported k_{inact}/K_i values) and thirteen type IV compounds (with reported k_{inact} and K_i values).

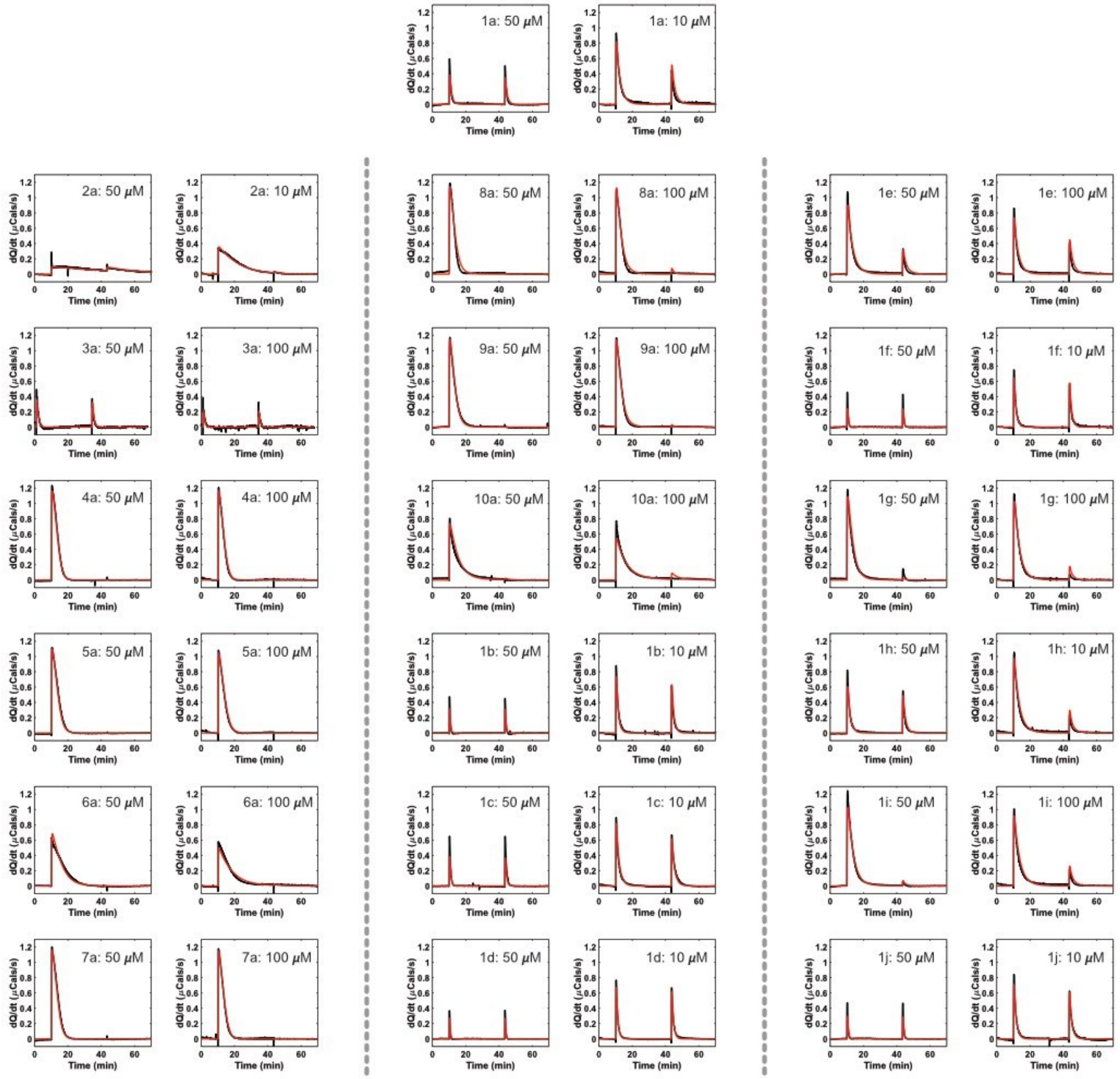


Figure 5.3: Baseline corrected experimental ITC traces for all compounds. Concentrations and compound names are shown on each plot individually. Black lines represent baseline corrected experimental data, red lines represent the best-fit parameters for the pre-equilibrium irreversible inhibition model described in the methods section.

Compounds	**k_{inact} (s^{-1})**	**K_i (μM)**	**$1/K_i$ (μM^{-1})**	**k_{inact}/K_i**
1a	3.0e-2 ± 0.5e-2	18 ± 4	6e4 ± 1e4	1600 ± 100
1a*	1.2e-2 ± 0.3e-2	14 ± 4	7e4 ± 2	900 ± 300
2a	-	1.9 ± 0.1	5.3e5 ± 0.3e5	-
3a	3.5e-2 ± 0.1e-2	27 ± 2	3.8e4 ± 0.3e4	1340 ± 70
3a*	7e-3 ± 1e-3	8 ± 2	1.3e5 ± 0.3e5	900 ± 300
4a	-	-	-	-
5a	-	450 ± 40	2.2e3 ± 0.2e3	-
6a	3e-4 ± 2e-4	31 ± 1	3.2e4 ± 0.1e4	11 ± 7
6a*	-	6 ± 1	-	
7a	-	-	-	-
8a	-	-	-	30 ± 10
9a	-	-	-	21 ± 9
10a	1.81e-3 ± 0.07e-3	39 ± 3	2.6e4 ± 0.2e4	46 ± 3
1b	7e-2 ± 1e-2	20 ± 5	5e4 ± 1e4	3700 ± 300
1b*	1.5e-2 ± 0.4e-2	6 ± 2	1.7e5 ± 0.6e5	3000 ± 1000
1c	4.1e-2 ± 0.4e-2	17 ± 3	6e4 ± 1e4	2400 ± 200
1d	9e-2 ± 1e-2	18 ± 4	6e4 ± 1e4	5200 ± 300
1e	1.9e-2 ± 0.2e-2	120 ± 20	9e3 ± 1e3	163 ± 8
1f	4.8e-2 ± 0.4e-2	10 ± 2	1.0e5 ± 0.1e4	4500 ± 300
1g	4e-2 ± 1e-2	700 ± 300	1.5e3 ± 0.6e3	48 ± 3
1h	2.8e-2 ± 0.4e-2	40 ± 7	2.6e4 ± 0.4e4	710 ± 30
1h*	1.3e-2 ± 0.8e-2	100 ± 77	1.0e4 ± 0.7e4	100 ± 100
1i	1.6e-2 ± 0.5e-2	240 ± 50	4.4e3 ± 0.9e3	66 ± 3
1j	5.6e-2 ± 0.8e-2	15 ± 3	7e4 ± 1e4	3600 ± 200

Table 5.1: Kinetic constants calculated for each inhibitor. Errors were calculated as described in the Materials and methods. Fitted parameters which had errors larger than their mean value are not reported. (*) Asterisk indicates parameters which were found from IDPC analysis.

To validate our results, we also performed traditional inhibitor-dependent progress curve (IDPC) analysis for five of the compounds in our library using a continuous FRET-based assay (*Figure 5.4*)[296, 496]. These values were largely aligned with the ITC-KC results (*Table 5.1*). The compound with largest k_{inact} and k_{inact}/K_i values according to ITC-KC (**1b**, 7×10^{-2} s^{-1}, 1600 M^{-1} s^{-1}) also had the largest k_{inact} and k_{inact}/K_i values according to IDPC (1.5×10^{-2} s^{-1}, 900 M^{-1} s^{-1}), while the compound with the smallest k_{inact} value (**6a**, 3×10^{-4} s^{-1}) showed essentially no covalent bond formation in the IDPC analysis. The compound with the largest K_i value according to ITC-KC (**1h**, 40 μM) also had the largest K_i value by IDPC (100 μM). The K_i values of the rest of the compounds varied between 18 and 31 μM according to ITC-KC and between 6 and 14 μM according to IDPC. The overall agreement of ITC-KC and IDPC k_{inact}/K_i values (r=0.99, *Supplementary Figure 5.2c*) was higher than the agreement of K_i (r=0.78, *Supplementary Figure 5.2b*) or k_{inact} (r=0.81, *Supplementary Figure 5.2b*) values obtained from the two methods, which is consistent with the results of simulations, described below, which showed that the accuracy of IDPC-derived k_{inact}/K_i values is much better than that of K_i or k_{inact}, alone.

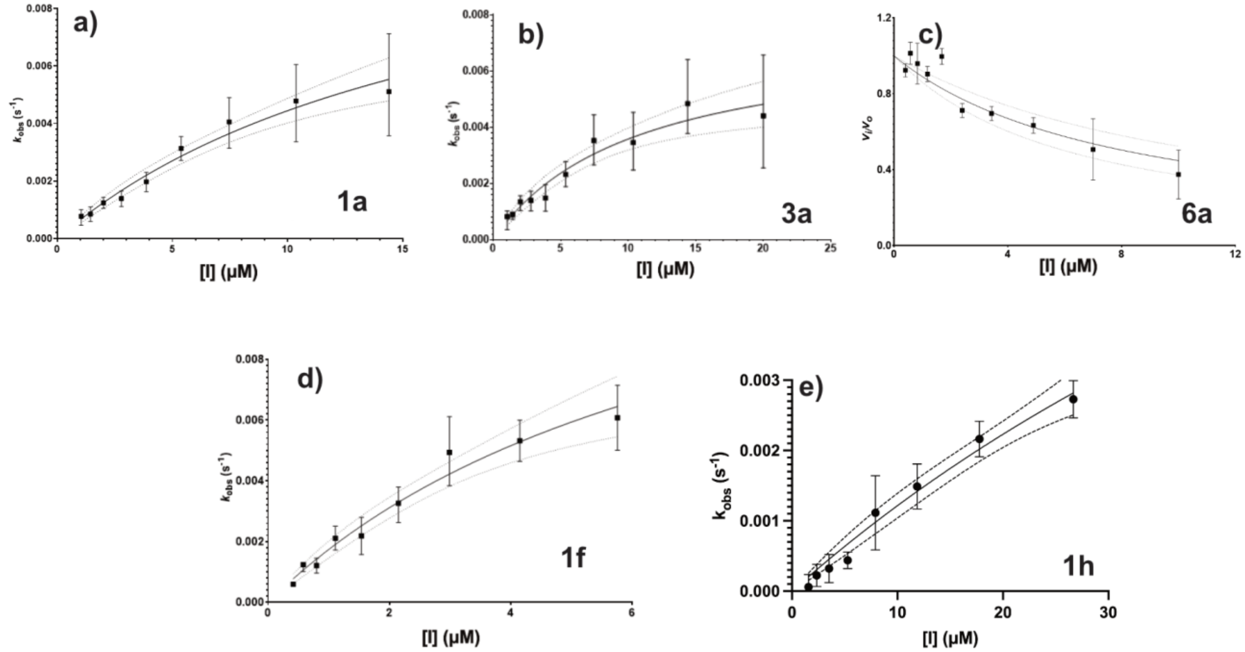


Figure 5.4: Experimental Inhibitor Dependent Progress Curve (IDPC) data. a) Compound 1a. b) Compound 3a. c) Compound 6a, IDPC analysis could not be run due to low k_{obs} so Dixon analysis was used instead[511]. d) Compound 1f. e) Compound 1h.

5.4.3 Comparison of ITC-KC with traditional methods

Mechanistic information on covalent inhibitors has historically been obtained by mixing enzyme, substrate, and various concentrations of inhibitor and monitoring the formation of product over time. When a continuous assay is available for the enzyme of interest, the reaction can easily be sampled at a large number of time points, and the data are analyzed in terms of inhibitor concentration-dependent progress curves (IDPCs)[296, 496]. When no continuous assay is available, typically many fewer time points are taken, and a series of time-dependent IC_{50}(t) values are calculated as input for a $TDIC_{50}$ analysis[296, 497]. To quantify the performance of our new ITC-KC method relative to these techniques, we have performed extensive simulations of the three approaches. We found that ITC-KC most accurately reproduced the correct K_i and k_{inact} values in all the simulations, sometimes by a large margin. IDPC produced slightly more accurate k_{inact}/K_i ratios under some, but not all, conditions tested. Furthermore, we found that the standard analysis method used for $TDIC_{50}$ datasets introduces large systematic errors in the extracted K_i, k_{inact}, and k_{inact}/K_i values. This can be corrected with a simple modification of the fitting procedure. However, even with the improved fitting, $TDIC_{50}$ simulations produced the least reliable values, which is unsurprising given the smaller number of data points collected in these experiments.

IDPC (and $TDIC_{50}$) experiments are performed under conditions where the uninhibited progress curves are linearly increasing, i.e. the rate of catalysis is constant throughout the experiment. However, in the presence of inhibitor, the enzyme gradually loses activity, leading to a curved profile that becomes horizontal after the enzyme is fully inactivated (Fig 3a). The shape of the progress curve measured by fluorescence is given by[296]

$$F_t = F_0 + \frac{v_i}{k_{obs}} * (1 - e^{-k_{obs}*t}) \qquad \textit{(Equation 5.2)}$$

where k_{obs} is the rate of formation of the E-I complex and v_i is the initial slope of the curve, F_t is the fluorescence at time t and F_0 is the background fluorescence. The value of k_{obs} depends on both the population of the EI intermediate (which is related to [I] and K_i) and the rate of covalent bond formation, k_{inact}, according to:

$$k_{obs} = \frac{[I]}{[I]+K_i\left(1+\frac{[S]}{K_m}\right)} k_{inact} \qquad \textit{(Equation 5.3)}$$

Thus a plot of k_{obs} versus [I] is hyperbolic with a maximum value of k_{inact} and is half-maximal when $[I]=K_i(1+[S]/K_m)$ (*Figure 5.4a,b,d,e, Figure 5.5b*). The values of k_{inact} and K_i can then be extracted from a Kitz-Wilson linearized plot[496] or by non-linear least-squares fitting[296]. A major technical challenge is that data must be obtained with $[I] > K_i(1+[S]/K_m)$ in order to define the hyperbola. However, under those conditions, the enzyme is already mostly in the EI inhibited state at t=0, meaning that very little product is formed at all and these data points suffer from low signal to noise ratios.

In a TDIC$_{50}$ analysis, the IC$_{50}$ depends only on K_i at t=0. At longer times, the value of IC$_{50}$ decreases in a K_i- and k_{inact}-dependent manner, reaching zero after an infinitely long incubation (*Figure 5.5g*). The shape of the time-dependent IC$_{50}$ profile is given by[497]:

$$IC_{50}(t) = K_i * \left(1 + \frac{[S]}{K_m}\right) * \left(\frac{2-2e^{-\eta_{IC_{50}} * k_{inact} * t}}{\eta_{IC_{50}} * k_{inact} * t} - 1\right) \qquad \textit{(Equation 5.4)}$$

$$\eta_{IC_{50}} = \frac{IC_{50}(t)}{K_i * \left(1+\frac{[S]}{K_m}\right) + IC_{50}(t)} \qquad \textit{(Equation 5.5)}$$

The value of IC$_{50}$(t) appears on both the left- and right-hand sides of *Equation 5.8*, thus its value cannot be computed directly from values of K_i and k_{inact}. In order to extract these parameters from a fit of experimental IC$_{50}$(t) values, the experimental values are used in place of IC$_{50}$(t) in *Equation 5.9*, and the values of K_i and k_{inact} are varied so that the left-hand side of *Equation 5.8* matches the experimental values as closely as possible[497]. A technical challenge here is that IC$_{50}$(t) values must be measured after very short incubations in order to confidently define the values of K_i and k_{inact}. However, very little product is formed after such short incubations, therefore these IC$_{50}$(t) curves are poorly defined and the extracted IC$_{50}$(t) values are inaccurate (*Figure 5.5f/g*).

We calculated 1000 sets of synthetic experimental data for IDPC,TDIC$_{50}$, and ITC-KC experiments, using simulated noise that matched the amplitude and time correlations of our actual data sets, setting K_i=10 µM and k_{inact}=0.1 (*Figure 5.5*), 0.01 (*Supplementary Figure 5.3*), and 0.001 (*Supplementary Figure 5.4*) s^{-1}. Representative synthetic raw data are shown in *Figure 5.5a,f,k.* The IDPC and TDIC$_{50}$ data sets were analyzed to yield 1000 sets of k_{obs} and IC$_{50}$(t) values, superimposed as thin black lines in *Figure 5.5b* and *Figure 5.5g*, respectively. Each set of k_{obs} and IC$_{50}$ values was then fitted to extract the apparent values of K_i, k_{inact}, and k_{inact}/K_i, with histograms of the resulting values shown in *Figure 5.5c-e* and *Figure 5.5h-j*. For comparison, a representative synthetic ITC-KC dataset is shown in Fig 3k, and histograms of the 1000 extracted parameters are shown in *Figure 5.5l-m*. The K_i and k_{inact} distributions obtained from IDPC and TDIC$_{50}$ simulations are far broader than those obtained using the ITC-KC method. For IDPC experiments, this is largely due to the large uncertainties in k_{obs} values obtained with large [I], visualized in the wide spread of the black lines in *Figure 5.5b*, and discussed above. For TDIC$_{50}$, the problem is two-fold. Firstly, the approximation of using experimental IC$_{50}$(t) values in the right-hand side of *Equation 5.8* leads to large systematic errors, such that extracted K_i

and k_{inact} parameters (yellow histograms in *Figure 5.5h,i*) are about 10-fold smaller than the true values. We find this can be partly corrected by a numerical, self-consistent solution of *Equation 5.8* (blue histograms in *Figure 5.5h,i*). Even then, most of the extracted K_i and k_{inact} are still far from the correct values, due to the sharp decay of the IC_{50}(t) curve and high uncertainty of the IC_{50}(t) values at short times, as discussed above (*Figure 5.5g*). In contrast, the distribution of IDPC-derived k_{inact}/K_i values (*Figure 5.5e*) is much narrower than those of the individual K_i and k_{inact} values, and slightly narrower than the corresponding ITC-KC distribution (*Figure 5.5n*). This is because the k_{inact}/K_i ratio is equal to slope of the k_{obs} vs [I] curve at low [I], which is much better defined than the asymptote at high [I] (*Figure 5.5b*). For $TDIC_{50}$, the approximate approach systematically underestimates k_{inact}/K_i, while the self-consistent calculation produces a distribution that is centred on the correct value (*Figure 5.5j*) and is only slightly wider than ITC-KC distribution (*Figure 5.5n*). Simulations with k_{inact}=0.01 s^{-1} produce similar results (*Supplementary Figure 5.3*). The IDPC and $TDIC_{50}$ K_i and k_{inact} distributions are narrower than with k_{inact}=0.1 s^{-1}, but still substantially more broad than those of ITC-KC. For $TDIC_{50}$, the approximate solution to Equation 6 leads to large underestimations of K_i, k_{inact}, and k_{inact}/K_i which are rectified by the self-consistent numerical approach. Finally, in simulations with k_{inact}=0.001 s^{-1}, ITC-KC outperformed IDPC and $TDIC_{50}$ in reproducing correct K_i, k_{inact}, and k_{inact}/K_i (*Supplementary Figure 5.4*). While extending the length of the progress curves past 15 minutes could improve the IDPC results, longer incubations push the system out of the linear regime and are thus not straightforward. Therefore, ITC-KC provides more accurate values of K_i and k_{inact} than does either the IDPC or $TDIC_{50}$ method over a wide range of reaction rates.

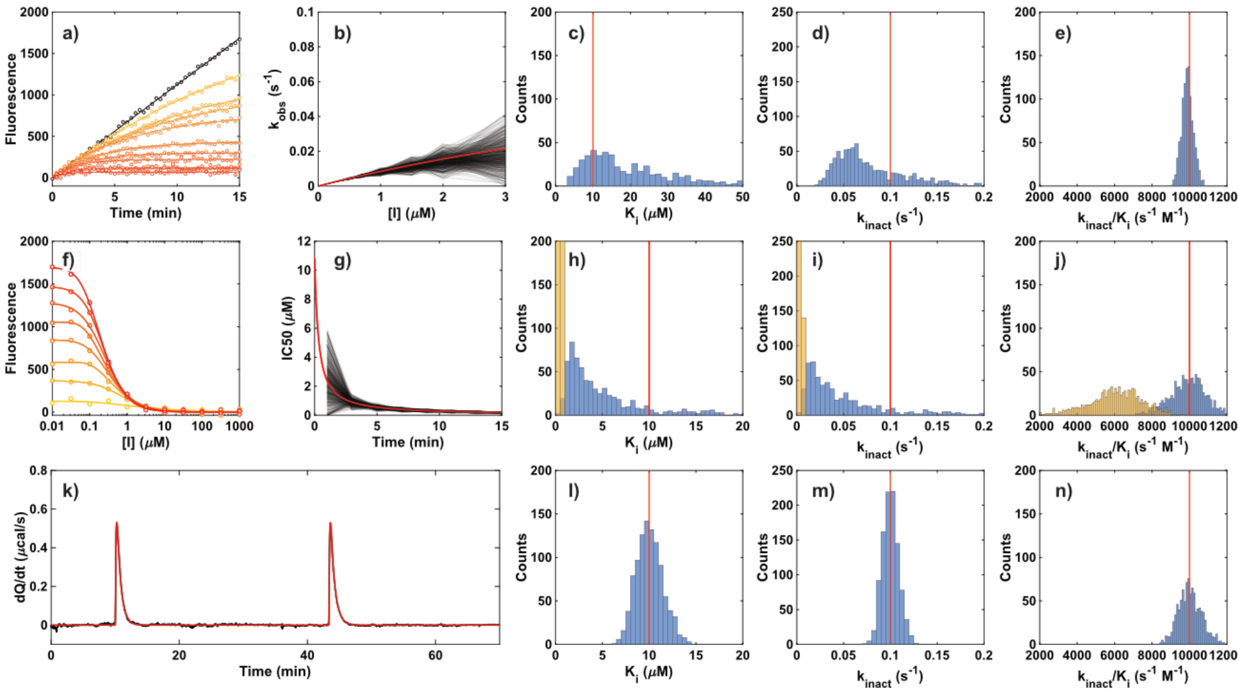


Figure 5.5: Comparison of IDPC, TDIC$_{50}$, and ITC-KC at K_i = 10μM and k_{inact} = 0.1 s^{-1}. a) Fluorescence traces at increasing inhibitor concentrations (yellow is low [I], red I high [I]). Dots indicate simulated noisy data, lines indicate fits to Equation 5.2. The black line indicates the no-inhibitor control. b) k_{obs} as a function of different inhibitor concentrations. Black lines represent 1000 samples from panel a, red line represents the true value. c-e) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value. f) TDIC$_{50}$ traces at different time points (fast times points are in yellow, long time points are in red). g) IC$_{50}$ values as a function of time, black lines represent 1000 samples from panel f, red line represents the true values. h-i) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value. k) ITC-KC plot at 10 μM inhibitor. i-n) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value.

5.4.4 Scope of the ITC-KC experiment

We then tested the scope of the ITC-KC experiments by analysing simulated noisy data generated with K_i values ranging from 10 nM to 1 mM and k_{inact} ranging from 10^{-5} to 1 s^{-1}, with 600 data sets generated for each *Ki*, k_{inact} pair (see Materials and methods). Relative RMSDs were calculated between the true and fitted values of K_i, k_{inact}, and k_{inact}/K_i, and are shown as contour plots in *Figure 5.6a-c*. The shaded areas indicate the values of K_i and k_{inact} for which ITC-KC can extract the true values to within a relative RMSD of 25%, and scatter plots of the true versus fitted values are shown in *Figure 5.6d-f*. The method can reliably measure all three parameters over about 4 orders of magnitude. The tightest K_i values that allowed reliable parameter extraction were about 100 nM. This limit is effectively set by the concentration of enzyme used in the assay, as it is not possible to accurately measure binding affinity when the enzyme is in large excess of the K_i. Thus, using lower concentrations of enzyme will enable one to characterize compounds with lower K_i values, but will reduce the magnitude of the heat signal. The optimal concentration will depend on the particular enzyme being used and the properties of the compounds of interest. The largest values of K_i that allowed K_i and k_{inact} to be extracted separately was about 100 μM. Increasing the concentration of inhibitor beyond that used in these simulations (50 μM) would extend this limit, but in practice would be constrained by compound solubility. The values of k_{inact}/K_i could be extracted from systems with even weaker K_i's, provided that k_{inact} was large enough to produce relatively rapid inhibition with a negligible population of EI (Type III curves). The smallest values of k_{inact} that could be measured were roughly 10^{-4} s^{-1}. In principle, this could be extended to even slower rates by increasing the amount of substrate and leaving longer delays between the enzyme injections. However, this would substantially reduce the throughput of the experiment. Values of K_i can be measured for arbitrarily slow k_{inact}, as the ITC curves simply become Type II as the rate of the covalent step approaches zero. The fastest measurable k_{inact} values were slightly above 0.1 s^{-1}. This upper limit is set by the response function of the calorimeter, which is effectively the dead time of the instrument.

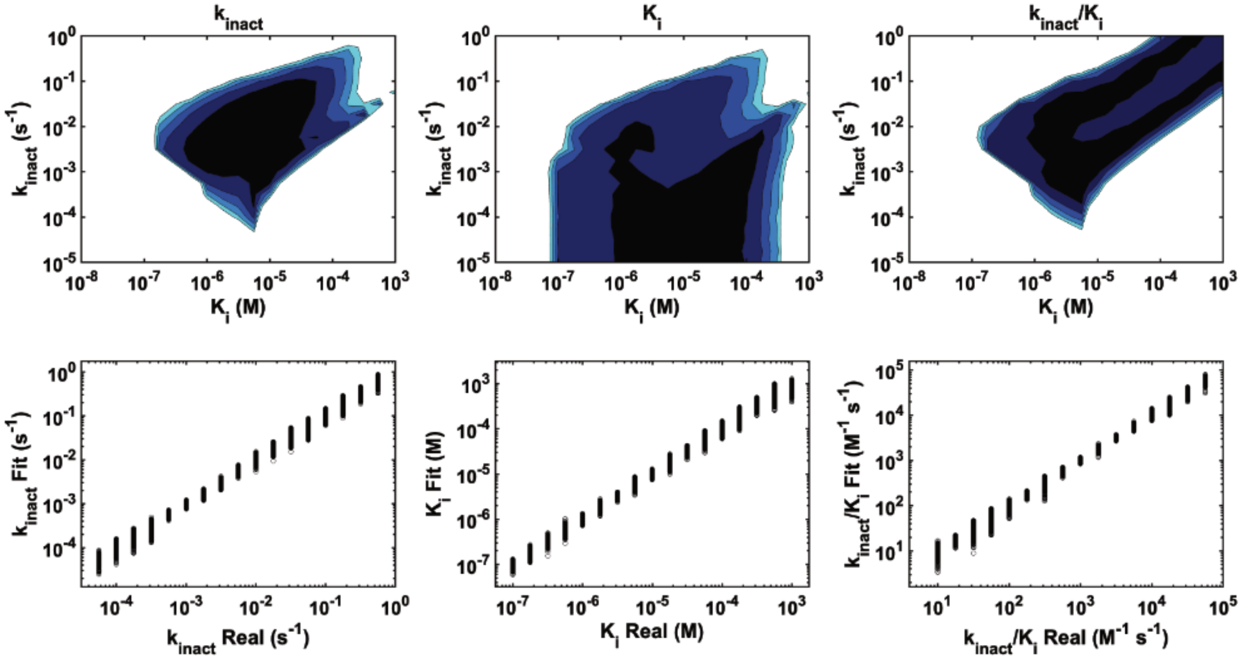


Figure 5.6: Root-mean squared deviation plots of K_i and k_{inact} for several orders of magnitude. All plots are on a log scale. The contours in a/b/c represent RMSD values of 5% (black) up to 25% (cyan). Panels d/e/f represent show the fitted parameters vs the real parameters used in the simulations.

5.5 Discussion

Our groups have previously developed approaches for using ITC to characterize the kinetics of inhibitor binding[307], however, ITC-KC has many capabilities that these previous methods did not. Firstly, and most importantly, ITC-KC allows one to determine the values of both K_i and k_{inact} for covalent inhibitors. Our previous approach only measured the net association and dissociation rates, meaning that crucial mechanistic detail was lost. Secondly, ITC-KC is suitable for automation, as the enzyme and substrate/inhibitor solutions used to fill the syringe and sample cell are stable for hours or days in 96-well plates. In contrast, our previous method required an enzyme/substrate mixture reacting at a constant rate that had to be prepared by hand immediately prior to starting the experiment, and therefore was not amenable to automation. More generally, ITC-KC offers many advantages over current approaches for quantifying the mechanisms of covalent and other slow inhibitors. In our simulations, ITC-KC performed much better than the commonly used inhibitor dependent progress curve (IDPC) approach or the time-dependent IC_{50} ($TDIC_{50}$) method at determining K_i and k_{inact} values, and nearly as well in determining the ratio k_{inact}/K_i. The availability of separate K_i and k_{inact} parameters is crucial for mechanism-based optimization of covalent inhibitors. Furthermore, both IDPC and $TDIC_{50}$ require tens of enzyme assays to be performed for each inhibitor of interest[296]. In contrast, ITC-KC yields K_i and k_{inact} in a single, hour-long experiment (plus a no-inhibitor control that is shared for all compounds tested). Finally, ITC is an essentially universal approach, in that nearly all enzymatic reactions release or absorb heat, which is detected by the calorimeter, meaning that this approach can be applied generally in drug development programs. In contrast, current approaches rely on pre-existing assays which are different for each enzyme. Continuous (real-time) assays are usually not possible with the native substrate alone; discontinuous assays are relatively time-consuming, costly, and necessitate the use of the $TDIC_{50}$ approach which performed by far the most poorly in our simulations. The main drawback of the ITC-KC method is that its sensitivity depends on the enthalpy of the reaction (ΔH_r) and the velocity of the enzyme (k_{cat}); slower enzymes catalyzing more isothermic reactions will need to be present at higher concentrations to generate sufficient quantities of heat. For example, our ITC-KC experiments utilized ~1.5 µM $3CL^{pro}$ during each injection (k_{cat}=2.2 ± 0.3 s^{-1}, ΔH_r = 3.6 ± 0.2 kcal mol^{-1}) while the fluorescence-based IDPC experiments were performed at 40 nM. To some extent, the fewer number of experiments needed for ITC-KC and the cheaper (non-fluorogenic) substrate balance the costs of the experiments. However, higher concentrations of enzyme place a lower limit on the values of K_i that can be measured. It should be noted that much lower concentrations would be needed for more active enzymes. For example, prolyl oligopeptidase (k_{cat}=42.19 s^{-1}, ΔH_r = -6.72 ± 0.06 kcal mol^{-1}) produces similar heat signals to the ones measured here at a concentration of just 40 nM[307].

Our method has a number of adjustable run parameters which can be tuned to optimize sensitivity, dynamic range, as well as material and time requirements. These are **i)** the amounts of substrate and enzyme used, **ii)** the concentration of inhibitor used, and **iii)** the delays between the injections. Firstly, the concentration of substrate used should be greater than or equal to the K_m. We find that this is the minimum requirement for determining K_m and k_{cat} from an ITC peak. The value of K_m must be known in order to account for competition between inhibitor and substrate during the reaction. The concentration of enzyme should be large enough to generate enough heat for accurate quantitation. We recommend peaks that are greater than or equal to 1 µcal s^{-1} in height. The amount of enzyme required to achieve this will vary with K_m, k_{cat}, ΔH_r, and [S]. An important consideration is that wider ITC peaks allow the experiment to characterize more slowly binding inhibitors. An inhibitor must be able to completely inactivate the enzyme before all of the substrate is consumed on order to reliably obtain K_i and k_{inact} values. The higher the initial concentration of substrate, the more time inhibitors will have to inactivate the enzyme. However, widening the peak also requires the injections to be further apart, which lengthens the experiment and reduces the throughput. Secondly, the concentration of inhibitor in the cells must be large enough to produce a clear effect on the ITC signal. For Type II (rapidly equilibrating) peaks, this means that the concentration of inhibitor must not be substantially lower than the dissociation constant; we find that values of [I] greater than about $K_i/4$ produce quantifiable slowing of catalysis. On the other hand, the inhibitor concentration must be low enough so that the enzyme can still consume all of the substrate in a reasonable length of time; we find that values up about 20-fold the K_i are suitable. Similar considerations hold for Types III and IV (slow, completely inhibited) peaks. In order to quantify the stability of the non-covalent intermediate EI, the inhibitor concentration must not be much lower than K_i. On the other hand, the concentration must still be low enough to maintain a population of active enzyme at the start of the experiment. Furthermore, the rate of E–I formation increases with increasing [I]; this rate should not exceed about 2 min^{-1} in order to be quantifiable by ITC-KC. We recommend performing an initial scan with a mid-range inhibitor concentration and a second scan with a higher or lower concentration, depending on the initial data, as was done in this study. Finally, the drawback to increasing the delay between injections is lengthening of the experiment time. Conversely, the benefit is increased tolerance to highly broadened peaks brought about by potent, yet incomplete inhibition of the enzyme. Ultimately, there is no single set of parameters that are optimized for every enzyme. We would recommend starting with [S] > K_m and sufficient [E] to fully convert S to P with 10 to 20 minutes. If experimental IC_{50} values are available, then [I] should be set close to the upper end of these values if the IC_{50} incubation time was short (<5 minutes). If the incubation time was long (>20 minutes), then [I] values of 5- to 10- fold greater than the IC_{50} would be appropriate. For a pilot experiment, leaving a generous delay between injections (60-90 minutes) is advantageous for ensuring peaks have time to return to the baseline. As parameters are optimized, the delay between injections can be reduced to slightly more than the width of the broadest peak. Fortunately, this optimization need only be performed once for a given enzyme and class of inhibitors. All subsequent inhibitors within a chemical series can be rapidly characterized using ITC-KC with no further optimization.

The ITC-KC screen has provided us with a comprehensive set of mechanistic data to better understand structure-activity relationships within our library of potential $3CL^{pro}$ inhibitors (Fig 2 and Table 1). We tested ten different warheads in this study, all with the same chemical scaffold. Four of them showed complete inhibition with a well-populated EI intermediate (Type IV peaks): **1a** (vinyl sulfonamide), **3a** (alkyne), **6a** (diacetyl), and **10a** (nitrile). Two showed complete inhibition with negligible EI formation (Type III peaks): **8a** (acrylate) and **9a** (ethyl ester), and two showed a rapid equilibrium with no slow step (Type II peaks): **2a** (the parent non-covalent compound X77) and **5a** (vinyl ketone)**.** As expected, different warheads produced very different reactivities with k_{inact} ranging over more than two orders of magnitude from unmeasurably slow (**5a**) to slow (0.0003 s^{-1}, **6a**), to relatively rapid (0.035 s^{-1}, **3a**). Interestingly, the stability of the non-covalent intermediate EI also varied widely. The vinyl sulfonamide (**1a**, K_i=18 μM) bound about 10-fold more weakly than the parent non-covalent compound (**2a**, K_i=1.9 μM), which has an imidazole group in place of a reactive group. This may be explained by the presence of an extra hydrogen bond formed between the enzyme and the imidazole group which has been seen in the crystal structure[502]**.** **5a** bound about 200-fold more weakly than **2a**, while the Type III compounds **8a** and **9a** bound so poorly that the population of EI was undetectable, indicating a highly destabilized non-covalent complex. Taken together, this points to an important role for the warhead in modulating the affinity of the non-covalent intermediate in addition to reacting to form the covalent bond. The warheads in this study differ in their stabilization of the non-covalent intermediate by more than an order of magnitude, and all are far less stabilizing than the imidazole of non-covalent **2a**.

We also characterized 10 different scaffolds with the same vinyl sulfonamide warhead. All of these compounds behaved as two-step, Type IV covalent inhibitors. Seven of these compounds gave similar K_i values in the 10 to 40 μM range, including **1a**. Three others bound 5- and 10-fold more weakly than **1a**. Interestingly, five of the compounds had larger k_{inact} values than **1a**, even though they all shared the same reactive group. In the case of **1c**, in which the cyclohexyl group on the scaffold was replaced with an *m*-chlorotolyl group, the non-covalent complex was destabilized by about 50%, but the warhead reacted more than twice as rapidly, leading to a slight increase in potency. The most potent compound we tested, **1d**, formed the third most stable non-covalent intermediate within the library (behind **2a** and **1f**) and reacted the most rapidly of all compounds studied. These results strongly suggest that the inhibitor scaffold plays a role not just in stabilizing the EI and E–I complexes through non-covalent interactions, but also helps to position the warhead to effectively react with the target.

5.6 Conclusion

Our ITC-KC method represents a powerful new tool for characterizing covalent inhibition in the context of pharmacological drug and probe development. As a calorimetry-based method, it offers an essentially universal way to screen potential covalent inhibitors without the need to develop or adapt new assays for each enzyme of interest. Furthermore, it promises to provide more robust mechanistic parameters than current methods, particularly the elucidation K_i and k_{inact}. As proof of principle, we applied ITC-KC to 19 rationally designed potential 3CL^{pro} inhibitors. We collected a complete dataset in under 48 hours, which we validated with traditional time-dependent IC_{50} measurements. Our dataset illuminated the complex interplay between the chemistries of the scaffold and warhead, K_i and k_{inact}. The scaffold not only stabilizes the target/inhibitor complex, but also influences the rate of covalent bond formation. The warhead not only reacts to form a covalent bond but also participates in stabilizing the non-covalent intermediate complex. This highlights the need to collect mechanistic information to guide covalent drug design. We believe that our new ITC-KC experiment will be a valuable asset in meeting this need.

5.7 Materials and methods

5.7.1 Protein production and purification

SARS-CoV-2 3-chymotrypsin-like cysteine protease (3CLpro) was expressed and purified as previously described and stored in aliquots at -80°C[502].

5.7.2 ITC Experimental Conditions

All experiments were carried out in potassium phosphate buffer with 68.5 mM NaCl, 1.35 mM KCl, 4 mM Na_2HPO_3, 1 mM KH_2PO_4, and 0.5% (*v/v*) DMSO. Inhibitors were dissolved into DMSO then diluted into buffer prior to the experiment. The sample cell contained 362 μM of the peptide Cbz-TSAVLQSGFRK (CanPeptide, Montreal, QC, Canada) dissolved in the phosphate buffer with either 10 μM inhibitor, 50 μM inhibitor, 100 μM inhibitor, or DMSO. The injection syringe contained 115 μM 3CLpro dissolved in the phosphate buffer with 0.5 mg/mL BSA, and 1mM DTT.

5.7.3 ITC Data Collection

ITC experiments were performed on either an automated PEAQ ITC instrument, using a 96-well plate which was held in a temperature-controlled chamber and set at 4 °C or run manually on an ITC200 instrument (Malvern Panalytical Ltd, UK). Each experiment was run at 25 °C in high-feedback mode with a 1s signal averaging window, a stirring rate of 750 rpm, pre-injection delay of either 600 s (automated PEAQ-ITC) or 120 s (ITC200), and a reference power of 7. The syringe was held at room temperature for 10 minutes to allow the solution temperature to equilibrate. Two sequential injections of 2.5 μL were injected over a period of 5 s and measured for 2000 s each. Each inhibitor was run at two different concentrations, 50 μM for the initial screen and then ran at either 10 μM or 100 μM depending on the level of inhibition. A negative control was run before each set of inhibitors. 3a was run on the ITC200 and all other inhibitors were run on the automated PEAQ-ITC.

5.7.4 Kinetic simulations

All simulations were performed with MATLAB 2023a. The differential equations shown below describe Michaelis-Menten kinetics and pre-equilibrium irreversible inhibition. These equations were numerically integrated using MATLAB's built in ODE solver *ode15s.*

Michaelis-Menten Kinetics

$$\frac{d[S]}{dt} = -\frac{k_{cat}[E][S]_t}{K_m+[S]_t}$$ *(Equation 5.6)*

$$\frac{d[P]}{dt} = \frac{k_{cat}[E][S]_t}{K_m+[S]_t}$$ *(Equation 5.7)*

Pre-equilibrium irreversible inhibition

$$[ES]_t = \frac{K_i[S]_t[E]_t}{K_mK_i+K_i[S]_t+K_m[I]_t}$$ *(Equation 5.8)*

$$[EI]_t = \frac{K_M[I]_t[ES]_t}{K_i[S]_t}$$ *(Equation 5.9)*

$$\frac{d[S]}{dt} = -\frac{k_{cat}[E]_t[S]_t}{K_m\left(1+\frac{[I]}{K_I}\right)+[S]_t}$$ *(Equation 5.10)*

$$\frac{d[P]}{dt} = \frac{k_{cat}[E]_t[S]_t}{K_m\left(1+\frac{[I]}{K_I}\right)+[S]_t}$$ *(Equation 5.11)*

$$\frac{d[I]}{dt} = -k_{inact}[EI]_t$$ *(Equation 5.12)*

$$\frac{d[E]}{dt} = -k_{inact}[EI]_t$$ *(Equation 5.13)*

$$\frac{d[E-I]}{dt} = k_{inact}[EI]_t$$ *(Equation 5.14)*

where $k_{cat}$and K_M are the catalytic rate and the Michaelis constant respectively. $[E]_t, [S]_t, [P]_t, [I]_t, [EI]_t, \text{and } [E-I]_t$ are the total concentrations of enzyme, substrate, product, inhibitor, non-covalently bound enzyme-inhibitor complex, and covalently bound enzyme inhibitor complex at time t. K_i and k_{inact} are the affinity of the inhibitor and the reactivity of the warhead respectively.

5.7.5 ITC Fitting procedure.

The automated PEAQ ITC had two no-inhibitor controls and the ITC200 had one. ΔH_r values were averaged from the no-inhibitor controls, inactive compounds, and competitive inhibitors of the automated PEAQ-ITC, and were found to be 3.6 ± 0.2 kcal mol^{-1}. The DMSO control for the ITC200 experiments had a lower ΔH_r than the PEAQ-ITC at 2.5 kcal mol^{-1}, which was likely due to measurement errors when weighing out the substrate. To account for this discrepancy the concentration of substrate for experiments run on the ITC200 was normalized by using the ratio of the ΔH_r measured in the ITC200 and automated PEAQ-ITC. All the no-inhibitor controls were fit to the Michaelis-Menten model to obtain the parameters k_{cat} = 2.2 ± 0.3 s^{-1} and K_m = 290 ± 60 μM. The pre-equilibrium irreversible inhibition model was used for all inhibitors. Fits were initiated with the starting parameters of k_{inact} = 0, 1e-3, 1e-2, 1e-1 s^{-1} and K_i = 1, 10, 100, 1000 μM.

Kinetics were simulated using *Equation 5.8* to *Equation 5.14* with the inclusion of *Equation 5.15* to account for dilution of the contents of the cell during the duration of the injection. Each species present in the model used the following equation to account for dilution, which was included alongside the rate equations for each mechanism.

$$\frac{d[cell]}{dt} = [syringe]_t * \frac{V_{inj}}{V_{cell}*t_{inj}} - [cell]_t * \frac{V_{cell}+V_{inj}}{t_{inj}} \ \langle For\ t = 0{:}\, t_{inj}\rangle \qquad \textit{(Equation 5.15)}$$

Where V_{inj} and V_{cell} are the volume of each injection and total volume of the cell respectively. t_{inj} is the length of the injection.

The instantaneous heat, $h(t)$, is calculated from the enthalpy of the reaction, ΔH_r, the total volume of the cell, V_{cell}, and the rate of product formation, $\frac{d[P]}{dt}$, according to

$$h(t) = \Delta H_r * V_{cell} * \frac{d[P]}{dt} \qquad \textit{(Equation 5.16)}$$

The instantaneous heat generated by the reaction, $h(t)$, was then convoluted with the instrument response function $f(t)$ according to obtain the theoretical experimental signal $g(t)$[306].

$$g(t) = h(t) \otimes f(t) \qquad \textit{(Equation 5.17)}$$

A second order polynomial $b(t)$, with coefficients $a, b, \text{and } c$ was used to account for curved baselines in the experimental ITC signal,

$$base(t) = at^2 + bt + c \qquad \textit{(Equation 5.18)}$$

Injection artifacts $inj(t)$ caused by mismatch between the solutions in the syringe and cell were determined by averaging the second injection from the no-inhibitor controls for each instrument and subtracted from each injection individually.

The kinetic parameters were globally fit to all replicates by minimizing the target function

$$RSS = \sum_{n=0}^{N} \left(\frac{dQ}{dt}(t_n) - g(t_n) - base(t_n) - inj(t_n) \right)^2 \qquad \textit{(Equation 5.19)}$$

where the RSS is the residual sum of squared differences, $\frac{dQ}{dt}(t_n)$ is the raw experimental data, and N is the total number of time points. The minimization of this function was done in MATLAB 2023a using *fminsearch.*

5.7.6 Statistical analysis

We found large errors in the amount of substrate present in the ITC cell, from our no-inhibitor and inactive compounds. This could come from pipetting errors, or from enzyme diffusing out of the syringe during ITC equilibration. In order to account for this, errors for fitted parameters were calculated using a Monte-Carlo approach, where the concentration of substrate in the cell was varied according to the mean (3.6 kcal mol^{-1}) and standard deviation (0.2 kcal mol^{-1}) of the ΔH_r values seen in the experiments. Each inhibitor was fit 1000 times using different substrate concentrations drawn from a normal distribution with the experimental mean and standard deviation, and resampling each trace with replacement, resulting in 1000 different sets of kinetic parameters. The K_i and k_{inact} values for each inhibitor were then taken as the mean of these fitted parameters, and the error was reported as the standard deviation.

5.7.7 ITC Noise generation

Distinctly non-random noise from the ITC signal was observed in experimental residuals. This was modelled by taking the mean absolute value of the Fourier transform of our inactive inhibitor and no-inhibitor control residuals. Half of this Fourier transform was taken and multiplied by random noise with a mean of zero and a standard deviation of one. This was joined with its reverse conjugate and an inverse Fourier transform was performed to give random noise with the same frequency spectrum as the original residuals. Finally, noise was weighted corresponding to the standard deviation of each time point, measured as the standard deviation of the experimental residuals at time t.

5.7.8 Inhibitor concentration-dependent progress curve simulations

Kinetics were generated using *Equation 5.8* to *Equation 5.14* with the modification of $\frac{d[S]}{dt} = 0$, which was necessary to produce linear product formation in the no-inhibitor control. Conversion from [P] was done with a factor of 277 Counts/μM, which was measured from experimental no-inhibitor controls. Random noise was added to each trace which had a standard deviation of 18 counts, once again taken from our no-inhibitor controls. Kinetics were simulated for 15 minutes in 20 second intervals, matching what was achieved experimentally. The concentration of substrate was 25 μM, the concentration of enzyme was 40 nM, and the concentrations of inhibitor was 1, 2, 3, 5, 8, 13, 21 ,34, 55, 89 μM.

1000 simulated fluorescence traces at each inhibitor concentration were fit to *Equation 5.2* to obtain values of k_{obs}. This gave 1000 values of k_{obs} at each inhibitor concentration, outliers were removed from each set of k_{obs} using MATLABs *isoutlier* function, fitting weights were used as the standard deviation (σ) of this set of k_{obs}. Finally, 1000 sets of k_{obs} values at each inhibitor concentration were randomly sampled with resampling to produce 1000 sets of k_{obs} at each inhibitor concentration. Traces were fit using MATLABs *fminsearch* function and *Equation 5.3* to according to

$$RSS = \sum_{n=0}^{N} \left(\frac{k_{obs(exp)_n} - k_{obs(calc)_n}}{\sigma_n} \right)^2 \qquad \textit{(Equation 5.20)}$$

Where $k_{obs(\mathrm{exp})}$ is the rate found from fitting to the *Equation 5.2* to simulated fluorescence traces. $k_{obs(\mathrm{calc})}$ is the rate given by *Equation 5.3* at a specific K_i and k_{inact}*, and* σ is the standard deviation of $k_{obs(\mathrm{exp})}$ at a specific inhibitor concentration (*n*).

5.7.9 Time-dependent IC$_{50}$ Simulations

Fluorescence data was generated as described above in the inhibitor concentration-dependent progress curves simulations, using concentrations of inhibitor of 0.01 to 1000 µM. Time points of 1, 3, 5, 7, 9, 11, 13, and 15 minutes were used. At each time point, IC$_{50}$ curves were fit the five-parameter asymmetric equation below

$$F_{([I])} = F_{max} - F_{min} + \frac{F_{max}-F_{min}}{\left(1+\left(2^{\frac{1}{S}}-1\right)*\left(\frac{IC_{50}}{[I]}\right)^{H}\right)^{S}} \qquad \textit{(Equation 5.21)}$$

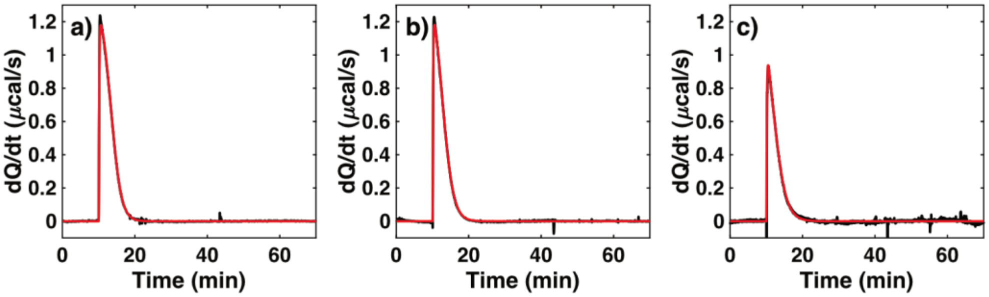
Where $F_{([I])}$ is the fluorescence at a specific inhibitor concentration ([I]), F_{max} and F_{min} are the maximum and minimum fluorescence values. *H* is the Hill slope, and *S* is the symmetry parameter. IC$_{50}$ is the inhibitor concentration where half of the product is formed when compared to the no-inhibitor control.

1000 simulated Time-dependent IC$_{50}$ experiments were ran. This gave 1000 values of $IC_{50(exp)}$ at each time point, outliers were removed from each set of $IC_{50(exp)}$ using MATLABs *isoutlier* function, fitting weights were used as the standard deviation (σ) of this set of $IC_{50(exp)}$. Finally, 1000 sets of $IC_{50(exp)}$ values at each inhibitor concentration were randomly sampled with resampling to produce 1000 sets of k_{obs} at each inhibitor concentration. In order to find K_i and k_{inact}, simulated IC$_{50(t)}$(sim) values were found by optimizing IC$_{50(calc)}$ in *Equation 5.4* and *Equation 5.5*, using a specific k_{inact} and K_i.

$$RSS = \sum_{t=0}^{N}\left(\frac{IC_{50(exp)_t}-IC_{50(calc)_t}}{\sigma_t}\right)^2 \qquad \textit{(Equation 5.22)}$$

Where $IC_{50(exp)t}$ is the IC_{50} value at time t from fitting to simulated fluorescence time points to *Equation 5.21*. $IC_{50(calc)_t}$ is the self-conisitent IC_{50} value at time *t* at a specific K_i and k_{inact} using *Equation 5.4* and *Equation 5.5*, and σ is the standard deviation of $IC_{50(exp)t}$ at a specific time *t.*

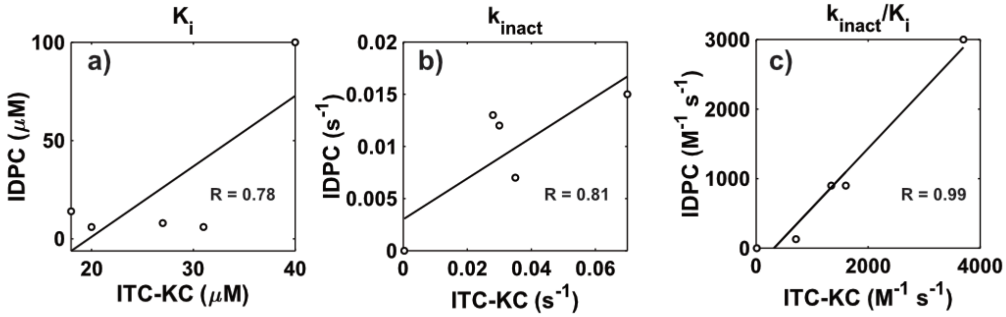
5.8 Supplementary information



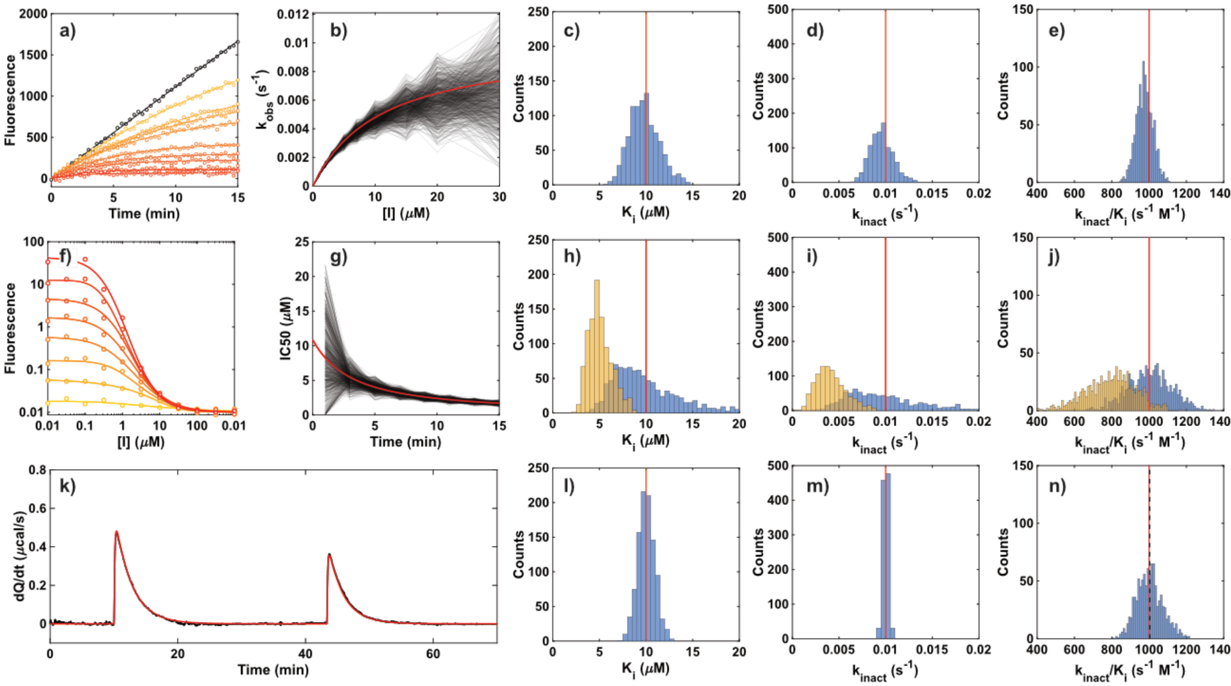
Supplementary Figure 5.1: ITC traces of no-inhibitor control. a) and b) were ran on the automated PEAQ-ITC over two different days. c) was ran on the ITC200 separately. The traces have been baseline corrected with a 2nd order polynomial baseline, and the red line show the best fit parameters of Michaelis-Menten kinetics as described in the methods section.

	k_{cat} (s^{-1})	K_m (μM)	k_{cat}/K_m (M^{-1} s^{-1})	ΔH (kcal/mol)
PEAK-ITC (day 1)	1.9	220	8600	3.8
PEAK-ITC (day 2)	2.5	320	8000	3.4
ITC200	2.2	330	6700	3.6* (2.5)
Average	2.2 ± 0.3	290 ± 60	8000 ± 1000	-

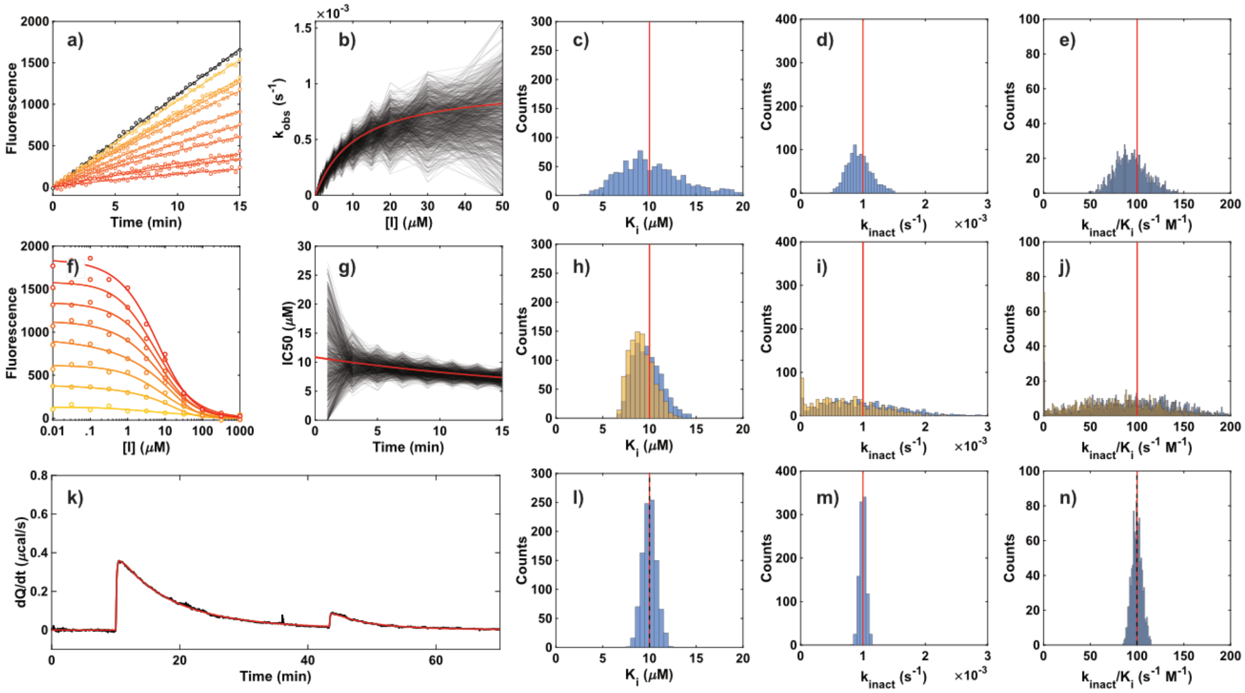
Supplementary Table 5.1: Best fit parameters from Supplementary Figure 5.1 for each no-inhibitor control. See Materials and methods for the fitting procedure. Errors are reported as the standard deviation of the three replicates. The ITC200 control had a lower ΔH than the PEAQ-ITC controls, likely due to errors in measuring the substrate. The k_{cat} and k_{inact} of the ITC200 control comes from fitting the data with the corrected substrate, detailed in the Materials and methods.



Supplementary Figure 5.2: Correlations between ITC-KC and IDPC parameters. a) K_i values. b) k_{inact} values. c) k_{inact}/K_i values. Correlation coefficients are shown on each plot.



Supplementary Figure 5.3: Comparison of IDPC, $TDIC_{50}$, and ITC-KC at K_i = 10μM and k_{inact} = 0.01 s^{-1}. a) Fluorescence traces at increasing inhibitor concentrations (yellow is low [I], red I high [I]). Dots indicate simulated noisy data, lines indicate fits to Equation 5.2. The black line indicates the no-inhibitor control. b) k_{obs} as a function of different inhibitor concentrations. Black lines represent 1000 samples from panel a, red line represents the true value. c-e) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value. f) $TDIC_{50}$ traces at different time points (fast times points are in yellow, long time points are in red). g) IC_{50} values as a function of time, black lines represent 1000 samples from panel f, red line represents the true values. h-i) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value. k) ITC-KC plot at 30 μM inhibitor. i-n) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value.



Supplementary Figure 5.4: Comparison of IDPC, TDIC$_{50}$, and ITC-KC at K_i = 10μM and k_{inact} = 0.1 s^{-1}. a) Fluorescence traces at increasing inhibitor concentrations (yellow is low [I], red I high [I]). Dots indicate simulated noisy data, lines indicate fits to Equation 5.2. The black line indicates the no-inhibitor control. b) k_{obs} as a function of different inhibitor concentrations. Black lines represent 1000 samples from panel a, red line represents the true value. c-e) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value. f) TDIC$_{50}$ traces at different time points (fast times points are in yellow, long time points are in red). g) IC$_{50}$ values as a function of time, black lines represent 1000 samples from panel f, red line represents the true values. h-i) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value. k) ITC-KC plot at 50 μM inhibitor. i-n) Histograms of kinetic parameters found from 1000 analyses, the red vertical line represents the true value.

Discussion

This thesis has described the development of multiple different biophysical assays, and has touched on the themes of kinetics, thermodynamics, bioinformatics, and drug discovery. This section will serve as a discussion of the results of each project, and the impact they have had on their fields. Some of the methods developed here have been used to answer other questions for collaborators or have been picked up by independent groups and applied to their research projects. Much of the work presented in this thesis was an advancement of the work previously done by members of the Mittermaier lab, and in a similar vein, these projects have begun to be carried on by new members of the lab. Each chapter's work will be discussed as it relates to the project as well as to the broader research community, and future directions for each project will give a foundation for new researchers to continue the work from this thesis.

6.1 Chapter 2 and 3

As described in the introduction of this thesis, guanine quadruplexes are a unique non-canonical DNA structure which exhibit a diverse array of dynamics[139]. They have been shown to fold into a multitude of different isomers. However, the physical and biological implications of this polymorphism are not fully understood. Chapters 2 and 3 were aimed at understanding these phenomena and used a mixture of different biophysical and bioinformatic techniques to give us fundamental insight into both the kinetic effects of polymorphism, and its prevalence in the human genome. Together with Robert Harkness, a previous graduate student from the Mittermaier lab and co-first author of the manuscript adapted in Chapter 2, we began a multifaceted approach to tease out the kinetics and assembly mechanism of a quadruplex located in the promoter region of the c-MYC oncogene. This quadruplex has been studied extensively and has been shown to regulate c-MYC expression in E. coli[199]. This particular sequence contains four G-tracts, two with four guanines each and two with three guanines. This asymmetry in G-tract length leads to the presence of four distinct isomers. Previously, Rob had shown that these isomers entropically stabilize the folded ensemble, causing the wild-type ensemble to be more stable than any of the individual isomers[210]. However, questions regarding the exact folding mechanism of these sequences remained. Did these sequences all fold through the same transition state? Or were there multiple pathways leading to different folded conformations?

We addressed these questions by creating eight distinct variations of the wild-type sequence by substituting select guanines to inosines. This effectively allowed us to trap the wild-type sequence into either a single isomer (referred to as the fully-trapped mutants) by substituting two guanines, or into two isomers (referred to as the half-trapped mutants) by substituting only a single guanine. Rob showed that these mutants were good structural mimics using a series of NMR experiments. This was further supported by CD experiments, showing that each isomer was in its native parallel conformation[512]. After confirming that our mutants were appropriate mimics, I ran thermal hysteresis experiments to determine the folding kinetics of each isomer.

My thermal hysteresis measurements were the key piece of evidence, which showed that all of the fully trapped mutants folded more slowly (lower k_{on}), and were less stable (lower k_{on}/k_{off}), than their half-trapped counterparts (i.e. 55 folded slower and was less stable than both 5X and X5, 33 folded more slowly and was less stable than both 3X and X3). Furthermore, all the mutated sequences folded more slowly than the wild-type sequence. This was a key piece of information, as it allowed us to better understand the folding mechanism of these sequences: if all of the GR isomers fold through the same transition state, then they should all share the same folding rate and differ only based on their unfolding rate. However, this was not the case and implies the presence of individual pathways to each of the distinct isomers. This type of mechanism would lead to a net increase in the folding rate of the wild-type ensemble. We were able to further support this conclusion by globally fitting a model with parallel pathways to all TH datasets simultaneously and obtained good agreement with experimental data.

The discovery of this net folding acceleration has important implications for the biological function of G4s. Firstly, it demonstrates that measuring the folding rates of individual G4 structures does not adequately describe the folding kinetics of an overall system. Secondly, our simulations show that the initial distribution of isomers is far from their equilibrium populations. We show that G4s initially fold into distributions based on their folding rates, not their thermodynamic stability. In the case of the c-MYC quadruplex we studied, this meant that two isomers which were populated at about 1% at equilibrium were actually both populated near 10% after initial folding. This is an important observation, as G4 interacting helicases have been shown to interact differently depending on the topology of the G4[370]. Furthermore, this type of redistribution was directly observed by our collaborators in the Schwalbe lab, who looked at the folding of the X3 G4 and saw fast folding followed by a slow exchange to equilibrium distributions.

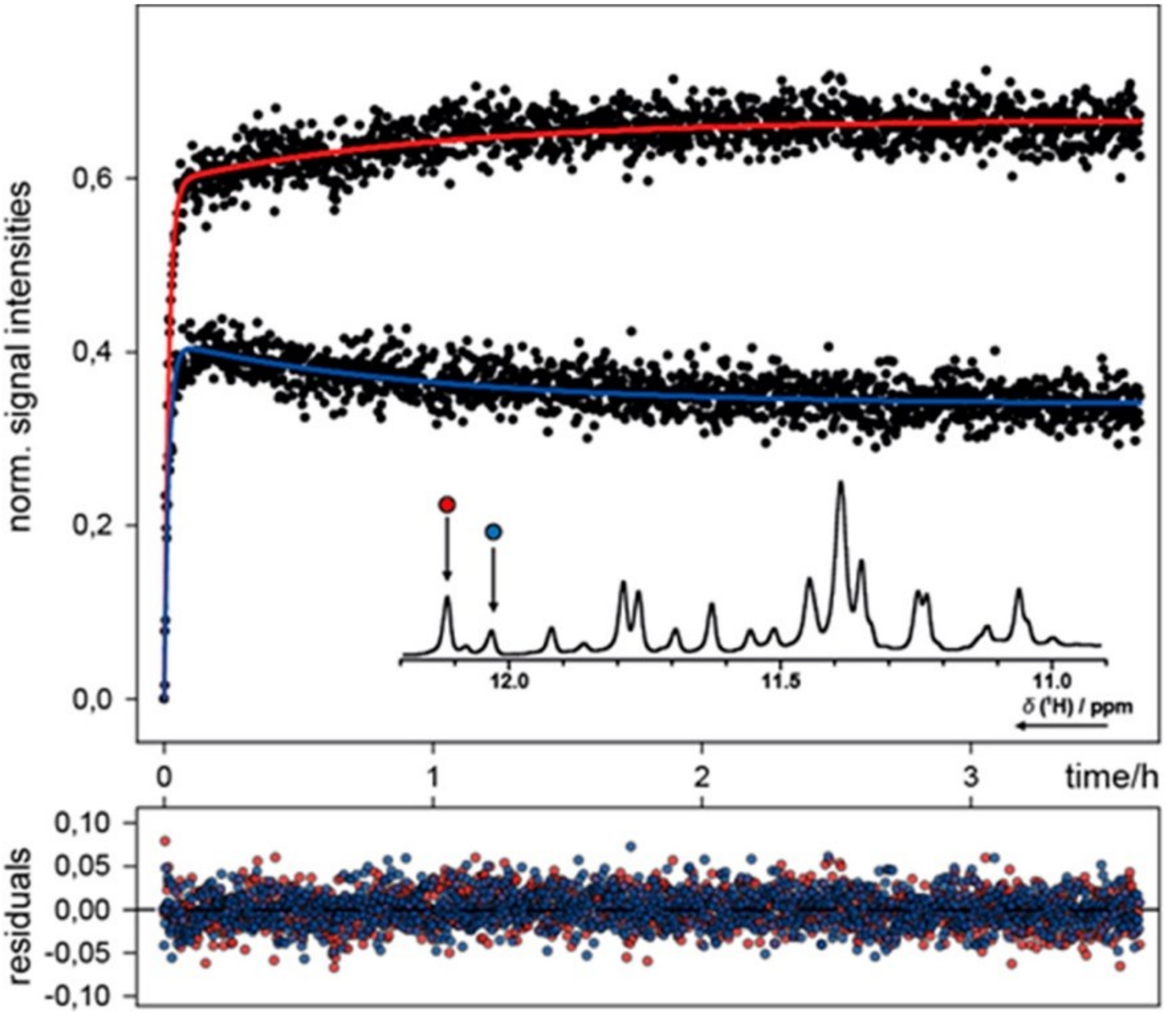


Figure 6.1: Isothermal NMR exchange experiments. An isothermal experiment for folding of a half-trapped c-MYC 3X G4. The NMR peak with the red fitted line corresponds to the 33 isomer, the NMR peak with the blue fitted line corresponds to the 53 isomer. Reprinted from Grün et al. with permission[216].

Our discovery that G4s containing multiple isomers leads to a net increase in the folding rate of the wild-type ensemble led us to ask the simple question: How polymorphic are quadruplexes in the human genome? The role polymorphism plays for G4s has been examined in a large number of papers published in recent years, where they are discussed as both a novel therapeutic target and important in a biological context[139, 401, 513]. However, many of the bioinformatic approaches used to identify quadruplexes simply rank quadruplexes based on their stability, and do not give an overall indication of how many quadruplexes could form. Furthermore, other students from the Mittermaier lab found that multimeric quadruplexes cause an inherent frustrated folding landscape which had implications in G4 folding[180], however the prevalence of these effects outside of the telomeric region has not been studied. To address this, we developed a new algorithm which was able to find G4 motifs in a sequence of DNA. We included only G4 motifs which contained relatively short loops, and contained no or few bulged residues.

Our results were surprising, previous members of our lab had attempted to analyze the human promoters and found that 5% of G4 sequences could form more than 20 G4s, in contrast, we found that this number is actually much higher. In fact, some of the previously unidentified G4CRs that we found could form thousands of different G4 isomers and be hundreds of nucleotides long. Furthermore, we found that these more polymorphic sequences were tightly clustered around the transcription start site (TSS), which suggests that they may be more functionally relevant than less polymorphic G4CRs. The reason behind this polymorphism is not understood. Some groups have reasoned that the presence of different isomers could be a mechanism to repair DNA damage[366], and as the Mittermaier lab has shown, these sequences can also increase the stability and folding rate of the G4CR[210]. We were able to identify several highly polymorphic G4CRs which were located in the promoters of oncogenes, which could provide useful therapeutic targets. The multimeric nature of these G4CRs may also allow them to be more selectively targeted when compared to a single G4 sequence. The fact that this polymorphism is much more prevalent than we previously thought, along with the physical implications we have shown them to have provide the groundwork for future scientists to understand how this affects the biological context of the G4CRs.

6.1.1 Future directions

Throughout Chapters 2 and 3 of this thesis, we have demonstrated both the prevalence of polymorphism and physical consequences on G4s in the human promoters. However, questions still remain such as:

1) Does the net folding acceleration scale to the degree we see from our bioinformatic approach? We only showed this acceleration on a sequence which could form four isomers, however we now know that there are sequences out there which can form many more. In order to analyze these effects, sequences with more G4 isomers should be studied. These types of sequences are often hard to synthesize, but new enzymatic techniques may be the key to overcoming these problems[514]. Running the TH experiments on a sequence containing more G4 isomers would provide more evidence that this is a phenomenon that occurs more generally. Furthermore, experiments could be performed on RNA G4s, as they have been found to exhibit some differences in folding[515].

2) We saw a distinct trend in polymorphic G4CRs being more clustered near the transcription start site, as well as being more prevalent in high-order genomes. This leads to questions as to the evolutionary pressure to forming a G4CR. Can we explain this increase just based on the mutations we expect to see, or is nature purposefully creating these regions and if so why? This is an ongoing project in the Mittermaier lab, currently being pursued by Amos Zhang, who along with the Blanchette lab here at McGill university, is using ancestral genomic data to understand how polymorphism evolved over millions of years.

6.2 Chapter 4

The TREQ method described in Chapter 4 represents a novel way of using existing experimental equipment to measure values that have been previously unobtainable. Prior to the development of this method, several groups have performed analyses which we have shown to be unreliable in practice[463, 464, 470-475]. In this Chapter, we described how we used the technique to determine the small molecule loading efficiency of polyA-CA fibres and made the discovery that about 33% of CA binding sites were unoccupied. This was initially a surprising result to us and in order to reconcile this with data taken from the original manuscript describing the formation of these fibres, we created a mathematical model to describe the multivalent assembly of these structures. In parallel with the work shown in this thesis, Felix Rizzuto and Casey Platnich used single molecule experiments to probe the assembly mechanism of these structures[245]. They found that isothermally assembled polyA-CA fibres were able to incorporate shorter strands of polyA even after assembling. This indicates that these fibres are, in fact, assembling with defects, which taken with our results, could explain why the lower occupancy is observed.

This, however, was not the only critical information that TREQ was able to provide for these fibres. My colleague and collaborator Dr. Christophe Lachance-Brais, used TREQ to understand how functionalizing one of the faces of the CA molecule would affect the stability of the fibres[241]. Christophe tested two series of molecules, where in each group the length of the alkyl chain increased from two to six carbons: One using a negatively charged hydroxyl group, and one using a positively charged amino group (*Figure 6.2a*). We found that all of the substitutions were destabilizing to the fibres which may be explained in part by the entropic cost of losing one of CA's binding faces. While some of the substitutions were too destabilizing to measure thermodynamics robustly, we were able to establish two trends in our data. 1) The hydroxyl substitutions showed decreasing stability with increasing chain length (*Figure 6.2b*). 2) The amines showed that C2 amine was the most stable, C3 was the least stable, and C4 and C5 had comparable stabilities (*Figure 6.2c*).

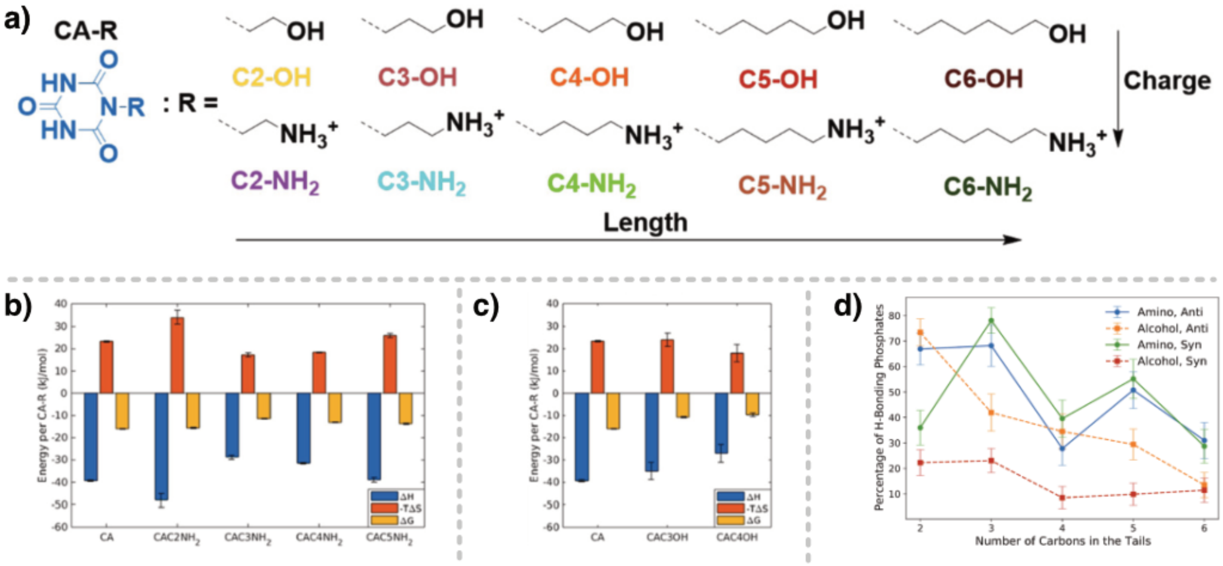


Figure 6.2: Thermodynamics of CA derivatives binding to polyA. a) chemical structures of the two series of molecules which were studied. Top, the more negatively charged hydroxyl series. Bottom, the more positively charged amine series. b) Thermodynamic parameters for amine series. c) Thermodynamic parameters for hydroxyl series. d) Percentage of h-binding phosphates determined by molecular dynamics simulations for each series as a function of the length of the alkyl chain, comparing expected results from the syn and anti conformations. Adapted from Lachance-Brais et al. with permission[241].

To further understand why this distinct trend was occurring in the amine group, Christophe turned to molecular dynamics simulations. Previously, it was impossible to determine the glycosidic bond angle of the adenosine molecules, as both structures were consistent with experimental data. However, Christophe was able to show that the syn angles predicted a low stability for the CAC2NH2 molecule, and the anti angles a high stability for this molecule, along with a valley of low stability for C4 (*Figure 6.2d*). When compared to the TREQ thermodynamic data, the trends for the anti conformation were more consistent with the thermodynamic results. This led to the conclusion that the polyA fibres form using syn glycosidic bond angles, showing once again how robust thermodynamic characterization of these systems can lead to a deeper understanding of their structure.

TREQ has not just been used to study the assembly of polyA-CA fibres. An independent group lead by Stefano Mezzasalma and Marek Grzelczak used the technique to look at the thermal reversible clustering of gold nanoparticles[516]. They had previously studied this clustering and noticed the presence of thermal hysteresis[517]. TREQ allowed them to measure the thermodynamic parameters (ΔH and ΔS) of this clustering for the first time, and they were able to use this characterization, along with physical chemical theory to develop a mathematical model to describe the behavior they observed. This demonstrates the applicability of TREQ to different systems, as a method to gain robust thermodynamic information.

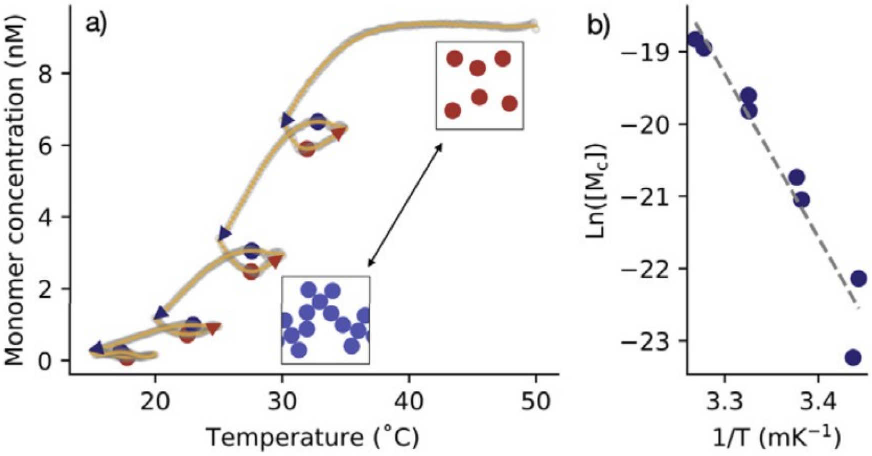


Figure 6.3: TREQ analysis on the reversible clustering of gold nanoparticles.TREQ trace, with heating minima shown in red, and cooling maxima shown in blue. Arrows indicate the direction of the temperature ramp. b) Van 't Hoff analysis of the critical monomer concentrations in panel a. Reproduced with permission from Mazzasalma et al. with permission[516].

6.2.1 Future directions

TREQ was developed as primarily a way to measure the thermodynamics of slowly assembling supramolecular systems. However, through our investigations it became clear that not only does TREQ provide a way to get these thermodynamics, but may also provide a robust kinetic profile of the system as well. For example, TH traces of the polyA-CA fibres can be described well by the GS model of nucleated self assembly (*Figure 6.4a*). This however, is not the case for a TREQ experiment (*Figure 6.4b*), where we see systematic deviations from the model itself. These fibres have been shown to assemble not only via monomer addition, but also by coagulation[245]. These multiple assembly pathways could be causing these systematic deviations from the GS fits, as they are not considered. It should be noted that these extra assembly mechanisms would not affect the thermodynamic stability of the fibres, and thus not the TREQ analysis itself. The benefits of running these heating and cooling cycles has been the subject of much of Masahiko Yamaguchi's research in recent years, who has observed many different types of hysteresis loops in self-catalytic reactions[518]. This diverse number of patterns emerging from these hysteresis loops, and inability to be reproduced with simple models, indicates that they contain rich kinetic information which could lead to more robust characterization of these slowly assembling systems.

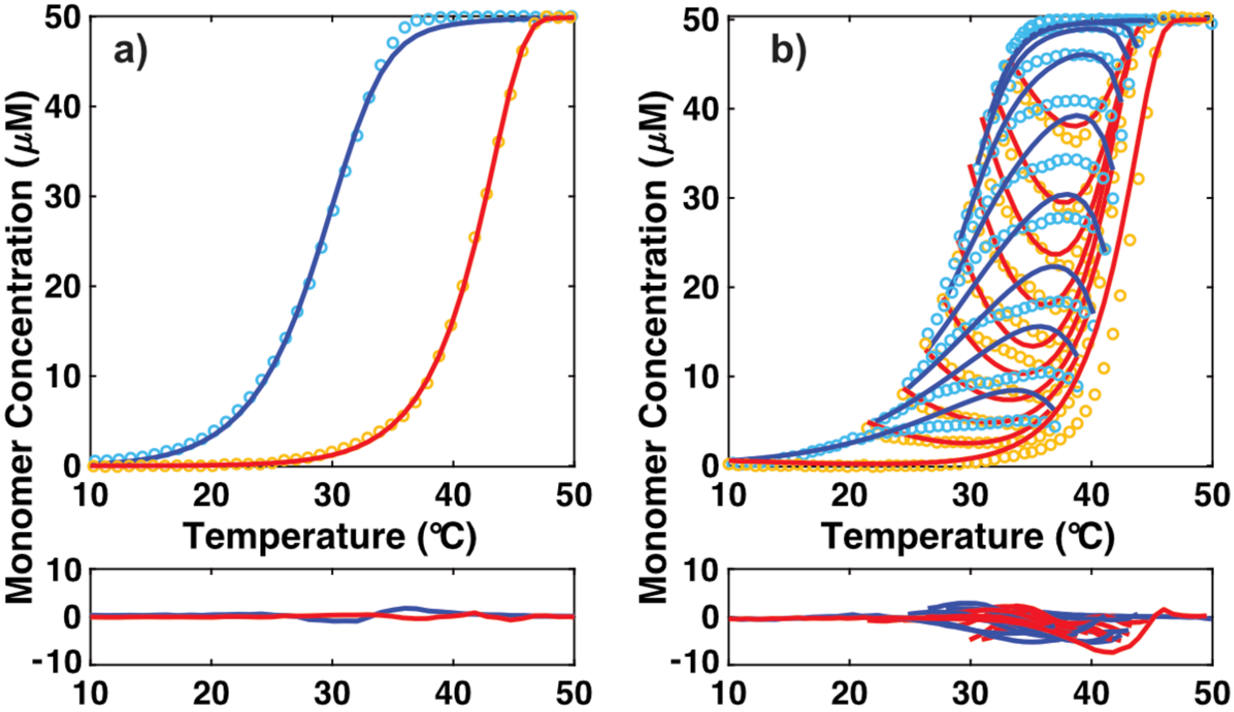


Figure 6.4: Example kinetics fits of TH and TREQ traces. a) Fits to polyA-CA TH traces at 15mM CA and 1 K/min. b) Fits to polyA-CA TREQ traces at 15mM CA and 1 K/min. Heating traces are shown in red/orange, cooling traces are shown in blue/cyan. Experimental data is shown as dots and curve fits are shown as the solid lines. Residuals are plotted in the bottom panel of each figure.

6.3 Chapter 5:

Chapter 5 represents the only unpublished research in this thesis, because of this no other groups have yet incorporated the KC-ITC technique into their drug discovery platforms. However, as shown in the chapter, the KC-ITC technique outperforms both IDPC analysis and Time-dependent IC_{50} analysis over a range of different K_i and k_{inact} values. It provides a robust characterization of covalent inhibitors in a single hour-long experiment and can be performed using either a standard or an automated-ITC instrument. Within this chapter we also discovered a systematic error in Time-dependent IC_{50} analysis where we showed it would underestimate the true values. We described a new way of fitting this data to account for this systematic deviation. Furthermore, while IDPC and Time-dependent IC_{50} analysis have been around for many years, drug discovery efforts still tend to only measure either the k_{inact}/K_i or a single IC_{50} value due to the time-consuming and costly nature of measuring the values individually[519]. Finally, both IDPC and Time-dependent IC_{50} analysis require the ability to measure the amount of product formation over time. Typically, this is done with fluorescently labelled substrates which can be expensive. KC-ITC measures the heat released by the enzymatic reaction and thus does not require spectroscopically active substrates, making it a near universal assay for measuring the k_{inact} and K_i of covalent inhibitors.

6.3.1 Future directions

In recent years, the Mittermaier lab has been collaborating with the Moitessier lab here at McGill University to create novel covalent inhibitors of the main proteases of coronaviruses. Now that we have demonstrated the power of KC-ITC we need to take advantage of the technique to see what insights it can give us into the structure-activity relationships of our inhibitors. Ongoing projects include the development of inhibitors for $3CL^{pro}$[502], which was the protease studied in Chapter 5, and PL^{pro} which is another protease from SARS-CoV-2 which has not been discussed in this thesis[520]. KC-ITC gives us the tool we need to start measuring the reactivity and affinity of our molecules in a relatively high-throughput manor, and this information will give important insight into developing new generations of inhibitors.

Conclusion

In conclusion, this dissertation describes the development and implementation of several new biophysical tools to measure biomacromolecular systems. Chapter 2 details the combined approach of mutagenesis, thermal hysteresis experiments, and global analysis to study the parallel folding pathways of the c-MYC quadruplex. This approach is not only of general use to the G4 community, but also allowed us to make the discovery that these parallel pathways accelerate the folding of this G4 by over 2.5-fold. This acceleration had been previously seen for proteins, but had not yet been discovered in G4 DNA, and has implications for the biological function of these structures. Chapter 3 followed along with the research from Chapter 2, and described the development of the GReg algorithm which is the first bioinformatic algorithm for finding and classifying G4CRs. We observed that more polymorphic G4CRs are tightly clustered around the transcription start site. Chapters 4 introduces TREQ, which is the first method to allow for robust thermodynamic characterization of slowly assembling supramolecular systems. It details best-practices for setting up a TREQ experiment and describes how to correctly analyze the data the method produces. We show how other methods are completely inadequate for characterizing slowly assembling supramolecular systems, and use our new method to study the small-molecule loading efficiency of polyA-CA fibres. We made the surprising discovery that nearly 33% of the CA binding sites were unoccupied and developed a generalized multivalent binding model to describe the experimental data we observed. Finally, Chapter 5 describes my part in the Mittermaier and Moitessier labs ongoing effort to develop novel covalent inhibitors for the main protease of SARS-CoV-2. We show how one can measure both the affinity (K_i) and reactivity (k_{inact}) of covalent inhibitors using a multi-injection isothermal titration calorimetry method. We showed that this new method provides a more robust characterization of these parameters compared with traditional methods and that our method can measure K_i and k_{inact} values over 4 orders of magnitude. We further pointed out a systematic flaw in time-dependent IC_{50} analysis, and gave a solution to fitting these types of data. After characterizing 19 molecules with different covalent warheads and scaffolds, we showed that changes in both of these features can have dramatic effects on both the affinity (K_i) and reactivity (k_{inact}) of covalent inhibitors.

This thesis shows the value of combining experimental approaches and mathematical modelling and gives three distinct new techniques to the scientific community. The GReg algorithm, TREQ, and ITC-KC represent three significant contributions to their respective fields, some of which are already being used by independent research groups.

Bibliography

(1) Neidle, S.; Sanderson, M. *Principles of nucleic acid structure*; Academic Press, 2021.

(2) Blackburn, G. M. *Nucleic acids in chemistry and biology*; Royal Society of Chemistry, 2006.

(3) Olson, W. K.; Sussman, J. L. How flexible is the furanose ring? 1. A comparison of experimental and theoretical studies. *Journal of the American Chemical Society* **1982**, *104* (1), 270-278.

(4) Foloppe, N.; Hartmann, B.; Nilsson, L.; MacKerell, A. D. Intrinsic conformational energetics associated with the glycosyl torsion in DNA: a quantum mechanical study. *Biophysical Journal* **2002**, *82* (3), 1554-1569.

(5) Vichier-Guerre, S.; POMPON, A.; LEFEBVRE, I.; IMBACH, J.-L. New Insights into the Resistance of α-Oligonucleotides to Nucleases. *Antisense Research and Development* **1994**, *4* (1), 9-18.

(6) MacKerell, A. D. Influence of magnesium ions on duplex DNA structural, dynamic, and solvation properties. *The Journal of Physical Chemistry B* **1997**, *101* (4), 646-650.

(7) Stellwagen, E.; Muse, J. M.; Stellwagen, N. C. Monovalent Cation Size and DNA Conformational Stability. *Biochemistry* **2011**, *50* (15), 3084-3094. DOI: 10.1021/bi1015524.

(8) Neidle, S.; Sanderson, M. Chapter 2 - The building blocks of DNA and RNA. In *Principles of Nucleic Acid Structure (Second Edition)*, Neidle, S., Sanderson, M. Eds.; Academic Press, 2022; pp 29-51.

(9) Dickerson, R. E. Definitions and nomenclature of nucleic acid structure parameters. *Journal of Biomolecular Structure and Dynamics* **1989**, *6* (4), 627-634.

(10) Nikolova, E. N.; Zhou, H.; Gottardo, F. L.; Alvey, H. S.; Kimsey, I. J.; Al ‐ Hashimi, H. M. A historical account of hoogsteen base ‐ pairs in duplex DNA. *Biopolymers* **2013**, *99* (12), 955-968.

(11) Nair, D. T.; Johnson, R. E.; Prakash, S.; Prakash, L.; Aggarwal, A. K. Replication by human DNA polymerase-ı occurs by Hoogsteen base-pairing. *Nature* **2004**, *430* (6997), 377-380.

(12) Ohyama, T. *DNA conformation and transcription*; Springer, 2005.

(13) Altun, A.; Garcia-Ratés, M.; Neese, F.; Bistoni, G. Unveiling the complex pattern of intermolecular interactions responsible for the stability of the DNA duplex. *Chemical Science* **2021**, *12* (38), 12785-12793.

(14) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic acids research* **2006**, *34* (2), 564-574.

(15) Watson, J. D.; Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171* (4356), 737-738. DOI: 10.1038/171737a0.

(16) Franklin, R. E.; Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **1953**, *171* (4356), 740-741. DOI: 10.1038/171740a0.

(17) Wilkins, M. H. F.; Stokes, A. R.; Wilson, H. R. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature* **1953**, *171* (4356), 738-740. DOI: 10.1038/171738a0.

(18) Cobb, M.; Comfort, N. What Rosalind Franklin truly contributed to the discovery of DNA's structure. *Nature* **2023**, *616* (7958), 657-660.

(19) Wemmer, D. E.; Dervan, P. B. Targeting the minor groove of DNA. *Current opinion in structural biology* **1997**, *7* (3), 355-361.

(20) Schleif, R. DNA binding by proteins. *Science* **1988**, *241* (4870), 1182-1187.

(21) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **2011**, *470* (7335), 498-502.

(22) Lilley, D. Kinking of DNA and RNA by base bulges. *Proceedings of the National Academy of Sciences* **1995**, *92* (16), 7140-7142.

(23) Carell, T.; Kurz, M. Q.; Müller, M.; Rossa, M.; Spada, F. Non‐canonical bases in the genome: the regulatory information layer in DNA. *Angewandte Chemie International Edition* **2018**, *57* (16), 4296-4312.

(24) Rich, A.; Nordheim, A.; Wang, A. H.-J. The chemistry and biology of left-handed Z-DNA. *Annual review of biochemistry* **1984**, *53* (1), 791-846.

(25) Arnott, S.; Chandrasekaran, R.; Birdsall, D.; Leslie, A.; Ratliff, R. Left-handed DNA helices. *Nature* **1980**, *283* (5749), 743-745.

(26) Wang, G.; Vasquez, K. M. Z-DNA, an active element in the genome. *Front Biosci* **2007**, *12* (4424), 38.

(27) Felsenfeld, G.; Davies, D. R.; Rich, A. Formation of a three-stranded polynucleotide molecule. *Journal of the American Chemical Society* **1957**, *79* (8), 2023-2024.

(28) Frank-Kamenetskii, M. D.; Mirkin, S. M. Triplex DNA structures. *Annual review of biochemistry* **1995**, *64* (1), 65-95.

(29) Han, H.; Dervan, P. B. Sequence-specific recognition of double helical RNA and RNA. DNA by triple helix formation. *Proceedings of the National Academy of Sciences* **1993**, *90* (9), 3806-3810.

(30) Macaya, R.; Wang, E.; Schultze, P.; Sklenář, V.; Feigon, J. Proton nuclear magnetic resonance assignments and structural characterization of an intramolecular DNA triplex. *Journal of molecular biology* **1992**, *225* (3), 755-773.

(31) Holland, J. A.; Hoffman, D. W. Structural features and stability of an RNA triple helix in solution. *Nucleic acids research* **1996**, *24* (14), 2841-2848.

(32) Grigoriev, M.; Praseuth, D.; Guieysse, A.; Robin, P.; Thuong, N.; Hélène, C.; Harel-Bellan, A. Inhibition of interleukin-2 receptor alpha-subunit gene expression by oligonucleotide-directed triple helix formation. *Comptes Rendus de L'academie des sciences. Serie III, Sciences de la vie* **1993**, *316* (5), 492-495.

(33) Faria, M.; Giovannangeli, C. Triplex‐forming molecules: from concepts to applications. *The Journal of Gene Medicine: A cross‐disciplinary journal for research on the science of gene transfer and its clinical applications* **2001**, *3* (4), 299-310.

(34) Faruqi, A. F.; Datta, H. J.; Carroll, D.; Seidman, M. M.; Glazer, P. M. Triple-helix formation induces recombination in mammalian cells via a nucleotide excision repair-dependent pathway. *Molecular and cellular biology* **2000**, *20* (3), 990-1000.

(35) Vasquez, K. M.; Narayanan, L.; Glazer, P. M. Specific mutations induced by triplex-forming oligonucleotides in mice. *Science* **2000**, *290* (5491), 530-533.

(36) Strobel, S. A.; Dervan, P. B. Site-specific cleavage of a yeast chromosome by oligonucleotide-directed triple-helix formation. *Science* **1990**, *249* (4964), 73-75.

(37) Neidle, S.; Balasubramanian, S. Quadruplex Nucleic Acids. **2006**. DOI: 10.1039/9781847555298.

(38) Varshney, D.; Spiegel, J.; Zyner, K.; Tannahill, D.; Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nature Reviews Molecular Cell Biology* **2020**, *21* (8), 459-474.
(39) Kotar, A.; Rigo, R.; Sissi, C.; Plavec, J. Two-quartet kit* G-quadruplex is formed via double-stranded pre-folded structure. *Nucleic acids research* **2019**, *47* (5), 2641-2653.
(40) Bartas, M.; Brázda, V.; Karlický, V.; Červeň, J.; Pečinka, P. Bioinformatics analyses and in vitro evidence for five and six stacked G-quadruplex forming sequences. *Biochimie* **2018**, *150*, 70-75.
(41) Rachwal, P. A.; Brown, T.; Fox, K. R. Effect of G-Tract Length on the Topology and Stability of Intramolecular DNA Quadruplexes. *Biochemistry* **2007**, *46* (11), 3036-3044. DOI: 10.1021/bi062118j.
(42) Smargiasso, N.; Rosu, F.; Hsia, W.; Colson, P.; Baker, E. S.; Bowers, M. T.; De Pauw, E.; Gabelica, V. G-quadruplex DNA assemblies: loop length, cation identity, and multimer formation. *Journal of the American Chemical Society* **2008**, *130* (31), 10208-10216.
(43) Liu, W.; Zhu, H.; Zheng, B.; Cheng, S.; Fu, Y.; Li, W.; Lau, T.-C.; Liang, H. Kinetics and mechanism of G-quadruplex formation and conformational switch in a G-quadruplex of PS2. M induced by Pb2+. *Nucleic Acids Research* **2012**, *40* (9), 4229-4236.
(44) Chen, F. M. Strontium (2+) facilitates intermolecular G-quadruplex formation of telomeric sequences. *Biochemistry* **1992**, *31* (15), 3769-3776.
(45) Lee, M. P.; Parkinson, G. N.; Hazel, P.; Neidle, S. Observation of the coexistence of sodium and calcium ions in a DNA G-quadruplex ion channel. *Journal of the American Chemical Society* **2007**, *129* (33), 10106-10107.
(46) Deng, H.; Braunlin, W. H. Kinetics of sodium ion binding to DNA quadruplexes. *Journal of molecular biology* **1996**, *255* (3), 476-483.
(47) Haider, S.; Parkinson, G. N.; Neidle, S. Crystal structure of the potassium form of an Oxytricha nova G-quadruplex. *Journal of Molecular Biology* **2002**, *320* (2), 189-200.
(48) Podbevšek, P.; Hud, N. V.; Plavec, J. NMR evaluation of ammonium ion movement within a unimolecular G-quadruplex in solution. *Nucleic acids research* **2007**, *35* (8), 2554-2563.
(49) Cai, M.; Shi, X.; Sidorov, V.; Fabris, D.; Lam, Y.-f.; Davis, J. T. Cation-directed self-assembly of lipophilic nucleosides: the cation's central role in the structure and dynamics of a hydrogen-bonded assembly. *Tetrahedron* **2002**, *58* (4), 661-671.
(50) Miyoshi, D.; Nakao, A.; Toda, T.; Sugimoto, N. Effect of divalent cations on antiparallel G-quartet structure of d (G4T4G4). *FEBS letters* **2001**, *496* (2-3), 128-133.
(51) Basu, S.; Szewczak, A. A.; Cocco, M.; Strobel, S. A. Direct detection of monovalent metal ion binding to a DNA G-quartet by 205Tl NMR. *Journal of the American Chemical Society* **2000**, *122* (13), 3240-3241.
(52) Bhattacharyya, D.; Mirihana Arachchilage, G.; Basu, S. Metal cations in G-quadruplex folding and stability. *Frontiers in chemistry* **2016**, *4*, 38.
(53) Largy, E.; Mergny, J.-L.; Gabelica, V. Role of alkali metal ions in G-quadruplex nucleic acid structure and stability. *The alkali metal ions: Their role for life* **2016**, 203-258.
(54) Wong, A.; Wu, G. Selective binding of monovalent cations to the stacking G-quartet structure formed by guanosine 5 ‘-monophosphate: a solid-state NMR study. *Journal of the American Chemical Society* **2003**, *125* (45), 13895-13905.
(55) Dai, J.; Carver, M.; Yang, D. Polymorphism of human telomeric quadruplex structures. *Biochimie* **2008**, *90* (8), 1172-1183.

(56) Marchand, A.; Gabelica, V. Folding and misfolding pathways of G-quadruplex DNA. *Nucleic Acids Research* **2016**, *44* (22), 10999-11012. DOI: 10.1093/nar/gkw970 (acccessed 9/5/2020).
(57) 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 research* **2012**, *40* (10), 4691-4700.
(58) Harrell Jr, W. A. *Quadruplex Nucleic Acids*; The Royal Society of Chemistry, 2006. DOI: 10.1039/9781847555298.
(59) Harkness, R. W. V.; Mittermaier, A. K. G-quadruplex dynamics. *Biochim Biophys Acta* **2017**, *1865* (11 Pt B), 1544-1554. DOI: 10.1016/j.bbapap.2017.06.012.
(60) Chung, W. J.; Heddi, B.; Schmitt, E.; Lim, K. W.; Mechulam, Y.; Phan, A. T. Structure of a left-handed DNA G-quadruplex. *Proceedings of the National Academy of Sciences* **2015**, *112* (9), 2729-2733.
(61) Wang, Y.; Patel, D. J. Solution structure of a parallel-stranded G-quadruplex DNA. *Journal of molecular biology* **1993**, *234* (4), 1171-1183.
(62) Lim, K. W.; Lacroix, L.; Yue, D. J.; Lim, J. K.; Lim, J. M.; Phan, A. T. Coexistence of two distinct G-quadruplex conformations in the hTERT promoter. *J Am Chem Soc* **2010**, *132* (35), 12331-12342. DOI: 10.1021/ja101252n.
(63) Lane, A. N.; Chaires, J. B.; Gray, R. D.; Trent, J. O. Stability and kinetics of G-quadruplex structures. *Nucleic acids research* **2008**, *36* (17), 5482-5515.
(64) Joachimi, A.; Benz, A.; Hartig, J. S. A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorganic & medicinal chemistry* **2009**, *17* (19), 6811-6815.
(65) Langridge, R.; Rich, A. Molecular structure of helical polycytidylic acid. *Nature* **1963**, *198* (4882), 725-728.
(66) Gehring, K.; Leroy, J.-L.; Guéron, M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature* **1993**, *363* (6429), 561-565.
(67) Lepper, C. P.; Williams, M. A.; Edwards, P. J.; Filichev, V. V.; Jameson, G. B. Effects of Pressure and pH on the Physical Stability of an I‐Motif DNA Structure. *ChemPhysChem* **2019**, *20* (12), 1567-1571.
(68) Cui, J.; Waltman, P.; Le, V. H.; Lewis, E. A. The effect of molecular crowding on the stability of human c-MYC promoter sequence I-motif at neutral pH. *Molecules* **2013**, *18* (10), 12751-12767.
(69) 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* **2012**, *51* (1), 250-253. DOI: 10.1002/anie.201104938.
(70) Abou Assi, H.; Garavís, M.; González, C.; Damha, M. J. i-Motif DNA: structural features and significance to cell biology. *Nucleic Acids Research* **2018**, *46* (16), 8038-8056.
(71) Wright, E. P.; Huppert, J. L.; Waller, Z. A. Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. *Nucleic acids research* **2017**, *45* (6), 2951-2959.
(72) Zhou, J.; Wei, C.; Jia, G.; Wang, X.; Feng, Z.; Li, C. Formation of i-motif structure at neutral and slightly alkaline pH. *Molecular BioSystems* **2010**, *6* (3), 580-586.
(73) Lieblein, A. L.; Fürtig, B.; Schwalbe, H. Optimizing the kinetics and thermodynamics of DNA i‐Motif folding. *ChemBioChem* **2013**, *14* (10), 1226-1230.

(74) Berger, I.; Egli, M.; Rich, A. Inter-strand CH... O hydrogen bonds stabilizing four-stranded intercalated molecules: stereoelectronic effects of O4'in cytosine-rich DNA. *Proceedings of the National Academy of Sciences* **1996**, *93* (22), 12116-12121.

(75) Školáková, P.; Gajarský, M.; Palacký, J.; Šubert, D.; Renčiuk, D.; Trantírek, L.; Mergny, J.-L.; Vorlíčková, M. DNA i-motif formation at neutral pH is driven by kinetic partitioning. *Nucleic Acids Research* **2023**, *51* (6), 2950-2962.

(76) Kumar, N.; Nielsen, J. T.; Maiti, S.; Petersen, M. i‐Motif Formation with Locked Nucleic Acid (LNA). *Angewandte Chemie International Edition* **2007**, *46* (48), 9220-9222.

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

(78) El-Khoury, R.; Damha, M. J. End-ligation can dramatically stabilize i-motifs at neutral pH. *Chemical Communications* **2023**, *59* (25), 3715-3718.

(79) Tang, W.; Niu, K.; Yu, G.; Jin, Y.; Zhang, X.; Peng, Y.; Chen, S.; Deng, H.; Li, S.; Wang, J. In vivo visualization of the i-motif DNA secondary structure in the Bombyx mori testis. *Epigenetics & Chromatin* **2020**, *13*, 1-12.

(80) Zeraati, M.; Langley, D. B.; Schofield, P.; Moye, A. L.; Rouet, R.; Hughes, W. E.; Bryan, T. M.; Dinger, M. E.; Christ, D. I-motif DNA structures are formed in the nuclei of human cells. *Nature chemistry* **2018**, *10* (6), 631-637.

(81) Yoga, Y. M.; Traore, D. A.; Sidiqi, M.; Szeto, C.; Pendini, N. R.; Barker, A.; Leedman, P. J.; Wilce, J. A.; Wilce, M. C. Contribution of the first K-homology domain of poly (C)-binding protein 1 to its affinity and specificity for C-rich oligonucleotides. *Nucleic acids research* **2012**, *40* (11), 5101-5114.

(82) Niu, K.; Zhang, X.; Deng, H.; Wu, F.; Ren, Y.; Xiang, H.; Zheng, S.; Liu, L.; Huang, L.; Zeng, B. BmILF and i-motif structure are involved in transcriptional regulation of BmPOUM2 in Bombyx mori. *Nucleic Acids Research* **2018**, *46* (4), 1710-1723.

(83) El-Khoury, R.; Roman, M.; Assi, H. A.; Moye, A. L.; Bryan, T. M.; Damha, M. J. Telomeric i-motifs and C-strands inhibit parallel G-quadruplex extension by telomerase. *Nucleic Acids Research* **2023**, *51* (19), 10395-10410.

(84) King, J. J.; Irving, K. L.; Evans, C. W.; Chikhale, R. V.; Becker, R.; Morris, C. J.; Peña Martinez, C. D.; Schofield, P.; Christ, D.; Hurley, L. H. DNA G-quadruplex and i-motif structure formation is interdependent in human cells. *Journal of the American Chemical Society* **2020**, *142* (49), 20600-20604.

(85) Dhakal, S.; Yu, Z.; Konik, R.; Cui, Y.; Koirala, D.; Mao, H. G-quadruplex and i-motif are mutually exclusive in ILPR double-stranded DNA. *Biophysical journal* **2012**, *102* (11), 2575-2584.

(86) Cui, Y.; Kong, D.; Ghimire, C.; Xu, C.; Mao, H. Mutually exclusive formation of G-quadruplex and i-motif is a general phenomenon governed by steric hindrance in duplex DNA. *Biochemistry* **2016**, *55* (15), 2291-2299.

(87) Dong, Y.; Yang, Z.; Liu, D. DNA nanotechnology based on i-motif structures. *Accounts of chemical research* **2014**, *47* (6), 1853-1860.

(88) Modi, S.; MG, S.; Goswami, D.; Gupta, G. D.; Mayor, S.; Krishnan, Y. A DNA nanomachine that maps spatial and temporal pH changes inside living cells. *Nature nanotechnology* **2009**, *4* (5), 325-330.

(89) Miao, D.; Yu, Y.; Chen, Y.; Liu, Y.; Su, G. Facile construction of i-Motif DNA-conjugated gold nanostars as near-infrared and pH dual-responsive targeted drug delivery systems for combined cancer therapy. *Molecular Pharmaceutics* **2020**, *17* (4), 1127-1138.

(90) Shu, W.; Liu, D.; Watari, M.; Riener, C. K.; Strunz, T.; Welland, M. E.; Balasubramanian, S.; McKendry, R. A. DNA molecular motor driven micromechanical cantilever arrays. *Journal of the American Chemical Society* **2005**, *127* (48), 17054-17060.

(91) Ellington, A. D.; Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *nature* **1990**, *346* (6287), 818-822.

(92) Tuerk, C.; Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *science* **1990**, *249* (4968), 505-510.

(93) Darmostuk, M.; Rimpelova, S.; Gbelcova, H.; Ruml, T. Current approaches in SELEX: An update to aptamer selection technology. *Biotechnology advances* **2015**, *33* (6), 1141-1161.

(94) Mendonsa, S. D.; Bowser, M. T. In vitro evolution of functional DNA using capillary electrophoresis. *Journal of the American Chemical Society* **2004**, *126* (1), 20-21.

(95) Bruno, J. G. In vitroselection of DNA to chloroaromatics using magnetic microbead-based affinity separation and fluorescence detection. *Biochemical and biophysical research communications* **1997**, *234* (1), 117-120.

(96) McKeague, M.; DeRosa, M. C. Challenges and opportunities for small molecule aptamer development. *Journal of nucleic acids* **2012**, *2012*.

(97) Keefe, A. D.; Cload, S. T. SELEX with modified nucleotides. *Current opinion in chemical biology* **2008**, *12* (4), 448-456.

(98) Kong, D.; Yeung, W.; Hili, R. In vitro selection of diversely functionalized aptamers. *Journal of the American Chemical Society* **2017**, *139* (40), 13977-13980.

(99) Slavkovic, S.; Churcher, Z. R.; Johnson, P. E. Nanomolar binding affinity of quinine-based antimalarial compounds by the cocaine-binding aptamer. *Bioorganic & Medicinal Chemistry* **2018**, *26* (20), 5427-5434. DOI: https://doi.org/10.1016/j.bmc.2018.09.017.

(100) Sun, A.; Gasser, C.; Li, F.; Chen, H.; Mair, S.; Krasheninina, O.; Micura, R.; Ren, A. SAM-VI riboswitch structure and signature for ligand discrimination. *Nature Communications* **2019**, *10* (1), 5728. DOI: 10.1038/s41467-019-13600-9.

(101) Mayer, G. The chemical biology of aptamers. *Angewandte Chemie International Edition* **2009**, *48* (15), 2672-2689.

(102) Zhou, J.; Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nature reviews Drug discovery* **2017**, *16* (3), 181-202.

(103) Famulok, M.; Mayer, G. n. Aptamer modules as sensors and detectors. *Accounts of chemical research* **2011**, *44* (12), 1349-1358.

(104) He, F.; Wen, N.; Xiao, D.; Yan, J.; Xiong, H.; Cai, S.; Liu, Z.; Liu, Y. Aptamer-based targeted drug delivery systems: current potential and challenges. *Current medicinal chemistry* **2020**, *27* (13), 2189-2219.

(105) Dhiman, A.; Kalra, P.; Bansal, V.; Bruno, J. G.; Sharma, T. K. Aptamer-based point-of-care diagnostic platforms. *Sensors and Actuators B: Chemical* **2017**, *246*, 535-553.

(106) Slavkovic, S.; Altunisik, M.; Reinstein, O.; Johnson, P. E. Structure–affinity relationship of the cocaine-binding aptamer with quinine derivatives. *Bioorganic & Medicinal Chemistry* **2015**, *23* (10), 2593-2597.

(107) Nahvi, A.; Sudarsan, N.; Ebert, M. S.; Zou, X.; Brown, K. L.; Breaker, R. R. Genetic control by a metabolite binding mRNA. *Chemistry & biology* **2002**, *9* (9), 1043-1049.
(108) Winkler, W.; Nahvi, A.; Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **2002**, *419* (6910), 952-956.
(109) Mironov, A. S.; Gusarov, I.; Rafikov, R.; Lopez, L. E.; Shatalin, K.; Kreneva, R. A.; Perumov, D. A.; Nudler, E. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *cell* **2002**, *111* (5), 747-756.
(110) Serganov, A.; Nudler, E. A decade of riboswitches. *Cell* **2013**, *152* (1), 17-24.
(111) Mandal, M.; Breaker, R. R. Gene regulation by riboswitches. *Nature reviews Molecular cell biology* **2004**, *5* (6), 451-463.
(112) Wachter, A. Riboswitch-mediated control of gene expression in eukaryotes. *RNA biology* **2010**, *7* (1), 67-76.
(113) Ellinger, E.; Chauvier, A.; Romero, R. A.; Liu, Y.; Ray, S.; Walter, N. G. Riboswitches as therapeutic targets: Promise of a new era of antibiotics. *Expert Opinion on Therapeutic Targets* **2023**, (just-accepted).
(114) Kilpatrick, M.; Torri, A.; Kang, D.; Engler, J.; Wells, R. Unusual DNA structures in the adenovirus genome. *Journal of Biological Chemistry* **1986**, *261* (24), 11350-11354.
(115) Evans, T.; Schon, E.; Gora-Maslak, G.; Patterson, J.; Efstratiadis, A. S1-hypersensitive sites in eukaryotic promoter regions. *Nucleic Acids Research* **1984**, *12* (21), 8043-8058.
(116) Blackburn, E.; Szostak, J. The molecular structure of centromeres and telomeres. *Annual review of biochemistry* **1984**, *53* (1), 163-194.
(117) Wang, Y.; Patel, D. J. Solution structure of the human telomeric repeat d [AG3 (T2AG3) 3] G-tetraplex. *Structure* **1993**, *1* (4), 263-282.
(118) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A. The sequence of the human genome. *science* **2001**, *291* (5507), 1304-1351.
(119) Huppert, J. L.; Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research* **2005**, *33* (9), 2908-2916. DOI: 10.1093/nar/gki609 (acccessed 4/19/2022).
(120) Hazel, P.; Huppert, J.; Balasubramanian, S.; Neidle, S. Loop-length-dependent folding of G-quadruplexes. *Journal of the American Chemical Society* **2004**, *126* (50), 16405-16415.
(121) 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. *Proceedings of the National Academy of Sciences* **2002**, *99* (18), 11593-11598.
(122) Gray, L. T.; Vallur, A. C.; Eddy, J.; Maizels, N. G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nature chemical biology* **2014**, *10* (4), 313-318.
(123) Fernando, H.; Sewitz, S.; Darot, J.; Tavare, S.; Huppert, J. L.; Balasubramanian, S. Genome-wide analysis of a G-quadruplex-specific single-chain antibody that regulates gene expression. *Nucleic acids research* **2009**, *37* (20), 6716-6722.
(124) Huppert, J. L.; Balasubramanian, S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **2007**, *35* (2), 406-413. DOI: gkl1057 [pii]
10.1093/nar/gkl1057.

(125) Parkinson, G. N.; Lee, M. P.; Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **2002**, *417* (6891), 876-880.
(126) Blackburn, E. H. Telomeres: no end in sight. *Cell* **1994**, *77* (5), 621-623.
(127) Smith, J. S.; Chen, Q.; Yatsunyk, L. A.; Nicoludis, J. M.; Garcia, M. S.; Kranaster, R.; Balasubramanian, S.; Monchaud, D.; Teulade-Fichou, M.-P.; Abramowitz, L. Rudimentary G-quadruplex–based telomere capping in Saccharomyces cerevisiae. *Nature structural & molecular biology* **2011**, *18* (4), 478-485.
(128) Wellinger, R.; Sen, D. The DNA structures at the ends of eukaryotic chromosomes. *European Journal of Cancer* **1997**, *33* (5), 735-749.
(129) Huang, H.; Zhang, J.; Harvey, S. E.; Hu, X.; Cheng, C. RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. *Genes & development* **2017**, *31* (22), 2296-2309.
(130) Beaudoin, J.-D.; Perreault, J.-P. 5′ -UTR G-quadruplex structures acting as translational repressors. *Nucleic acids research* **2010**, *38* (20), 7022-7036.
(131) Bugaut, A.; Balasubramanian, S. 5′ -UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic acids research* **2012**, *40* (11), 4727-4741.
(132) Kumari, S.; Bugaut, A.; Huppert, J. L.; Balasubramanian, S. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. *Nature chemical biology* **2007**, *3* (4), 218-221.
(133) Bedrat, A.; Lacroix, L.; Mergny, J.-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic acids research* **2016**, *44* (4), 1746-1759.
(134) Puig Lombardi, E.; Londoño-Vallejo, A. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Research* **2019**, *48* (1), 1-15. DOI: 10.1093/nar/gkz1097 (acccessed 12/2/2022).
(135) Chambers, V. S.; Marsico, G.; Boutell, J. M.; Di Antonio, M.; Smith, G. P.; Balasubramanian, S. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature Biotechnology* **2015**, *33* (8), 877-881. DOI: 10.1038/nbt.3295.
(136) Kwok, C. K.; Marsico, G.; Sahakyan, A. B.; Chambers, V. S.; Balasubramanian, S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nature methods* **2016**, *13* (10), 841-844.
(137) Bentley, D. R.; Balasubramanian, S.; Swerdlow, H. P.; Smith, G. P.; Milton, J.; Brown, C. G.; Hall, K. P.; Evers, D. J.; Barnes, C. L.; Bignell, H. R. Accurate whole human genome sequencing using reversible terminator chemistry. *nature* **2008**, *456* (7218), 53-59.
(138) Ewing, B.; Hillier, L.; Wendl, M. C.; Green, P. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research* **1998**, *8* (3), 175-185.
(139) Grün, J. T.; Schwalbe, H. Folding dynamics of polymorphic G ‐ quadruplex structures. *Biopolymers* **2021**, e23477.
(140) Marsico, G.; Chambers, V. S.; Sahakyan, A. B.; McCauley, P.; Boutell, J. M.; Antonio, M. D.; Balasubramanian, S. Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic acids research* **2019**, *47* (8), 3862-3874.
(141) Sahakyan, A. B.; Chambers, V. S.; Marsico, G.; Santner, T.; Di Antonio, M.; Balasubramanian, S. Machine learning model for sequence-driven DNA G-quadruplex formation. *Scientific reports* **2017**, *7* (1), 14535.

(142) Rocher, V.; Genais, M.; Nassereddine, E.; Mourad, R. DeepG4: a deep learning approach to predict cell-type specific active G-quadruplex regions. *PLOS Computational Biology* **2021**, *17* (8), e1009308.

(143) Barshai, M.; Engel, B.; Haim, I.; Orenstein, Y. G4mismatch: Deep neural networks to predict G-quadruplex propensity based on G4-seq data. *PLOS Computational Biology* **2023**, *19* (3), e1010948.

(144) Biffi, G.; Tannahill, D.; McCafferty, J.; Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature chemistry* **2013**, *5* (3), 182-186.

(145) Summers, P. A.; Lewis, B. W.; Gonzalez-Garcia, J.; Porreca, R. M.; Lim, A. H.; Cadinu, P.; Martin-Pintado, N.; Mann, D. J.; Edel, J. B.; Vannier, J. B. Visualising G-quadruplex DNA dynamics in live cells by fluorescence lifetime imaging microscopy. *Nature communications* **2021**, *12* (1), 162.

(146) Di Antonio, M.; Ponjavic, A.; Radzevičius, A.; Ranasinghe, R. T.; Catalano, M.; Zhang, X.; Shen, J.; Needham, L.-M.; Lee, S. F.; Klenerman, D.; et al. Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nature Chemistry* **2020**, *12* (9), 832-837. DOI: 10.1038/s41557-020-0506-4.

(147) Monchaud, D. Chapter Five - Quadruplex detection in human cells. In *Annual Reports in Medicinal Chemistry*, Neidle, S. Ed.; Vol. 54; Academic Press, 2020; pp 133-160.

(148) Paeschke, K.; Juranek, S.; Simonsson, T.; Hempel, A.; Rhodes, D.; Lipps, H. J. Telomerase recruitment by the telomere end binding protein-β facilitates G-quadruplex DNA unfolding in ciliates. *Nature structural & molecular biology* **2008**, *15* (6), 598-604.

(149) Paeschke, K.; Simonsson, T.; Postberg, J.; Rhodes, D.; Lipps, H. J. Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nature structural & molecular biology* **2005**, *12* (10), 847-854.

(150) Castillo Bosch, P.; Segura ‐ Bayona, S.; Koole, W.; van Heteren, J. T.; Dewar, J. M.; Tijsterman, M.; Knipscheer, P. FANCJ promotes DNA synthesis through G ‐ quadruplex structures. *The EMBO journal* **2014**, *33* (21), 2521-2533.

(151) Valton, A. L.; Hassan ‐ Zadeh, V.; Lema, I.; Boggetto, N.; Alberti, P.; Saintomé, C.; Riou, J. F.; Prioleau, M. N. G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *The EMBO journal* **2014**, *33* (7), 732-746.

(152) Besnard, E.; Babled, A.; Lapasset, L.; Milhavet, O.; Parrinello, H.; Dantec, C.; Marin, J.-M.; Lemaitre, J.-M. Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature structural & molecular biology* **2012**, *19* (8), 837-844.

(153) Mao, S.-Q.; Ghanbarian, A. T.; Spiegel, J.; Martínez Cuesta, S.; Beraldi, D.; Di Antonio, M.; Marsico, G.; Hänsel-Hertsch, R.; Tannahill, D.; Balasubramanian, S. DNA G-quadruplex structures mold the DNA methylome. *Nature structural & molecular biology* **2018**, *25* (10), 951-957.

(154) Hänsel-Hertsch, R.; Beraldi, D.; Lensing, S. V.; Marsico, G.; Zyner, K.; Parry, A.; Di Antonio, M.; Pike, J.; Kimura, H.; Narita, M. G-quadruplex structures mark human regulatory chromatin. *Nature genetics* **2016**, *48* (10), 1267-1272.

(155) Sarkies, P.; Reams, C.; Simpson, L. J.; Sale, J. E. Epigenetic instability due to defective replication of structured DNA. *Molecular cell* **2010**, *40* (5), 703-713.

(156) David, A. P.; Margarit, E.; Domizi, P.; Banchio, C.; Armas, P.; Calcaterra, N. B. G-quadruplexes as novel cis-elements controlling transcription during embryonic development. *Nucleic acids research* **2016**, *44* (9), 4163-4173.
(157) Cogoi, S.; Xodo, L. E. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Research* **2006**, *34* (9), 2536-2549. DOI: 10.1093/nar/gkl286 (acccessed 4/28/2022).
(158) Kwok, C. K.; Ding, Y.; Shahid, S.; Assmann, S. M.; Bevilacqua, P. C. A stable RNA G-quadruplex within the 5′ -UTR of Arabidopsis thaliana ATR mRNA inhibits translation. *Biochemical Journal* **2015**, *467* (1), 91-102.
(159) Wieland, M.; Hartig, J. S. RNA quadruplex-based modulation of gene expression. *Chemistry & biology* **2007**, *14* (7), 757-763.
(160) Lopez, C. R.; Singh, S.; Hambarde, S.; Griffin, W. C.; Gao, J.; Chib, S.; Yu, Y.; Ira, G.; Raney, K. D.; Kim, N. Yeast Sub1 and human PC4 are G-quadruplex binding proteins that suppress genome instability at co-transcriptionally formed G4 DNA. *Nucleic acids research* **2017**, *45* (10), 5850-5862.
(161) Paeschke, K.; Bochman, M. L.; Garcia, P. D.; Cejka, P.; Friedman, K. L.; Kowalczykowski, S. C.; Zakian, V. A. Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature* **2013**, *497* (7450), 458-462.
(162) Rodriguez, R.; Miller, K. M.; Forment, J. V.; Bradshaw, C. R.; Nikan, M.; Britton, S.; Oelschlaegel, T.; Xhemalce, B.; Balasubramanian, S.; Jackson, S. P. Small-molecule–induced DNA damage identifies alternative DNA structures in human genes. *Nature chemical biology* **2012**, *8* (3), 301-310.
(163) Balasubramanian, S.; Hurley, L. H.; Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nature Reviews Drug Discovery* **2011**, *10* (4), 261-275. DOI: 10.1038/nrd3428.
(164) Patel, D. J.; Phan, A. T.; Kuryavyi, V. Human telomere, oncogenic promoter and 5 ′ -UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic acids research* **2007**, *35* (22), 7429-7455.
(165) Han, H.; Hurley, L. H. G-quadruplex DNA: a potential target for anti-cancer drug design. *Trends in pharmacological sciences* **2000**, *21* (4), 136-142.
(166) Yadav, P.; Kim, N.; Kumari, M.; Verma, S.; Sharma, T. K.; Yadav, V.; Kumar, A. G-quadruplex structures in bacteria: Biological relevance and potential as an antimicrobial target. *Journal of Bacteriology* **2021**, *203* (13), 10.1128/jb. 00577-00520.
(167) Ruggiero, E.; Richter, S. N. G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. *Nucleic acids research* **2018**, *46* (7), 3270-3283.
(168) Métifiot, M.; Amrane, S.; Litvak, S.; Andreola, M.-L. G-quadruplexes in viruses: function and potential therapeutic applications. *Nucleic acids research* **2014**, *42* (20), 12352-12366.
(169) Xu, H.; Hurley, L. H. A first-in-class clinical G-quadruplex-targeting drug. The bench-to-bedside translation of the fluoroquinolone QQ58 to CX-5461 (Pidnarulex). *Bioorganic & Medicinal Chemistry Letters* **2022**, *77*, 129016. DOI: https://doi.org/10.1016/j.bmcl.2022.129016.
(170) Spiegel, J.; Adhikari, S.; Balasubramanian, S. The Structure and Function of DNA G-Quadruplexes. *Trends in Chemistry* **2020**, *2* (2), 123-136. DOI: 10.1016/j.trechm.2019.07.002 (acccessed 2020/04/27).

(171) Tran, P. L. T.; Mergny, J.-L.; Alberti, P. Stability of telomeric G-quadruplexes. *Nucleic acids research* **2011**, *39* (8), 3282-3294.

(172) Brázda, V.; Hároníková, L.; Liao, J. C.; Fojta, M. DNA and RNA quadruplex-binding proteins. *International journal of molecular sciences* **2014**, *15* (10), 17493-17517.

(173) Makarov, V. L.; Hirose, Y.; Langmore, J. P. Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. *Cell* **1997**, *88* (5), 657-666.

(174) Wright, W. E.; Tesmer, V. M.; Huffman, K. E.; Levene, S. D.; Shay, J. W. Normal human chromosomes have long G-rich telomeric overhangs at one end. *Genes & development* **1997**, *11* (21), 2801-2809.

(175) Stewart, S. A.; Ben-Porath, I.; Carey, V. J.; O'Connor, B. F.; Hahn, W. C.; Weinberg, R. A. Erosion of the telomeric single-strand overhang at replicative senescence. *Nature genetics* **2003**, *33* (4), 492-496.

(176) Dai, J.; Punchihewa, C.; Ambrus, A.; Chen, D.; Jones, R. A.; Yang, D. Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation. *Nucleic acids research* **2007**, *35* (7), 2440-2450.

(177) Yu, H.-Q.; Miyoshi, D.; Sugimoto, N. Characterization of structure and stability of long telomeric DNA G-quadruplexes. *Journal of the American Chemical Society* **2006**, *128* (48), 15461-15468.

(178) Renčiuk, D.; Kejnovská, I.; Školáková, P.; Bednářová, K.; Motlová, J.; Vorlíčková, M. Arrangements of human telomere DNA quadruplex in physiologically relevant K+ solutions. *Nucleic acids research* **2009**, *37* (19), 6625-6634.

(179) Monsen, R. C.; Chakravarthy, S.; Dean, W. L.; Chaires, J. B.; Trent, J. O. The solution structures of higher-order human telomere G-quadruplex multimers. *Nucleic Acids Research* **2021**, *49* (3), 1749-1768. DOI: 10.1093/nar/gkaa1285 (acccessed 11/27/2023).

(180) Carrino, S.; Hennecker, C. D.; Murrieta, A. C.; Mittermaier, A. Frustrated folding of guanine quadruplexes in telomeric DNA. *Nucleic acids research* **2021**, *49* (6), 3063-3076.

(181) Zahler, A. M.; Williamson, J. R.; Cech, T. R.; Prescott, D. M. Inhibition of telomerase by G-quartet DMA structures. *Nature* **1991**, *350* (6320), 718-720.

(182) Sun, D.; Thompson, B.; Cathers, B. E.; Salazar, M.; Kerwin, S. M.; Trent, J. O.; Jenkins, T. C.; Neidle, S.; Hurley, L. H. Inhibition of human telomerase by a G-quadruplex-interactive compound. *Journal of medicinal chemistry* **1997**, *40* (14), 2113-2116.

(183) Vannier, J.-B.; Pavicic-Kaltenbrunner, V.; Petalcorin, M. I.; Ding, H.; Boulton, S. J. RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. *Cell* **2012**, *149* (4), 795-806.

(184) Clynes, D.; Jelinska, C.; Xella, B.; Ayyub, H.; Scott, C.; Mitson, M.; Taylor, S.; Higgs, D. R.; Gibbons, R. J. Suppression of the alternative lengthening of telomere pathway by the chromatin remodelling factor ATRX. *Nature communications* **2015**, *6* (1), 7538.

(185) Gowan, S. M.; Heald, R.; Stevens, M. F.; Kelland, L. R. Potent inhibition of telomerase by small-molecule pentacyclic acridines capable of interacting with G-quadruplexes. *Molecular pharmacology* **2001**, *60* (5), 981-988.

(186) Han, H.; Hurley, L. H.; Salazar, M. A DNA polymerase stop assay for G-quadruplex-interactive compounds. *Nucleic acids research* **1999**, *27* (2), 537-542.

(187) Valton, A.-L.; Prioleau, M.-N. G-quadruplexes in DNA replication: a problem or a necessity? *Trends in Genetics* **2016**, *32* (11), 697-706.

(188) Cheung, I.; Schertzer, M.; Rose, A.; Lansdorp, P. M. Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. *Nature genetics* **2002**, *31* (4), 405-409.

(189) Lopes, J.; Piazza, A.; Bermejo, R.; Kriegsman, B.; Colosio, A.; Teulade - Fichou, M. P.; Foiani, M.; Nicolas, A. G - quadruplex - induced instability during leading - strand replication. *The EMBO journal* **2011**, *30* (19), 4033-4046.

(190) Puget, N.; Miller, K. M.; Legube, G. Non-canonical DNA/RNA structures during transcription-coupled double-strand break repair: roadblocks or bona fide repair intermediates? *DNA repair* **2019**, *81*, 102661.

(191) Lemmens, B.; van Schendel, R.; Tijsterman, M. Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers. *Nature Communications* **2015**, *6* (1), 8909. DOI: 10.1038/ncomms9909.

(192) Kruisselbrink, E.; Guryev, V.; Brouwer, K.; Pontier, D. B.; Cuppen, E.; Tijsterman, M. Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective C. elegans. *Current Biology* **2008**, *18* (12), 900-905.

(193) Crabbe, L.; Verdun, R. E.; Haggblom, C. I.; Karlseder, J. Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. *Science* **2004**, *306* (5703), 1951-1953.

(194) Tsuge, K.; Shimamoto, A. Research on Werner Syndrome: Trends from Past to Present and Future Prospects. *Genes* **2022**, *13* (10), 1802.

(195) Tian, Y.; Wang, W.; Lautrup, S.; Zhao, H.; Li, X.; Law, P. W. N.; Dinh, N.-D.; Fang, E. F.; Cheung, H. H.; Chan, W.-Y. WRN promotes bone development and growth by unwinding SHOX-G-quadruplexes via its helicase activity in Werner Syndrome. *Nature Communications* **2022**, *13* (1), 5456. DOI: 10.1038/s41467-022-33012-6.

(196) Ravichandran, S.; Kim, Y.-E.; Bansal, V.; Ghosh, A.; Hur, J.; Subramani, V. K.; Pradhan, S.; Lee, M. K.; Kim, K. K.; Ahn, J.-H. Genome-wide analysis of regulatory G-quadruplexes affecting gene expression in human cytomegalovirus. *PLoS Pathogens* **2018**, *14* (9), e1007334.

(197) Rawal, P.; Kummarasetti, V. B. R.; Ravindran, J.; Kumar, N.; Halder, K.; Sharma, R.; Mukerji, M.; Das, S. K.; Chowdhury, S. Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome research* **2006**, *16* (5), 644-655.

(198) Holder, I. T.; Hartig, J. S. A matter of location: influence of G-quadruplexes on Escherichia coli gene expression. *Chemistry & biology* **2014**, *21* (11), 1511-1521.

(199) Hurley, L. H.; Von Hoff, D. D.; Siddiqui-Jain, A.; Yang, D. Drug targeting of the c-MYC promoter to repress gene expression via a G-quadruplex silencer element. In *Seminars in oncology*, 2006; Elsevier: Vol. 33, pp 498-512.

(200) Agarwal, T.; Roy, S.; Kumar, S.; Chakraborty, T. K.; Maiti, S. In the sense of transcription regulation by G-quadruplexes: asymmetric effects in sense and antisense strands. *Biochemistry* **2014**, *53* (23), 3711-3718.

(201) Halder, K.; Wieland, M.; Hartig, J. S. Predictable suppression of gene expression by 5′ -UTR-based RNA quadruplexes. *Nucleic acids research* **2009**, *37* (20), 6811-6817.

(202) Armas, P.; David, A.; Calcaterra, N. B. Transcriptional control by G-quadruplexes: In vivo roles and perspectives for specific intervention. *Transcription* **2017**, *8* (1), 21-25.

(203) Neidle, S.; Read, M. A. G‐quadruplexes as therapeutic targets. *Biopolymers: Original Research on Biomolecules* **2000**, *56* (3), 195-208.

(204) Balasubramanian, S.; Neidle, S. G-quadruplex nucleic acids as therapeutic targets. *Current opinion in chemical biology* **2009**, *13* (3), 345-353.

(205) Neidle, S. Quadruplex nucleic acids as targets for anticancer therapeutics. *Nature Reviews Chemistry* **2017**, *1* (5), 0041.

(206) Hilton, J.; Gelmon, K.; Bedard, P. L.; Tu, D.; Xu, H.; Tinker, A. V.; Goodwin, R.; Laurie, S. A.; Jonker, D.; Hansen, A. R.; et al. Results of the phase I CCTG IND.231 trial of CX-5461 in patients with advanced solid tumors enriched for DNA-repair deficiencies. *Nature Communications* **2022**, *13* (1), 3607. DOI: 10.1038/s41467-022-31199-2.

(207) McLuckie, K. I. E.; Di Antonio, M.; Zecchini, H.; Xian, J.; Caldas, C.; Krippendorff, B.-F.; Tannahill, D.; Lowe, C.; Balasubramanian, S. G-Quadruplex DNA as a Molecular Target for Induced Synthetic Lethality in Cancer Cells. *Journal of the American Chemical Society* **2013**, *135* (26), 9640-9643. DOI: 10.1021/ja404868t.

(208) Wang, K.-B.; Elsayed, M. S.; Wu, G.; Deng, N.; Cushman, M.; Yang, D. Indenoisoquinoline topoisomerase inhibitors strongly bind and stabilize the MYC promoter G-quadruplex and downregulate MYC. *Journal of the American Chemical Society* **2019**, *141* (28), 11059-11070.

(209) Šponer, J.; Bussi, G.; Stadlbauer, P.; Kührová, P.; Banáš, P.; Islam, B.; Haider, S.; Neidle, S.; Otyepka, M. Folding of guanine quadruplex molecules–funnel-like mechanism or kinetic partitioning? An overview from MD simulation studies. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2017**, *1861* (5, Part B), 1246-1263. DOI: https://doi.org/10.1016/j.bbagen.2016.12.008.

(210) Harkness, R. W.; Mittermaier, A. K. G-register exchange dynamics in guanine quadruplexes. *Nucleic acids research* **2016**, *44* (8), 3481-3494.

(211) Grün, J. T.; Blümler, A.; Burkhart, I.; Wirmer-Bartoschek, J.; Heckel, A.; Schwalbe, H. Unraveling the Kinetics of Spare-Tire DNA G-Quadruplex Folding. *Journal of the American Chemical Society* **2021**, *143* (16), 6185-6193.

(212) Koirala, D.; Ghimire, C.; Bohrer, C.; Sannohe, Y.; Sugiyama, H.; Mao, H. Long-loop G-quadruplexes are misfolded population minorities with fast transition kinetics in human telomeric sequences. *Journal of the American Chemical Society* **2013**, *135* (6), 2235-2241.

(213) Gray, R. D.; Buscaglia, R.; Chaires, J. B. Populated intermediates in the thermal unfolding of the human telomeric quadruplex. *Journal of the American Chemical Society* **2012**, *134* (40), 16834-16844.

(214) Bončina, M.; Lah, J.; Prislan, I.; Vesnaver, G. Energetic basis of human telomeric DNA folding into G-quadruplex structures. *Journal of the American Chemical Society* **2012**, *134* (23), 9657-9663.

(215) 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. *Angewandte Chemie* **2015**, *127* (29), 8564-8568.

(216) Grün, J. T.; Hennecker, C.; Klötzner, D.-P.; Harkness, R. W.; Bessi, I.; Heckel, A.; Mittermaier, A. K.; Schwalbe, H. Conformational Dynamics of Strand Register Shifts in DNA G-Quadruplexes. *Journal of the American Chemical Society* **2020**, *142* (1), 264-273. DOI: 10.1021/jacs.9b10367.

(217) Ma, L.; Iezzi, M.; Kaucher, M. S.; Lam, Y.-F.; Davis, J. T. Cation Exchange in Lipophilic G-Quadruplexes: Not All Ion Binding Sites Are Equal. *Journal of the American Chemical Society* **2006**, *128* (47), 15269-15277. DOI: 10.1021/ja064878n.

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

(219) Mergny, J.-L.; De Cian, A.; Ghelab, A.; Saccà, B.; Lacroix, L. Kinetics of tetramolecular quadruplexes. *Nucleic Acids Research* **2005**, *33* (1), 81-94. DOI: 10.1093/nar/gki148 (acccessed 12/5/2023).

(220) Pörschke, D.; Eigen, M. Co-operative non-enzymatic base recognition III. Kinetics of the helix—coil transition of the oligoribouridylic· oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic pH. *Journal of molecular biology* **1971**, *62* (2), 361-381.

(221) Rougée, M.; Faucon, B.; Mergny, J.; Barcelo, F.; Giovannangeli, C.; Garestier, T.; Hélène, C. Kinetics and thermodynamics of triple-helix formation: Effects of ionic strength and mismatched. *Biochemistry* **1992**, *31* (38), 9269-9278.

(222) Mergny, J.-L.; Lacroix, L. Kinetics and thermodynamics of i-DNA formation: phosphodiester versus modified oligodeoxynucleotides. *Nucleic acids research* **1998**, *26* (21), 4797-4803.

(223) Seeman, N. C.; Sleiman, H. F. DNA nanotechnology. *Nature Reviews Materials* **2017**, *3* (1), 1-23.

(224) Iqbal, P.; Preece, J. A.; Mendes, P. M. Nanotechnology: the “top‐down” and “bottom‐up” approaches. *Supramolecular chemistry: from molecules to nanomaterials* **2012**.

(225) Holliday, R. A mechanism for gene conversion in fungi. *Genetics Research* **1964**, *5* (2), 282-304.

(226) Seeman, N. C. Nucleic acid junctions and lattices. *Journal of theoretical biology* **1982**, *99* (2), 237-247.

(227) Fu, T. J.; Seeman, N. C. DNA double-crossover molecules. *Biochemistry* **1993**, *32* (13), 3211-3220.

(228) Dey, S.; Fan, C.; Gothelf, K. V.; Li, J.; Lin, C.; Liu, L.; Liu, N.; Nijenhuis, M. A.; Saccà, B.; Simmel, F. C. DNA origami. *Nature Reviews Methods Primers* **2021**, *1* (1), 13.

(229) Rothemund, P. W. Folding DNA to create nanoscale shapes and patterns. *Nature* **2006**, *440* (7082), 297-302.

(230) Douglas, S. M.; Dietz, H.; Liedl, T.; Högberg, B.; Graf, F.; Shih, W. M. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* **2009**, *459* (7245), 414-418.

(231) Andersen, E. S.; Dong, M.; Nielsen, M. M.; Jahn, K.; Subramani, R.; Mamdouh, W.; Golas, M. M.; Sander, B.; Stark, H.; Oliveira, C. L. Self-assembly of a nanoscale DNA box with a controllable lid. *Nature* **2009**, *459* (7243), 73-76.

(232) Benson, E.; Mohammed, A.; Gardell, J.; Masich, S.; Czeizler, E.; Orponen, P.; Högberg, B. DNA rendering of polyhedral meshes at the nanoscale. *Nature* **2015**, *523* (7561), 441-444.

(233) Thubagere, A. J.; Li, W.; Johnson, R. F.; Chen, Z.; Doroudi, S.; Lee, Y. L.; Izatt, G.; Wittman, S.; Srinivas, N.; Woods, D. A cargo-sorting DNA robot. *Science* **2017**, *357* (6356), eaan6558.

(234) Chidchob, P.; Sleiman, H. F. Recent advances in DNA nanotechnology. *Current opinion in chemical biology* **2018**, *46*, 63-70.

(235) Platnich, C. M.; Hariri, A. A.; Sleiman, H. F.; Cosa, G. Advancing Wireframe DNA Nanostructures Using Single-Molecule Fluorescence Microscopy Techniques. *Accounts of Chemical Research* **2019**, *52* (11), 3199-3210. DOI: 10.1021/acs.accounts.9b00424.
(236) Saliba, D.; Luo, X.; Rizzuto, F. J.; Sleiman, H. F. Programming rigidity into size-defined wireframe DNA nanotubes. *Nanoscale* **2023**, *15* (11), 5403-5413.
(237) Hamblin, G. D.; Hariri, A. A.; Carneiro, K. M.; Lau, K. L.; Cosa, G.; Sleiman, H. F. Simple design for DNA nanotubes from a minimal set of unmodified strands: rapid, room-temperature assembly and readily tunable structure. *ACS nano* **2013**, *7* (4), 3022-3028.
(238) Li, Q.; Zhao, J.; Liu, L.; Jonchhe, S.; Rizzuto, F. J.; Mandal, S.; He, H.; Wei, S.; Sleiman, H. F.; Mao, H. A poly (thymine)–melamine duplex for the assembly of DNA nanomaterials. *Nature Materials* **2020**, *19* (9), 1012-1018.
(239) Avakyan, N.; Greschner, A. A.; Aldaye, F.; Serpell, C. J.; Toader, V.; Petitjean, A.; Sleiman, H. F. Reprogramming the assembly of unmodified DNA with a small molecule. *Nature Chemistry* **2016**, *8*, 368, Article. DOI: 10.1038/nchem.2451 https://www.nature.com/articles/nchem.2451#supplementary-information.
(240) Roy, B.; Bairi, P.; Nandi, A. K. Supramolecular assembly of melamine and its derivatives: nanostructures to functional materials. *RSC advances* **2014**, *4* (4), 1708-1734.
(241) Lachance-Brais, C.; Hennecker, C. D.; Alenaizan, A.; Luo, X.; Toader, V.; Taing, M.; Sherrill, C. D.; Mittermaier, A. K.; Sleiman, H. F. Tuning DNA Supramolecular Polymers by the Addition of Small, Functionalized Nucleobase Mimics. *Journal of the American Chemical Society* **2021**.
(242) Lachance‐Brais, C.; Rammal, M.; Asohan, J.; Katolik, A.; Luo, X.; Saliba, D.; Jonderian, A.; Damha, M. J.; Harrington, M. J.; Sleiman, H. F. Small Molecule‐Templated DNA Hydrogel with Record Stiffness Integrates and Releases DNA Nanostructures and Gene Silencing Nucleic Acids. *Advanced Science* **2023**, 2205713.
(243) Alenaizan, A.; Fauché, K.; Krishnamurthy, R.; Sherrill, C. D. Noncovalent Helicene Structure between Nucleic Acids and Cyanuric Acid. *Chemistry–A European Journal* **2021**, *27* (12), 4043-4052.
(244) Hennecker, C. D.; Lachance-Brais, C.; Sleiman, H.; Mittermaier, A. Using transient equilibria (TREQ) to measure the thermodynamics of slowly assembling supramolecular systems. *Science Advances* **2022**, *8* (14), eabm8455.
(245) Rizzuto, F. J.; Platnich, C. M.; Luo, X.; Shen, Y.; Dore, M. D.; Lachance-Brais, C.; Guarné, A.; Cosa, G.; Sleiman, H. F. A dissipative pathway for the structural evolution of DNA fibres. *Nature Chemistry* **2021**, *13* (9), 843-849.
(246) Harkness V, R. W.; Avakyan, N.; Sleiman, H. F.; Mittermaier, A. K. Mapping the energy landscapes of supramolecular assembly by thermal hysteresis. *Nature Communications* **2018**, *9* (1), 3152. DOI: 10.1038/s41467-018-05502-z.
(247) Oosawa, F.; Kasai, M. A theory of linear and helical aggregations of macromolecules. *Journal of Molecular Biology* **1962**, *4* (1), 10-21. DOI: https://doi.org/10.1016/S0022-2836(62)80112-0.
(248) Goldstein, R. F.; Stryer, L. Cooperative polymerization reactions. Analytical approximations, numerical examples, and experimental strategy. *Biophysical Journal* **1986**, *50* (4), 583-599. DOI: https://doi.org/10.1016/S0006-3495(86)83498-1.

(249) Zhao, D.; Moore, J. S. Nucleation–elongation: a mechanism for cooperative supramolecular polymerization. *Organic & Biomolecular Chemistry* **2003**, *1* (20), 3471-3491, 10.1039/B308788C. DOI: 10.1039/B308788C.

(250) Luo, Y.; Granzhan, A.; Verga, D.; Mergny, J. L. FRET ‐ MC: A fluorescence melting competition assay for studying G4 structures in vitro. *Biopolymers* **2021**, *112* (4), e23415.

(251) Guo, Q.; Lu, M.; Kallenbach, N. R. Adenine affects the structure and stability of telomeric sequences. *Journal of Biological Chemistry* **1992**, *267* (22), 15293-15300.

(252) Mergny, J.-L.; Phan, A.-T.; Lacroix, L. Following G ‐ quartet formation by UV ‐ spectroscopy. *FEBS letters* **1998**, *435* (1), 74-78.

(253) Pagano, B.; Randazzo, A.; Fotticchia, I.; Novellino, E.; Petraccone, L.; Giancola, C. Differential scanning calorimetry to investigate G-quadruplexes structural stability. *Methods* **2013**, *64* (1), 43-51.

(254) Pal, S. *Fundamentals of molecular structural biology*; Academic Press, 2019.

(255) Keller, B. Structural cell wall proteins. *Plant physiology* **1993**, *101* (4), 1127.

(256) Parkinson, J. S.; Kofoid, E. C. Communication modules in bacterial signaling proteins. *Annual review of genetics* **1992**, *26* (1), 71-112.

(257) de Souza Vandenberghe, L. P.; Karp, S. G.; Pagnoncelli, M. G. B.; von Linsingen Tavares, M.; Junior, N. L.; Diestra, K. V.; Viesser, J. A.; Soccol, C. R. Classification of enzymes and catalytic properties. In *Biomass, biofuels, biochemicals*, Elsevier, 2020; pp 11-30.

(258) Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **1995**, *80* (2), 225-236.

(259) Reyes-Turcu, F. E.; Ventii, K. H.; Wilkinson, K. D. Regulation and cellular roles of ubiquitin-specific deubiquitinating enzymes. *Annual review of biochemistry* **2009**, *78*, 363-397.

(260) Benhar, M.; Forrester, M. T.; Stamler, J. S. Protein denitrosylation: enzymatic mechanisms and cellular functions. *Nature reviews Molecular cell biology* **2009**, *10* (10), 721-732.

(261) Capaldi, R. A.; Aggeler, R. Mechanism of the F1F0-type ATP synthase, a biological rotary motor. *Trends in biochemical sciences* **2002**, *27* (3), 154-160.

(262) Savojardo, C.; Baldazzi, D.; Babbi, G.; Martelli, P. L.; Casadio, R. Mapping human disease-associated enzymes into Reactome allows characterization of disease groups and their interactions. *Scientific Reports* **2022**, *12* (1), 17963. DOI: 10.1038/s41598-022-22818-5.

(263) Wohlgemuth, R. Biocatalysis—key to sustainable industrial chemistry. *Current opinion in biotechnology* **2010**, *21* (6), 713-724.

(264) Choi, J.-M.; Han, S.-S.; Kim, H.-S. Industrial applications of enzyme biocatalysis: Current status and future aspects. *Biotechnology Advances* **2015**, *33* (7), 1443-1454. DOI: https://doi.org/10.1016/j.biotechadv.2015.02.014.

(265) Miller, D. C.; Athavale, S. V.; Arnold, F. H. Combining chemistry and protein engineering for new-to-nature biocatalysis. *Nature synthesis* **2022**, *1* (1), 18-23.

(266) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. How enzymes work: analysis by modern rate theory and computer simulations. *Science* **2004**, *303* (5655), 186-195.

(267) Zhang, M.; Zhou, M.; Van Etten, R. L.; Stauffacher, C. V. Crystal structure of bovine low molecular weight phosphotyrosyl phosphatase complexed with the transition state analog vanadate. *Biochemistry* **1997**, *36* (1), 15-23.

(268) Kienhöfer, A.; Kast, P.; Hilvert, D. Selective stabilization of the chorismate mutase transition state by a positively charged hydrogen bond donor. *Journal of the American Chemical Society* **2003**, *125* (11), 3206-3207.

(269) Shurki, A.; Štrajbl, M.; Villa, J.; Warshel, A. How much do enzymes really gain by restraining their reacting fragments? *Journal of the American Chemical Society* **2002**, *124* (15), 4097-4107.

(270) Agarwal, P. K.; Billeter, S. R.; Rajagopalan, P. R.; Benkovic, S. J.; Hammes-Schiffer, S. Network of coupled promoting motions in enzyme catalysis. *Proceedings of the National Academy of Sciences* **2002**, *99* (5), 2794-2799.

(271) Hammes-Schiffer, S. Catalytic efficiency of enzymes: a theoretical analysis. *Biochemistry* **2013**, *52* (12), 2012-2020.

(272) Briggs, G. E.; Haldane, J. B. S. A note on the kinetics of enzyme action. *Biochemical journal* **1925**, *19* (2), 338.

(273) Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology* **2011**, *162* (6), 1239-1249.

(274) Cook, M. A.; Wright, G. D. The past, present, and future of antibiotics. *Science Translational Medicine* **2022**, *14* (657), eabo7793. DOI: doi:10.1126/scitranslmed.abo7793.

(275) Kausar, S.; Said Khan, F.; Ishaq Mujeeb Ur Rehman, M.; Akram, M.; Riaz, M.; Rasool, G.; Hamid Khan, A.; Saleem, I.; Shamim, S.; Malik, A. A review: Mechanism of action of antiviral drugs. *International journal of immunopathology and pharmacology* **2021**, *35*, 20587384211002621.

(276) Shaker, B.; Ahmad, S.; Lee, J.; Jung, C.; Na, D. In silico methods and tools for drug discovery. *Computers in biology and medicine* **2021**, *137*, 104851.

(277) Jager, S.; Brand, L.; Eggeling, C. New fluorescence techniques for high-throughput drug discovery. *Current Pharmaceutical Biotechnology* **2003**, *4* (6), 463-476.

(278) Simon, R. P.; Winter, M.; Kleiner, C.; Ries, R.; Schnapp, G.; Heimann, A.; Li, J.; Zuvela-Jelaska, L.; Bretschneider, T.; Luippold, A. H. MALDI-TOF mass spectrometry-based high-throughput screening for inhibitors of the cytosolic DNA sensor cGAS. *SLAS DISCOVERY: Advancing the Science of Drug Discovery* **2020**, *25* (4), 372-383.

(279) Umscheid, C. A.; Margolis, D. J.; Grossman, C. E. Key concepts of clinical trials: a narrative review. *Postgraduate medicine* **2011**, *123* (5), 194.

(280) Smietana, K.; Siatkowski, M.; Møller, M. Trends in clinical success rates. *Nat Rev Drug Discov* **2016**, *15* (6), 379-380.

(281) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics* **2016**, *47*, 20-33.

(282) Guce, A. I.; Mortimer, S. E.; Yoon, T.; Painter, C. A.; Jiang, W.; Mellins, E. D.; Stern, L. J. HLA-DO acts as a substrate mimic to inhibit HLA-DM by a competitive mechanism. *Nature structural & molecular biology* **2013**, *20* (1), 90-98.

(283) Cornish-Bowden, A. Why is uncompetitive inhibition so rare?: A possible explanation, with implications for the design of drugs and pesticides. *FEBS Letters* **1986**, *203* (1), 3-6. DOI: https://doi.org/10.1016/0014-5793(86)81424-7.

(284) Hoylaerts, M. F.; Manes, T.; Millán, J. L. Molecular mechanism of uncompetitive inhibition of human placental and germ-cell alkaline phosphatase. *Biochemical Journal* **1992**, *286* (1), 23-30. DOI: 10.1042/bj2860023 (acccessed 1/2/2024).

(285) Shyur, L. F.; Poland, B. W.; Honzatko, R. B.; Fromm, H. J. Major changes in the kinetic mechanism of AMP inhibition and AMP cooperativity attend the mutation of Arg49 in fructose-1,6-bisphosphatase. *Journal of Biological Chemistry* **1997**, *272* (42), 26295-26299, Article. DOI: 10.1074/jbc.272.42.26295 Scopus.

(286) Pallanca, A.; Mazzaracchio, R.; Brigotti, M.; Carnicelli, D.; Alvergna, P.; Sperti, S.; Montanaro, L. Uncompetitive inhibition by adenine of the RNA-N-glycosidase activity of ribosome-inactivating proteins. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1998**, *1384* (2), 277-284. DOI: https://doi.org/10.1016/S0167-4838(98)00019-3.

(287) Sebaugh, J. Guidelines for accurate EC50/IC50 estimation. *Pharmaceutical statistics* **2011**, *10* (2), 128-134.

(288) Gubler, H.; Schopfer, U.; Jacoby, E. Theoretical and experimental relationships between percent inhibition and IC50 data observed in high-throughput screening. *Journal of biomolecular screening* **2013**, *18* (1), 1-13.

(289) Yung-Chi, C.; Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical pharmacology* **1973**, *22* (23), 3099-3108.

(290) Johnson, D. S.; Weerapana, E.; Cravatt, B. F. Strategies for discovering and derisking covalent, irreversible enzyme inhibitors. *Future medicinal chemistry* **2010**, *2* (6), 949-964.

(291) De Cesco, S.; Kurian, J.; Dufresne, C.; Mittermaier, A. K.; Moitessier, N. Covalent inhibitors design and discovery. *European Journal of Medicinal Chemistry* **2017**, *138*, 96-114.

(292) Zhang, H.; Amunugama, H.; Ney, S.; Cooper, N.; Hollenberg, P. F. Mechanism-based inactivation of human cytochrome P450 2B6 by clopidogrel: involvement of both covalent modification of cysteinyl residue 475 and loss of heme. *Molecular pharmacology* **2011**, *80* (5), 839-847.

(293) Roth, G. J.; Machuga, E. T.; Ozols, J. Isolation and covalent structure of the aspirin-modified, active-site region of prostaglandin synthetase. *Biochemistry* **1983**, *22* (20), 4672-4675.

(294) Robertson, J. G. Mechanistic basis of enzyme-targeted drugs. *Biochemistry* **2005**, *44* (15), 5561-5571.

(295) Thorarensen, A.; Balbo, P.; Banker, M. E.; Czerwinski, R. M.; Kuhn, M.; Maurer, T. S.; Telliez, J.-B.; Vincent, F.; Wittwer, A. J. The advantages of describing covalent inhibitor in vitro potencies by IC50 at a fixed time point. IC50 determination of covalent inhibitors provides meaningful data to medicinal chemistry for SAR optimization. *Bioorganic & Medicinal Chemistry* **2021**, *29*, 115865.

(296) Mons, E.; Roet, S.; Kim, R. Q.; Mulder, M. P. A comprehensive guide for assessing covalent inhibition in enzymatic assays illustrated with kinetic simulations. *Current Protocols* **2022**, *2* (6), e419.

(297) Exnowitz, F.; Meyer, B.; Hackl, T. NMR for direct determination of Km and Vmax of enzyme reactions based on the Lambert W function-analysis of progress curves. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2012**, *1824* (3), 443-449.

(298) Copeland, R. A. Kinetics of single-substrate enzyme reactions. *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis* **2000**, *2*, 137-139.

(299) Schwabe, C. A fluorescent assay for proteolytic enzymes. *Analytical biochemistry* **1973**, *53* (2), 484-490.

(300) Van Oers, T. J.; Piercey, A.; Belovodskiy, A.; Reiz, B.; Donnelly, B. L.; Vuong, W.; Lemieux, M. J.; Nieman, J. A.; Auclair, K.; Vederas, J. C. Deuteration for Metabolic Stabilization of SARS-CoV-2 Inhibitors GC373 and Nirmatrelvir. *Organic Letters* **2023**, *25* (31), 5885-5889.

(301) Scriba, G. K.; Belal, F. Advances in capillary electrophoresis-based enzyme assays. *Chromatographia* **2015**, *78*, 947-970.

(302) Lambeth, D. O.; Muhonen, W. W. High-performance liquid chromatography-based assays of enzyme activities. *Journal of Chromatography B: Biomedical Sciences and Applications* **1994**, *656* (1), 143-157.

(303) Todd, M. J.; Gomez, J. Enzyme kinetics determined using calorimetry: a general assay for enzyme activity? *Analytical biochemistry* **2001**, *296* (2), 179-187.

(304) Goličnik, M. Exact and approximate solutions for the decades ‐ old Michaelis ‐ Menten equation: Progress ‐ curve analysis through integrated rate equations. *Biochemistry and Molecular Biology Education* **2011**, *39* (2), 117-125.

(305) Di Trani, J. M.; Moitessier, N.; Mittermaier, A. K. Complete kinetic characterization of enzyme inhibition in a single isothermal titration calorimetric experiment. *Analytical chemistry* **2018**, *90* (14), 8430-8435.

(306) Di Trani, J. M.; Moitessier, N.; Mittermaier, A. K. Measuring rapid time-scale reaction kinetics using isothermal titration calorimetry. *Analytical chemistry* **2017**, *89* (13), 7022-7030.

(307) Di Trani, J. M.; De Cesco, S.; O'Leary, R.; Plescia, J.; do Nascimento, C. J.; Moitessier, N.; Mittermaier, A. K. Rapid measurement of inhibitor binding kinetics by isothermal titration calorimetry. *Nature communications* **2018**, *9* (1), 1-7.

(308) Mukundan, V. T.; Phan, A. T. Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. *Journal of the American Chemical Society* **2013**, *135* (13), 5017-5028. DOI: 10.1021/ja310251r.

(309) Heddi, B.; Martín-Pintado, N.; Serimbetov, Z.; Kari, T. M. A.; 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.* **2016**, *44* (2), 910-916. DOI: 10.1093/nar/gkv1357 PubMed.

(310) 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.* **2010**, *38* (3), 1009-1021. DOI: 10.1093/nar/gkp1029 PMC.

(311) Petraccone, L.; Erra, E.; Esposito, V.; Randazzo, A.; Mayol, L.; Nasti, L.; Barone, G.; Giancola, C. Stability and Structure of Telomeric DNA Sequences Forming Quadruplexes Containing Four G-Tetrads with Different Topological Arrangements. *Biochemistry* **2004**, *43* (16), 4877-4884. DOI: 10.1021/bi0300985.

(312) Ding, Y.; Fleming, A. M.; Burrows, C. J. Case studies on potential G-quadruplex-forming sequences from the bacterial orders Deinococcales and Thermales derived from a survey of published genomes. *Scientific Reports* **2018**, *8* (1), 15679. DOI: 10.1038/s41598-018-33944-4.

(313) Kaplan, O. I.; Berber, B.; Hekim, N.; Doluca, O. G-quadruplex prediction in E. coli genome reveals a conserved putative G-quadruplex-Hairpin-Duplex switch. *Nucleic Acids Res.* **2016**, *44* (19), 9083-9095. DOI: 10.1093/nar/gkw769 (acccessed 4/27/2020).

(314) Falabella, M.; Fernandez, R. J.; Johnson, F. B.; Kaufman, B. A. Potential Roles for G-Quadruplexes in Mitochondria. *Curr. Med. Chem.* **2019**, *26* (16), 2918-2932. DOI: 10.2174/0929867325666180228165527 PubMed.

(315) Murat, P.; Zhong, J.; Lekieffre, L.; Cowieson, N. P.; Clancy, J. L.; Preiss, T.; Balasubramanian, S.; Khanna, R.; Tellam, J. G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. *Nat. Chem. Biol.* **2014**, *10* (5), 358-364. DOI: 10.1038/nchembio.1479.

(316) Rhodes, D.; Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* **2015**, *43* (18), 8627-8637. DOI: 10.1093/nar/gkv862.

(317) Kim, N. The Interplay between G-quadruplex and Transcription. *Current medicinal chemistry* **2019**, *26* (16), 2898-2917. DOI: 10.2174/0929867325666171229132619 PubMed.

(318) Arora, A.; Dutkiewicz, M.; Scaria, V.; Hariharan, M.; Maiti, S.; Kurreck, J. Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. *RNA* **2008**, *14* (7), 1290-1296. DOI: rna.1001708 [pii]
10.1261/rna.1001708.

(319) Song, J.; Perreault, J.-P.; Topisirovic, I.; Richard, S. RNA G-quadruplexes and their potential regulatory roles in translation. *Translation* **2016**, *4* (2), e1244031. DOI: 10.1080/21690731.2016.1244031 PMC.

(320) Endoh, T.; Kawasaki, Y.; Sugimoto, N. Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor alpha. *Nucleic Acids Res.* **2013**, *41* (12), 6222-6231. DOI: 10.1093/nar/gkt286.

(321) Harkness, R. W. t.; Mittermaier, A. K. G-register exchange dynamics in guanine quadruplexes. *Nucleic Acids Res.* **2016**, *44* (8), 3481-3494. DOI: 10.1093/nar/gkw190.

(322) 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* **2002**, *99* (18), 11593-11598. DOI: 10.1073/pnas.182256799
182256799 [pii].

(323) Simonsson, T.; Kubista, M.; Pecinka, P. DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.* **1998**, *26* (5), 1167-1172. DOI: 10.1093/nar/26.5.1167 (acccessed 4/27/2020).

(324) González, V.; Guo, K.; Hurley, L.; Sun, D. Identification and Characterization of Nucleolin as a c-myc G-quadruplex-binding Protein. *Journal of Biological Chemistry* **2009**, *284* (35), 23622-23635. DOI: 10.1074/jbc.M109.018028.

(325) Li, W.; Wu, P.; Ohmichi, T.; Sugimoto, N. Characterization and thermodynamic properties of quadruplex/duplex competition. *FEBS Lett.* **2002**, *526* (1), 77-81. DOI: https://doi.org/10.1016/S0014-5793(02)03118-6.

(326) Zhang, A. Y. Q.; Balasubramanian, S. The Kinetics and Folding Pathways of Intramolecular G-Quadruplex Nucleic Acids. *J. Am. Chem. Soc.* **2012**, *134* (46), 19297-19308. DOI: 10.1021/ja309851t.

(327) Endoh, T.; Sugimoto, N. Conformational Dynamics of the RNA G-Quadruplex and its Effect on Translation Efficiency. *Molecules (Basel, Switzerland)* **2019**, *24* (8), 1613. DOI: 10.3390/molecules24081613 PubMed.

(328) Kreig, A.; Calvert, J.; Sanoica, J.; Cullum, E.; Tipanna, R.; Myong, S. G-quadruplex formation in double strand DNA probed by NMM and CV fluorescence. *Nucleic Acids Res.* **2015**, *43* (16), 7961-7970. DOI: 10.1093/nar/gkv749 PubMed.

(329) Shirude, P. S.; Balasubramanian, S. Single molecule conformational analysis of DNA G-quadruplexes. *Biochimie* **2008**, *90* (8), 1197-1206. DOI: 10.1016/j.biochi.2008.01.015 PubMed.

(330) Cramer, P. Structure and Function of RNA Polymerase II. In *Advances in Protein Chemistry*, Vol. 67; Academic Press, 2004; pp 1-42.

(331) Kugel, J. F.; Goodrich, J. A. A Kinetic Model for the Early Steps of RNA Synthesis by Human RNA Polymerase II. *J. Biol. Chem.* **2000**, *275* (51), 40483-40491.

(332) Jonkers, I.; Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* **2015**, *16* (3), 167-177. DOI: 10.1038/nrm3953.

(333) Hamdan, S. M.; Loparo, J. J.; Takahashi, M.; Richardson, C. C.; van Oijen, A. M. Dynamics of DNA replication loops reveal temporal control of lagging-strand synthesis. *Nature* **2009**, *457* (7227), 336-339. DOI: 10.1038/nature07512 PubMed.

(334) Yates, L. A.; Aramayo, R. J.; Pokhrel, N.; Caldwell, C. C.; Kaplan, J. A.; Perera, R. L.; Spies, M.; Antony, E.; Zhang, X. A structural and dynamic model for the assembly of Replication Protein A on single-stranded DNA. *Nature Communications* **2018**, *9* (1), 5447. DOI: 10.1038/s41467-018-07883-7.

(335) Chen, R.; Subramanyam, S.; Elcock, A. H.; Spies, M.; Wold, M. S. Dynamic binding of replication protein a is required for DNA repair. *Nucleic Acids Res.* **2016**, *44* (12), 5758-5772. DOI: 10.1093/nar/gkw339 (acccessed 4/27/2020).

(336) Murat, P.; Marsico, G.; Herdy, B.; Ghanbarian, A.; Portella, G.; Balasubramanian, S. RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs. *Genome Biology* **2018**, *19* (1), 229. DOI: 10.1186/s13059-018-1602-2.

(337) Guo, J. U.; Bartel, D. P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **2016**, *353* (6306), aaf5371. DOI: 10.1126/science.aaf5371.

(338) Espah Borujeni, A.; Salis, H. M. Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism. *J. Am. Chem. Soc.* **2016**, *138* (22), 7016-7023. DOI: 10.1021/jacs.6b01453.

(339) Hou, X.-M.; Fu, Y.-B.; Wu, W.-Q.; Wang, L.; Teng, F.-Y.; Xie, P.; Wang, P.-Y.; Xi, X.-G. Involvement of G-triplex and G-hairpin in the multi-pathway folding of human telomeric G-quadruplex. *Nucleic Acids Res.* **2017**, *45* (19), 11401-11412. DOI: 10.1093/nar/gkx766 (acccessed 4/27/2020).

(340) Limongelli, V.; De Tito, S.; Cerofolini, L.; Fragai, M.; Pagano, B.; Trotta, R.; Cosconati, S.; Marinelli, L.; Novellino, E.; Bertini, I.; et al. The G-Triplex DNA. *Angewandte Chemie International Edition* **2013**, *52* (8), 2269-2273. DOI: 10.1002/anie.201206522.

(341) Gray, R. D.; Buscaglia, R.; Chaires, J. B. Populated intermediates in the thermal unfolding of the human telomeric quadruplex. *J. Am. Chem. Soc.* **2012**, *134* (40), 16834-16844. DOI: 10.1021/ja307543z.

(342) Boncina, M.; Lah, J.; Prislan, I.; Vesnaver, G. Energetic basis of human telomeric DNA folding into G-quadruplex structures. *J. Am. Chem. Soc.* **2012**, *134* (23), 9657-9663. DOI: 10.1021/ja300605n.

(343) 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* **2015**, *54* (29), 8444-8448. DOI: 10.1002/anie.201502286.

(344) Gray, R. D.; Chaires, J. B. Analysis of multidimensional G-quadruplex melting curves. *Current protocols in nucleic acid chemistry* **2011**, *Chapter 17*, Unit17.14-Unit17.14. DOI: 10.1002/0471142700.nc1704s45 PubMed.

(345) Wang, H.; Nora, G. J.; Ghodke, H.; Opresko, P. L. Single Molecule Studies of Physiologically Relevant Telomeric Tails Reveal POT1 Mechanism for Promoting G-quadruplex Unfolding. *J. Biol. Chem.* **2011**, *286* (9), 7479-7489.

(346) Long, X.; Parks, J. W.; Bagshaw, C. R.; Stone, M. D. Mechanical unfolding of human telomere G-quadruplex DNA probed by integrated fluorescence and magnetic tweezers spectroscopy. *Nucleic Acids Res.* **2013**, *41* (4), 2746-2755. DOI: 10.1093/nar/gks1341 PubMed.

(347) Ying, L.; Green, J. J.; Li, H.; Klenerman, D.; Balasubramanian, S. Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences* **2003**, *100* (25), 14629. DOI: 10.1073/pnas.2433350100.

(348) Aksel, T.; Barrick, D. Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. *Biophysical journal* **2014**, *107* (1), 220-232. DOI: 10.1016/j.bpj.2014.04.058 PubMed.

(349) Seenisamy, J.; Rezler, E. M.; Powell, T. J.; Tye, D.; Gokhale, V.; Joshi, C. S.; Siddiqui-Jain, A.; Hurley, L. H. The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. *J. Am. Chem. Soc.* **2004**, *126* (28), 8702-8709. DOI: 10.1021/ja040022b.

(350) Chen, H.; Liu, H.; Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduction and Targeted Therapy* **2018**, *3* (1), 5. DOI: 10.1038/s41392-018-0008-7.

(351) Harkness, R. W.; Hennecker, C.; Grün, J. T.; Blümler, A.; Heckel, A.; Schwalbe, H.; Mittermaier, A. K. Parallel reaction pathways accelerate folding of a guanine quadruplex. *Nucleic acids research* **2021**, *49* (3), 1247-1262.

(352) 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* **2005**, *44* (6), 2048-2058. DOI: 10.1021/bi048242p.

(353) You, H.; Wu, J.; Shao, F.; Yan, J. Stability and Kinetics of c-MYC Promoter G-Quadruplexes Studied by Single-Molecule Manipulation. *Journal of the American Chemical Society* **2015**, *137* (7), 2424-2427. DOI: 10.1021/ja511680u.

(354) 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* **2010**, *49* (43), 9152-9160.

(355) Gray, R. D.; Trent, J. O.; Arumugam, S.; Chaires, J. B. Folding Landscape of a Parallel G-Quadruplex. *The journal of physical chemistry letters* **2019**, *10* (5), 1146-1151. DOI: 10.1021/acs.jpclett.9b00227 PubMed.

(356) Koehler, E.; Brown, E.; Haneuse, S. J.-P. On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician* **2009**, *63* (2), 155-162.

(357) Alberty, R. A. Principle of Detailed Balance in Kinetics. *Journal of Chemical Education* **2004**, *81* (8), 1206. DOI: 10.1021/ed081p1206.

(358) Bardin, C.; Leroy, J. L. The formation pathway of tetramolecular G-quadruplexes. *Nucleic Acids Research* **2007**, *36* (2), 477-488. DOI: 10.1093/nar/gkm1050 (acccessed 5/4/2020).

(359) Langlois, V. CHAPTER 2 - Laboratory Evaluation at Different Ages. In *Comprehensive Pediatric Nephrology*, Geary, D. F., Schaefer, F. Eds.; Mosby, 2008; pp 39-54.

(360) Wright, C. F.; Lindorff-Larsen, K.; Randles, L. G.; Clarke, J. Parallel protein-unfolding pathways revealed and mapped. *Nature Structural & Molecular Biology* **2003**, *10* (8), 658-662. DOI: 10.1038/nsb947.

(361) Dobson, C. M.; Šali, A.; Karplus, M. Protein Folding: A Perspective from Theory and Experiment. *Angewandte Chemie International Edition* **1998**, *37* (7), 868-893. DOI: 10.1002/(SICI)1521-3773(19980420)37:7<868::AID-ANIE868>3.0.CO;2-H (acccessed 2020/04/27).

(362) Baldwin, R. L. The nature of protein folding pathways: The classical versus the new view. *Journal of Biomolecular NMR* **1995**, *5* (2), 103-109. DOI: 10.1007/BF00208801.

(363) Wildegger, G.; Kiefhaber, T. Three-state model for lysozyme folding: triangular folding mechanism with an energetically trapped intermediate. *Journal of molecular biology* **1997**, *270* (2), 294-304. DOI: 10.1006/jmbi.1997.1030 PubMed.

(364) Dinner, A. R.; Šali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in Biochemical Sciences* **2000**, *25* (7), 331-339. DOI: 10.1016/S0968-0004(00)01610-8 (acccessed 2020/04/27).

(365) Zaidi, F. N.; Nath, U.; Udgaonkar, J. B. Multiple intermediates and transition states during protein unfolding. *Nature Structural Biology* **1997**, *4* (12), 1016-1024. DOI: 10.1038/nsb1297-1016.

(366) 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* **2015**, *1* (5), 226-233.

(367) Dickerhoff, J.; Onel, B.; Chen, L.; Chen, Y.; Yang, D. Solution structure of a MYC promoter G-quadruplex with 1: 6: 1 loop length. *ACS omega* **2019**, *4* (2), 2533-2539.

(368) Sengupta, P.; Bhattacharya, A.; Sa, G.; Das, T.; Chatterjee, S. Truncated G-quadruplex isomers cross-talk with the transcription factors to maintain homeostatic equilibria in c-MYC transcription. *Biochemistry* **2019**, *58* (15), 1975-1991.

(369) Gonzalez-Pena, V.; Sun, D.; Hurley, L. Differential binding of Nucleolin to G-quadruplex structures. *Cancer Research* **2008**, *68* (9 Supplement), 171.

(370) Tippana, R.; Hwang, H.; Opresko, P. L.; Bohr, V. A.; Myong, S. Single-molecule imaging reveals a common mechanism shared by G-quadruplex–resolving helicases. *Proceedings of the National Academy of Sciences* **2016**, *113* (30), 8448-8453.

(371) Miyoshi, D.; Karimata, H.; Sugimoto, N. Hydration Regulates Thermodynamics of G-Quadruplex Formation under Molecular Crowding Conditions. *Journal of the American Chemical Society* **2006**, *128* (24), 7957-7963. DOI: 10.1021/ja061267m.
(372) Sun, D.; Hurley, L. H. The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. *Journal of medicinal chemistry* **2009**, *52* (9), 2863-2874. DOI: 10.1021/jm900055s PubMed.
(373) 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 Research* **2008**, *36* (12), 4191-4203. DOI: 10.1093/nar/gkn379 (acccessed 4/28/2020).
(374) Cantor, C. R.; Warshaw, M. M.; Shapiro, H. Oligonucleotide interactions. III. Circular dichroism studies of the conformation of deoxyoligonucleolides. *Biopolymers* **1970**, *9* (9), 1059-1077. DOI: 10.1002/bip.1970.360090909 (acccessed 2020/04/27).
(375) Box G. E. P., H., W. G., Hunter, J. S. . *Statistics for experimenters: an introduction to design, data analysis, and model building.* ; Wiley, 1978.
(376) Tellinghuisen, J. Statistical Error Propagation. *The Journal of Physical Chemistry A* **2001**, *105* (15), 3917-3921. DOI: 10.1021/jp003484u.
(377) Tian, T.; Chen, Y.-Q.; Wang, S.-R.; Zhou, X. G-Quadruplex: a regulator of gene expression and its chemical targeting. *Chem* **2018**, *4* (6), 1314-1344.
(378) Huppert, J. L.; Bugaut, A.; Kumari, S.; Balasubramanian, S. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Research* **2008**, *36* (19), 6260-6268. DOI: 10.1093/nar/gkn511 (acccessed 11/28/2022).
(379) Hänsel-Hertsch, R.; Di Antonio, M.; Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nature Reviews Molecular Cell Biology* **2017**, *18* (5), 279-284. DOI: 10.1038/nrm.2017.3.
(380) Nishio, M.; Tsukakoshi, K.; Ikebukuro, K. G-quadruplex: Flexible conformational changes by cations, pH, crowding and its applications to biosensing. *Biosensors and Bioelectronics* **2021**, *178*, 113030. DOI: https://doi.org/10.1016/j.bios.2021.113030.
(381) Mathad, R. I.; Hatzakis, E.; Dai, J.; Yang, D. c-MYC promoter G-quadruplex formed at the 5′ -end of NHE III 1 element: insights into biological relevance and parallel-stranded G-quadruplex stability. *Nucleic Acids Research* **2011**, *39* (20), 9023-9033. DOI: 10.1093/nar/gkr612 (acccessed 4/22/2022).
(382) Trajkovski, M.; Endoh, T.; Tateishi-Karimata, H.; Ohyama, T.; Tanaka, S.; Plavec, J.; Sugimoto, N. Pursuing origins of (poly)ethylene glycol-induced G-quadruplex structural modulations. *Nucleic Acids Research* **2018**, *46* (8), 4301-4315. DOI: 10.1093/nar/gky250 (acccessed 4/22/2022).
(383) Lim, K. W.; Lacroix, L.; Yue, D. J. E.; Lim, J. K. C.; Lim, J. M. W.; Phan, A. T. Coexistence of Two Distinct G-Quadruplex Conformations in the hTERT Promoter. *Journal of the American Chemical Society* **2010**, *132* (35), 12331-12342. DOI: 10.1021/ja101252n.
(384) Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. Coexistence of two main folded G-quadruplexes within a single G-rich domain in the EGFR promoter. *Nucleic acids research* **2017**, *45* (17), 10132-10142. DOI: 10.1093/nar/gkx678 PubMed.
(385) 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* **2005**, *44* (6), 2048-2058. DOI: 10.1021/bi048242p.

(386) Yang, D.; Hurley, L. H. Structure of the Biologically Relevant G-Quadruplex in The c-MYC Promoter. *Nucleosides, Nucleotides & Nucleic Acids* **2006**, *25* (8), 951-968. DOI: 10.1080/15257770600809913.

(387) Kerkour, A.; Marquevielle, J.; Ivashchenko, S.; Yatsunyk, L. A.; Mergny, J.-L.; Salgado, G. F. High-resolution three-dimensional NMR structure of the <em>KRAS</em> proto-oncogene promoter reveals key features of a G-quadruplex involved in transcriptional regulation. *Journal of Biological Chemistry* **2017**, *292* (19), 8082-8091. DOI: 10.1074/jbc.M117.781906 (acccessed 2022/04/27).

(388) De Nicola, B.; Lech, C. J.; Heddi, B.; Regmi, S.; Frasson, I.; Perrone, R.; Richter, S. N.; Phan, A. T. Structure and possible function of a G-quadruplex in the long terminal repeat of the proviral HIV-1 genome. *Nucleic Acids Research* **2016**, *44* (13), 6442-6451. DOI: 10.1093/nar/gkw432 (acccessed 4/27/2022).

(389) Phan, A. T.; Modi, Y. S.; Patel, D. J. Propeller-Type Parallel-Stranded G-Quadruplexes in the Human c-myc Promoter. *Journal of the American Chemical Society* **2004**, *126* (28), 8710-8716. DOI: 10.1021/ja048805k.

(390) Ducani, C.; Bernardinelli, G.; Högberg, B.; Keppler, B. K.; Terenzi, A. Interplay of Three G-Quadruplex Units in the KIT Promoter. *Journal of the American Chemical Society* **2019**, *141* (26), 10205-10213. DOI: 10.1021/jacs.8b12753.

(391) Amrane, S.; Adrian, M.; Heddi, B.; Serero, A.; Nicolas, A.; Mergny, J.-L.; Phan, A. T. n. Formation of pearl-necklace monomorphic G-quadruplexes in the human CEB25 minisatellite. *Journal of the American Chemical Society* **2012**, *134* (13), 5807-5816.

(392) Adrian, M.; Ang, D. J.; Lech, C. J.; Heddi, B.; Nicolas, A.; Phan, A. T. Structure and Conformational Dynamics of a Stacked Dimeric G-Quadruplex Formed by the Human CEB1 Minisatellite. *Journal of the American Chemical Society* **2014**, *136* (17), 6297-6305. DOI: 10.1021/ja4125274.

(393) Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Research* **2006**, *34* (19), 5402-5415. DOI: 10.1093/nar/gkl655 (acccessed 4/22/2022).

(394) Schnarr, L.; Jana, J.; Preckwinkel, P.; Weisz, K. Impact of a Snap-Back Loop on Stability and Ligand Binding to a Parallel G-Quadruplex. *The Journal of Physical Chemistry B* **2020**, *124* (14), 2778-2787. DOI: 10.1021/acs.jpcb.0c00700.

(395) Amrane, S.; Kerkour, A.; Bedrat, A.; Vialet, B.; Andreola, M.-L.; Mergny, J.-L. Topology of a DNA G-Quadruplex Structure Formed in the HIV-1 Promoter: A Potential Target for Anti-HIV Drug Development. *Journal of the American Chemical Society* **2014**, *136* (14), 5249-5252. DOI: 10.1021/ja501500c.

(396) Ngoc Nguyen, T. Q.; Lim, K. W.; Phan, A. T. Duplex formation in a G-quadruplex bulge. *Nucleic Acids Research* **2020**, *48* (18), 10567-10575. DOI: 10.1093/nar/gkaa738 (acccessed 4/22/2022).

(397) Onel, B.; Carver, M.; Wu, G.; Timonina, D.; Kalarn, S.; Larriva, M.; Yang, D. A new G-quadruplex with hairpin loop immediately upstream of the human BCL2 P1 promoter modulates transcription. *Journal of the American Chemical Society* **2016**, *138* (8), 2563-2570.

(398) Ghimire, C.; Park, S.; Iida, K.; Yangyuoru, P.; Otomo, H.; Yu, Z.; Nagasawa, K.; Sugiyama, H.; Mao, H. Direct Quantification of Loop Interaction and π–π Stacking for G-

Quadruplex Stability at the Submolecular Level. *Journal of the American Chemical Society* **2014**, *136* (44), 15537-15544. DOI: 10.1021/ja503585h.

(399) Martadinata, H.; Phan, A. T. n. Structure of human telomeric RNA (TERRA): stacking of two G-quadruplex blocks in K+ solution. *Biochemistry* **2013**, *52* (13), 2176-2183.

(400) Jana, J.; Mohr, S.; Vianney, Y. M.; Weisz, K. Structural motifs and intramolecular interactions in non-canonical G-quadruplexes. *RSC Chemical Biology* **2021**, *2* (2), 338-353.

(401) Rigo, R.; Groaz, E.; Sissi, C. Polymorphic and Higher-Order G-Quadruplexes as Possible Transcription Regulators: Novel Perspectives for Future Anticancer Therapeutic Applications. *Pharmaceuticals* **2022**, *15* (3), 373.

(402) Monsen, R. C.; DeLeeuw, L. W.; Dean, William L.; Gray, Robert D.; Chakravarthy, S.; Hopkins, Jesse B.; Chaires, Jonathan B.; Trent, John O. Long promoter sequences form higher-order G-quadruplexes: an integrative structural biology study of c-Myc, k-Ras and c-Kit promoter sequences. *Nucleic Acids Research* **2022**, *50* (7), 4127-4147. DOI: 10.1093/nar/gkac182 (acccessed 4/22/2022).

(403) Farhath, M. M.; Thompson, M.; Ray, S.; Sewell, A.; Balci, H.; Basu, S. G-Quadruplex-Enabling Sequence within the Human Tyrosine Hydroxylase Promoter Differentially Regulates Transcription. *Biochemistry* **2015**, *54* (36), 5533-5545. DOI: 10.1021/acs.biochem.5b00209.

(404) Mendoza, O.; Bourdoncle, A.; Boulé, J.-B.; Brosh, R. M., Jr; Mergny, J.-L. G-quadruplexes and helicases. *Nucleic Acids Research* **2016**, *44* (5), 1989-2006. DOI: 10.1093/nar/gkw079 (acccessed 4/22/2022).

(405) Palumbo, S. L.; Ebbinghaus, S. W.; Hurley, L. H. Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. *Journal of the American Chemical Society* **2009**, *131* (31), 10878-10891.

(406) Monsen, R. C.; DeLeeuw, L.; Dean, W. L.; Gray, R. D.; Sabo, T. M.; Chakravarthy, S.; Chaires, J. B.; Trent, J. O. The hTERT core promoter forms three parallel G-quadruplexes. *Nucleic acids research* **2020**, *48* (10), 5720-5734.

(407) Berselli, M.; Lavezzo, E.; Toppo, S. QPARSE: searching for long-looped or multimeric G-quadruplexes potentially distinctive and druggable. *Bioinformatics* **2020**, *36* (2), 393-399.

(408) Adrian, M.; Heddi, B.; Phan, A. T. NMR spectroscopy of G-quadruplexes. *Methods* **2012**, *57* (1), 11-24.

(409) Li, J.; Correia, J. J.; Wang, L.; Trent, J. O.; Chaires, J. B. Not so crystal clear: the structure of the human telomere G-quadruplex in solution differs from that present in a crystal. *Nucleic acids research* **2005**, *33* (14), 4649-4659.

(410) Matsugami, A.; Ouhashi, K.; Kanagawa, M.; Liu, H.; Kanagawa, S.; Uesugi, S.; Katahira, M. 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. *Journal of molecular biology* **2001**, *313* (2), 255-269.

(411) Guedin, A.; Gros, J.; Alberti, P.; Mergny, J.-L. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic acids research* **2010**, *38* (21), 7858-7868.

(412) Fersht, A. R. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus

mechanism. *Proceedings of the National Academy of Sciences* **2000**, *97* (4), 1525-1529. DOI: doi:10.1073/pnas.97.4.1525.
(413) Dreos, R.; Ambrosini, G.; Groux, R.; Cavin Périer, R.; Bucher, P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research* **2017**, *45* (D1), D51-D55.
(414) Qin, Y.; Hurley, L. H. Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. *Biochimie* **2008**, *90* (8), 1149-1171.
(415) Weidner, N.; Folkman, J.; Pozza, F.; Bevilacqua, P.; Allred, E. N.; Moore, D. H.; Meli, S.; Gasparini, G. Tumor angiogenesis: a new significant and independent prognostic indicator in early-stage breast carcinoma. *JNCI: Journal of the National Cancer Institute* **1992**, *84* (24), 1875-1887.
(416) Ferrara, N. Vascular endothelial growth factor. *European Journal of Cancer* **1996**, *32* (14), 2413-2422.
(417) Sun, D.; Guo, K.; Shin, Y.-J. Evidence of the formation of G-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene. *Nucleic acids research* **2011**, *39* (4), 1256-1265.
(418) 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 research* **2013**, *41* (22), 10584-10592.
(419) Finkenzeller, G.; Sparacio, A.; Technau, A.; Marmé, D.; Siemeister, G. Sp1 recognition sites in the proximal promoter of the human vascular endothelial growth factor gene are essential for platelet-derived growth factor-induced gene expression. *Oncogene* **1997**, *15* (6), 669-676.
(420) Raiber, E.-A.; Kranaster, R.; Lam, E.; Nikan, M.; Balasubramanian, S. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic acids research* **2012**, *40* (4), 1499-1508.
(421) Sun, D.; Liu, W.-J.; Guo, K.; Rusche, J. J.; Ebbinghaus, S.; Gokhale, V.; Hurley, L. H. The proximal promoter region of the human vascular endothelial growth factor gene has a G-quadruplex structure that can be targeted by G-quadruplex–interactive agents. *Molecular cancer therapeutics* **2008**, *7* (4), 880-889.
(422) 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 Research* **2006**, *34* (18), 5133-5144. DOI: 10.1093/nar/gkl610 (acccessed 4/22/2022).
(423) Heckman, C.; Mochon, E.; Arcinas, M.; Boxer, L. M. The WT1 Protein Is a Negative Regulator of the Normalbcl-2 Allele in t (14; 18) Lymphomas. *Journal of Biological Chemistry* **1997**, *272* (31), 19609-19614.
(424) 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. *Journal of the American Chemical Society* **2014**, *136* (5), 1750-1753.
(425) Ashman, L. K.; Griffith, R. Therapeutic targeting of c-KIT in cancer. *Expert opinion on investigational drugs* **2013**, *22* (1), 103-115.
(426) Rankin, S.; Reszka, A. P.; Huppert, J.; Zloh, M.; Parkinson, G. N.; Todd, A. K.; Ladame, S.; Balasubramanian, S.; Neidle, S. Putative DNA quadruplex formation within

the human c-kit oncogene. *Journal of the American Chemical Society* **2005**, *127* (30), 10584-10589.
(427) McLuckie, K. I.; Waller, Z. A.; Sanders, D. A.; Alves, D.; Rodriguez, R.; Dash, J.; McKenzie, G. J.; Venkitaraman, A. R.; Balasubramanian, S. G-quadruplex-binding benzo [a] phenoxazines down-regulate c-KIT expression in human gastric carcinoma cells. *Journal of the American Chemical Society* **2011**, *133* (8), 2658-2663.
(428) Kuryavyi, V.; Phan, A. T.; Patel, D. J. Solution structures of all parallel-stranded monomeric and dimeric G-quadruplex scaffolds of the human c-kit2 promoter. *Nucleic Acids Research* **2010**, *38* (19), 6757-6773. DOI: 10.1093/nar/gkq558 (acccessed 4/27/2022).
(429) Cogoi, S.; Ferino, A.; Miglietta, G.; Pedersen, E. B.; Xodo, L. E. The regulatory G4 motif of the Kirsten ras (KRAS) gene is sensitive to guanine oxidation: implications on transcription. *Nucleic acids research* **2018**, *46* (2), 661-676.
(430) Kaiser, C. E.; Van Ert, N. A.; Agrawal, P.; Chawla, R.; Yang, D.; Hurley, L. H. Insight into the complexity of the i-motif and G-quadruplex DNA structures formed in the KRAS promoter and subsequent drug-induced gene repression. *Journal of the American Chemical Society* **2017**, *139* (25), 8522-8536.
(431) Morgan, R. K.; Batra, H.; Gaerig, V. C.; Hockings, J.; Brooks, T. A. Identification and characterization of a new G-quadruplex forming region within the kRAS promoter as a transcriptional regulator. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **2016**, *1859* (2), 235-245.
(432) Xodo, L. E. Quadruplex nucleic acids in KRAS targeted-cancer therapy. In *Annual Reports in Medicinal Chemistry*, Vol. 54; Elsevier, 2020; pp 325-359.
(433) Jeffreys, A. J.; Wilson, V.; Thein, S. L. Hypervariable ‘minisatellite’regions in human DNA. *Nature* **1985**, *314* (6006), 67-73.
(434) Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **1999**, *27* (2), 573-580.
(435) Adair, D.; Worsham, P.; Hill, K.; Klevytska, A.; Jackson, P.; Friedlander, A.; Keim, P. Diversity in a variable-number tandem repeat from Yersinia pestis. *Journal of clinical microbiology* **2000**, *38* (4), 1516-1519.
(436) Maizels, N. G4 ‐ associated human diseases. *EMBO reports* **2015**, *16* (8), 910-922.
(437) Piazza, A.; Adrian, M.; Samazan, F.; Heddi, B.; Hamon, F.; Serero, A.; Lopes, J.; Teulade-Fichou, M.-P.; Phan, A. T.; Nicolas, A. Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. *The EMBO Journal* **2015**, *34* (12), 1718-1734. DOI: https://doi.org/10.15252/embj.201490702.
(438) Piazza, A.; Cui, X.; Adrian, M.; Samazan, F.; Heddi, B.; Phan, A.-T.; Nicolas, A. G. Non-Canonical G-quadruplexes cause the hCEB1 minisatellite instability in Saccharomyces cerevisiae. *Elife* **2017**, *6*, e26884.
(439) Kotur, N.; Stankovic, B.; Kassela, K.; Georgitsi, M.; Vicha, A.; Leontari, I.; Dokmanovic, L.; Janic, D.; Krstovski, N.; Klaassen, K. 6-mercaptopurine influences TPMT gene transcription in a TPMT gene promoter variable number of tandem repeats-dependent manner. *Pharmacogenomics* **2012**, *13* (3), 283-295.
(440) Catasti, P.; Chen, X.; Moyzis, R. K.; Bradbury, E. M.; Gupta, G. Structure–function correlations of the insulin-linked polymorphic region. *Journal of molecular biology* **1996**, *264* (3), 534-545.

(441) Yoon, S.-L.; Roh, Y.-G.; Lee, S.-H.; Kim, S.-H.; Kim, M. C.; Kim, S. J.; Leem, S.-H. Analysis of Promoter Methylation and Polymorphic Minisatellites of BORIS and Lack of Association with Gastric Cancer. *DNA and Cell Biology* **2011**, *30* (9), 691-698. DOI: 10.1089/dna.2011.1248 (acccessed 2022/04/23).

(442) Bhavsar, P. K.; Brand, N. J.; Yacoub, M. H.; Barton, P. J. Isolation and characterization of the human cardiac troponin I gene (TNNI3). *Genomics* **1996**, *35* (1), 11-23.

(443) 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 Research* **2017**, *45* (6), 2951-2959. DOI: 10.1093/nar/gkx090 (acccessed 4/23/2022).

(444) Lan, C.; Tang, H.; Liu, S.; Ma, L.; Li, J.; Wang, X.; Hou, Y. Comprehensive analysis of prognostic value and immune infiltration of calpains in pancreatic cancer. *Journal of Gastrointestinal Oncology* **2021**, *12* (6), 2600.

(445) Ocejo ‐ Garcia, M.; Baokbah, T. A.; Louise Ashurst, H.; Cowlishaw, D.; Soomro, I.; Coulson, J. M.; Woll, P. J. Roles for USF ‐ 2 in lung cancer proliferation and bronchial carcinogenesis. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **2005**, *206* (2), 151-159.

(446) Tan, Y.; Chen, Y.; Du, M.; Peng, Z.; Xie, P. USF2 inhibits the transcriptional activity of Smurf1 and Smurf2 to promote breast cancer tumorigenesis. *Cellular Signalling* **2019**, *53*, 49-58.

(447) Yang, S.; Liang, Y.; Qian, H.; Li, Q. TTLL12 expression in ovarian cancer correlates with a poor outcome. *International Journal of Clinical and Experimental Pathology* **2020**, *13* (2), 239.

(448) Kaikkonen, E.; Rantapero, T.; Zhang, Q.; Taimen, P.; Laitinen, V.; Kallajoki, M.; Jambulingam, D.; Ettala, O.; Knaapila, J.; Boström, P. J. ANO7 is associated with aggressive prostate cancer. *International journal of cancer* **2018**, *143* (10), 2479-2487.

(449) Marx, A.; Koopmann, L.; Höflmayer, D.; Büscheck, F.; Hube-Magg, C.; Steurer, S.; Eichenauer, T.; Clauditz, T. S.; Wilczak, W.; Simon, R. Reduced anoctamin 7 (ANO7) expression is a strong and independent predictor of poor prognosis in prostate cancer. *Cancer Biology & Medicine* **2021**, *18* (1), 245.

(450) Diefenbach, A.; Raulet, D. H. The innate immune response to tumors and its role in the induction of T ‐ cell immunity. *Immunological reviews* **2002**, *188* (1), 9-21.

(451) Zhou, J.; Liu, M.; Fleming, A. M.; Burrows, C. J.; Wallace, S. S. Neil3 and NEIL1 DNA glycosylases remove oxidative damages from quadruplex DNA and exhibit preferences for lesions in the telomeric sequence context. *Journal of Biological Chemistry* **2013**, *288* (38), 27263-27272.

(452) Zhou, J.; Fleming, A. M.; Averill, A. M.; Burrows, C. J.; Wallace, S. S. The NEIL glycosylases remove oxidized guanine lesions from telomeric and promoter quadruplex DNA structures. *Nucleic acids research* **2015**, *43* (8), 4039-4054.

(453) Oganesian, L.; Bryan, T. M. Physiological relevance of telomeric G ‐ quadruplex formation: a potential drug target. *Bioessays* **2007**, *29* (2), 155-165.

(454) Busseron, E.; Ruff, Y.; Moulin, E.; Giuseppone, N. Supramolecular self-assemblies as functional nanomaterials. *Nanoscale* **2013**, *5* (16), 7098-7140.

(455) Elemans, J. A.; Rowan, A. E.; Nolte, R. J. Mastering molecular matter. Supramolecular architectures by hierarchical self-assembly. *Journal of Materials Chemistry* **2003**, *13* (11), 2661-2670.

(456) Corbett, P. T.; Leclaire, J.; Vial, L.; West, K. R.; Wietor, J.-L.; Sanders, J. K. M.; Otto, S. Dynamic Combinatorial Chemistry. *Chemical Reviews* **2006**, *106* (9), 3652-3711. DOI: 10.1021/cr020452p.

(457) Medrano, M.; Fuertes, M. A. n.; Valbuena, A.; Carrillo, P. J.; Rodríguez-Huete, A.; Mateu, M. G. Imaging and quantitation of a succession of transient intermediates reveal the reversible self-assembly pathway of a simple icosahedral virus capsid. *Journal of the American Chemical Society* **2016**, *138* (47), 15385-15396.

(458) Pinotsi, D.; Buell, A. K.; Galvagnion, C.; Dobson, C. M.; Kaminski Schierle, G. S.; Kaminski, C. F. Direct Observation of Heterogeneous Amyloid Fibril Growth Kinetics via Two-Color Super-Resolution Microscopy. *Nano Letters* **2014**, *14* (1), 339-345. DOI: 10.1021/nl4041093.

(459) 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* **2017**, *56* (5), 692-703.

(460) Bellot, M.; Bouteiller, L. Thermodynamic description of bis-urea self-assembly: competition between two supramolecular polymers. *Langmuir* **2008**, *24* (24), 14176-14182.

(461) Pasternack, R. F.; Goldsmith, J. I.; Szép, S.; Gibbs, E. J. A spectroscopic and thermodynamic study of porphyrin/DNA supramolecular assemblies. *Biophysical journal* **1998**, *75* (2), 1024-1031.

(462) De Greef, T. F.; Smulders, M. M.; Wolffs, M.; Schenning, A. P.; Sijbesma, R. P.; Meijer, E. Supramolecular polymerization. *Chemical Reviews* **2009**, *109* (11), 5687-5754.

(463) Greciano, E. E.; Alsina, S.; Ghosh, G.; Fernández, G.; Sánchez, L. Alkyl Bridge Length to Bias the Kinetics and Stability of Consecutive Supramolecular Polymerizations. *Small Methods* **2020**, *4* (2), 1900715. DOI: https://doi.org/10.1002/smtd.201900715.

(464) Osypenko, A.; Moulin, E.; Gavat, O.; Fuks, G.; Maaloum, M.; Koenis, M. A. J.; Buma, W. J.; Giuseppone, N. Temperature Control of Sequential Nucleation–Growth Mechanisms in Hierarchical Supramolecular Polymers. *Chemistry – A European Journal* **2019**, *25* (56), 13008-13016. DOI: https://doi.org/10.1002/chem.201902898.

(465) Singh, S.; Zlotnick, A. J. J. o. B. C. Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly. **2003**, *278* (20), 18249-18255.

(466) Sambe, L.; de La Rosa, V. R.; Belal, K.; Stoffelbach, F.; Lyskawa, J.; Delattre, F.; Bria, M.; Cooke, G.; Hoogenboom, R.; Woisel, P. Programmable Polymer-Based Supramolecular Temperature Sensor with a Memory Function. *Angewandte Chemie International Edition* **2014**, *53* (20), 5044-5048. DOI: 10.1002/anie.201402108.

(467) Mizuno, K.; Boudko, S. P.; Engel, J.; Bächinger, H. P. Kinetic hysteresis in collagen folding. *Biophysical journal* **2010**, *98* (12), 3004-3014.

(468) Dastan, A.; Frith, W. J.; Cleaver, D. J. Thermal Hysteresis and Seeding of Twisted Fibers Formed by Achiral Discotic Particles. *The Journal of Physical Chemistry B* **2017**, *121* (42), 9920-9928. DOI: 10.1021/acs.jpcb.7b05316.

(469) Yamaguchi, M. Thermal Hysteresis Involving Reversible Self-Catalytic Reactions. *Accounts of Chemical Research* **2021**, 10226-10234.

(470) Fukushima, T.; Tamaki, K.; Isobe, A.; Hirose, T.; Shimizu, N.; Takagi, H.; Haruki, R.; Adachi, S.-i.; Hollamby, M. J.; Yagai, S. Diarylethene-Powered Light-Induced Folding of Supramolecular Polymers. *Journal of the American Chemical Society* **2021**, *143* (15), 5845-5854. DOI: 10.1021/jacs.1c00592.

(471) Kar, H.; Ghosh, G.; Ghosh, S. Solvent Geometry Regulated Cooperative Supramolecular Polymerization. *Chemistry – A European Journal* **2017**, *23* (44), 10536-10542. DOI: https://doi.org/10.1002/chem.201701299.

(472) Fernández, Z.; Fernández, B.; Quiñoá, E.; Freire, F. The Competitive Aggregation Pathway of an Asymmetric Chiral Oligo(p-phenyleneethynylene) Towards the Formation of Individual P and M Supramolecular Helical Polymers. *Angewandte Chemie International Edition* **2021**, *60* (18), 9919-9924. DOI: https://doi.org/10.1002/anie.202100162.

(473) Xu, F.; Pfeifer, L.; Crespi, S.; Leung, F. K.-C.; Stuart, M. C. A.; Wezenberg, S. J.; Feringa, B. L. From Photoinduced Supramolecular Polymerization to Responsive Organogels. *Journal of the American Chemical Society* **2021**, *143* (15), 5990-5997. DOI: 10.1021/jacs.1c01802.

(474) Sarkar, S.; Sarkar, A.; George, S. J. Stereoselective Seed-Induced Living Supramolecular Polymerization. *Angewandte Chemie International Edition* **2020**, *59* (45), 19841-19845. DOI: https://doi.org/10.1002/anie.202006248.

(475) Haedler, A. T.; Meskers, S. C. J.; Zha, R. H.; Kivala, M.; Schmidt, H.-W.; Meijer, E. W. Pathway Complexity in the Enantioselective Self-Assembly of Functional Carbonyl-Bridged Triarylamine Trisamides. *Journal of the American Chemical Society* **2016**, *138* (33), 10539-10545. DOI: 10.1021/jacs.6b05184.

(476) Ogi, S.; Stepanenko, V.; Sugiyasu, K.; Takeuchi, M.; Würthner, F. Mechanism of Self-Assembly Process and Seeded Supramolecular Polymerization of Perylene Bisimide Organogelator. *Journal of the American Chemical Society* **2015**, *137* (9), 3300-3307. DOI: 10.1021/ja511952c.

(477) Wang, H.; Zhang, Y.; Chen, Y.; Pan, H.; Ren, X.; Chen, Z. Living Supramolecular Polymerization of an Aza‐BODIPY Dye Controlled by a Hydrogen‐Bond‐Accepting Triazole Unit Introduced by Click Chemistry. *Angewandte Chemie* **2020**, *132* (13), 5223-5230.

(478) Mukhopadhyay, R. D.; Ajayaghosh, A. Living supramolecular polymerization. *Science* **2015**, *349* (6245), 241-242. DOI: doi:10.1126/science.aac7422.

(479) Jonkheijm, P.; van der Schoot, P.; Schenning, A. P.; Meijer, E. Probing the solvent-assisted nucleation pathway in chemical self-assembly. *Science* **2006**, *313* (5783), 80-83.

(480) van der Schoot, P. Supramolecular polymers. *Taylor & Francis Group, London* **2005**.

(481) Arisawa, M.; Iwamoto, R.; Yamaguchi, M. Unstable and Stable Thermal Hysteresis Under Thermal Triangle Waves. *ChemistrySelect* **2021**, *6* (18), 4461-4465.

(482) Efron, B.; Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science* **1986**, 54-75.

(483) Markham, N. R.; Zuker, M. DINAMelt web server for nucleic acid melting prediction. *Nucleic acids research* **2005**, *33* (suppl_2), W577-W581.

(484) Mikulecky, P. J.; Feig, A. L. Heat capacity changes associated with nucleic acid folding. **2006**, *82* (1), 38-58. DOI: 10.1002/bip.20457.

(485) Self-Processes — Programmed Supramolecular Systems. In *Supramolecular Chemistry*, 1995; pp 139-197.

(486) Hamacek, J. J. M. Self-Assembly Principles of Helicates. Wiley Online Library: 2013; pp 91-124.

(487) Paneerselvam, A. P.; Mishra, S. S.; Chand, D. K. J. J. o. C. S. Linear and circular helicates: A brief review. **2018**, *130* (7), 96.

(488) Hamacek, J.; Borkovec, M.; Piguet, C. J. C. A. E. J. A Simple Thermodynamic Model for Quantitatively Addressing Cooperativity in Multicomponent Self ‐ Assembly Processes—Part 1: Theoretical Concepts and Application to Monometallic Coordination Complexes and Bimetallic Helicates Possessing Identical Binding Sites. **2005**, *11* (18), 5217-5226.

(489) Dill, K. A.; Bromberg, S.; Stigter, D. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*; Garland Science, 2010.

(490) Lehn, J.-M.; Rigault, A.; Siegel, J.; Harrowfield, J.; Chevrier, B.; Moras, D. Spontaneous assembly of double-stranded helicates from oligobipyridine ligands and copper (I) cations: structure of an inorganic double helix. *Proceedings of the National Academy of Sciences* **1987**, *84* (9), 2565-2569.

(491) Mulder, A.; Huskens, J.; Reinhoudt, D. N. Multivalency in supramolecular chemistry and nanofabrication. *Organic & biomolecular chemistry* **2004**, *2* (23), 3409-3424.

(492) Lin, M.; Dai, Y.; Xia, F.; Zhang, X. Advances in non-covalent crosslinked polymer micelles for biomedical applications. *Materials Science and Engineering: C* **2020**, 111626.

(493) van der Weegen, R.; Teunissen, A. J.; Meijer, E. J. C. A. E. J. Directing the Self ‐ Assembly Behaviour of Porphyrin ‐ Based Supramolecular Systems. **2017**, *23* (15), 3773-3783.

(494) Ghosh, A. K.; Samanta, I.; Mondal, A.; Liu, W. R. Covalent inhibition in drug discovery. *ChemMedChem* **2019**, *14* (9), 889-906.

(495) Bauer, R. A. Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies. *Drug Discovery Today* **2015**, *20* (9), 1061-1073. DOI: https://doi.org/10.1016/j.drudis.2015.05.005.

(496) Kitz, R.; Wilson, I. B. Esters of methanesulfonic acid as irreversible inhibitors of acetylcholinesterase. *Journal of Biological Chemistry* **1962**, *237* (10), 3245-3249.

(497) Krippendorff, B.-F.; Neuhaus, R.; Lienau, P.; Reichel, A.; Huisinga, W. Mechanism-based inhibition: deriving KI and kinact directly from time-dependent IC50 values. *SLAS Discovery* **2009**, *14* (8), 913-923.

(498) Dueñas, M. E.; Peltier ‐ Heap, R. E.; Leveridge, M.; Annan, R. S.; Büttner, F. H.; Trost, M. Advances in high ‐ throughput mass spectrometry in drug discovery. *EMBO Molecular Medicine* **2023**, *15* (1), e14850.

(499) Wang, G.; Moitessier, N.; Mittermaier, A. K. Computational and biophysical methods for the discovery and optimization of covalent drugs. *Chemical Communications* **2023**, *59* (73), 10866-10882, 10.1039/D3CC03285J. DOI: 10.1039/D3CC03285J.

(500) Wang, Y.; Wang, G.; Moitessier, N.; Mittermaier, A. K. Enzyme kinetics by isothermal titration calorimetry: allostery, inhibition, and dynamics. *Frontiers in Molecular Biosciences* **2020**, *7*, 583826.

(501) Olsen, S. N. Applications of isothermal titration calorimetry to measure enzyme kinetics and activity in complex solutions. *Thermochimica Acta* **2006**, *448* (1), 12-18.

(502) Stille, J. K.; Tjutrins, J.; Wang, G.; Venegas, F. A.; Hennecker, C.; Rueda, A. M.; Sharon, I.; Blaine, N.; Miron, C. E.; Pinus, S. Design, synthesis and in vitro evaluation of novel SARS-CoV-2 3CLpro covalent inhibitors. *European Journal of Medicinal Chemistry* **2022**, *229*, 114046.

(503) Owen, D. R.; Allerton, C. M.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J. An oral SARS-CoV-2 Mpro inhibitor clinical candidate for the treatment of COVID-19. *Science* **2021**, *374* (6575), 1586-1593.

(504) Unoh, Y.; Uehara, S.; Nakahara, K.; Nobori, H.; Yamatsu, Y.; Yamamoto, S.; Maruyama, Y.; Taoda, Y.; Kasamatsu, K.; Suto, T. Discovery of S-217622, a noncovalent oral SARS-CoV-2 3CL protease inhibitor clinical candidate for treating COVID-19. *Journal of medicinal chemistry* **2022**, *65* (9), 6499-6512.

(505) John, S. E. S.; Mesecar, A. D. Broad-spectrum non-covalent coronavirus protease inhibitors. Google Patents: 2018.

(506) Tripathi, P. K.; Upadhyay, S.; Singh, M.; Raghavendhar, S.; Bhardwaj, M.; Sharma, P.; Patel, A. K. Screening and evaluation of approved drugs as inhibitors of main protease of SARS-CoV-2. *International Journal of Biological Macromolecules* **2020**, *164*, 2622-2631. DOI: https://doi.org/10.1016/j.ijbiomac.2020.08.166.

(507) Rut, W.; Groborz, K.; Zhang, L.; Sun, X.; Zmudzinski, M.; Pawlik, B.; Wang, X.; Jochmans, D.; Neyts, J.; Młynarski, W.; et al. SARS-CoV-2 Mpro inhibitors and activity-based probes for patient-sample imaging. *Nature Chemical Biology* **2021**, *17* (2), 222-228. DOI: 10.1038/s41589-020-00689-z.

(508) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **2020**, *368* (6489), 409-412. DOI: doi:10.1126/science.abb3405.

(509) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582* (7811), 289-293. DOI: 10.1038/s41586-020-2223-y.

(510) Wang, G.; Venegas, F. A.; Rueda, A. M.; Weerasinghe, N. W.; Uggowitzer, K. A.; Thibodeaux, C. J.; Moitessier, N.; Mittermaier, A. K. A naturally occurring G11S mutation in the 3C-like protease from the SARS-CoV-2 virus dramatically weakens the dimer interface. *Protein Science* **2024**, *33* (1), e4857. DOI: https://doi.org/10.1002/pro.4857.

(511) Dixon, M. The graphical determination of Km and Ki. *Biochemical Journal* **1972**, *129* (1), 197-202.

(512) del Villar-Guerra, R.; Trent, J. O.; Chaires, J. B. G-Quadruplex Secondary Structure Obtained from Circular Dichroism Spectroscopy. *Angewandte Chemie International Edition* **2018**, *57* (24), 7171-7175. DOI: 10.1002/anie.201709184 (acccessed 2020/04/27).

(513) Monsen, R. C.; Trent, J. O.; Chaires, J. B. G-quadruplex DNA: a longer story. *Accounts of Chemical Research* **2022**, *55* (22), 3242-3252.

(514) Wang, X.; Yu, B.; Sakurabayashi, S.; Paz-Villatoro, J. M.; Iwahara, J. Robust Enzymatic Production of DNA G-Quadruplex, Aptamer, DNAzyme, and Other Oligonucleotides: Applications for NMR. *Journal of the American Chemical Society* **2024**. DOI: 10.1021/jacs.3c11219.

(515) Müller, D.; Bessi, I.; Richter, C.; Schwalbe, H. The Folding Landscapes of Human Telomeric RNA and DNA G‐Quadruplexes are Markedly Different. *Angewandte Chemie* **2021**, *133* (19), 10990-10996.
(516) Mezzasalma, S. A.; Kruse, J.; Ibarra, A. I.; Arbe, A.; Grzelczak, M. Statistical thermodynamics in reversible clustering of gold nanoparticles. A first step towards nanocluster heat engines. *Journal of Colloid and Interface Science* **2022**, *628*, 205-214. DOI: https://doi.org/10.1016/j.jcis.2022.07.037.
(517) Kruse, J.; Merkens, S.; Chuvilin, A.; Grzelczak, M. Kinetic and Thermodynamic Hysteresis in Clustering of Gold Nanoparticles: Implications for Nanotransducers and Information Storage in Dynamic Systems. *ACS Applied Nano Materials* **2020**, *3* (9), 9520-9527.
(518) Yamaguchi, M. Thermal Hysteresis Involving Reversible Self-Catalytic Reactions. *Accounts of Chemical Research* **2021**, *54* (11), 2603-2613. DOI: 10.1021/acs.accounts.1c00090.
(519) Strelow, J. M. A perspective on the kinetics of covalent and irreversible inhibition. *SLAS Discovery: Advancing Life Sciences R&D* **2017**, *22* (1), 3-20.
(520) Huang, H. Y.; Pinus, S.; Zhang, X. C.; Wang, G.; Rueda, A. M.; Souaibou, Y.; Huck, S.; Huot, M.; Vlaho, D.; Pottel, J. Integration of Computational and Experimental Techniques for the Discovery of SARS-CoV-2 PLpro Covalent Inhibitors. **2023**.