ABSTRACT

Title of Dissertation: EXPLORING THE STRUCTURAL DIVERSITY OF DNA BEYOND THE DOUBLE HELIX

Betty Chu, Doctor of Philosophy, 2020

Dissertation directed by: Dr. Paul J. Paukstelis, Associate Professor
Department of Chemistry and Biochemistry

The iconic model of DNA is the Watson-Crick double helix, but it can form other types of structures. The extent of DNA structural diversity is not well understood. We are interested in discovering new types of DNA structure through the crystal screening and structure determination of a library of oligonucleotide sequences. Through this, we aim to identify previously unobserved motifs that may be biologically relevant and investigate the correlation between the sequence and the structures solved from the library. Furthermore, we can better understand the structural diversity of DNA and sample the different types of motifs formed, as well as the frequency of them. Moreover, from a nanotechnological standpoint, determining new DNA motifs can expand the structure space for rational DNA crystal design to create more precise nanostructures targeted for specific applications.

In this dissertation, I will discuss two new crystal structures of single oligonucleotides that interact via noncanonical base pairing. d(CGTAAGGCG) forms

a non-G-quadruplex fold-back structure through both Watson-Crick and noncanonical interactions. The tetrameric assembly encloses a central cation binding pocket and features a hexad base pairing arrangement through two C─G─G base triples. We have also determined three variant sequences that form the same structure, suggesting that there is a large number of potential fold-back sequences in genomes. This is of particular biological relevance since fold-back structures have been observed in promoter regions of developmental genes in humans. d(CCAGGCTGCAA) features a barium-stabilized G-quadruplex, which is flanked on either side by a base triple formed through noncanonical interactions and a peripheral i-motif. This structure suggests the necessity of a spacer region to bridge the geometric differences between the G-quadruplex and i-motif. This is the first structure of a hybrid DNA G-quadruplex/i-motif and demonstrates the possibility of the coexistence of G-quadruplexes and i-motifs in a single strand of DNA in genomes.

The fold-back quadruplex and hybrid G-quadruplex/i-motif highlight the growing structural diversity of DNA and suggest greater biological roles for non-duplex structures. These structures demonstrate that DNA assemblies beyond the traditional double helix exist and suggest that DNA can form even more diverse structures.

EXPLORING THE STRUCTURAL DIVERSITY OF DNA BEYOND THE DOUBLE HELIX

by

Betty Chu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Paul J. Paukstelis, Chair
Professor David Fushman
Professor Douglas Julin
Professor John Orban
Professor Ian White, Dean's Representative

Acknowledgements

I would like to thank Dr. Paul J. Paukstelis for his outstanding guidance throughout the past five years. Paul has been a supportive advisor, an extraordinary scientist, and a compassionate friend; he has been my rock throughout all this time. His door has always been open for me to share my joys in research, to brainstorm new research directions, to provide advice on failed experiments, to lend an ear during difficult times, and to enjoy friendly conversation and banter. My graduate career has been such an edifying experience because of Paul's support, compassion, and encouragement.

I am grateful to past and present members of the Paukstelis Lab, particularly Dr. Diana Zhang, Emily Luteran, and undergraduates Seth Cohen and Logan Robeck for making graduate school an enjoyable experience.

Thank you to Dr. John Orban, Dr. Douglas Julin, and Dr. David Fushman for serving as my dissertation committee and for continually providing advice and encouragement. Thank you to Dr. Ian White for serving as my Dean's Representative.

I am thankful to my collaborators for teaching me new instrumentation and for providing technical support. Thank you to Dr. Daoning Zhang for helping me with the NMR experiments and to Dr. Wonseok Hwang for providing assistance with the SAXS experiments. Thank you to Dr. Jason D. Kahn for allowing to run my experiments on the UV-Vis spectrophotometer at my convenience.

I would like to especially acknowledge Dr. Jeffery T. Davis for his recruitment efforts amidst the snowstorm on Visit Maryland Day in 2015. If not for you, I would

not have decided to pursue my doctoral degree here at UMD. Thank you for keeping in contact and checking in with me throughout my time here.

I am grateful to the rest of the Biochemistry Faculty Members, particularly Dr. Dorothy Beckett, Dr. Kwaku Theodore Dayie, and Dr. Jinwoo Lee, for insightful conversations and valuable advice on research, career paths, and just life in general.

Thank you to Carl Womack and Judith A. Mahaffy for providing administrative support and putting up with all my questions on orders, reimbursements, and paychecks.

I would also like to thank my friends from the program, especially Songjun Xiao, Chenlu He, Nishat Seraj, Owen Becette, and Andrew Boughton, for engaging in productive and enjoyable discussions on research, post-graduate careers, and just random things in general.

Thank you to my parents and sister, Fanny, for supporting me in my journey to pursue my doctoral degree. And last but not least, I would like to thank my significant other, Shiny Yang, for encouraging me and supporting me throughout my dissertation process and pushing me to the finish line.

The work presented in this document is adapted from the following research articles:

- Crystal Structure of a Tetrameric DNA Fold-back Quadruplex
  Betty Chu, Daoning Zhang, Wonseok Hwang, and Paul J. Paukstelis
  Journal of the American Chemical Society 2018

- A DNA G-quadruplex/i-motif Hybrid
  Betty Chu, Daoning Zhang, and Paul J. Paukstelis
  Nucleic Acids Research 2019

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Abbreviations and Symbols

A	Adenine/Adenosine
C	Cytosine/Cytidine
CD	Circular Dichroism
DNA	Deoxyribonucleic acid
FOM	Figure of Merit
FOM, DM	Figure of Merit after Density Modification
G	Guanine/Guanosine
G4	G-quadruplex
I / σ I	Signal-to-noise ratio
MPD	2-methyl-2,4-pentanediol
N	Any nucleobase/nucleotide (A, C, G, T)
NaCaC	Sodium cacodylate
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect Spectroscopy
nt	nucleotide
PEG400	Polyethylene glycol 400
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
SAD	Single-wavelength Anomalous Dispersion
SAXS	Small Angle X-ray Scattering
SIRAS	Single Isomorphous Replacement with Anomalous Scattering
T	Thymine/Thymidine

TOCSY	Total Correlation Spectroscopy
U	Uracil/Uridine
UV	Ultraviolet
Y	pyrimidine (C, T)
Å	Angstroms
α	alpha
χ	chi
χ^2	chi-squared
λ	lambda
σ	sigma
ζ	zeta

Chapter 1: Introduction

Chapter 1.1: The Beginnings of DNA Structure

It has been known since the work of Oswald Avery and colleagues in 1944 that DNA encodes the genetic blueprint of all organisms. However, DNA structure was not determined until about a decade later. In 1953, Rosalind Franklin crystallized nucleic acid fibers from the calf thymus and obtained X-ray diffraction images of the B-form double helical structure,[1] the best of which was known as Photo 51 (Figure 1.1a). At the same time, James Watson and Francis Crick proposed a structural analysis for how the double helix self-duplicates to act as a genetic material,[2, 3] which led to their being credited for the discovery of the B-DNA structure. Their proposal also resulted in the naming of the complementary A─T and G─C hydrogen bonding interactions as Watson-Crick base pairs (Figure 1.1b). This A─T pairing utilizes the A(N1)─T(N3) and A(N6)─T(O4) atoms, while the G─C pairing interacts through the G(N1)─C(N3), G(N2)─C(O2), and G(O6)─C(N4) atoms.

Following the discovery of the right-handed B-form duplex as our genetic blueprint, in 1956, Jerry Donohue began to explore the possibility of other DNA structures.[4] He conducted a systematic study on the geometries of noncanonical base pairs, which are defined as any hydrogen bonding interactions between the four nucleobases that are alternative to the Watson-Crick A─T and G─C base pairs. Although Donohue identified 24 geometrically acceptable base pairs between the four nucleotides, including the non-Watson-Crick A─A, C─C^+, G─G, T─T, G─A, G─T, T─C, and A─C interactions, he expressed doubt that all of these noncanonical pairings

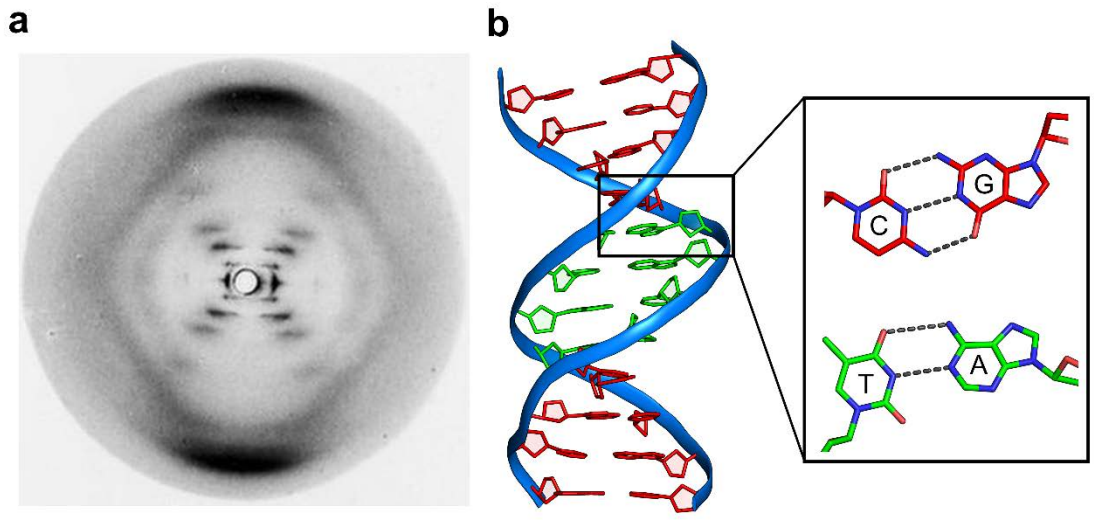

Figure 1.1. B-form DNA. (a) X-ray diffraction image of B-DNA known as Photo 51 collected by Rosalind Franklin. This image was obtained from https://www.bbc.com/news/health-18041884 (Copyright: King's College London). **(b)** Cartoon representation of the Watson-Crick double helix (PDB: 1BNA). The inset shows the hydrogen bonding interactions between the Watson-Crick G─C (red) and A─T (green) base pairs.

could be biologically relevant. Furthermore, in 1963, Karst Hoogsteen observed a new type of base pairing between modified A and T residues interacting through the A(N3)─T(O4) and the A(N7)─T(N3) atoms.[5] This resulted in the naming of the N7 face of purines as the Hoogsteen face (Figure 1.2a). By incorporating additional hydrogen bond donors and acceptors from the Hoogsteen and sugar faces, many more pairings between the four nucleotides are possible. These alternative interactions may provide new geometries that can accommodate different types of DNA structure. Several examples of observed non-traditional base pairs that utilize the additional Hoogsteen and sugar faces are shown in Figure 1.2.

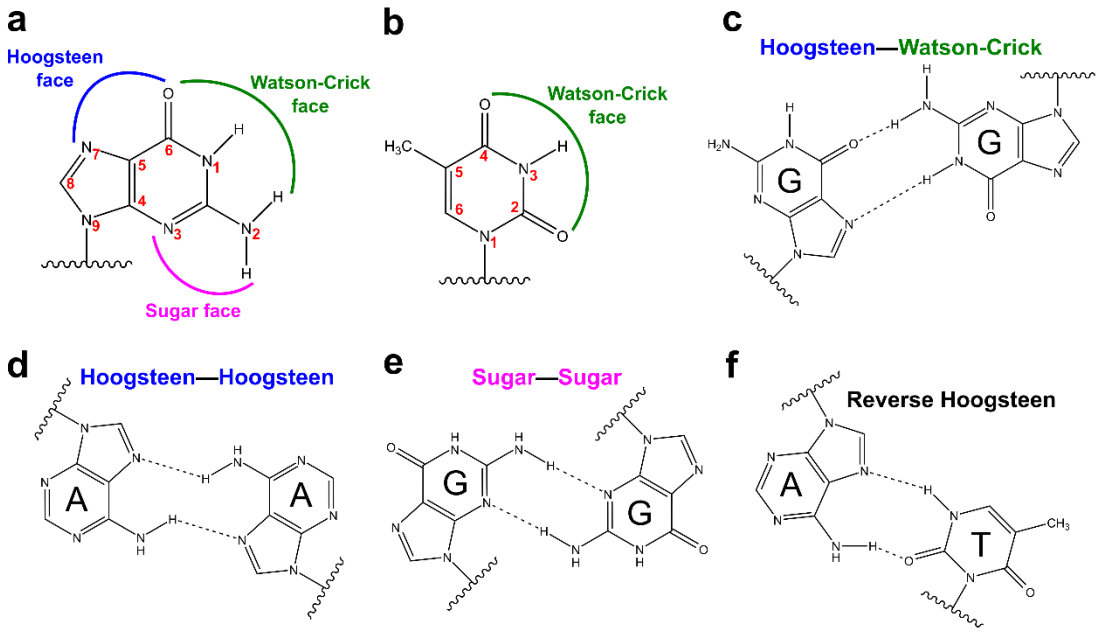

Figure 1.2. Noncanonical Base Pairs. (a-b) The **(a)** purine and **(b)** pyrimidine nucleobases are shown with atom numbers and labeled faces. **(c)** A Hoogsteen─Watson-Crick base pair is shown between two G residues that interact through the O6─N2 and the N7─N1 atoms. **(d)** A symmetric A─A base pair forms between the N6─N7 atoms of the Hoogsteen face. **(e)** A symmetric G─G base pair forms between the N2─N3 atoms of the sugar face. **(f)** A Reverse Hoogsteen A─T base pair interacts through the A(N6)─T(O2) and A(N7)─T(N3) atoms.

Chapter 1.2: DNA structure is more than the double helix

DNA structures incorporating noncanonical base pairing interactions were discovered as early as the 1960s. In 1962, Martin Gellert and colleagues crystallized fibers of guanylic acid salts and observed a unique diffraction pattern.[6] They proposed that four guanosine residues are arranged on the same plane and that each one acts as a donor and acceptor of two hydrogen bonds. Further, they proceeded to hypothesize that each plane is stacked on another and stabilized through van der Waals forces, resulting in an overall cylindrical structure with a hole in the center. Although Gellert and colleagues described the first crystal structure of the G-quadruplex, they did not immediately conduct additional studies on this assembly. Little did they know at the

time that this non-Watson-Crick structure would become one of great significance to biology many years later.

Since the discovery of the G-quadruplex, a wide array of other non-Watson-Crick structural motifs has been reported, including i-motifs,[7] fold-back structures,[8] and parallel-stranded duplexes.[9, 10] In this section, I will give an overview of several types of structures that utilize noncanonical base pairing motifs. This is, by no means, a complete list of noncanonical DNA structures that have been identified to date, but simply a brief overview to demonstrate the breadth of structures that can form beyond the traditional double helix.

Chapter 1.2.1. G-quadruplexes and i-motifs

The G-quadruplex and the i-motif are two noncanonical structures that have been studied extensively, and each is characterized by specific types of noncanonical interactions. G-quadruplexes (G4s) are formed from G-rich sequences and contain stacked guanosine tetrads, each of which is organized in a cyclic hydrogen bonding arrangement between the Hoogsteen and Watson-Crick faces of neighboring nucleobases (Figure 1.3a).[11, 12] G4s can be formed through inter- or intramolecular interactions in a variety of topologies and are stabilized by central cations.[13-15] The DNA i-motif is characterized by the formation of hemiprotonated C─C^+ parallel-stranded base pairs (Figure 1.3b), which are organized to allow two duplexes to intercalate in an antiparallel fashion to form a quadruplex structure.[7, 16] Both G4s and i-motifs can form as unimolecular, bimolecular, or tetramolecular assemblies, leading to diverse folding topologies.[17, 18]

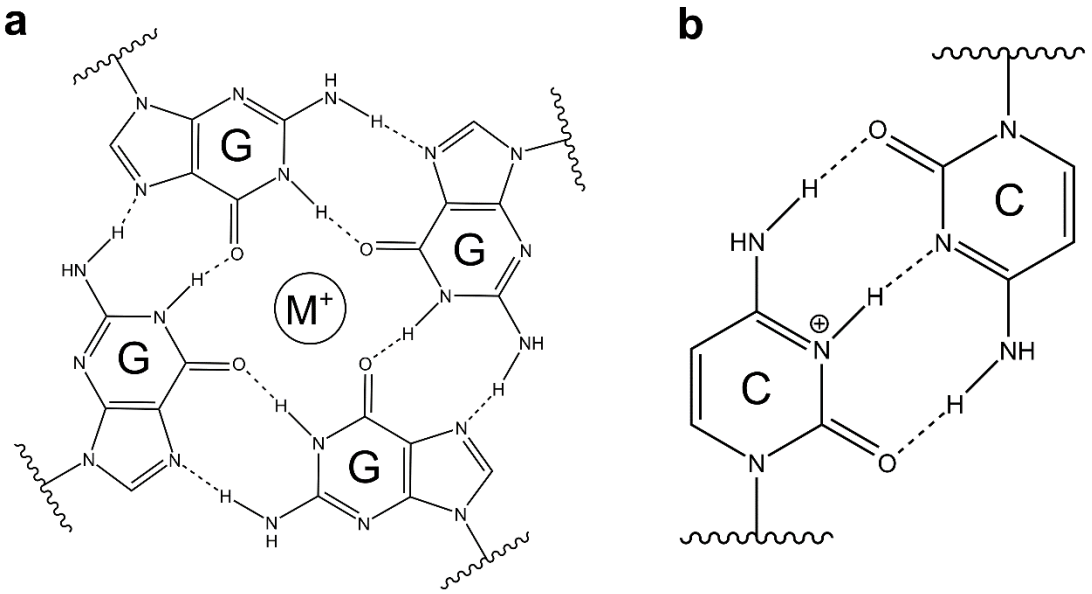

Figure 1.3. Fundamental Unit of the G-quadruplex and i-motif. (a) The G-tetrad is comprised of four G residues that are arranged on the same plane and held together by Hoogsteen─Watson-Crick hydrogen bonds. It is stabilized by a central metal ion, which coordinates with the oxygen atom of each G residue. **(b)** A hemiprotonated homocytosine base pair is stabilized by three hydrogen bonds through the N3─N3 and symmetric O2─N4 atoms.

Although G4 and i-motif structures tend to form from sequences that contain contiguous stretches of G's or C's, respectively, structural characterization has revealed a relatively wide distribution of sequences capable of forming these and similar noncanonical motifs. A unimolecular G4 consensus motif, $G_{3–5}N_{1–7}G_{3–5}N_{1–7}G_{3–5}N_{1–7}G_{3–5}$, was initially used for G4 identification,[19] leading to initial estimates of ~300,000 possible G4-forming structures in the human genome.[20] However, mounting structural evidence indicated that the sequences capable of forming G4s and the G4 structures, themselves, were more diverse than originally thought. Structural variations of G4 structures include motifs that incorporate non-G-tetrads,[21] bulged residues,[22] G-triads,[23, 24] G-tetrads as part of pentad assemblies,[25] and hybrid G4/duplexes.[26, 27] This

sequence and structural diversity led to the doubling of the predicted G4-forming sequences in the human genome to >700,000.[28]

Similarly, a unimolecular i-motif folding rule was formulated based on experimental evidence.[29] This specified five cytosine residues for each of the four C-tracts, but allowed for greater variation in the length and sequence of the loop regions. Based on this, a preliminary search predicted >5,000 i-motif-forming sequences in the human genome.[29] However, isolated i-motif structures with shorter or longer C-tracts have been reported,[30-32] and the characteristic C─C^+ base pair of i-motifs is prevalent in a variety of other noncanonical DNA structures,[9, 33-36] suggesting that they can serve as building blocks or structural units for other types of structures. Additionally, the structural topology of i-motifs is not limited to only C─C^+ base pairs. Even the earliest i-motif structures incorporated other noncanonical base pairs[7, 37-40] or base triples[41, 42] that stabilize the motif through stacking on the hemiprotonated cytosine base pairs.[43] As a result, the number of sequences in the human genome with the potential to form i-motifs or related structures is likely much greater than previously predicted.

Chapter 1.2.2. Fold-back Structures

DNA fold-back motifs are a general class of quadruplex structure formed from the assembly of two similar or identical oligonucleotides. The 180° fold-back nature of both strands positions the four backbone segments to form base paired tetrads or coplanar base pairs (Figure 1.4a). Although the sequences that adopt fold-back structures vary, they share several commonly observed features around the fold-back loop. This fold-back loop is comprised of two nucleotides: a 5′ stacking nucleotide and a 3′ bulged nucleotide. The stacking nucleotide is most commonly, though not

exclusively, a pyrimidine. As a pyrimidine, it does not form base pairing interactions between partner strands, but stacks with the adjacent tetrad base pair to form a terminal cap on the core tetrad stacking interactions. In all examples for which structures have been reported, a large ζ torsion angle (113.8–175.9°) between the preceding nucleotide and the capping nucleotide is characteristic of the structural transition point that begins the fold-back. The bulged nucleotide exits the core of base pairing and stacking interactions. In all reported structures, the bulged nucleotide is observed as a pyrimidine, though evidence suggests that any nucleotide can be accommodated at this position with different thermodynamic stabilities.[44]

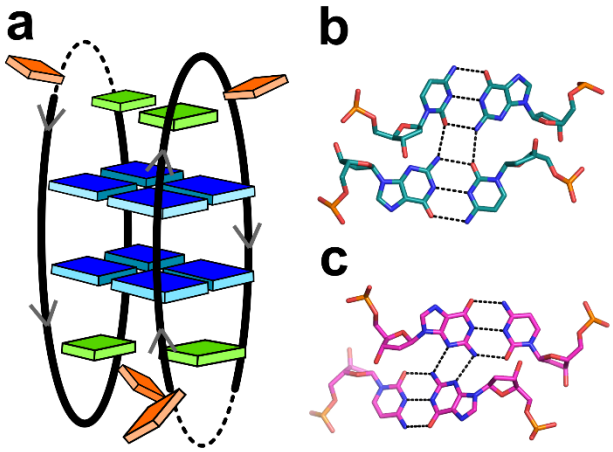

Figure 1.4. DNA Fold-back Motif. (a) Schematic representation of the dimeric fold-back. Sequences can be linear (solid) or cyclic (dashes). Gray arrows indicate the directionality of each strand. Four backbone segments of two DNA strands associate to form base paired tetrads or coplanar base pairs (blue). The loop residues are the 5′ stacking (green) and the 3′ bulged (orange) nucleotides. **(b-c)** Stick representation of **(b)** direct (PDB: 1EU2) and **(c)** slipped (PDB: 2HK4) minor groove tetrads. Gray dashes represent hydrogen bonding interactions.

The nucleobase identities flanking these two loop nucleotides dictate the type of tetrad base pairing interactions that occur between the two fold-back strands. Sequences containing C-**YN**-G (bold letters indicate the loop nucleotides described

above) form “direct” minor groove tetrads[45] from intermolecular Watson-Crick G─C base pairs and intramolecular G(N2)─C(O2) hydrogen bonds (Figure 1.4b). Interactions between G-**YN**-Y sequences form “slipped” minor groove tetrads[46] characterized by intermolecular G─G sugar-edge base pairs and Watson-Crick face G─Y (G─C or G─T) interactions (Figure 1.4c).

The sequences on either side of the tetrad provide further sequence and structural diversity. A bulk of the structural characterization of this motif has come from solution studies of cyclized oligonucleotides that generate two-fold symmetric fold-backs on either end of the dimer pair to create two stacked tetrads in what has been termed a biloop (Figure 1.5a,b).[8] Linear fold-back sequences have similar overall

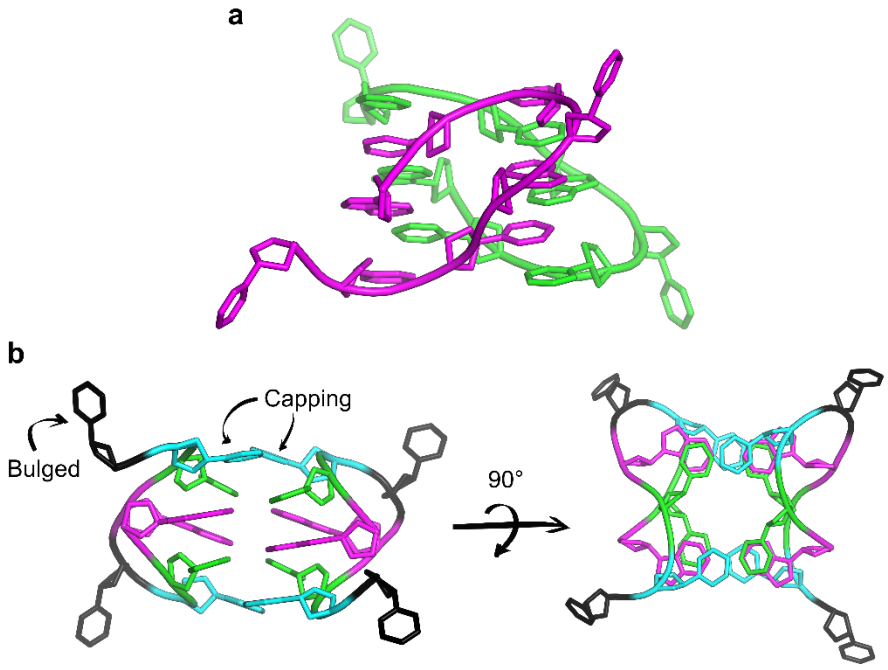

Figure 1.5. d<pATTCATTC> Fold-back Structure (PDB: 284D). (a) The fold-back dimer (with one chain colored magenta and the other green) is shown as a cartoon representation and oriented to highlight the fold-back architecture in the magenta chain. **(b)** Residues in the fold-back dimer are colored differently to distinguish between the different groups. The capping and bulged residues are indicated and colored in cyan and black, respectively. The core interactions are made up of four A─T base pairs, with the purine colored in magenta and the pyrimidine in green.

structures, but generally lack the bulged nucleotide on one side of each monomer. These linear sequences can also have diverse flanking sequences and structures. These include extended i-motif structures,[47] homopyrimidine and homopurine base pairs,[48-50] mini-i-motif-like structures,[34] and unimolecular fold-back structures from sequence repeats that contain the mini-i-motif.[51] The similar structural characteristics of this type of fold-back quadruplex motif in different sequence contexts suggest that it is a robust motif with the potential for involvement in biological function.

Chapter 1.2.3. Parallel-stranded Duplexes

Distinct from the antiparallel B-form duplex, DNA can adopt an alternative type of double helix in which both strands of the same sequence are oriented in the same direction and held together through non-Watson-Crick symmetric homo base pairs (i.e., A─A, C─C^+, G─G, T─T). The observation of self-pairings in a number of crystal structures of nucleosides and nucleotides led to the proposal of such interactions in longer nucleic acid sequences. Indeed, short oligonucleotide sequences can assemble into parallel-stranded duplexes in which Watson-Crick pairings are completely absent (Figure 1.6a), resulting in assemblies with different geometric properties relative to B-DNA.[9, 10] Parallel-stranded double helices, especially those with many cytosine residues, like the d(CGACGAC) heptamer, have typically been observed in acidic conditions, since low pH facilitates protonation and promotes the formation of C─C^+ base pairs.[10] Another example of a parallel-stranded structure was observed in the d(ACTCGGATGAT) oligonucleotide, which is the first crystal structure containing all four different symmetric homo base pairs.[36] This assembly is rather unique because

two parallel-stranded helices are coaxially stacked and intercalated in a head-to-head fashion, forming a tetraplex structure (Figure 1.6b).

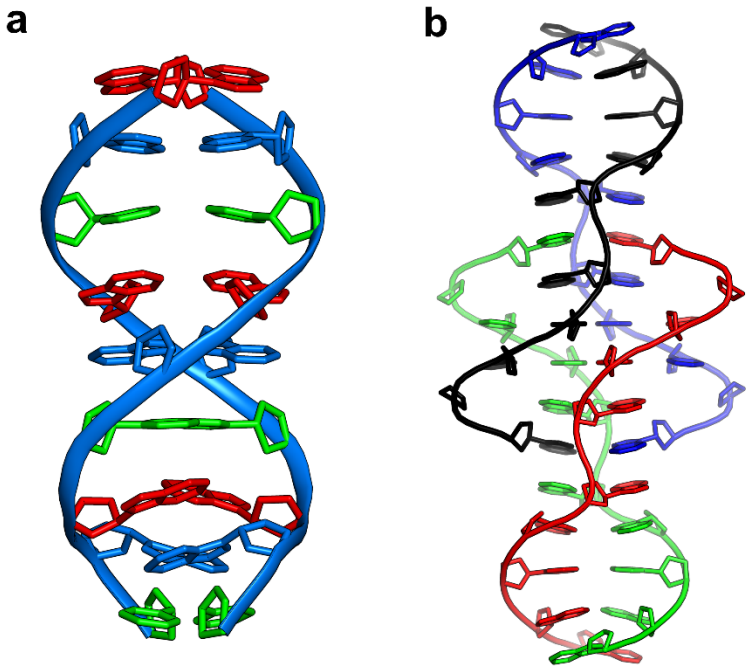

Figure 1.6. Parallel-stranded Motif. (a) Two strands of the d(GAC)$_3$ oligonucleotide (PDB: 1NP5) interact through noncanonical symmetric homo base pairs to form a parallel-stranded duplex. **(b)** The 11-mer, d(ACTCGGATGAT), oligonucleotide (4RIM) associates with a symmetry-related strand to form a dimer (red/green). This dimer then intercalates with a crystallographically-equivalent dimer (black/blue) in a head-to-head fashion to form a parallel-stranded tetraplex.

Chapter 1.2.4. Hybrid Quadruplexes

Beyond individual motifs, noncanonical DNA consisting of a combination of different motifs have been reported. For example, the d(GCGAAAGCT) nonamer associates with three symmetry-related strands to form a quadruplex structure containing a Watson-Crick and parallel-stranded duplex region, which are oriented orthogonally to one another.[33] The $C_2G_3A_4A_5$ residues are involved in symmetric homo base pairs, forming the parallel-stranded region, while the self-complementary $A_6G_7C_8T_9$ residues assemble into a short antiparallel Watson-Crick double helical

region (Figure 1.7a). Further, there have been reports of hybrid structures juxtaposing the biologically relevant G4 and Watson-Crick duplex motifs.[26, 27] Interestingly, two different crystal structures both show a unimolecular assembly of this hybrid in which the wide groove of the G4 extends continuously into the duplex minor groove, orienting the two motifs one above the other without the need for additional residues to facilitate the transition (Figure 1.7b).

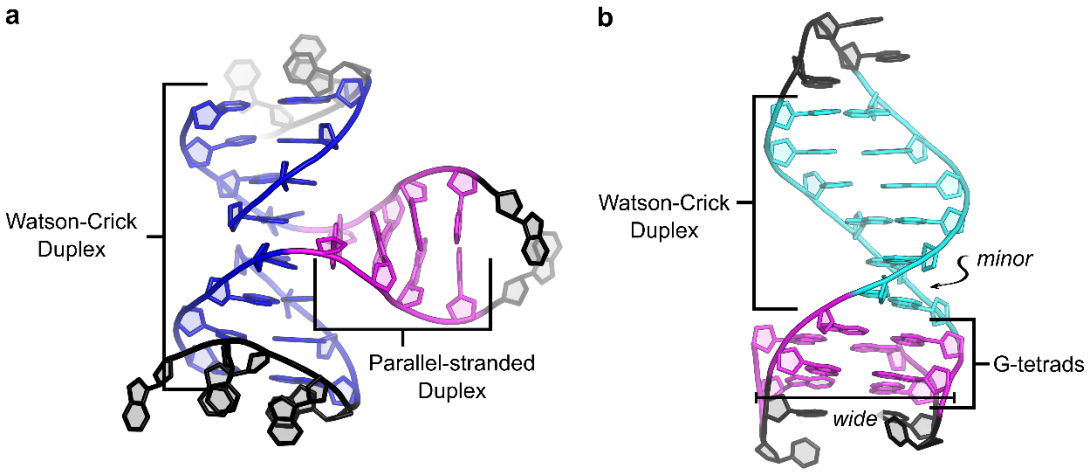

Figure 1.7. Hybrid Quadruplexes. (a) The d(GCGAAAGCT) oligonucleotide (PDB: 1IXJ) associates with three symmetry-related strands to form a hybrid structure juxtaposing a Watson-Crick duplex (blue) and parallel-stranded duplex (magenta) region. **(b)** The unimolecular 27-mer oligonucleotide (PDB: 2M8Z) forms a hybrid structure in which the Watson-Crick duplex region (cyan) transitions continuously into the G-quadruplex region (magenta). The duplex minor groove and the G-quadruplex wide groove are indicated.

Chapter 1.3: Noncanonical DNA structures are biologically relevant

Both G4s and i-motifs are present in cellular DNA, though their roles in biological processes are just beginning to be understood. G4s have been implicated in a wide variety of normal cellular processes, including DNA replication and transcription, as well as a number of disease states.[52] Telomeric G4 structures have been visualized using specific antibodies.[53] The active formation of G4s,[54, 55] as well as

their stabilization by small molecule ligands,[54] in human cells have also been confirmed. With a predicted 50% of human genes containing G4s at or around promoter regions, DNA G4 structures are predicted to have widespread roles in gene expression.[56] In particular, the significant enrichment of the G4 motif in a wide range of oncogene promoters suggests its functional importance in cancer.[57] Examples of G4s modulating gene transcription have been found in the c-MYC,[58] bcl-2,[59] and KRAS[60] oncogene promoters. Additionally, the stabilization of G4s by small molecule ligands at the hTERT[61] and PDGFR-β[62] oncogene promoters has been associated with downregulated activity. Nonetheless, the highly thermostable G4s can be detrimental to biological processes and lead to genome instabilities.[52, 63]

DNA i-motifs have long been implicated in biological processes,[29, 57, 64] but have now been observed *in vivo*. In-cell NMR identified characteristic i-motif signals in HeLa extracts with transfected i-motif DNAs, providing direct evidence that i-motif structures are stable in cellular environments.[65] Furthermore, the antibody-mediated observation of i-motifs in the nuclei of human cells[66] and the discovery of i-motif binding proteins that regulate gene activity[67] demonstrate that i-motifs can have biological function. The sequence and structural diversity of G4s and i-motifs and their growing importance in cellular DNA transactions open the possibility of new variations of these motifs with distinct biological functions.

The biological relevance of other non-Watson-Crick DNA structures is less clear. While they have been implicated in biological settings, we have yet to directly visualize these motifs in physiological conditions. Recent studies have, however, shown that tandem repeats of a fold-back motif containing the mini i-motif are clustered

near promoter regions of human developmental genes, indicating their potential role in gene regulation.[51] Further, parallel-stranded duplex structures may have roles in recombination processes.[10] The sequence-dependent structural arrangement of the parallel-stranded motif may be useful in precisely aligning homologous sequences to form intermediate structures during DNA recombination.[9, 10] This motif may also be present in highly folded RNA molecules like tRNA, which contain segments of the RNA backbone oriented in a parallel fashion.[9] Moreover, hybrid structures containing motifs of known physiological relevance may present more opportunities for biological applications.[26] For example, if hybrid structures juxtaposing the Watson-Crick duplex and G4 motifs are known to be associated with diseases, we can then design more specific therapeutic drugs to target the localized area. While further studies are necessary to elucidate the *in vivo* function of these noncanonical DNA structures, the identification of such motifs in physiological conditions is a significant initial step in gaining an understanding of their potential biological roles.

Chapter 1.4: Discovering New DNA Structures

Chapter 1.4.1. Motivation for This Study

As we continue to discover new noncanonical DNA structures, it is clear that we have yet to reach the limits of DNA's structural diversity. Many other non-Watson-Crick base pairing motifs may, perhaps, exist in biological systems, but have just not been discovered yet. Historically, the biological relevance of noncanonical DNA has been questioned since the earliest studies on non-traditional base pairs. In 1956, Donohue expressed doubt that all geometrically acceptable noncanonical base pairs

have significance in nature.[4] Further, when Gellert and colleagues discovered the G-quadruplex in 1962, they were unaware that it was relevant to biology. And even in 1994, after subsequent studies identified G4-forming sequences in telomeric DNA, the biological roles were still unclear.[14] However, over the course of 50 years, continuous research on the structural and biophysical properties of the G4 motif has resulted in the recent visualization of such structures in human cancer cells, indicating their potential as therapeutic targets.[68]

This example clearly demonstrates merit in the structure-before-function approach. Through this strategy, we can first uncover unprecedented knowledge on DNA structure and then investigate their potential biological roles. Because we are not limited to searching for structures that we are already familiar with, this approach can greatly expand the likelihood of identifying novel structural assemblies.

Chapter 1.4.2. A Structural Biology Approach

We are using a structural biology approach to discover new DNA structures and to probe the limits of its structural diversity. I am screening a large number (500+) of short DNA oligonucleotide sequences ranging between 10 and 13 nucleotides and determining the structures of those that crystallize. This sequence library was generated from a customized computer script that randomly assigned a discrete DNA sequence to the specified length. The script was not programmed to omit identical sequences because the presence of such sequences would serve as internal positive control samples. The only constraint of the script was sequence length and this was done intentionally to minimize bias in the generating of the randomized sequences.

Each DNA sample was synthesized and screened against a set of 96 homemade conditions that vary in pH and in the type and concentration of precipitants, cations, and polyamines (Table 1.1). Not only is this approach advantageous because of the cost-effectiveness of DNA synthesis, the components of the crystallization conditions are commercially available and relatively inexpensive. Furthermore, DNA is easy to synthesize and purify, which significantly decreases the length of time to obtain samples ready for experiments.

Table 1.1. Components of the 96 Homemade Cation-rich Conditions.

Component	Type	Range
Precipitant	MPD, PEG 400	0–20%
Monovalent Cation	K^+, Li^+, Na^+, NH_4^+	0–100 mM
Divalent Cation	Ba^{2+}, Ca^{2+}, Mg^{2+}, Sr^{2+}	0–120 mM
Polyamine	Cobalt hexammine(III), Spermidine	0–20 mM
pH	Sodium cacodylate (30 mM)	5.5–7.4

Our lab has crystallized a large number of DNA oligonucleotides from our sequence library. The initial criteria for identifying positive hits from the screen were based on qualitatively observing birefringent solid material. Preliminary data show that 20% of sequences crystallize in some form, which is an incredible result. Not only does this provide us many targets to work with, it also indicates that there is a lot of order in the library and suggests that there is potential diversity in the structures formed by these DNA. We have obtained crystals in a variety of morphologies, such as rods, trapezoids, plates, and needles (Figure 1.8). While some of these are large, individual, sharp-edged crystals and appear to be diffraction-quality, others are small, circular, and clustered together. For crystals that fall in the latter group, conditions are modified in an attempt

to yield individual crystals with well-defined edges. Lastly, we have observed DNA sequences crystallizing in multiple conditions of the homemade solutions, indicating that these crystals are not artifacts.

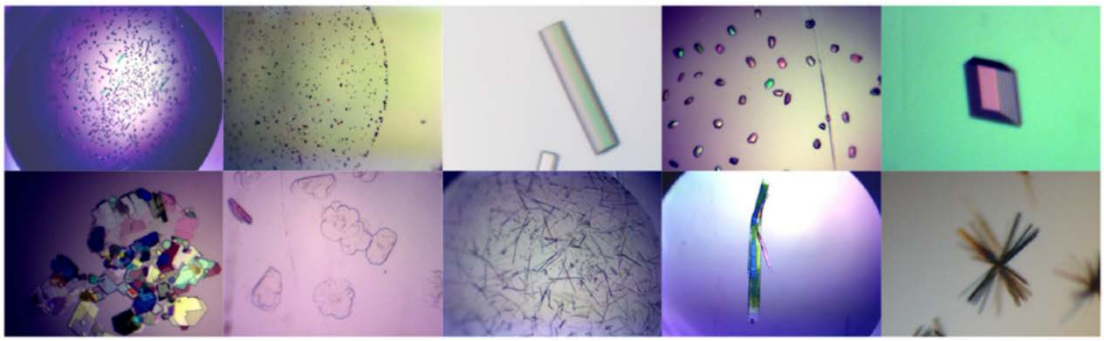

Figure 1.8. DNA Crystal Images from Screening. Initial screening against the homemade 96 conditions yielded a broad array of crystals and crystalline material, including microcrystals, rounded crystals, needle clusters, and single crystals. Each picture represents a different DNA sequence.

Chapter 1.4.3. Goals

From the screen, we aim to discover new DNA structures and identify previously unobserved motifs that may be biologically relevant. Some of these may be short imperfect repeats, which would suggest the potential existence in physiological conditions since more than 50% of the human genome is comprised of repetitive DNA.[69] Through this project, we can also investigate the correlation between the sequence and the structures solved from the library. Once we have solved a parent structure, we can then examine point mutations to identify residues that contribute to structural importance. Moreover, we can better understand the structural diversity of DNA and sample the different types of motifs formed, as well as the frequency of them.

Beyond the biological perspective, we can expand the potential impacts of this project to the DNA nanotechnology field. Currently, the field primarily relies on the

predictable nature of Watson-Crick base pairing. Through this project, we aim to complement this by uncovering the predictability of noncanonical motifs. If we can accurately predict what situations favor noncanonical base pairing, we would have greater control over the types of desired DNA structures. This would greatly expand the structure space for rational crystal design, which would lead to the creation of more precise structures targeted for specific nanotechnological applications, including drug delivery vehicles,[70] biocatalysts,[71] and molecular sieves.[72]

Our lab has crystallized a number of short oligonucleotides in an effort to discover new DNA structures. In this dissertation, I will discuss two types of structures, the fold-back quadruplex and the hybrid G-quadruplex/i-motif, which were determined from the crystallization screening. Both structures contain previously unobserved features and interesting structural characteristics with significant implications. I will first provide an analysis of the crystal structures and then describe results from the solution studies, which will reveal insight into their potential biological relevance.

Chapter 2: d(CGTAAGGCG) Crystal Structure Analysis

This chapter is adapted from "Chu, B.; Zhang, D.; Hwang, W.; Paukstelis, P.J., Crystal Structure of a Tetrameric DNA Fold-back Quadruplex. *J. Am. Chem. Soc. 140* (47): 16291 – 16298."

Chapter 2.1: Results and Discussion

Chapter 2.1.1. Overview

In this chapter, I describe the 1.05 Å crystal structure of d(CGTAAGGCG) and several related oligonucleotide structures. This linear DNA oligonucleotide forms a dimeric fold-back motif with distinct features, including a G-**YN**-A fold-back core that arises from an A(*syn*):G:G:A(*syn*) minor groove slipped tetrad. Further, this structure represents the first observation of interactions between two fold-back dimers through their flanking sequences to create a tetrameric structure. The end-to-end interaction of these dimers creates a central cavity that contains a divalent cation. Structure determination of all variations of the fold-back bulged residue with different divalent cations indicates predictable assembly and suggests the potential for designed ligand binding in the central cavity.

Chapter 2.1.2. Initial Crystallization

As part of the screen to probe the structural diversity of DNA oligonucleotides, I crystallized d(CGTAAGGCGTA) and a 5-bromo-deoxy-U3 derivative, d(CGU^{Br}AAGGCGTA), for phasing. Despite reasonable resolution limits and strong anomalous signal from the incorporated bromine, both single-wavelength anomalous dispersion (SAD) and single isomorphous replacement with anomalous scattering (SIRAS) phasing methods indicated more heavy atom positions (14 bromines) than

were expected based on the Matthews coefficient.[73] The heavy atom sites had a number of positions with both reasonable B-factors and occupancy values that were less than 3.0 Å apart, indicating that they arose from multiple conformations. The resulting electron density maps were adequate to manually build a model of 4 dimeric fold-back units (8 chains) in the asymmetric unit, with two dimers having relatively high-quality density and two of lower quality. But, despite high local density fit correlations and reasonable model geometry, the refinement R-factors remained high, at 0.38–0.42, for both the native and derivative data sets. There were no signs of crystal lattice pathologies such as twinning or translocation defects. However, one common feature of all eight chains that were modeled from these data was disorder of the two 3′ most residues, T10 and A11.

In an attempt to mitigate the overall structural disorder, I crystallized a truncated oligomer lacking these two nucleotides. This truncation crystallized under the same conditions (Figure 2.1a), but in a different space group with a much higher diffraction limit (Figure 2.1b), and contained a fold-back dimer in the asymmetric unit.

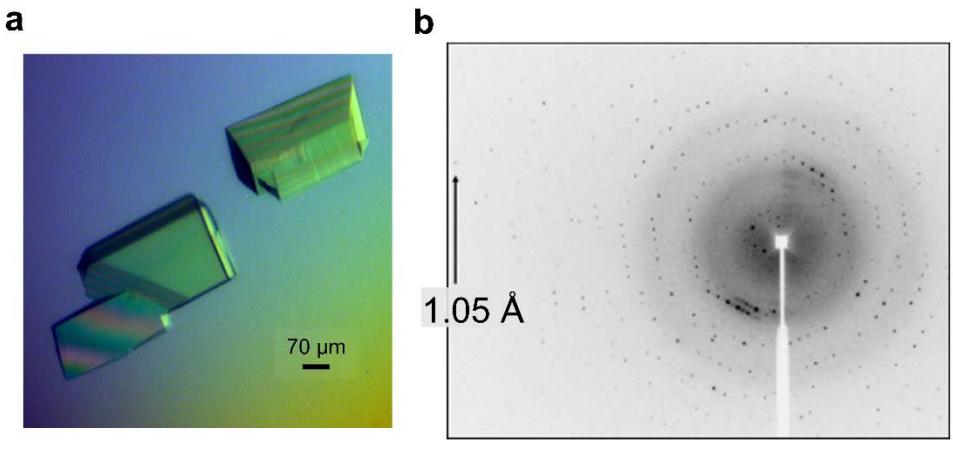

Figure 2.1. d(CGTAAGGCG) Crystals and Diffraction Pattern. (a) d(CGTAAGGCG) crystals grew as rectangular crystals, with the longest dimension measuring 340 µm. **(b)** Diffraction spots were observed to 1.05 Å.

Chapter 2.1.3. Dimeric Foldback

The asymmetric unit of d(CGTAAGGCG) contains two chains (A and B) that interact to form a fold-back dimer (Figure 2.2a,b). Over the first six residues, the two chains are nearly identical (RMSD = 0.230 Å for all atoms; Figure 2.3a), with the most variation coming from several locations that adopt well-defined alternate conformations. One of these multiple conformations occurs in the T3 capping nucleotide (Figure 2.3b). The nucleobase of T3 from Chain A is in two different

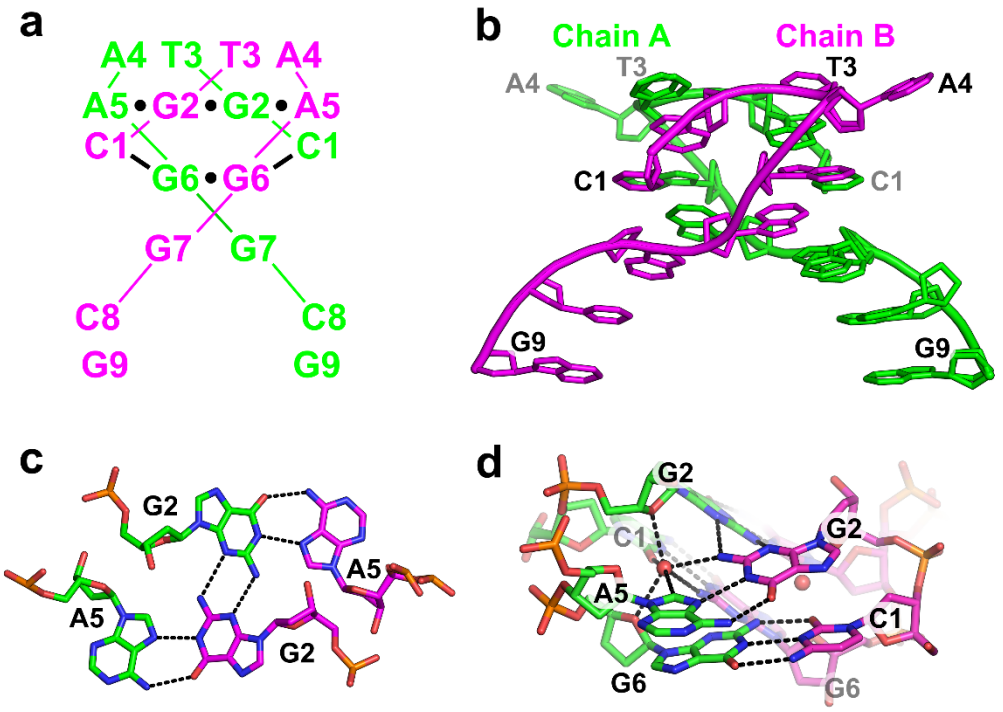

Figure 2.2. Asymmetric Unit. (a) Secondary structure of interactions formed between two DNA strands. Black dashes represent Watson-Crick base pairs. Black circles represent noncanonical base pairs. **(b)** Cartoon representation of the dimeric fold-back, with Chain A in green and Chain B in magenta. **(c)** Stick representation of the A(*syn*):G:G:A(*syn*) minor groove slipped tetrad. N2─N3 sugar-edge interactions are observed for the G2─G2 homo base pair. The G2─A5 base pair hydrogen bonds through the Watson-Crick and Hoogsteen face, respectively. **(d)** Water molecules (red spheres) are positioned within hydrogen bonding distance of O2 of C1, N2 of G2, and N2 of G6 of the partner strand (magenta), and C8 of the *syn* A5. Stick representation of the slipped C:G:G:C tetrad. Interactions are only shown for one of the two waters in the dimeric structure. Tetrads are named in the order of linear hydrogen bond connectivities.

conformations, originating at the glycosidic bond with refined occupancies of 0.53 and 0.47. This was the derivatized position in the original full-length crystals, and the apparent multiple conformations in both this and the heavy atom substructure suggests some inherent conformational flexibility at this position. Consistent with other fold-back structures, the transitions between G2 and T3 have large ζ angles of 136.0° for Chain A and 135.4° for Chain B.

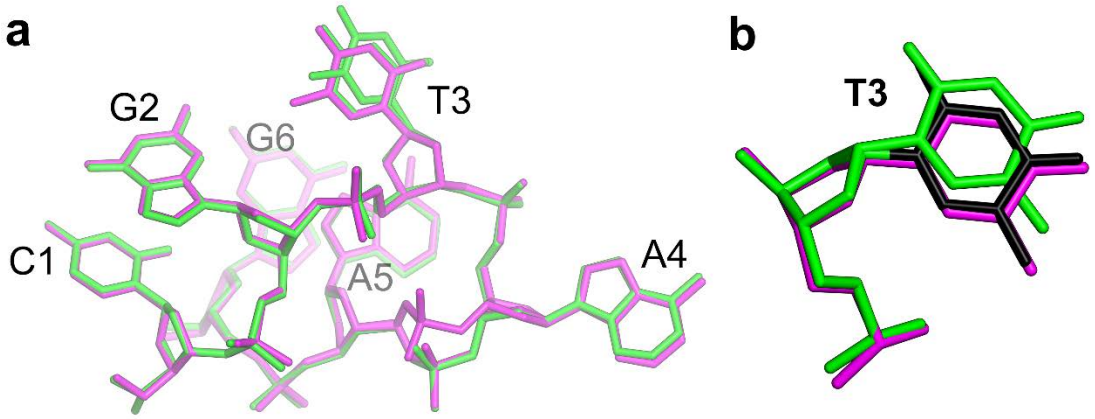

Figure 2.3. Structural Comparison of the d(CGTAAGGCG) Monomers. (a) Residues 1–6 from Chains A (green) and B (magenta) are structurally similar. The phosphate group of C1 and the T3 nucleobase from Chain A adopt a second conformation. The A4 sugar and A5 phosphate groups from Chain B occupy a second conformation. **(b)** Residue T3 from Chain A exists in two conformations (green and black) with occupancies of 0.47 and 0.53.

Distinct from previously reported fold-back structures for which coordinates are available, the bulged nucleotide in this structure is a purine, A4. As the bulged nucleotide, it does not make any hydrogen bonding contacts with the rest of the structural core and makes only a single hydrogen bond with residue A5 from a symmetry-related molecule in the unit cell. The two purines are oriented such that the N6 hydrogen bond donor is within 3.3 Å of the N3 hydrogen bond acceptor, resulting in a Hoogsteen─sugar face base pair (Figure 2.4a,b). The bulged A4 also forms base

stacking interactions with the indole ring of G9 from a symmetry-related molecule (Figure 2.4b). Lastly, the crystal lattice is further stabilized through stacking interactions between A4 symmetry mates that adopt reciprocal sugar/lone pair-base stacking interactions (Figure 2.4c).

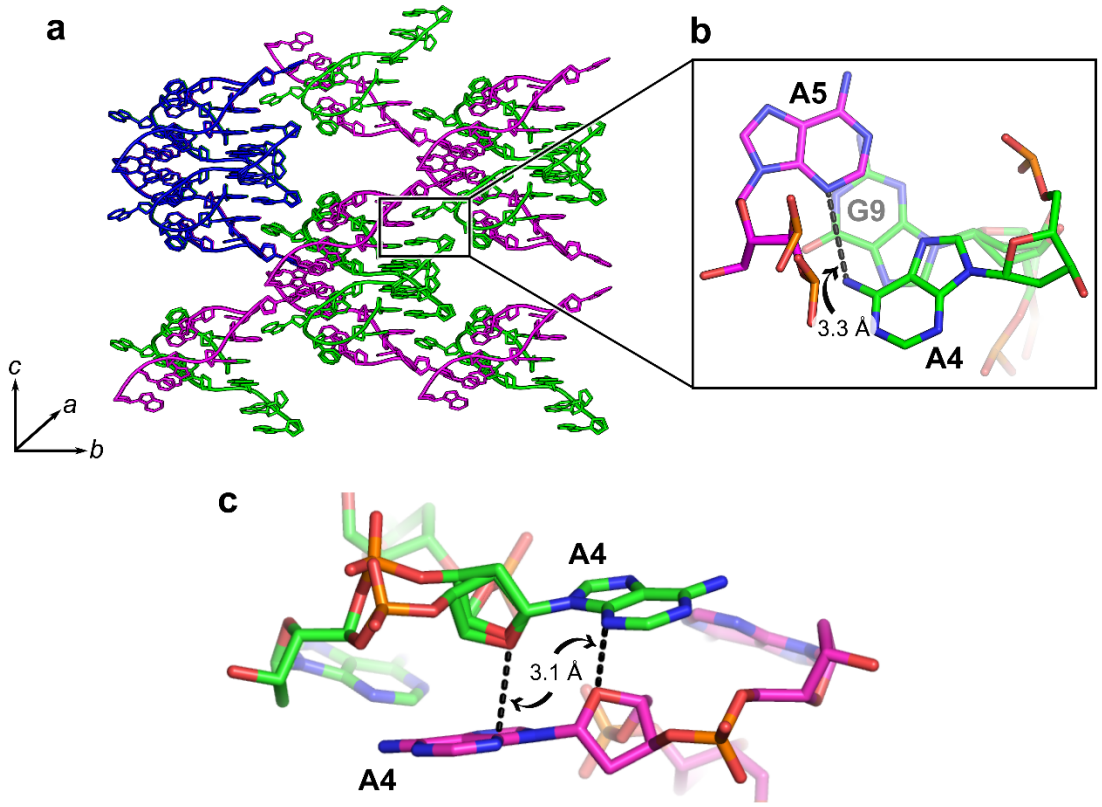

Figure 2.4. Native Crystal Contacts. (a) Tetramer units that make up the crystal lattice along the *b* and *c* dimensions of the monoclinic unit cell are shown. One unit is colored in blue to distinguish the tetrameric assembly. **(b)** The inset shows interactions between A4 and residues from symmetry-related molecules. Residue A4 forms a single hydrogen bond with an A5 residue through the N6─N3 Hoogsteen/sugar faces, respectively. The indole ring of A4 also forms base stacking interactions with the G9 indole ring. **(c)** The bulged A4 nucleobase stacks with the O4′ atom of a crystallographically equivalent A4 residue with a stacking distance of ~3.1 Å.

d(CGTAAGGCG) forms two tetrads as the core structural feature of the fold-back architecture. As expected with a guanosine 5′ of the capping nucleotide, the two G2 residues interact through a sugar-edge G─G base pair to establish the first slipped

tetrad in the structure. The tetrad is completed by base pairing between the Watson-Crick face of G2 of one strand and the Hoogsteen face of A5 from the partner strand, making this the first example of a fold-back motif containing this tetrad (Figure 2.2c). The G(N1)─A(N7) and G(O6)─A(N6) hydrogen bonds are facilitated by A5 adopting a *syn* glycosidic torsion angle. Relative to a pyrimidine at this position, the *syn*-adopting adenosine allows the binding of water molecules within each minor groove side (Figure 2.2d). These waters serve as bridging donors for intramolecular interactions between O4′ of G2 and G6, likely stabilizing the tight turn that brings the sugars into proximity. Simultaneously, they are positioned within hydrogen bonding distance of O2 of the C1 nucleobase (3.1 Å), G2 and G6 N2 amino positions from the partner strand (2.9 and 3.5 Å, respectively), and only 3.1 Å from the C8 position of the *syn* A5 nucleobase, suggesting a possible CH─O hydrogen bond. These waters are present in all A4 variant structures and appear tightly bound based on B-factor analysis (average B-factor = 6.7 vs. 25.9 $Å^2$ for all nonmetal-associated waters). The second tetrad of the fold-back core is formed by C1 and G6 of the dimer pair. This is a slipped C:G:G:C tetrad, formed through G6─G6 sugar-edge base pairs and Watson-Crick face pairing with C1 (Figure 2.2d).

Interestingly, the transitions between G6 and G7 in both strands of the dimer have the same large ζ torsion angle (151.8° for Chain A and 147.0° for Chain B) usually observed in the capping nucleotide. Although G7 is at a position in the sequence typically adopted by a capping nucleotide in a biloop structure, it has minimal stacking interactions with the tetrad core of the fold-back motif. This large ζ angle may reflect a structural preference at this position for the motif itself, but may also be influenced

by additional base pairing interactions made by G6 and G7 through crystal contacts (see Chapter 2.1.4).

Chapter 2.1.4. Higher-Order Assembly through Crystal Symmetry

The last two nucleotides, C8 and G9, are not part of the continuous stacking of the fold-back dimer. Instead, these nucleotides base pair with G7 and G6, respectively, of a crystallographically equivalent dimer to form a tetramer (Figure 2.5a,b). Residue C8 forms a Watson-Crick base pair with G7, while the G9 Watson-Crick face hydrogen bond donors base pair with the Hoogsteen face acceptors of G6. Because G6 is base paired through its Watson-Crick face to C1 of the partner strand, each half of the C1:G6:G6:C1 tetrad is converted to a base triple interaction. Thus, the G6's sugar-edge hydrogen bonding contacts effectively create a base hexad involving all four strands of the tetramer, with G6 hydrogen bond valences fully occupied (Figure 2.5c). One face

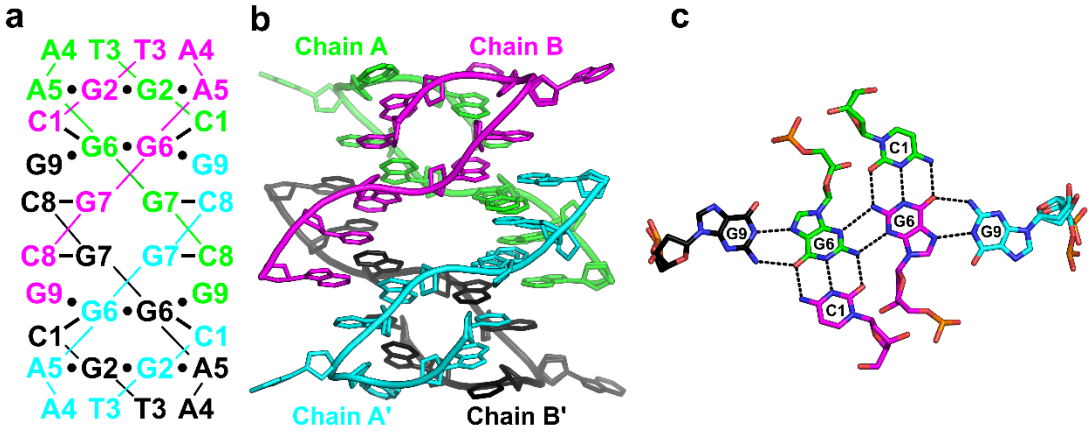

Figure 2.5. Tetrameric Assembly. (a) Secondary structure of interactions formed between two symmetry-related dimers. Black dashes represent Watson-Crick base pairs. Black circles represent noncanonical base pairs. Chains A and A′ are in green and cyan, respectively. Chains B and B′ are in magenta and black, respectively. **(b)** Cartoon representation of the tetramer. Two dimers interact in an end-to-end fashion. **(c)** Stick representation of the C:G:G:G:G:C hexad. N2─N3 sugar-edge interactions are observed for the G6─G6 homo base pair. The G6─G9 base pair hydrogen bonds through the Hoogsteen and Watson-Crick faces, respectively.

of the C:G:G:G:G:C hexad stacks directly with the A5:G2:G2:A5 tetrad and the other face stacks with the tertiary G7─C8 base pairs. Along with the stacking between symmetry-related G7─C8 base pairs, this joins the two dimers in an end-to-end fashion. The orientations of the two hexads provide continuous stacking interactions throughout six adjacent planes of the tetramer, which contribute to the stabilization of the overall fold. Although other DNA hexads[74] have been reported in the context of G4 structures, to our knowledge, this is the first example of a DNA hexad involving four strands with this base pairing arrangement.

The G7─C8 base pairs are formed between equivalent symmetry mates and multiple conformations of these two residues from Chain A were clearly distinguishable. The nucleobases retain their base pairing partner in the multiple conformations (Figure 2.6), suggesting that they move in a concerted fashion. Moreover, weak electron density suggested a possible third conformation of the G7─C8 base pair, which was not modeled, as well as multiple conformations of the G9 phosphate.

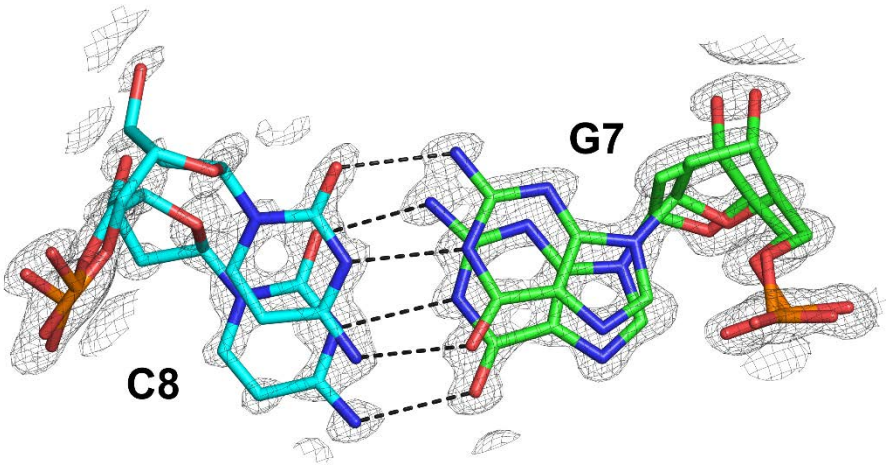

Figure 2.6. Multiple Conformations of the G7─C8 Base Pair. All atoms of the G7─C8 base pair between two symmetry-related chains are modeled in two conformations and are shown with $2mF_o$-DF_c electron density contoured at 1 σ.

Chapter 2.1.5. Cation-Binding Pocket

The interface between the two dimers generates a cavity in the center of the tetrameric structure (Figure 2.7a). The cavity has a volume of 232.9 $Å^3$, as calculated by Hollow,[75] and contains a hexahydrated magnesium ion and at least three additional water molecules. The keto group of each of the four G7 residues is directed towards the central cavity and coordinates directly with Mg^{2+}-bound waters with distances of 2.4, 2.6, 2.7, and 3.1 Å. The proximity of the Mg^{2+} to the crystallographic two-fold axis and its interactions with G7, which itself adopts multiple conformations (see Chapter 2.1.4), results in this ion occupying two positions, 1.8 Å apart with apparent equal occupancy (Figure 2.7b). The Mg^{2+} is coordinated to both conformations of the G7 nucleobase from Chain A, suggesting coordinated movement.

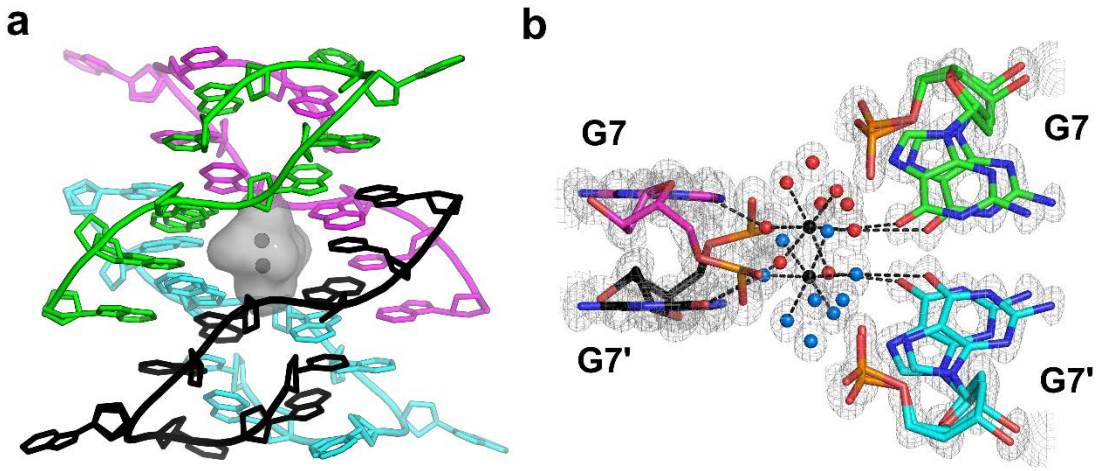

Figure 2.7. Divalent Cation Binding. (a) Cartoon representation of the tetramer with a surface representation of the central cavity (gray). A Mg^{2+} ion and its symmetry mate are shown as black spheres in the cavity. **(b)** Mg^{2+} coordination between DNA strands shown with $2mF_o$-DF_c electron density contoured at 1 σ. A Mg^{2+} (black sphere) exists at two positions at 50% occupancy, with each Mg^{2+} coordinating six water molecules (red or blue spheres). The O6 atom of each G7 nucleobase coordinates with the Mg^{2+}-bound waters.

Chapter 2.1.6. A4 Variants: Overview

As part of this study, we crystallized and determined structures of the other three variants containing modifications at the bulged position (A4C, A4G, A4T). These variants crystallized under different conditions (Figure 2.8), adopted different space groups, and had different numbers of molecules in their asymmetric units (Table 2.1). A4C crystallized with eight molecules that form two tetramers in the asymmetric unit, while A4G crystallized with four molecules that assemble into one tetramer. Like the native oligonucleotide, A4T crystallized with two molecules in the asymmetric unit and associates with symmetry-related chains to generate the tetramer.

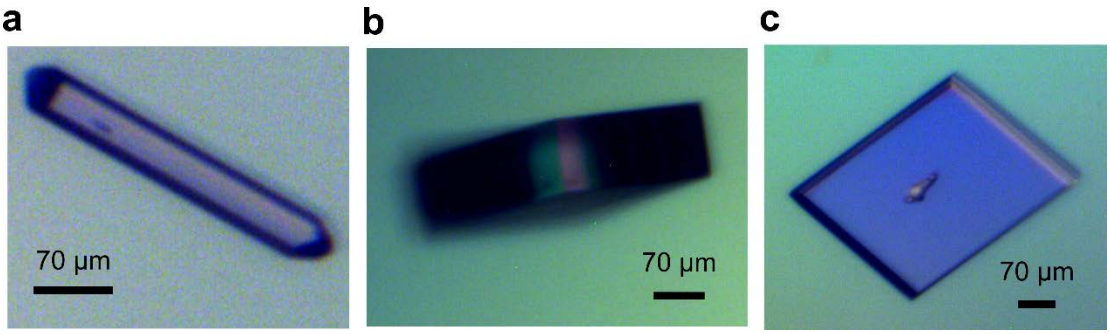

Figure 2.8. Crystals of A4 Variant Oligonucleotides. (a) A4C crystallized as a long rod with a triangular base, with the longest dimension measuring 280 μm. **(b)** A4G crystallized as a cubic crystal, with a thickness of ~90 μm. **(c)** A4T crystallized as a plate-like crystal in the shape of a parallelogram, with the longest dimension measuring 350 μm.

However, all the variants contained identical base pairing interactions throughout the structures, indicating that the differences in crystal characteristics arose predominantly from the different conditions and nucleobase identity at this position. Structural alignment of the fold-back tetramer unit from each variant with the native tetramer showed similarities in the overall structure, as seen in the RMSD values (Table 2.2). As expected, the largest variation occurred at the fourth residue because of the

differing nucleobase identities (Figure 2.9). Importantly, these results confirm that the bulged nucleotide is variable in the fold-back motif.

Table 2.1. Data Collection and Refinement Statistics for d(CGTAAGGCG) and Variants.

	Native (6MC2)	**A4C (6MC4)**	**A4G (6MC3)**	**A4T (6N4G)**	**U3-Br***
Sequence	d(CGTAAGGCG)	d(CGTCAGGCG)	d(CGTGAGGCG)	d(CGTTAGGCG)	d(CGU^{Br}AAGGCGTA)
Beamline	SER-CAT BM 22	GMCA 23-ID-D	GMCA 23-ID-D	GMCA 23-ID-D	NE-CAT 24-ID-C
Data Collection					
Space Group	I121	$P3_121$	$P222_1$	I121	$P3_221$
Cell Dimensions					
a, b, c (Å)	25.96, 46.96, 35.96	53.74, 53.74, 99.39	34.14, 46.43, 71.89	26.33, 53.71, 34.23	46.51, 46.51, 170.87
α, β, γ (°)	90, 101.12, 90	90, 90, 120	90, 90, 90	90, 102.04, 90	90, 90, 120
Resolution (Å)	23.48 – 1.05	26.99 – 2.25	28.42 – 1.57	28.41 – 1.40	170.87 – 2.36
R_{meas} (within I+/I-)**	0.076 (0.228)	0.134 (1.249)	0.090 (3.496)	0.087 (0.355)	0.042
R_{meas} (all I+ and I-)**	0.082 (0.280)	0.134 (1.265)	0.090 (3.981)	0.096 (0.405)	0.056
R_{pim} (within I+/I-)**	0.051 (0.151)	0.058 (0.557)	0.048 (1.843)	0.059 (0.240)	0.022
R_{pim} (all I+ and I-)**	0.043 (0.160)	0.043 (0.412)	0.036 (1.531)	0.052 (0.224)	0.022
No. of unique reflections**	19770 (1947)	8347 (737)	16520 (802)	9187 (452)	9480
I / σ I**	9.2 (2.3)	11.0 (1.8)	8.3 (0.4)	8.1 (2.2)	21.4
Completeness (%)**	99.85 (99.5)	99.9 (99.4)	99.6 (99.2)	99.5 (97.2)	99.8
Multiplicity**	3.5 (2.7)	9.6 (9.2)	6.4 (6.5)	3.3 (3.0)	6.3
Wavelength (Å)	0.9187	0.9200	0.9200	0.9200	0.9198
Phasing					
Atom/Sites					Br/14
FOM					0.33
FOM, DM					0.66
Refinement					
Resolution (Å)**	28.21 – 1.05 (1.07 – 1.05)	46.54 – 2.25 (2.31 – 2.25)	46.43 – 1.57 (1.61 – 1.57)	28.41 – 1.40 (1.44 – 1.40)	
No. reflections	18965	7906	15701	8753	
R_{work}	0.1069	0.2282	0.2004	0.1469	
R_{free}	0.1372	0.2682	0.2434	0.2192	
Total No. of atoms	592	1582	912	517	
No. of molecules in asymmetric unit	2	8	4	2	
Average B-factors (Å^2)	10.177	39.844	37.605	18.634	
RMSD					
Bond lengths (Å)	0.019	0.007	0.013	0.013	
Bond angles (°)	2.405	1.502	2.086	1.806	

*Phasing only.
**Values in parentheses correspond to the high-resolution shell.

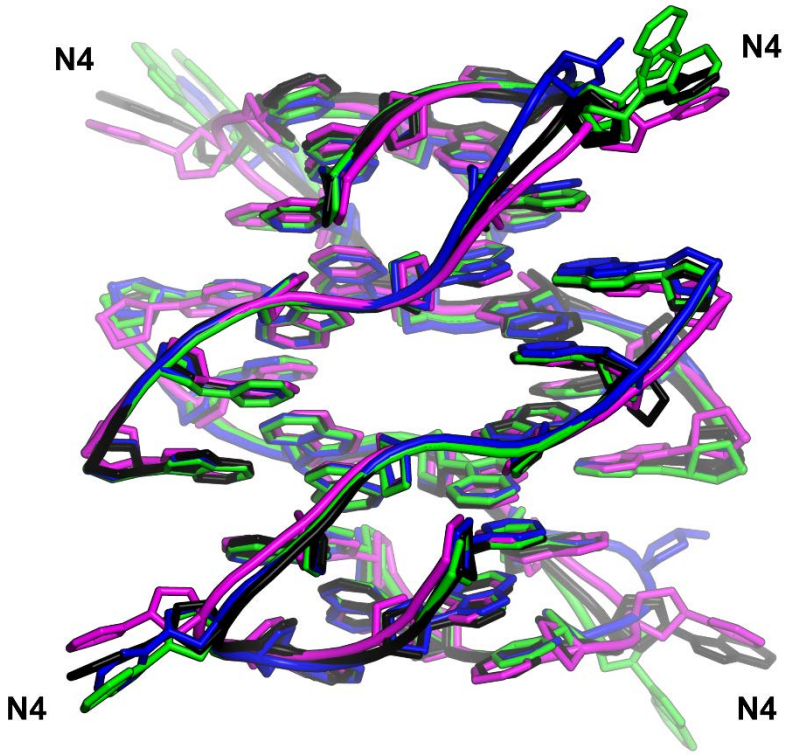

Figure 2.9. Structural Alignment of the Native and Variant Oligonucleotides. The A4C (blue), A4G (green), A4T (magenta) tetramers are aligned with the native (black) tetramer and are shown as cartoon representations. The fourth residue is denoted as N4 and contains a large degree of variation across all four structures. Three C4 residues from the A4C tetramer are truncated. Two G4 residues from the A4G tetramer are modeled with a second conformation.

Table 2.2. Comparison of Several Structural Features of the A4 Variants.

	A4C	**A4G**	**A4T**
RMSD from Native Tetramer	0.484 (452 atoms)	0.386 (444 atoms)	0.648 (498 atoms)
Divalent Cation	Ca^{2+}	Mg^{2+}	Ba^{2+}
Alternate Conformations	C4	G4	Ba^{2+}
Truncated Residues	C4	None	None

Chapter 2.1.7. A4 Variants: Cation Binding

In Chapter 2.1.5, I described the importance of the magnesium ion in stabilizing the tetrameric fold-back formed by d(CGTAAGGCG). Similarly, the three A4 variant

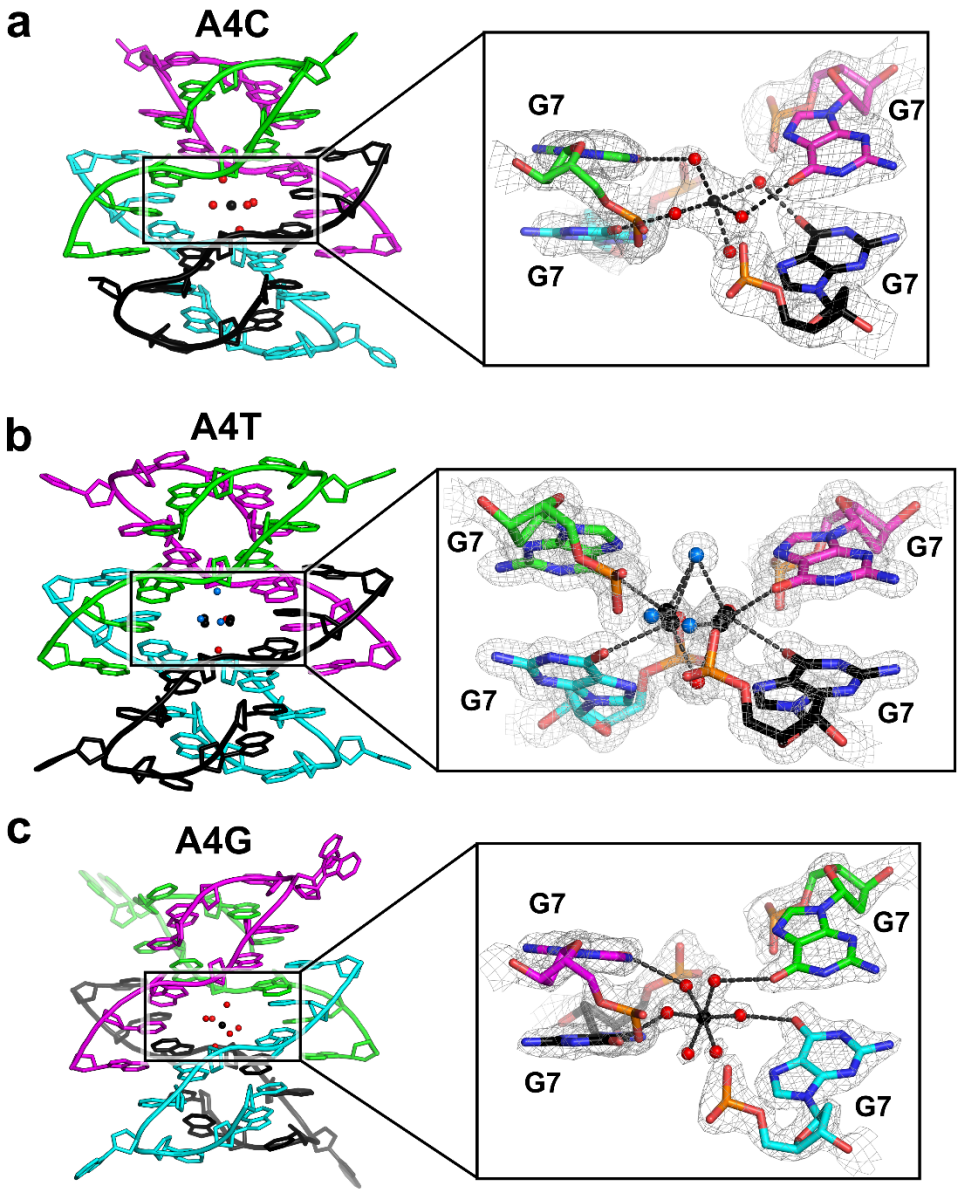

Figure 2.10. A4 Variant Tetramers and Metal Coordination. (a-c) Cartoon representation of **(a)** A4C, **(b)** A4T, and **(c)** A4G tetramers. Insets show the coordination between metal ions (black spheres) and the O6 atom of the G7 nucleobase and/or water molecules (red or blue spheres) in the central cavity with $2mF_o$-DF_c electron density contoured at 1 σ. **(a)** Two A4 residues of the A4C tetramer are truncated. A Ca^{2+} ion exists in one conformation and coordinates five water molecules. The O6 atom of each G7 nucleobase coordinates with Ca^{2+}-bound waters. **(b)** A Ba^{2+} ion exists at 4 positions, with each at 25% occupancy and each coordinating the keto group of one G7 and two water molecules. **(c)** The A4 residue of the A4G tetramer occupies two conformations. A Mg^{2+} ion occupies one position and coordinates six water molecules. The O6 atom of each G7 nucleobase coordinates with Mg^{2+}-bound waters.

structures also enclosed a divalent cation in the central cavity, but because they were crystallized in different conditions, the identity of the cation differed. The A4C and A4T variants were crystallized in the presence of Ca^{2+} and Ba^{2+}, respectively. Both structures showed clear density for a divalent cation at the interface of the tetrameric units (Figure 2.10a,b). Interestingly, the A4C Ca^{2+} appears to occupy a single position (Figure 2.10a), while the A4T Ba^{2+} adopts four positions in proximity to each G7 O6 atom in the pocket (Figure 2.10b). Similar to the native structure, the A4G variant was crystallized in Mg^{2+}, but this ion occupies a single position instead of two (Figure 2.10c). There are no apparent alternate conformations for G7 in these structures, though they were determined at lower resolution (Table 2.1). The presence of a divalent cation in the central pocket of all the structures we examined suggests that these cations may be necessary for the formation of the overall assembly.

Chapter 2.1.8. A4 Variants: Crystal Lattice-stabilizing Interactions

As mentioned in Chapter 2.1.6, the variant oligonucleotides formed crystals of different morphologies and space groups, as compared to each other. Only the A4T variant crystallized in the same space group as the native oligonucleotide (Table 2.1). As a result, the A4T fold-back tetramer units in the crystal lattice are organized in the same way as that of the native (Figure 2.11a). Similar to the bulged A4 residue in the native tetramer, the bulged T4 nucleobase in the A4T variant forms reciprocal lone pair/base stacking interactions with a crystallographically equivalent T4 residue (Figure 2.11b). This is the primary lattice-stabilizing interaction since the point mutation from an adenosine to a thymidine results in the loss of a hydrogen bond donor that could otherwise form a hydrogen bond with residue A5.

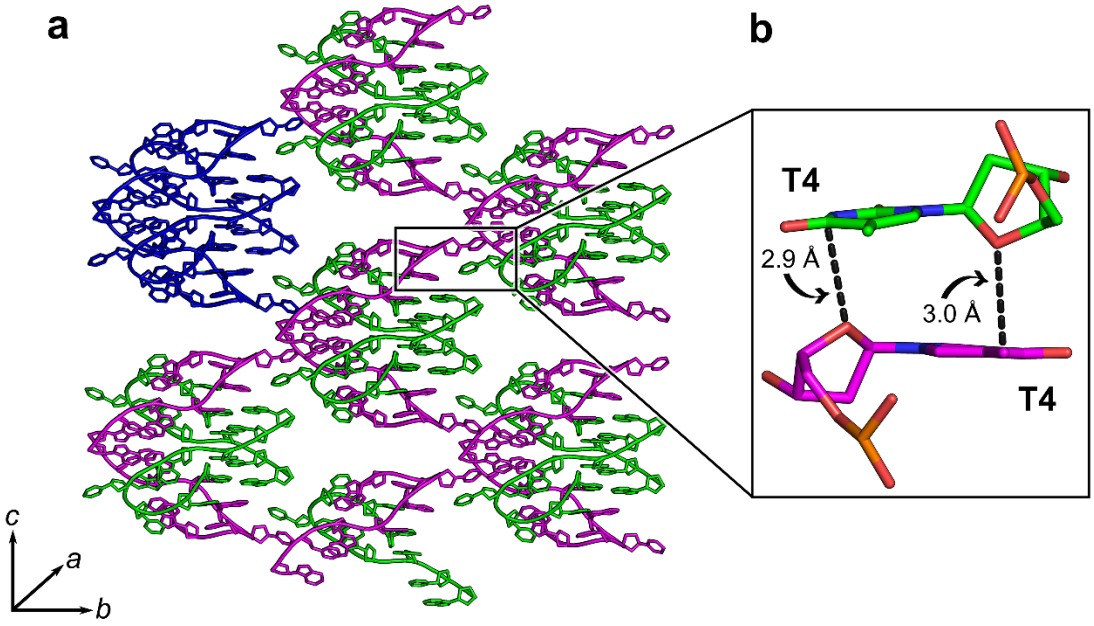

Figure 2.11. A4T Crystal Contacts. (a) Tetramer units that make up the crystal lattice along the *b* and *c* dimensions of the monoclinic unit cell are shown. One unit is colored in blue to distinguish the tetrameric assembly. **(b)** The inset shows interactions between T4 and residues from symmetry-related molecules. Residue T4 forms reciprocal lone pair/base stacking interactions between the O4′ atom and the nucleobase with a crystallographically-equivalent T4.

On the contrary, the mutation to cytidine revives the single hydrogen bond with the A5 residue, which is formed between the N4 hydrogen bond donor of the bulged C4 and the N3 hydrogen bond acceptor of A5 (Figure 2.12a,b). This bulged pyrimidine is also stabilized through base stacking interactions with residue G9 from a symmetry-related molecule (Figure 2.12b). Distinct from the native oligonucleotide, the A4C variant was crystallized in a trigonal, instead of a monoclinic, space group. The three-fold symmetry down the *c* axis of the unit cell can be seen in the organization of the A4C tetramer units (Figure 2.12a).

Differing from the native oligonucleotide, the A4G variant was crystallized in an orthorhombic space group, which indicates two-fold symmetry along all three axes

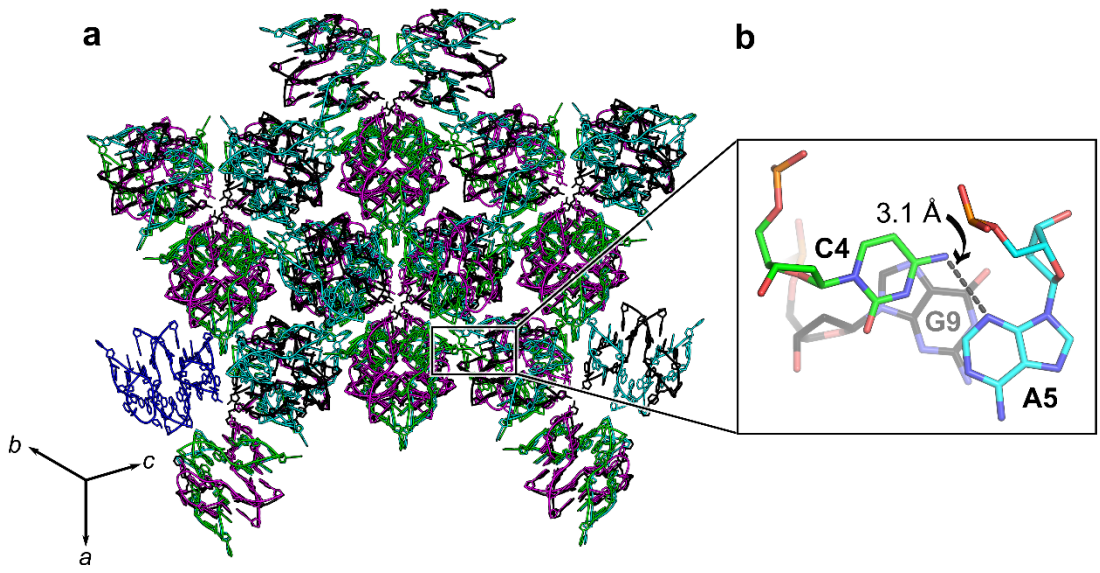

Figure 2.12. A4C Crystal Contacts. (a) Tetramer units that make up the crystal lattice along the *a* and *b* dimensions of the trigonal unit cell are shown. One unit is colored in blue to distinguish the tetrameric assembly. **(b)** The inset shows interactions between C4 and residues from symmetry-related molecules. Residue C4 forms a single hydrogen bond with an A5 residue through the N4─N3 Watson-Crick/sugar faces, respectively. The pyrimidine ring also forms base stacking interactions with the indole ring of G9.

of the unit cell. The two-fold symmetry down the *b* axis of the unit cell is observed in the A4G oligonucleotides that make up the crystal lattice (Figure 2.13a). The primary crystal lattice-stabilizing interaction is observed in the base stacking of the bulged guanosine residue. Interestingly, the A4G tetramer contains four G4 residues that are each involved in different stacking interactions. The G4 residues from Chains A and D are involved in continuous base stacking interactions through four layers (Figure 2.13b). In Chain A, the entire guanosine residue, including the sugar and phosphate groups, is modeled in two conformations, each at 50% occupancy. These two conformations are positioned on parallel planes with one above the other, but the nucleobases are shifted so that they do not stack directly with each other. One of these

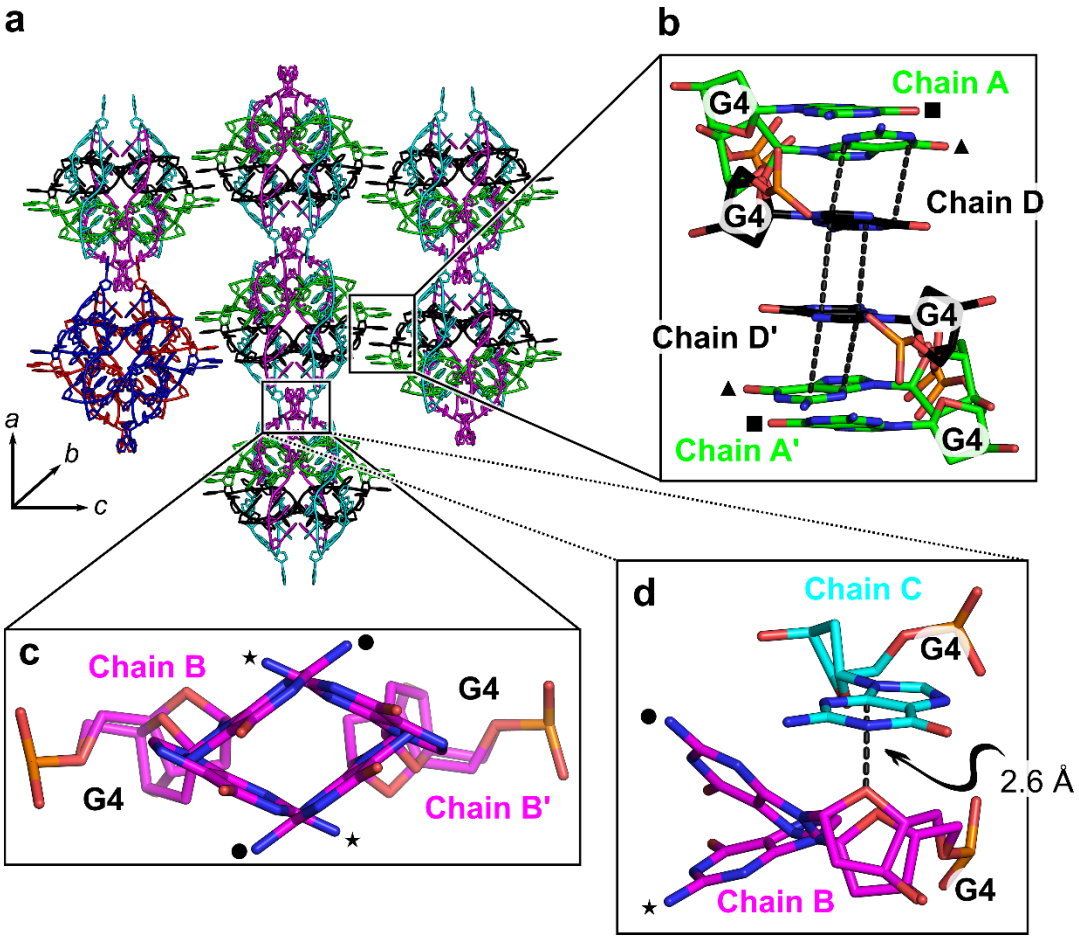

Figure 2.13. A4G Crystal Contacts. (a) Tetramer units that make up the crystal lattice along the *a* and *c* dimensions of the orthorhombic unit cell are shown. One unit is colored in blue to distinguish the tetrameric assembly. **(b)** Base stacking interactions between four layers of G4 residues from symmetry-related molecules are observed. Residue G4 from Chain A (green) adopts two conformations, each at 50% occupancy, and are denoted as a black square or triangle. The conformation denoted as a triangle stacks with G4 from Chain D (black). Chain D's G4 forms base stacking interactions with a crystallographically equivalent G4 from Chain D′. **(c)** Chain B's G4 residue (magenta) is modeled with two conformations, each at 50% occupancy, and are distinguished by a black circle or star. Each conformation stacks with its equivalent from Chain B′. **(d)** The nucleobase of G4 from Chain C (cyan) stacks with the lone pair from the O4′ atom of the G4 conformation denoted as a black circle in Chain B.

conformations is involved in base stacking interactions with residue G4 from Chain D at ~3.2 Å. This residue, in turn, stacks with an equivalent G4 from Chain D′ at ~3.5 Å, which is within 3.2 Å of the G4 conformation designated by a triangle from Chain A′.

Chain B's guanosine residue also adopts two conformations, but they are starkly different from those in Chain A. Here, the two conformations, which are denoted by a black circle or star, begin to diverge at the O5′ atom and the nucleobase planes are oriented 45° to each other (Figure 2.13c). Although these two conformations seem to clash, each has 50% occupancy and are mutually exclusive. For example, when G4 adopts the conformation designated by the black circle, the space occupied by the conformation indicated by the black star is not in use, eliminating the apparent steric clash. Each conformation is stabilized through base stacking interactions with its equivalent from Chain B′ at ~3.1 Å. Lastly, Chain C's G4 nucleobase exists in one conformation and forms lone pair/base stacking interactions with the O4′ atom of one of Chain B's G4 conformations (Figure 2.13d). But distinct from the crystal contacts in the native and A4T oligonucleotides, the lone pair/stacking interaction in the bulged guanosine residue is not reciprocal, which could be due to the different space group and packing of the oligonucleotide in the crystal lattice.

Chapter 2.2: Summary and Implications

In this chapter, I have provided a structural analysis on the d(CGTAAGGCG) and the A4 variant oligonucleotides. The structures described here expand the diversity of DNA fold-back motifs in several key ways. First, they indicate that there is significant sequence diversity in the fold-back core. The A:G:G:A slipped tetrad demonstrates a previously unknown base pair at the position 3′ of the bulged nucleotide, suggesting that an even larger number of potential fold-back sequences may be present in genomes. Second, the variant structures confirm that the bulged nucleotide position can accommodate any nucleobase identity. Perhaps most

interesting is that this is the first description of a fold-back motif capable of forming a tetrameric structure through the sequences flanking the fold-back core motif. This suggests that other diverse tertiary structures can be mediated by the fold-back quadruplex and opens the possibility of fold-back motifs being involved in processes that require the formation of DNA ternary structures, such as recombination or repair.

In the next chapter, I will discuss the structural characteristics of d(CGTAAGGCG) and the three A4 variant oligonucleotides in the solution state to provide an additional perspective on these structures.

Chapter 2.3: Experimental Procedures

Chapter 2.3.1. Oligonucleotide Synthesis and Purification

d(CGTAAGGCGTA), d(CGU^{Br}AAGGCGTA), and d(CGTAAGGCG) were synthesized using standard phosphoramidite chemistry on an Expedite 8909 Nucleic Acid Synthesizer (PerSeptive Biosystems, Inc.), with reagents from Glen Research (Sterling, VA). DNA oligonucleotides were purified using the Glen-Pak cartridges according to the manufacturer's protocol. The A4C, A4G, and A4T truncation variants were each purchased on the 1 μmol scale from Integrated DNA Technologies (Coralville, IA).

Chapter 2.3.2. Crystallization

d(CGTAAGGCGTA), d(CGU^{Br}AAGGCGTA), and d(CGTAAGGCG) were each crystallized by mixing 2 μL of 500 μM DNA solution with 1 μL of crystallization solution (100 mM magnesium chloride, 6 mM hexamminecobalt(III) chloride, and 30 mM sodium cacodylate buffer at pH 6.4). A4C, d(CGTCAGGCG), was crystallized by

mixing 2 μL of 500 μM DNA solution with 2 μL of crystallization solution (10% 2-methyl-2,4-pentanediol (MPD), 60 mM calcium chloride, 10 mM potassium chloride, and 30 mM sodium cacodylate at pH 7.0). A4G, d(CGTGAGGCG), was crystallized by mixing 1 μL of 500 μM DNA solution with 1 μL of crystallization solution (15% polyethylene glycol 400 (PEG400), 120 mM magnesium chloride, and 30 mM sodium cacodylate at pH 6.4). A4T, d(CGTTAGGCG), was crystallized by mixing 2 μL of 500 μM DNA solution with 1 μL of crystallization solution (15% MPD, 120 mM barium chloride, and 30 mM sodium cacodylate at pH 6.4). Crystallization was performed in sitting drops, which equilibrated against 300 μL of the corresponding 20% MPD or PEG400 in the well reservoir and were incubated at 22°C. Crystals were observed 2 days after plating.

Chapter 2.3.3. Data Collection

Diffraction data for d(CGTAAGGCGTA) and d(CGU^{Br}AAGGCGTA) were collected at the Advanced Photon Source (APS) 24-ID-C beamline at Argonne National Laboratory. Diffraction data for d(CGTAAGGCG) were collected at the APS 22-BM, using 0.5° rotation angles with an exposure time of 0.5 seconds. Diffraction data for A4C, A4G, and A4T variants were collected at the APS 23-ID-D.

Chapter 2.3.4. Structure Determination

Diffraction data for d(CGTAAGGCGTA), and its derivative, d(CGU^{Br}AAGGGCTA), were indexed and integrated using iMosflm.[76] From the derivative data, initial phases were determined by SAD phasing using AutoSol in Phenix.[77] Diffraction data for the truncated oligonucleotide, d(CGTAAGGCG), were

indexed and integrated using iMosflm, and truncated in Aimless.[78] A fold-back monomer generated from the full-length structural refinement was used as a molecular replacement search model in Phaser.[79] Further refinement was carried out in Phenix and additional model building was performed in Coot.[80] Data processing for A4C, A4G, and A4T was carried out in XDS[81] and Aimless, followed by molecular replacement with Phaser using the refined truncated structure as a search model. Subsequent refinement was performed in Refmac.[82-84] Final refinement statistics are shown in Table 2.1. The final models were analyzed using Pymol and are deposited in the Protein Data Bank (PDB).

Chapter 3: d(CGTAAGGCG) Solution Characterization

This chapter is adapted from "Chu, B.; Zhang, D.; Hwang, W.; Paukstelis, P.J., Crystal Structure of a Tetrameric DNA Fold-back Quadruplex. *J. Am. Chem. Soc. 140* (47): 16291 – 16298." D. Zhang acquired the nuclear magnetic resonance spectroscopy data and conducted the initial partial peak assignment. W. Hwang collected the small angle X-ray scattering data.

Chapter 3.1: Motivation for Solution State Characterization

In Chapter 2, I described the structural characteristics of the crystal structures, including the crystal packing interactions, of d(CGTAAGGCG) and the three variant oligonucleotides. These crystal structures have captured just one of many possible conformations of these oligonucleotides. It is possible that the structures adopted in the solid state could deviate from those in solution since additional crystal lattice-stabilizing interactions are necessary for ordered assembly in the crystal lattice, whereas the oligonucleotides can be oriented randomly in solution. While we observed the fold-back tetramer in the solid state, we are interested in probing to see if the same structural assembly is observed in the solution state. The formation of the fold-back tetramer in solution could hint at its potential biological significance.

Chapter 3.2: Results and Discussion

Chapter 3.2.1: Divalent Cation Dependence

To explore the importance of divalent cations in structure formation, I conducted circular dichroism (CD) and one-dimensional proton nuclear magnetic resonance (1D ^{1}H-NMR) spectroscopy. CD spectra from a Mg^{2+} titration (up to 0.5 M) showed the appearance of a negative band at ~245 nm and positive bands at 260 and

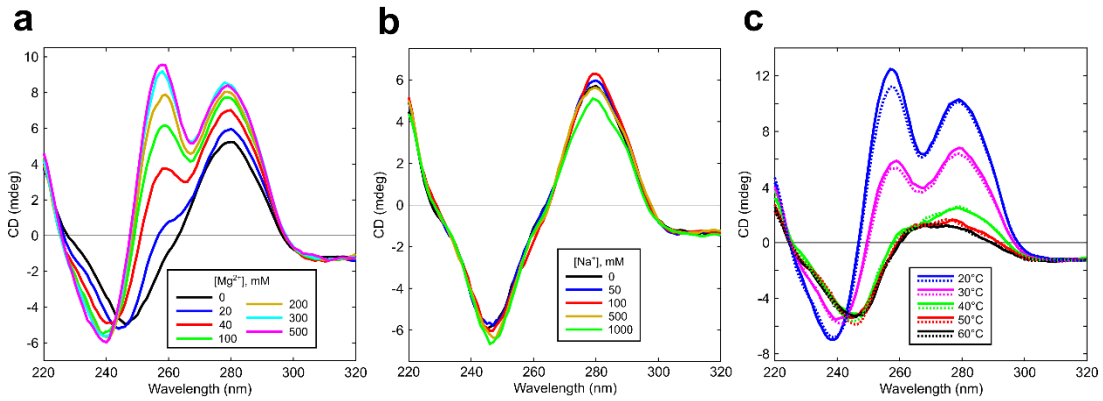

Figure 3.1. Cation Titrations and Thermal Denaturation of d(CGTAAGGCG) Monitored by CD. (a) Mg^{2+} titration up to 0.5 M and **(b)** Na^{+} titration up to 1 M at room temperature. **(c)** Heating (solid) and cooling (dotted) CD melting curves at 100 mM Mg^{2+} collected at 10°C increments from 20°C to 60°C. All samples were prepared with a final DNA concentration of 100 µM in sodium cacodylate buffer at pH 6.4 and/or supplemented with the specified cation.

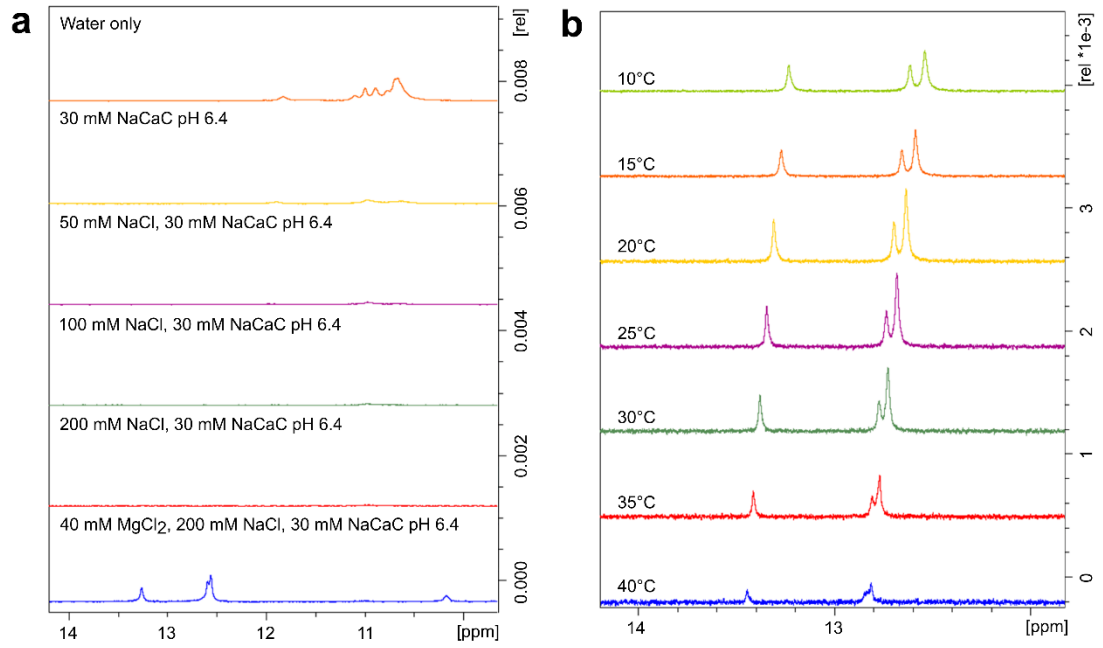

Figure 3.2. Imino Regions of 1D ^{1}H-NMR. (a) Na^{+} titration of 350 µM A4T up to 200 mM at 10°C. Signals observed at 10–12 ppm (orange) likely correspond to weak noncanonical base pairing, which are disturbed upon the addition of Na^{+}. Three imino signals appeared upon the addition of Mg^{2+} (blue). **(b)** Spectra of 500 µM d(CGTAAGGCG) in sodium cacodylate buffer at pH 6.4 containing 40 mM $MgCl_2$ collected at 5°C increments from 10°C to 40°C.

280 nm (Figure 3.1a). These characteristic bands appeared only in the presence of divalent cations and were not observed in Na^+ concentrations up to 1 M (Figure 3.1b). Similarly, 1D ^{1}H-NMR spectra revealed the absence of imino signals at 0.2 M Na^+, but upon addition of 0.04 M Mg^{2+}, imino signals were clearly observed at 12.5–13.5 ppm (Figure 3.2a). This transition indicates that monovalent cation (even at high concentrations) does not induce higher order assembly. CD melting analysis suggested that the characteristic bands were due to the formation of a specific structure, with nearly identical forward and reverse temperature dependence spectra (Figure 3.1c). Likewise, decreased intensity of imino signals in 1D ^{1}H-NMR spectra at elevated temperatures up to 40°C (Figure 3.2c) indicated a loss of higher order structure with increasing temperature.

CD spectra of the variant oligonucleotides showed the same characteristic bands in Mg^{2+} (Figure 3.3a), consistent with their structural similarity. The other divalent cations used for crystallizing the A4 variants, Ca^{2+} and Ba^{2+}, also induced the

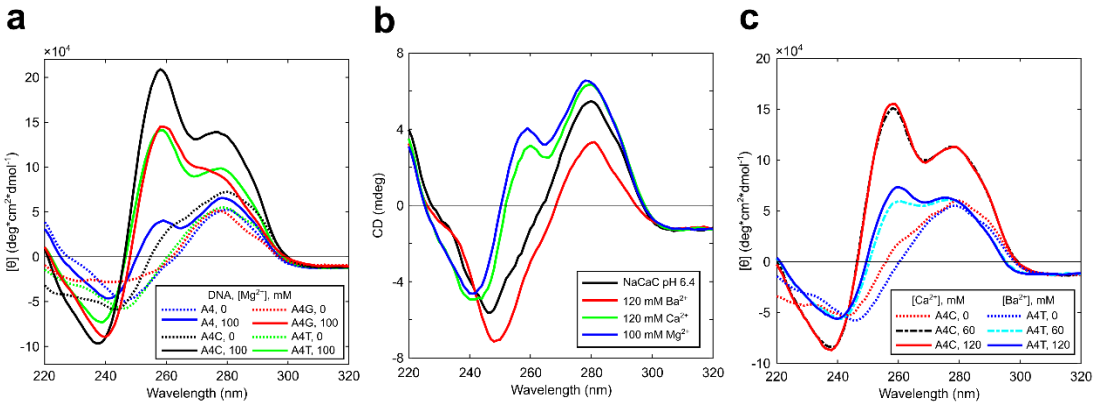

Figure 3.3 Comparison of Divalent Cation Effects on CD Spectra of d(CGTAAGGCG) and Variant Oligonucleotides. (a) Profile of d(CGTAAGGCG), A4C, A4G, and A4T in 0 or 100 mM Mg^{2+}. **(b)** Profile of d(CGTAAGGCG) in Ba^{2+}, Ca^{2+}, or Mg^{2+}. **(c)** Ca^{2+} titration of A4C and Ba^{2+} titration of A4T, each up to 120 mM. All spectra were collected in sodium cacodylate buffer at pH 6.4 at room temperature.

characteristic CD spectra. However, band intensity was reduced relative to Mg^{2+} (Figure 3.3b,c), while Ba^{2+} showed a preference for the A4T variant (Figure 3.3c). These results demonstrate the importance of divalent cations in the formation of CD-observable structure in solution. This appears to be a unique characteristic of this fold-back structure, as all previous CD characterization of fold-back motifs show structural transitions in the presence of only monovalent cations.[34, 44, 46, 48, 51, 85]

Chapter 3.2.2: Oligomeric Solution State Analysis via SAXS

CD and NMR clearly showed a structural transition upon the addition of divalent cations, but it was not clear if this transition reflected the dimerization or tetramerization observed in the crystal structure. I conducted small angle X-ray

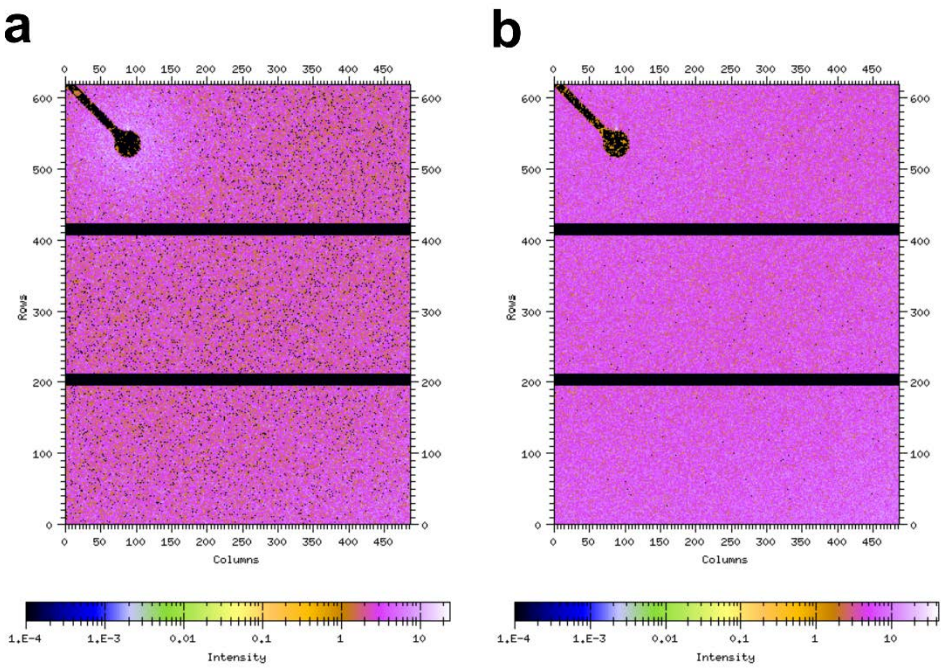

Figure 3.4. SAXS Scattering Images. (a) The average of three frames from the sample containing 1 mM d(CGTAAGGCG) in 100 mM $MgCl_2$ in 30 mM Bis-Tris pH 6.4 is shown. **(b)** The average of six frames from the reference sample, which contains 30 mM Bis-Tris pH 6.4 supplemented with 100 mM $MgCl_2$ is shown. The shadow of the beamstop is located at the upper left-hand corner of both images.

scattering (SAXS) experiments to determine the radius of gyration (R_g) of the DNA assemblies as a means of assessing the oligomeric state. Qualitative observation of the scattering images indicated that the concentration of the DNA in the prepared sample was sufficient for SAXS analysis. As expected, the white brilliance around the beamstop was observed only in the DNA sample (Figure 3.4a) and not in the buffer sample (Figure 3.4b).

The shape of the scattering curve indicated a monodisperse DNA solution free of large aggregates. We used the Guinier model for globular particles (qR_g<1.3) to fit the observed scattering curve in Igor,[86] which gave an R_g of 13.10 Å with a χ^2 of 2.49 (Figure 3.5), indicating that the particles have similar spherical shape. Analysis of the experimental scattering curve using Primus[87] estimated the R_g to be 12.83±0.71 Å. Both

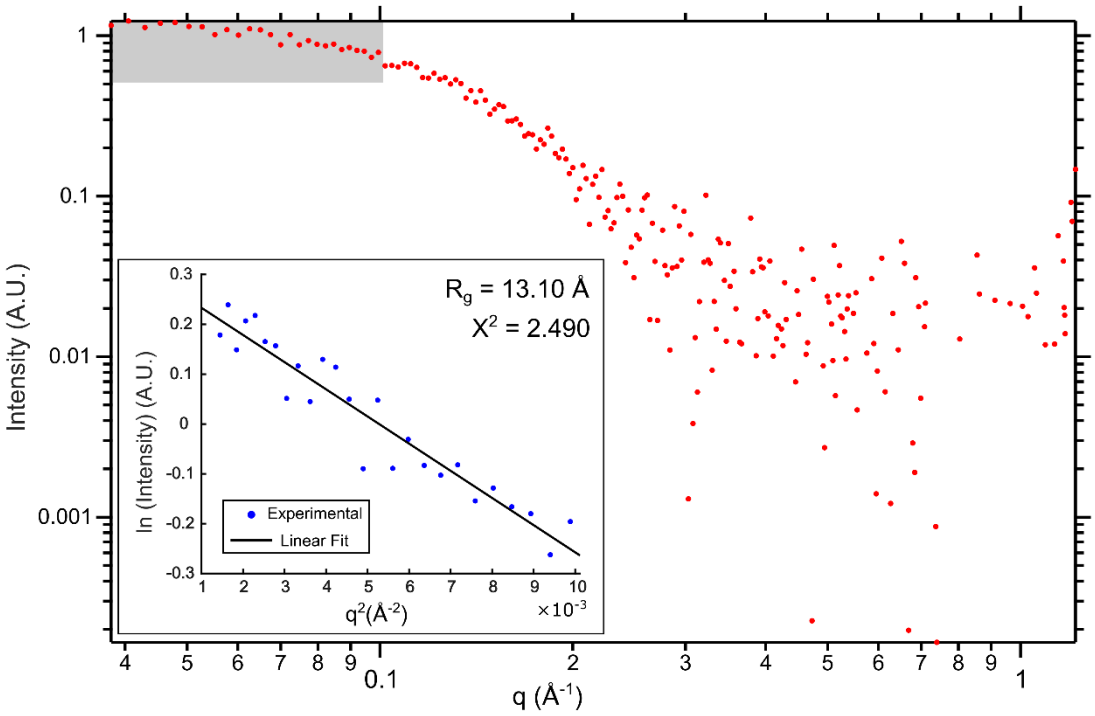

Figure 3.5. Experimental SAXS Scattering Profile of d(CGTAAGGCG). Raw values of I(*q*) vs. *q* are shown as red points. A Guinier Plot (inset) with a linear fit (solid black) was generated using values of *q* from the gray box.

of these values are in strong agreement with the calculated R_g of the tetrameric assembly at 13.06 Å (Figure 3.6a), as modeled by Crysol.[88] However, because of the large degree of structural overlap between the two dimers in the tetramer, the calculated R_g of the dimeric unit was not substantially different at 11.26 Å (Figure 3.6b). Potential ambiguity may also arise in this case because of the strong solvent and ion interactions with the polyanionic DNA backbone. It has been previously noted that it can be difficult to separate macromolecular scattering from solvent scattering around nucleic acids, which can hamper interpretation of the SAXS profile.[89] Given these caveats, however, the scattering curve is most consistent with the tetramer.

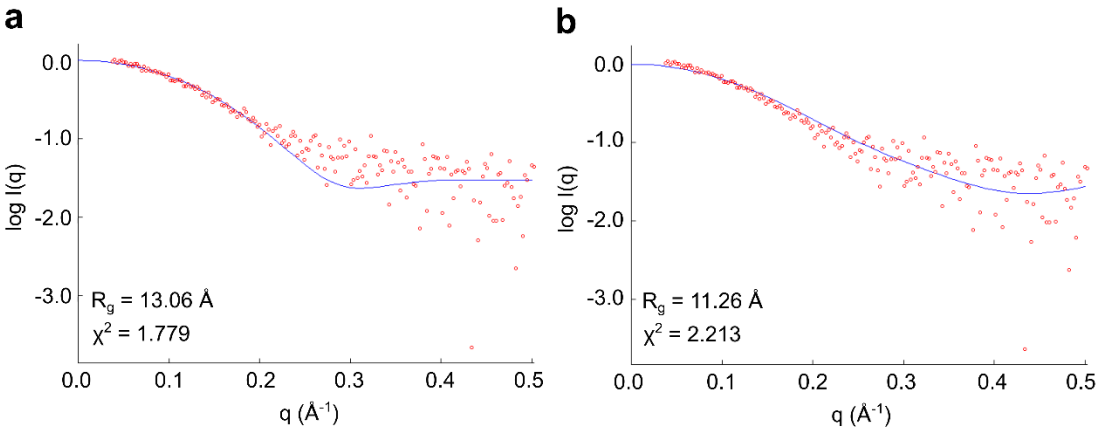

Figure 3.6. Modeling of d(CGTAAGGCG) Assemblies to Experimental SAXS Data. The **(a)** tetrameric and **(b)** dimeric assemblies were fitted to the experimental SAXS data. The CRYSOL algorithm calculated the goodness of fit (χ^2) and estimated R_g values for each assembly.

Chapter 3.2.3: Oligomeric Solution State Analysis via 2D NMR

To further address the apparent oligomeric state in solution, we conducted 2D-NOESY and 2D-TOCSY experiments on the native A4 and the A4T variant. Sequential assignment of the A4T spectra allowed us to identify most of the nucleobase and H1′

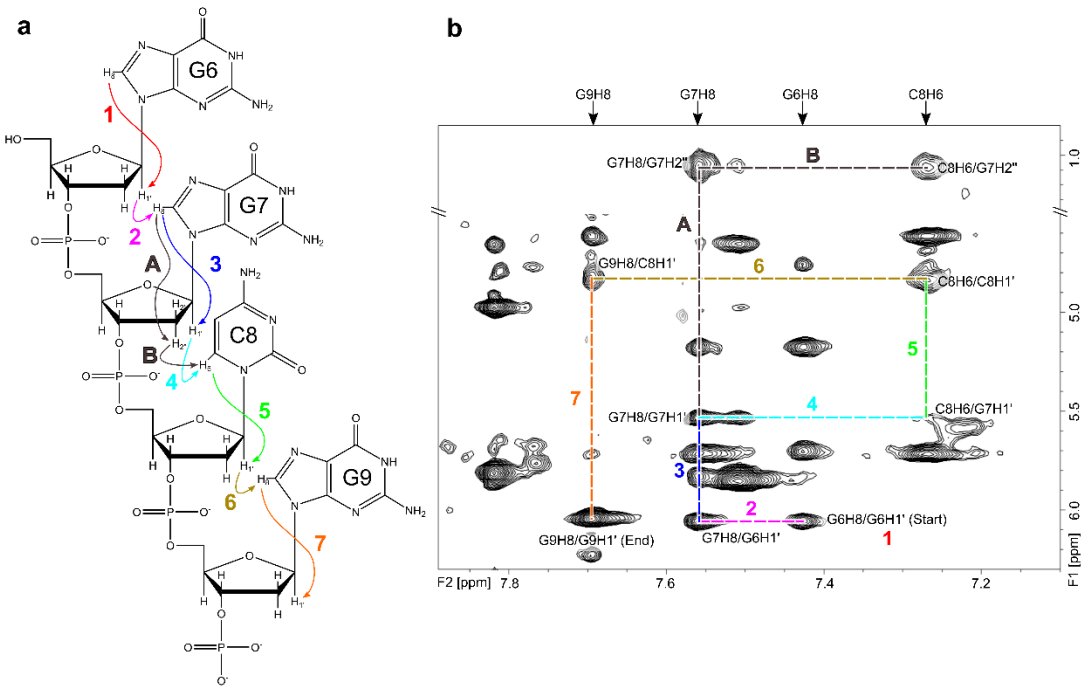

Figure 3.7. Sequential Intra-strand Connectivities for G6→G7→C8→G9. (a) Predicted resonances between non-exchangeable hydrogens are indicated by arrows. Each numbered arrow is shown in a different color. Arrows A and B, both colored gray, indicate predicted resonances of G7H8 with G7H2″ and G7H2″ with C8H6, respectively. **(b)** Non-exchangeable proton regions of the 2D-NOESY NMR spectra for A4T. Cross-peaks are labeled and indicated by intersecting lines. The colors of the lines and numbers correspond to those of the arrows in **(a)**. The G7H8/G7H2″ and C8H6/G7H2″ cross-peaks are specifically labeled to demonstrate the proximity of G7 and C8, since the C8H6/G7H1′ cross-peak is weak.

and H2′ protons. Similar to other linear DNA fold-back NMR spectra,[85] additional proton resonance signals were present, indicating multiple conformations. While this complicated performing a complete assignment, a substantial number of cross-peaks could be unambiguously assigned.

Sugar-to-base sequential connectivities were observed for the residues, G6→G7→C8→G9 (Figure 3.7), allowing for the identification of G9 nucleobase protons, including the assignment of the imino signal at 12.5 ppm as G9H1. Imino

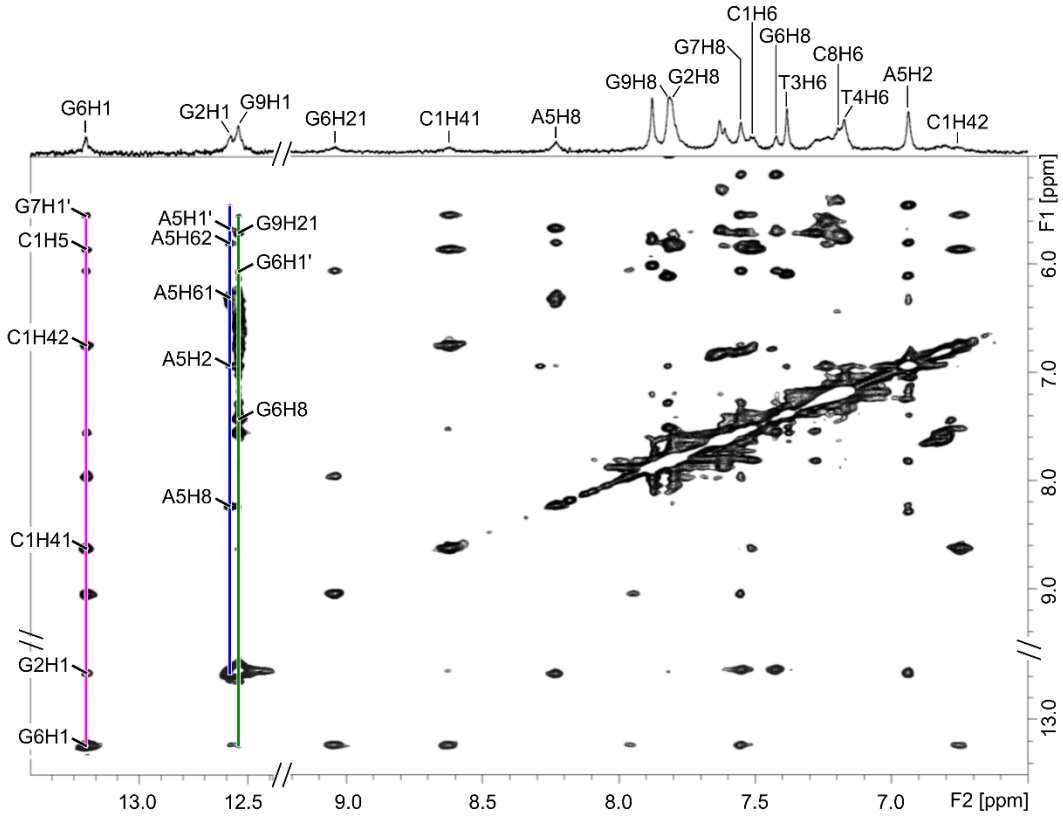

Figure 3.8. 2D-NOESY ^{1}H-NMR Spectra of A4T. The 1D projection of the peaks are labeled at the top of the spectra. Peak assignments indicate interactions between G6 and C1 (magenta), G2 and A5 (blue), and G6 and G9 (green), which are consistent with the crystal structure.

proton signals at 12.6 and 13.3 ppm were assigned to G2H1 and G6H1, respectively (Figure 3.8). These resonances verified key structural features and strongly suggest the formation of the tetramer in solution.

Multiple intrastrand resonances, C1H41/G7H1′, C1H42/G7H1′, and C1H5/G7H1′, confirmed the fold-back nature of the DNA (Figure 3.9a). These cross-peaks are consistent with the crystal structure in which the fold-back loop orients C1 and G7 close together, such that the nucleobase of C1 directly stacks above the sugar of G7 (Figure 3.9b). Additionally, the presence of the interstrand NOE cross-peak

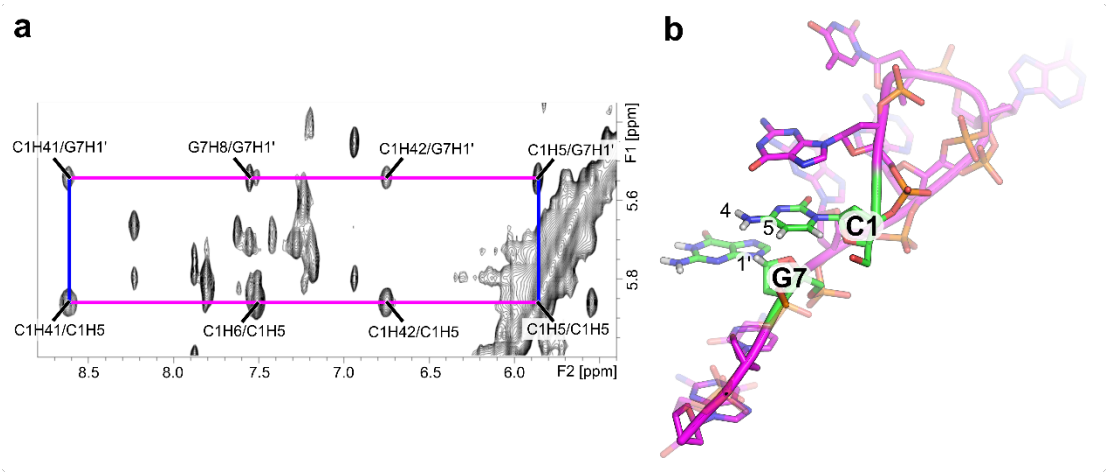

Figure 3.9. Intra-strand Cross-peaks Confirm the Fold-back Motif. (a) The three intra-strand cross-peaks, C1H41/G7H1′, C1H42/G7H1′, C1H5/G7H1′, and the intramolecular resonances, G7 H8 and H1′, C1 H41 and H5, C1 H42 and H5, and C1 H6 and H5, are labeled and indicated by intersecting lines. **(b)** Cartoon and stick representation of one chain of d(CGTAAGGCG). Residues C1 and G7 are within 5 Å of each other as a result of the fold-back architecture.

between H1 of G2 and H1 of G6 confirmed the overall arrangement of nucleotides (Figure 3.8, magenta). The base pairing arrangement of the fold-back core brings G2 and G6 within 5 Å of each other, accounting for this cross-peak. The G2 imino proton H1 showed cross-peaks with A5 amino protons (Figure 3.8, blue), indicating that G2 and A5 base pair through the Watson-Crick and Hoogsteen faces, respectively (Figure 3.10a). Consistent with the *syn* χ torsion angle of A5, the peak height of the A5H8/A5H1′ cross-peak was comparable to that of the C8H5/C8H6. Additionally, NOEs between A5H2 and both H2′ protons (Figure 3.10b) are consistent with A5 adopting the *syn* conformation. As expected, G6H1 showed cross-peaks with C1 amino protons (Figure 3.8, magenta), confirming the G6 and C1 base pair. Together, these demonstrate that many of the interactions observed in the crystal structure are present in solution.

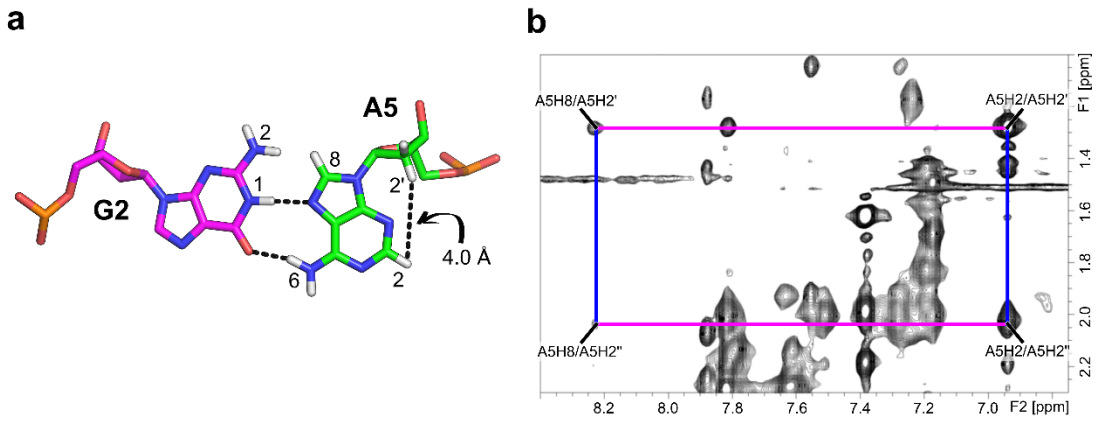

Figure 3.10. Intramolecular Cross-peaks Confirm the A5 *Syn* Conformation. (a) Stick representation of the G2─A5 base pair. The A5 H2 and H2′ protons are ~4.0 Å apart. **(b)** The intramolecular A5 H2 and H2′, A5 H2 and H2″, A5 H8 and H2′, and A5 H8 and H2″ cross-peaks are labeled and indicated by intersecting lines.

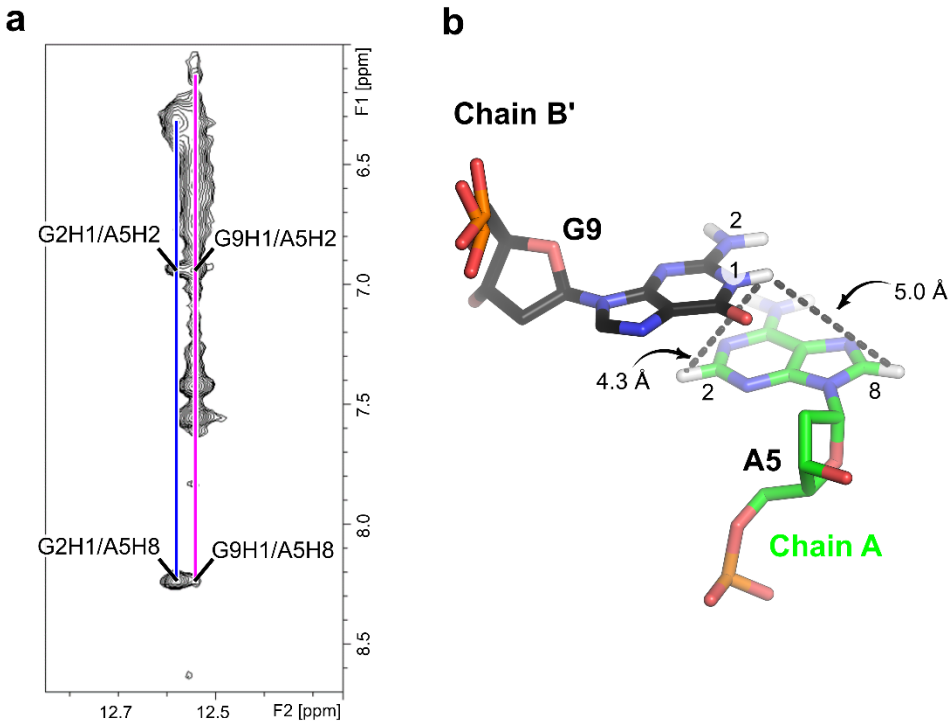

Figure 3.11. Inter-strand Cross-peaks Demonstrate Proximity Between the G9 and A5 Residues. (a) The inter-strand cross-peaks, G9H1/A5H2, G9H1/A5H8, G2H1/A5H2, and G2H1/A5H8, are labeled. **(b)** Residue A5 from Chain A (green) and residue G9 from Chain B′ (black) are shown as stick representations. Distance measurements from G9H1 to A5H2 and to A5H8 are labeled as 4.3 Å and 5.0 Å, respectively. These values are consistent with the relative intensities of their respective cross-peaks shown in **(a)**.

An intense cross-peak between the G9 imino proton H1 and H8 of G6 (Figure 3.8, green) provides direct evidence for the G6 and G9 base pair, supporting the formation of the tetrameric structure. This cross-peak would not be observed in the dimeric structure since the two guanosine residues are separated by a distance greater than 5 Å. Furthermore, the A5H2/G9H1 and A5H8/G9H1 interstrand NOEs (Figure 3.11a) support tetrameric formation since the partial stacking between the pyrimidine rings of A5 and G9 in the crystal structure orients G9H1 in proximity to A5H2, but farther from A5H8 (Figure 3.11b). Thus, the relative intensities are consistent with the *syn* conformation of A5.

Chapter 3.2.4: Preliminary Studies on Structural Dynamics and Equilibrium

To begin to investigate the structural dynamics of d(CGTAAGGCG), I conducted a time course study monitored by CD. The characteristic CD profile showing a negative band at ~245 nm and positive bands at 260 and 280 nm was observed even

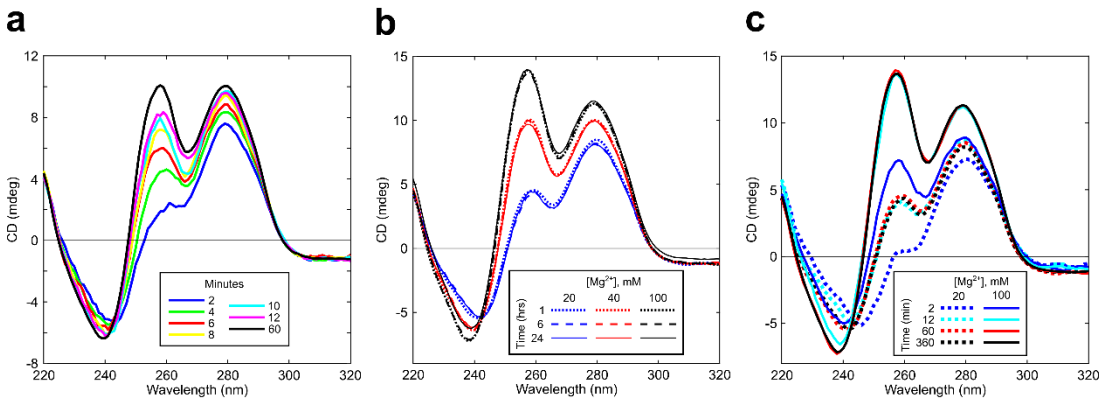

Figure 3.12. Time Course Study of d(CGTAAGGCG) Monitored by CD. (a) After adding 100 µM DNA to sodium cacodylate buffer at pH 6.4 containing 40 mM $MgCl_2$, CD profiles were collected at 2-minute intervals for the first 12 minutes and then at 60 minutes. **(b)** CD profiles corresponding to at 1, 6, and 24 hours are shown in three buffer conditions (20, 40, and 100 mM Mg^{2+}). **(c)** CD spectra corresponding to 2, 12, 60, and 360 minutes are shown for 20 mM (dotted) and 100 mM (solid) Mg^{2+} conditions.

after 2 minutes following the addition of 40 mM Mg^{2+} to the DNA (Figure 3.12a). Although weak at 2 minutes, the intensity of the positive 260 nm peak gradually increased in the first 12 minutes and reached a maximum at 60 minutes. Spectra collected at 6 and 24 hours showed no further increase in peak intensity (Figure 3.12b), This shows that saturation was reached after one hour and suggests that the self-assembly into the fold-back tetramer is a relatively slow process. The same trend was observed at two other Mg^{2+} concentrations, 20 mM and 100 mM (Figure 3.12c),

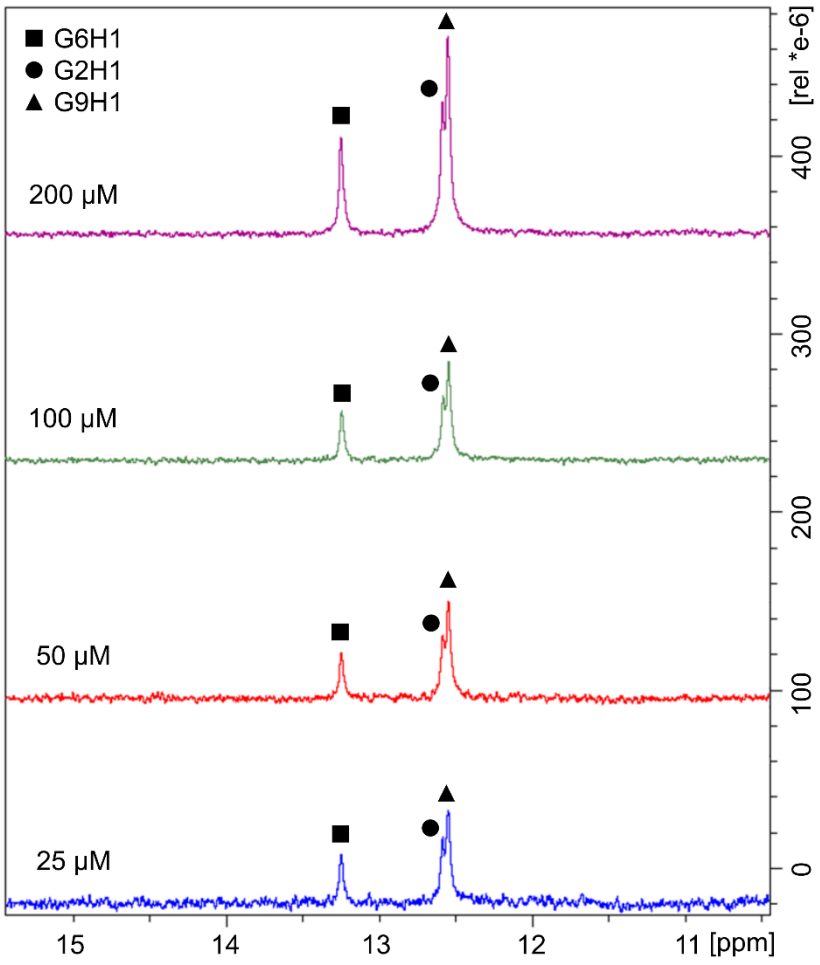

Figure 3.13. Dilution of A4T Oligonucleotide Monitored by 1D ^{1}H-NMR. A4T was prepared at 200 µM in sodium cacodylate buffer at pH 6.4 containing 40 mM $MgCl_2$. As DNA concentration is diluted to 25 µM, the intensities of all three imino signals weaken.

demonstrating that cation concentration does not affect the rate of structure formation, but does affect the intensity of the characteristic 260 nm peak, since higher Mg^{2+} corresponds to greater peak intensity.

In an attempt to distinguish the dimeric and tetrameric populations, I monitored the imino regions of 1D ^{1}H-NMR spectra of the A4T oligonucleotide as I lowered the DNA concentration. We hypothesized that the peak assigned to G9H1 would be absent or have decreased intensity if populations of the dimeric assembly were present. However, we observed all three imino peaks at low (25 μM) oligonucleotide concentrations (Figure 3.13), suggesting that the tetramer can form well below crystallization concentrations. Thus, the equilibrium between the dimeric and tetrameric populations remains unclear. Though we cannot entirely rule out the possibility of an alternate dimeric structure in solution that brings G6 and G9 in proximity, all available evidence from SAXS, CD, and NMR analyses strongly suggest that d(CGTAAGGCG) forms a tetramer in solution.

Chapter 3.3: Significance and Biological Implications

There is mounting evidence that non-Watson-Crick DNA structures have important roles in biology. Once considered a "DNA oddity", G4-type structures now clearly have a role in the regulation of gene expression at both the DNA and RNA level and in maintaining genome stability.[90, 91] Though genomic repeats of varying complexity have long been considered a potential source for the formation of noncanonical DNA structures,[92-94] the types of DNA structure they form are largely unknown. The demonstration that sequences capable of forming fold-back structures are clustered near promoters of developmental genes has highlighted the possible

biological importance for these structures.[51] The results described here adds to this idea, as I have shown four DNA sequences that self-assemble into the fold-back tetramer. This significantly expands the diversity of sequences that can form fold-back structures and suggests that many more sequences in genomes can form this motif.

Finally, CD analysis showed that only divalent cations induced structural assembly into the fold-back tetramer. The cation dependence of d(CGTAAGGCG) and variant oligonucleotides may offer a new tool for rationally designing DNA nanoarchitectures or sensors. Modification of the nucleobases oriented toward the cation binding pocket may allow for tuning of size and shape to bind specific cations, or to design specific interactions to enable binding and detection of small molecule ligands. This motif may also be useful in creating rationally designed DNA nanostructures. The ability to control the formation of a junctional motif by changing environmental conditions, in this case by the addition of specific cations, would allow for stepwise self-assembly processes with the potential to both diversify the types of structures that could be made and improve overall yield.

Chapter 3.4: Experimental Procedures

Chapter 3.4.1: Circular Dichroism (CD) Spectroscopy

CD spectra were acquired using the Jasco J-810 spectropolarimeter fitted with a thermostated cell holder. Samples were prepared with a DNA concentration of 100 μM in 30 mM sodium cacodylate buffer at pH 6.4 containing varying concentrations of divalent cations. Details for each specific experiment are provided in the figure captions. Samples were equilibrated for 15–20 hours at 4°C prior to the acquisition of

the spectra. All spectra were collected in a 1.0 mm path length cuvette at room temperature from 220 to 320 nm with a data pitch of 1.0 nm and a rate of 50 nm/minute.

For melting experiments, the sample was prepared at 100 mM $MgCl_2$ and spectra were collected at 10°C increments from 20°C to 60°C. The sample was allowed to dwell for 10 minutes at each temperature set point. For the time course studies, the DNA and buffer samples containing 20, 40, or 100 mM $MgCl_2$ were prepared in separate tubes. At the start of the experiment, the DNA was added to the buffer and quickly transferred to the cuvette. Spectra were continuously collected for the first 12 minutes at 2-minute intervals since the time lag in data collection was ~2 minutes. Spectra were also collected at 1, 6, and 24 hours.

Chapter 3.4.2: Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR data were acquired on a Bruker Avance III HD 800 MHz spectrometer equipped with a Cryo-QCI probe and a Bruker Avance III 600 MHz spectrometer equipped with a Cryo-TCI probe. 1D ^{1}H-NMR spectra were collected for both the native and A4T oligonucleotides (at differing concentrations) in sodium cacodylate buffer at pH 6.4, containing 7% D_2O and varying amounts of cations. All 1D spectra were collected at 10°C and analyzed in TopSpin. Details for each specific experiment are provided in the figure captions. For 2D experiments, the native A4 and A4T oligonucleotide samples were prepared at 500 and 233 μM, respectively, both in 30 mM sodium cacodylate buffer at pH 6.4, containing 40 mM $MgCl_2$, and 7% D_2O. A combination of 2D-NOESY and 2D-TOCSY experiments were performed at 10°C, in which the mixing time was set to 300 and 80 ms, respectively. For both experiments, multiple mixing times were initially tested and the ones that yielded spectra with the

best resolution were selected. The oligonucleotide sequential assignment was conducted using the Computer Aided Resonance Assignment (CARA) program.[95]

2D-NOESY and 2D-TOCSY spectra for the native d(CGTAAGGCG) and A4T oligonucleotides were used to generate proton assignments. Identification of NOEs between nucleobase protons and sugar H1′ and H2′ protons in the 7–9 ppm region established sequential connectivity between multiple residues. Nucleobase protons with characteristic chemical shifts could be assigned. The cytosine H5/H6 NOE is a characteristic peak in the 2D-TOCSY spectrum, allowing cytosines to be distinguished from the other three nucleotides. Comparison between the native and A4T spectra was particularly helpful in the assignment of several A5 nucleobase protons due to the differences in the number of peaks in distinct chemical shift regions. Further, proton signals in the 5–6 ppm range were used to verify the initial assignments in the nucleobase proton region. Imino signals were then identified based on NOEs between the protons on its own ring and those from their base pairing partner.

Chapter 3.4.3: Small Angle X-ray Scattering (SAXS)

SAXS experiments were performed on the Xenocs Xeuss system with a CuKα X-ray source (λ = 1.5418 Å, GeniX3D Cu ULD, Xenocs, SA, France) at 23°C. The oligonucleotide was prepared at 1 mM in 30 mM Bis-Tris buffer at pH 6.4, containing 100 mM $MgCl_2$. Samples were recorded in a 1.0 mm capillary flow cell under vacuum. Scattering data were collected on a Pilatus 300 K (DECTRIS, Switzerland) over 6 frames with a 10-minute acquisition time for each frame at a distance of 370 mm. Measurements were corrected for background air scattering, and beam intensity.

Scattering images were analyzed using Igor Pro software (ver. 6.37) with the Irena package to obtain circular averaged 1D plots of intensity vs. scattering wave vector q.[86] The ATSAS package was used to evaluate the experimental solution scattering data.[96]

Chapter 4: d(CCAGGCTGCAA) Crystal Structure Analysis

This chapter is adapted from "Chu, B.; Zhang, D.; Paukstelis, P.J., A DNA G-quadruplex/i-motif hybrid. *Nucleic Acids Res. 47* (22): 11921 – 11930."

Chapter 4.1: Results and Discussion

Chapter 4.1.1. Overview

As part of a screen to probe the structural diversity of DNA, I have crystallized many short DNA oligonucleotides, including d(CCAGGCTGCAA). Its structure was determined by SAD phasing using a 5-bromo-deoxyuridine substitution at the T7 position. Both the native, d(CCAGGCTGCAA), and U7-Br, d(CCAGGCU^{Br}GCAA) oligonucleotides were crystallized in the shape of a diamond (Figure 4.1a,b). Initial phases from the U7-Br derivative were used to create electron density maps for the higher resolution native structure (Table 4.1). Refined native and derivative structures were virtually identical, with an RMSD of 0.377 Å for all DNA atoms of the asymmetric unit.

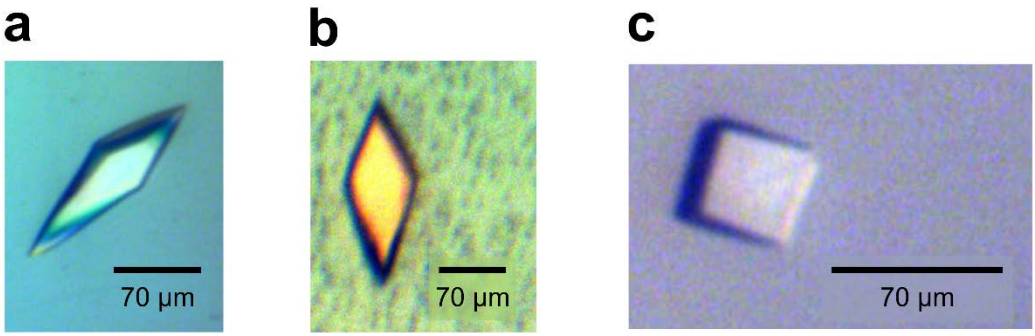

Figure 4.1. Crystals of d(CCAGGCTGCAA) and Derivative Oligonucleotides. (a-b) The native **(a)** and U7-Br **(b)** oligonucleotides grew as crystals in a diamond-shaped morphology, with the longest dimension measuring 160 µm and 150 µm, respectively. **(c)** The C9-Br oligonucleotide crystallized as cubic crystals, with each side measuring 40 µm.

Table 4.1. Data Collection and Refinement Statistics for d(CCAGGCTGCAA) and Derivatives.

	Native	**U7-Br Derivative**	**C9-Br Derivative**
PDB ID	6TZQ	6TZR	6TZS
Sequence	d(CCAGGCTGCAA)	d(CCAGGCU^{Br}GCAA)	d(CCAGGCTGC^{Br}AA)
Beamline	NE-CAT 24-ID-C	NE-CAT 24-ID-C	SER-CAT 22-BM
Data Collection			
Space Group	$P3_221$	$P3_221$	$I4_122$
Cell Dimensions			
a, b, c (Å)	37.38, 37.38, 98.65	37.10, 37.10, 98.31	51.52, 51.52, 112.95
α, β, γ (°)	90, 90, 120	90, 90, 120	90, 90, 120
Resolution (Å)*	49.33 – 2.29 (2.37 – 2.29)	32.77 – 2.40 (2.49 – 2.40)	55.80 – 2.60 (2.72 – 2.60)
R_{meas} (within I+/I-)*	0.142 (1.708)	0.083 (0.917)	0.177 (1.889)
R_{meas} (all I+ and I-)*	0.142 (1.702)	0.090 (0.917)	0.182 (1.895)
R_{pim} (within I+/I-)*	0.045 (0.543)	0.037 (0.399)	0.072 (0.712)
R_{pim} (all I+ and I-)*	0.035 (0.400)	0.032 (0.298)	0.0058 (0.526)
No. of unique *	3955 (373)	3399 (358)	2540 (294)
I / σ I*	8.2 (1.4)	11.0 (1.8)	11.5 (1.2)
Completeness (%)*	99.9 (100.0)	99.9 (100.0)	100.0 (100.0)
Multiplicity*	17.4 (18.0)	8.9 (9.4)	12.2 (12.7)
Wavelength (Å)	0.9196	0.9196	0.9187
Phasing			
Atom/Sites		Br/2	Br/2
CFOM from SHELX[97]		72.7	84.4
Refinement			
Resolution (Å)*	32.88 – 2.29 (2.35 – 2.29)	32.15 – 2.40 (2.46 – 2.40)	46.78 – 2.60 (2.67 – 2.60)
No. reflections	3560	3040	2272
No. reflections used in R_{free} Test Set	373	338	254
R_{work}**	0.2223	0.2121	0.2665
R_{free}**	0.2551	0.2538	0.3165
$R_{complete}$**	0.2551	0.2516	0.3156
Total No. of atoms	446	445	453
Average B-factors (Å^2)	71.434	79.866	46.873
RMS deviations			
Bond lengths (Å)	0.0055	0.0062	0.0076
Bond angles (°)	1.5652	1.6400	1.6148

*Values in parentheses correspond to the high-resolution shell.
**R_{work}, R_{free}, and $R_{complete}$ values are from 10-fold cross-validation in PDB-REDO.[98]

The asymmetric unit contains two molecules (Chains A and B) that interact as a dimer. The two monomers show a large degree of structural similarity in the first five residues (RMSD = 0.757 Å for 84 atoms), with the largest deviation arising from the

differing conformations of the A3 nucleobase (Figure 4.2a). However, the latter half of the chain contains significant conformational differences in both the backbone and nucleobase atoms (Figure 4.2b). Two dimers interact through crystal symmetry (symmetry molecules designated as Chains A′ and B′) to form a tetramer. This tetramer contains a number of distinct structural motifs, including a central G-quadruplex, a base triple interaction, a structurally variable spacer region, and a terminal i-motif (Figure 4.3).

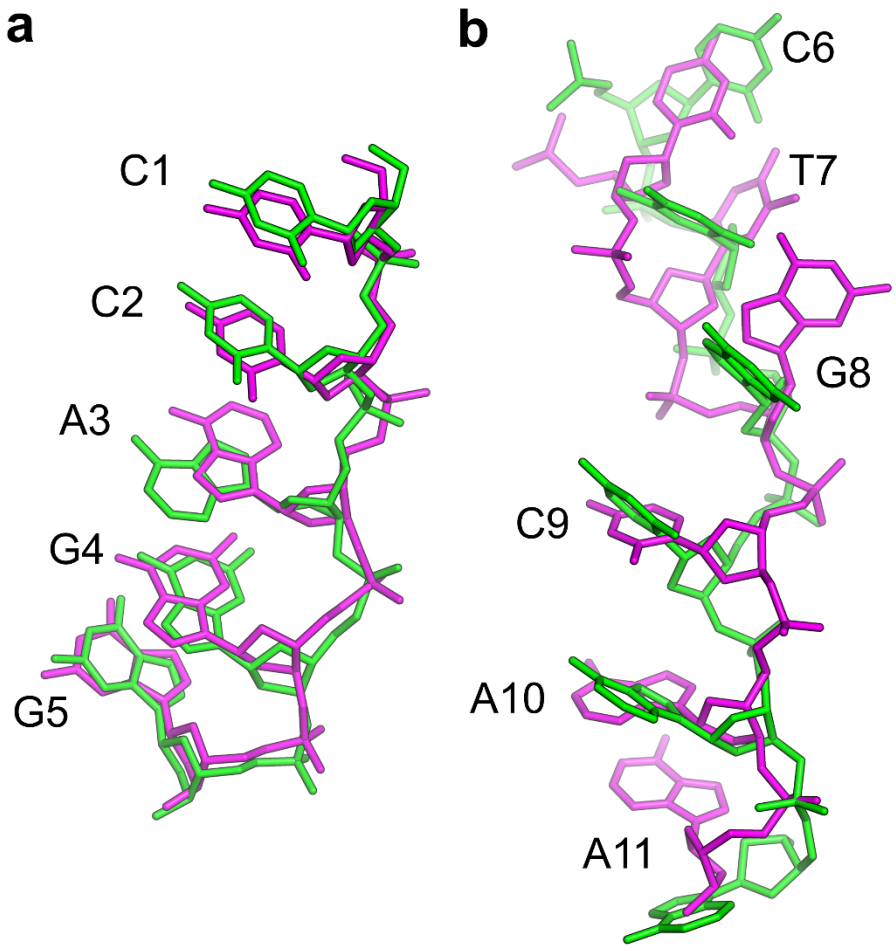

Figure 4.2. Structural Comparison of d(CCAGGCTGCAA) Monomers. Stick representation of the alignment between Chains A (green) and B (magenta) reveal structural similarity in residues 1–5 **(a)** and significant differences in residues 6-11 **(b)**.

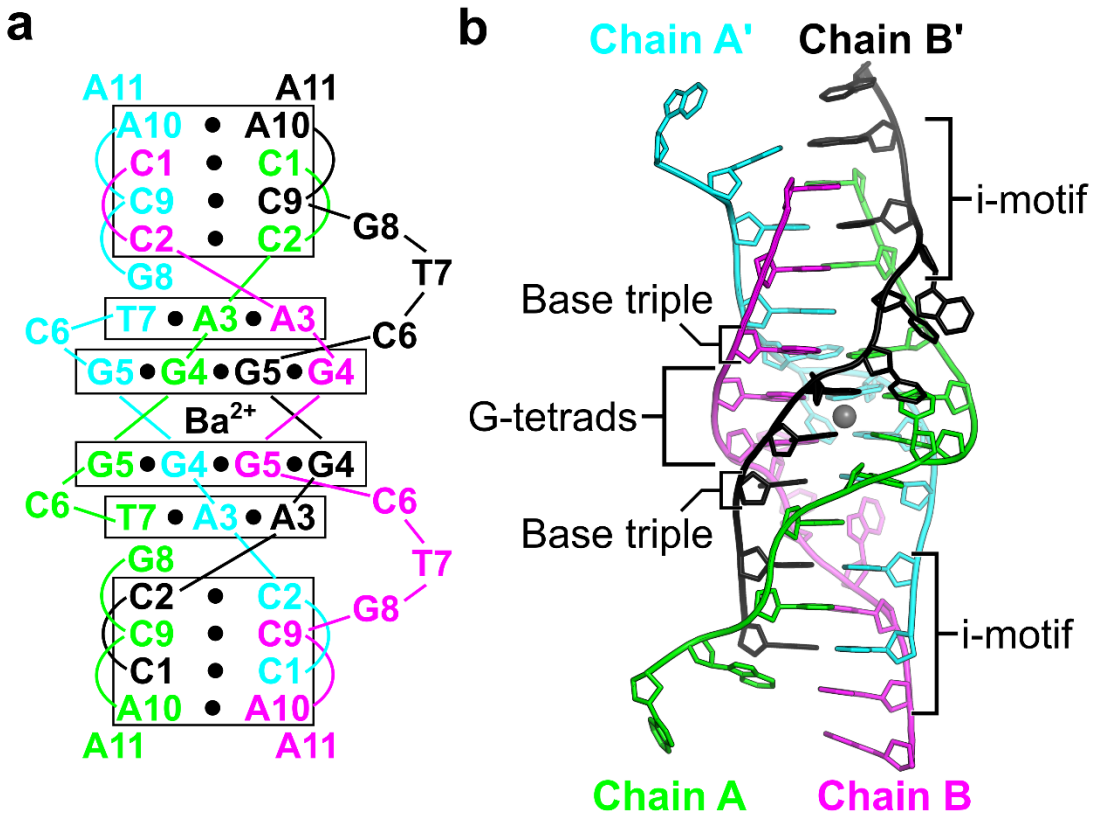

Figure 4.3. A G-quadruplex/i-motif Hybrid Structure Formed from d(CCAGGCTGCAA). (a) Secondary structure of interactions formed between two symmetry-related dimers. Black circles represent hydrogen bonding interactions. Chains A and A′ are in green and cyan, respectively. Chains B and B′ are in magenta and black, respectively. **(b)** Cartoon representation of the hybrid quadruplex with labeled features. The gray sphere represents a barium ion.

Chapter 4.1.2. A Barium-stabilized G-quadruplex

The central G-quadruplex is composed of two symmetrically equivalent G-tetrads, each of which is formed through two G4 and two G5 residues (Figure 4.4). The two dimers are antiparallel with respect to each other, with G4─G5 dinucleotide steps along each strand, leading to heteropolar stacking between the two G-tetrads. The G-tetrads are arranged in the *abab* topology.[17, 99, 100] Like other antiparallel G4s, the tetrad

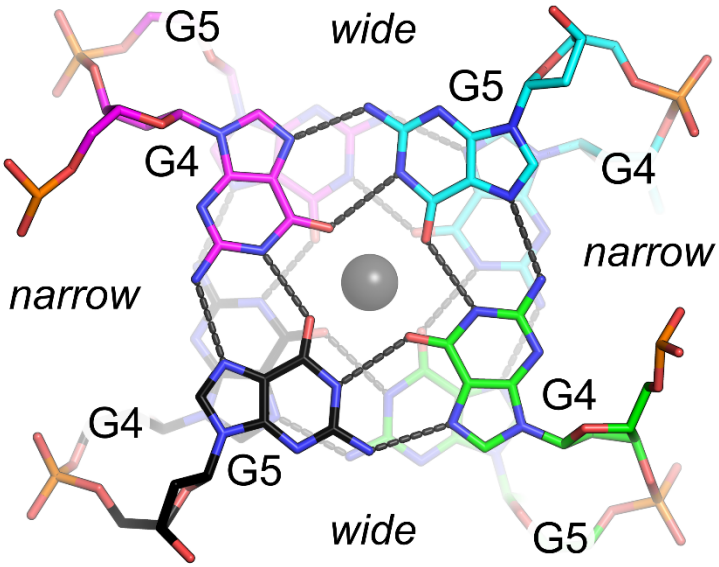

Figure 4.4. G-quadruplex. Top view of two stacked G-tetrads. Hydrogen bonds are indicated by gray dashes. The gray sphere represents a barium ion. The wide and narrow grooves are indicated.

adopts *syn-anti-syn-anti* glycosidic angles with residue G4 in *syn* and G5 in *anti* for both chains. The observed base pair and base step geometries are comparable to other quadruplex structures containing only two G-tetrads with the same topology.[101-103] This arrangement gives rise to two grooves of distinct widths.[99] The G5─G5 phosphate distances across the narrow and wide grooves are 12.78 and 19.15 Å, respectively.

The eight guanosines coordinate directly with a central cation that is located between the two G-tetrad planes. Both the native and U7-Br oligonucleotides were crystallized in the presence of barium chloride, and a strong (11 σ) anomalous difference electron density peak between the two G-tetrads was observed in both native and derivative structures. This peak is most consistent with Ba^{2+}, given the crystallization conditions and data collection energy (Figure 4.5). The Ba^{2+} ion lies on or very near a crystallographic symmetry axis with a refined final occupancy of 0.50

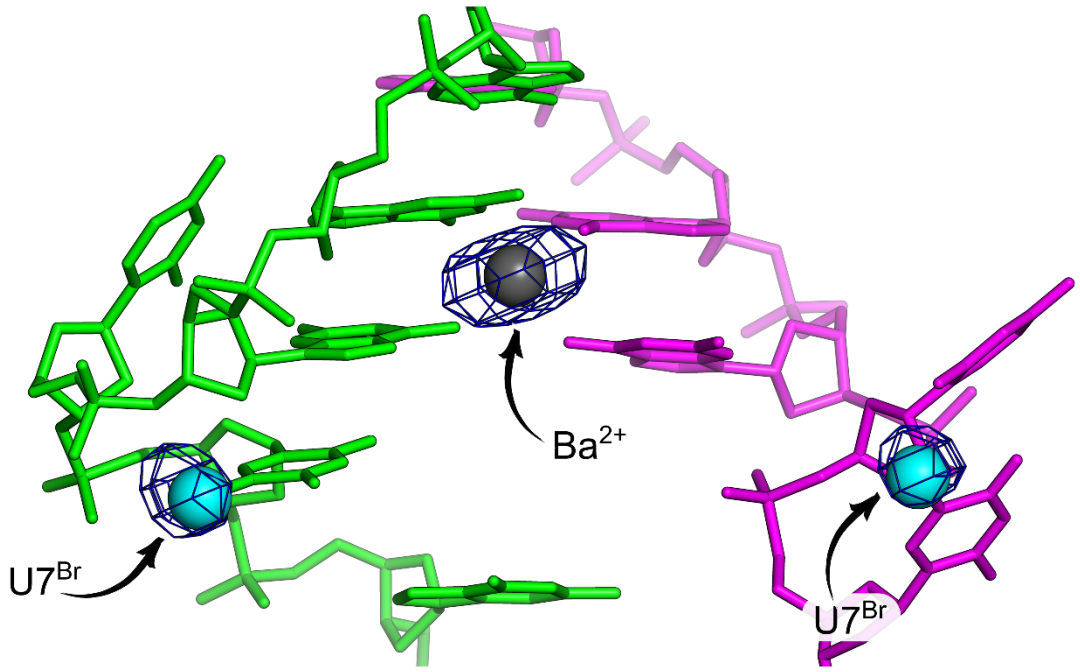

Figure 4.5. Anomalous Differences of the U7-Br Derivative. Anomalous difference electron density (dark blue) contoured at 6.2 σ corresponds to the bromine atoms (cyan spheres) of the $U7^{Br}$ residue that were used for phasing. A Ba^{2+} ion is shown as a gray sphere in anomalous electron density.

and B-factors of 73.27 $Å^2$. The coordination distances between the cation and guanosine O6 positions range from 2.5 to 2.8 Å, with an average distance of 2.63 Å. This is slightly shorter than the ~2.75 Å average coordination distance observed in previous examples of G-tetrads stabilized by Ba^{2+}.[104, 105] The apparent shorter metal-oxygen coordination may be the result of several factors, including difficulty in refining the cation residing near a special position. Alternatively, this more compact arrangement of guanosine residues could be a structural preference arising from the fewer base stacking interactions on either side of the two G-tetrads.

Chapter 4.1.3. Reverse-Hoogsteen Base Triple

Flanking each side of the G-quadruplex is an A─A─T base triple. This noncanonical base triple involves both A3 residues from the dimer and T7 from Chain A′ (Figure 4.6a). The A3─A3 base pair is formed through the Watson-Crick face of Chain A and the Hoogsteen face of Chain B, which adopts a *syn* glyosidic torsion angle to facilitate the N1─N6 and N6─N7 hydrogen bonds. The base triple is completed by interactions between A3 of Chain A and T7 from Chain A′. This is a reverse Hoogsteen

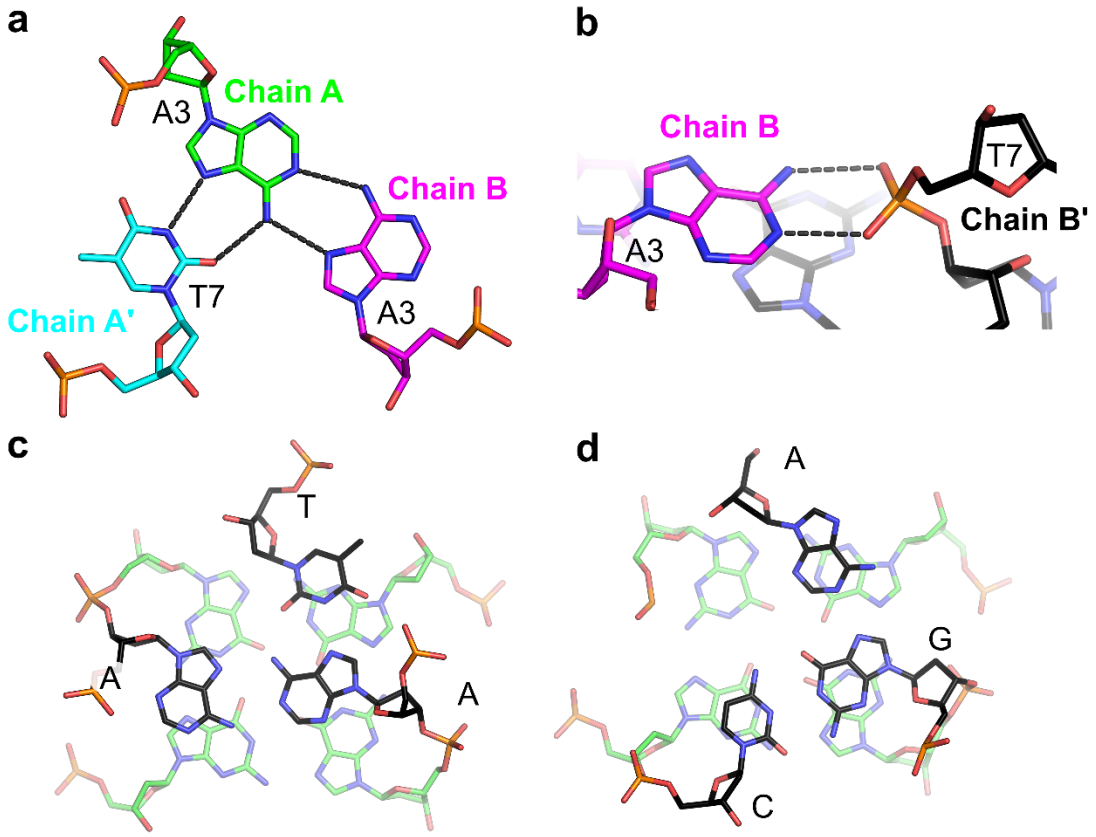

Figure 4.6. Base Triple Interactions. (a) Residue A3 from Chain A (green) forms a reverse Hoogsteen base pair with T7 from Chain A′ (cyan) through the A(N6)─T(O2) and A(N7)─T(N3) hydrogen bonds and interacts with A3 from Chain B (magenta) through the Watson-Crick/Hoogsteen faces. **(b)** Residue A3 from Chain B hydrogen bonds with the phosphate oxygens of T7 from Chain B′. Hydrogen bonds are indicated by gray dashes. A base triple (black) stacks above a G-tetrad (green) with the *abab* topology of guanosine residues in **(c)** the 11-mer and **(d)** PDB 2KM3.

base pair through the N6─O2 and N7─N3 hydrogen bonds. The *syn* glycosidic angle of A3 from Chain B allows the Watson-Crick face to make direct hydrogen bonding contacts with phosphate oxygens of T7 from Chain B′ of the tetramer (Figure 4.6b). With both N1 and N6 of A3 in hydrogen bonding distance with the non-bridging phosphate oxygens, this arrangement suggests protonation of the N1 position to serve as a hydrogen bond donor. Similar to observations in RNA structures, the electrostatic stabilization between the localized positive charge following N1 protonation and the negatively charged phosphate would facilitation this pK_a perturbation.[106]

Surprisingly, there is little direct nucleobase stacking between the G-tetrad and the A─A─T base triple. Rather, the adenosine and thymidine residues are largely positioned between the tetrad guanosines (Figure 4.6c). This is in contrast to the only other example of a base triple flanking one side of a G-quadruplex structure that contains two G-tetrads of the same topology.[107] In this case, the 22-nt d[AGGG(CTAGGG)$_3$] contains a C─G─A base triple that forms significant stacking interactions with the G-tetrad (Figure 4.6d). The large differences in stacking interactions between the triples and the tetrads suggest significant structural variability in these types of interactions based on intrinsic sequence differences and local structural constraints.

Chapter 4.1.4. Variable Spacer Region

The most distinct structural differences between the two molecules of the asymmetric unit are in residues C6, T7, and G8 that collectively make up the spacer regions between the central G-tetrads and the peripheral i-motif. Interestingly, these residues have contrasting functional roles in the overall architecture of the tetramer.

C6 of Chain A forms a single hydrogen bond with the G5 (N3─N2) from Chain B′ and is tucked into the quartet's wide groove. C6 in Chain B does not form any base pairing interactions within the tetrameric structure. It is bulged from the tetrameric core and serves primarily in mediating crystal contacts through base stacking interactions with the sugar of C6 from Chain A′ and with the nucleobase of A11 from a symmetry-related dimer. As described above, T7 of Chain A is involved in base triple interactions. In contrast, T7 of Chain B is not involved in any base pairing interactions within the tetramer. Instead, this bulged residue base pairs with A11 from a symmetry-related molecule through standard Watson-Crick pairing interactions to stabilize crystal packing. The G8 residues of both molecules are unpaired. In Chain A, G8 is positioned within the nucleobase core, stacking with A3 of the base triple on one face and with the C2─C2$^+$ base pair on the other (Figure 4.7a). However, the G8 residue in Chain B is flipped out from the core, where it stacks with A11 of Chain A from an adjacent symmetry-related molecule to serve in crystal lattice packing contacts (Figure 4.7b).

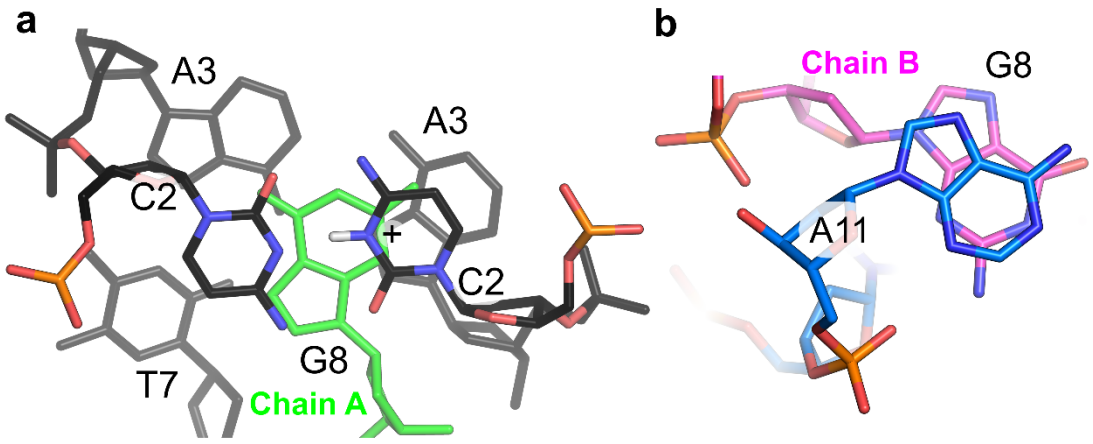

Figure 4.7. G8 Interactions. (a) The unpaired G8 from Chain A (green) stacks between the C2─C2$^+$ base pair and the A─A─T base triple. **(b)** The bulged G8 from Chain B (magenta) stacks with A11 from a symmetry-related molecule (blue).

This stacking is facilitated by A11 adopting a *syn* glycosidic angle, leading to partial stacking of both the pyrimidine and indole rings of the two purines.

These three residues from the parallel-stranded dimer have distinct functions within the structure. In Chain A, they form an integral part of the tetrameric structure, while the same residues in Chain B serve primarily as a bulged spacer that mediates crystal contacts. Because they have the same sequence, either strand could presumably take the role of the structural or bulged strand in solution. Though we cannot rule out the possibility of dynamic switching of these roles within the tetramer, there are several structural clues that suggest that this strand preference may arise at the time of assembly. The base triple interaction provides asymmetry between the parallel strands. This is seen in both the base pairing interactions with T7 and the *syn* A3 hydrogen bonding interactions with the phosphate from an antiparallel partner. These interactions bring the phosphate toward the stacked tetramer core and bias that partner strand toward bulging its nucleobases outward as found in the spacer. Additionally, the sequestration of the structural T7 in the base triple interaction would strongly bias the following nucleotide, G8, toward being stacked within the tetramer core.

Chapter 4.1.5. i-motif and 3′-terminal Nucleotides

The d(CCAGGCTGCAA) tetramer is capped at either end by i-motifs (Figure 4.3). The i-motif is comprised of three C─C^+ base pairs between C1, C2, and C9 residues of the dimers. The terminal C1─$C1^+$ base pair gives the i-motif a 5′-E topology.[18] Residues C1 and C2 of both chains adopt C3′-*endo* sugar puckers, allowing the sugar-phosphate to stretch to a helical rise of 6.5 Å. This provides the necessary space to allow the C9─$C9^+$ base pair from the symmetry-related dimer to intercalate

between them (Figure 4.8a). The geometries of the three hemiprotonated base pairs are similar, with the largest variation in the buckle and propeller angles (Table 4.2), consistent with what has been observed in other i-motifs.[37, 108] Complete base pair and base step parameters are listed in Table 4.2. Like the G-tetrads, the i-motif creates two grooves of dramatically different widths. The wide grooves are generated by the backbones of the parallel base paired strands and the narrow grooves are formed between one parallel-stranded dimer and the intercalated dimer (Figure 4.8b).

Table 4.2. Base Pair and Base Pair Step Parameters for the Helical Core Region of the d(CCAGGCTGCAA) Tetramer.

Local Base Pair Parameters*

	Base Pair	Shear (Å)	Stretch (Å)	Stagger (Å)	Buckle (°)	Propeller (°)	Opening (°)
1	C1─C1	1.97	1.50	-0.04	-2.24	-0.02	177.33
2	C9─C9	2.09	1.33	-0.01	-2.36	6.52	177.39
3	C2─C2	2.04	1.44	0.16	-0.24	-3.42	179.23
4	A3─A3	-4.16	1.39	0.69	8.05	10.39	-112.03
5	G4─G5	-1.45	-3.52	0.55	-16.11	10.74	87.68
6	G5─G4	1.65	3.38	-0.05	8.71	0.68	-89.11
7	A3─A3	-4.16	1.39	0.69	8.05	10.40	-112.04
8	C2─C2	2.04	1.44	0.16	-0.24	-3.42	179.23
9	C9─C9	2.09	1.33	-0.01	-2.36	6.51	177.40
10	C1─C1	1.97	1.50	-0.04	-2.25	-0.02	177.32

Local Base Pair Step Parameters*

	Step	Shift (Å)	Slide (Å)	Rise (Å)	Tilt (°)	Roll (°)	Twist (°)
1	CC/CC	-2.48	2.23	0.09	126.28	126.29	-136.14
2	CC/CC	1.70	-2.59	0.02	148.83	95.33	-82.42
3	CA/AC	-3.99	-5.66	1.62	-169.56	54.42	21.15
4	AG/GA	1.98	3.58	-3.52	-0.89	-2.11	-95.65
5	GG/GG	-2.39	-2.26	-1.95	86.27	-155.46	39.34
6	GA/AG	-0.05	3.29	3.15	-1.71	-0.38	-85.84
7	AC/CA	3.99	5.66	-1.62	169.56	-54.41	-21.14
8	CC/CC	-1.70	2.59	-0.02	-148.82	-95.33	82.37
9	CC/CC	2.48	-2.23	-0.08	-126.28	-126.28	136.23

*All values were calculated with x3DNA-DSSR.[109]

Along with the C─C$^+$ interactions, a noncanonical A10─A10 base pair caps the i-motif. The capping of the 5′-E i-motif by noncanonical (i.e., A─A, T─T) base pairs

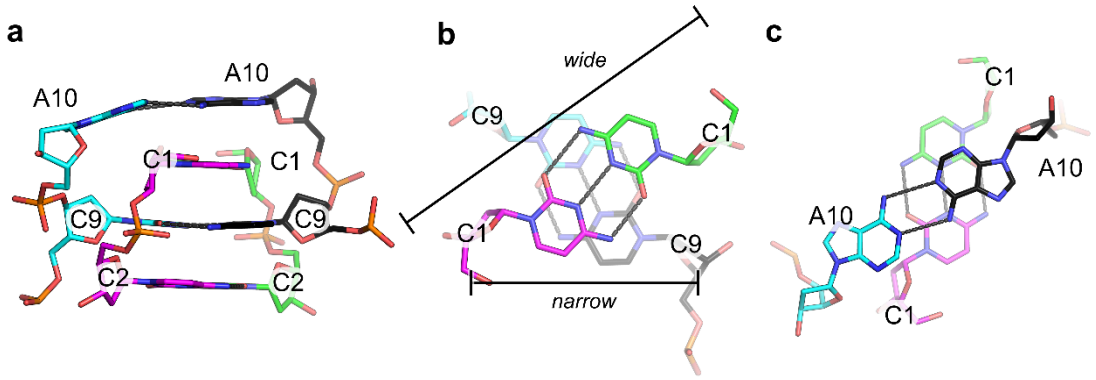

Figure 4.8. Terminal i-motif. (a) Side view of three C─C$^+$ base pairs capped by an A10─A10 base pair. The C9─C9$^+$ base pair (cyan/black) is intercalated between the dimeric (green/magenta) C1─C1$^+$ and C2─C2$^+$ base pairs. **(b)** Top view of the C1─C1$^+$ base pair stacking above the intercalated C9─C9$^+$ base pair. The wide and narrow grooves are indicated. **(c)** Top view of the A10─A10 base pair stacking above the C1─C1$^+$ base pair. Hydrogen bonds are indicated by gray dashes.

has been observed in previous examples of i-motif structures.[37, 40, 110] A large ζ angle between C9 and A10 in Chain A moves residue A10 away from the helical core, preventing direct stacking interactions between A10 and the intercalated C1 (Figure 4.8c). This creates a strong asymmetry with respect to the neighboring C1─C1$^+$ base pair. This asymmetry is likely induced by crystal contacts, most notably those made by the subsequent A11 nucleotides. These A11 residues are not involved in i-motif-like interactions, but form stabilizing contacts with the variable bulged region of another tetrameric assembly (see Chapter 4.1.4).

Chapter 4.1.6. An Alternative Hybrid Motif

We determined that d(CCAGGCTGCAA) can also assemble into an alternative hybrid structure. I crystallized a 5-bromo-deoxycytidine substitution at the C9 position and determined its structure (Table 4.1). The bromine substitution at this position and different crystallization conditions resulted in crystals of a different morphology

(Figure 4.1c), space group (Table 4.1), and an overall different structure, but with some similar features. The two molecules in the asymmetric unit interact with symmetry-related strands to form a hybrid quadruplex structure, in this case juxtaposing an i-motif at the 5′ end and a partial antiparallel duplex at the 3′ end (Figure 4.9).

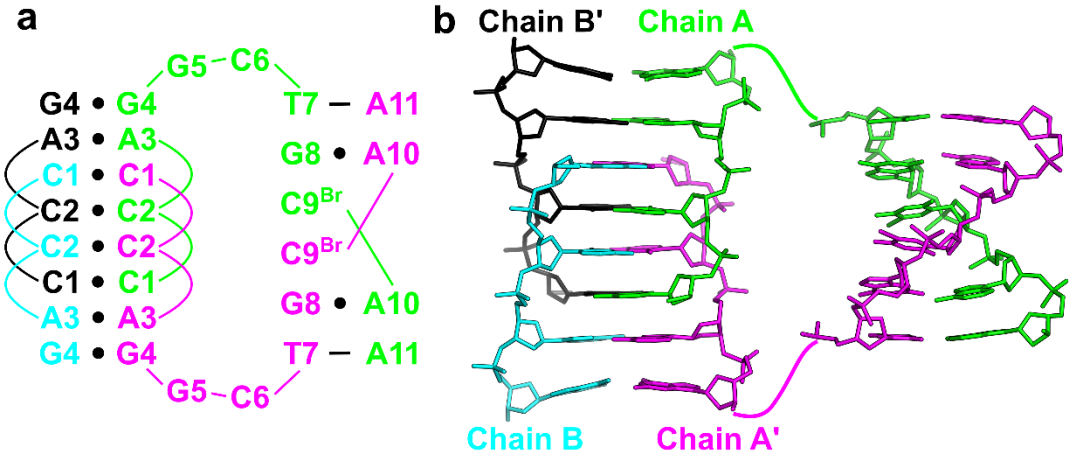

Figure 4.9. An i-motif/Duplex Hybrid Structure Formed from the C9^{Br} Derivative, d(CCAGGCTGC^{Br}AA). (a) Secondary structure representation of interactions formed between symmetry-related strands. Black dashes represent Watson-Crick base pairs. Black circles represent noncanonical base pairs. **(b)** Stick representation of the 5′ i-motif containing CCAG homo base stacking motif linked to the 3′ antiparallel duplex. The i-motif and duplex are rotated 70° about the x-axis with respect to each other. Chains A and A′ are in green and magenta, respectively. Chains B and B′ are in cyan and black, respectively.

In this structure, the i-motif C─C$^+$ base pairs are formed exclusively from residues C1 and C2. The four strands create a 5′-E topology, with the symmetry axis between the intercalated C2─C2$^+$ base pairs. This i-motif region is extended on either side by a homo base pairing region that includes symmetric A3─A3 (N6─N7) and G4─G4 (N1─O6) base pairs. The base stacking interactions provided by these noncanonical base pairs stabilize the i-motif tertiary interactions. The brominated C9 residues are no longer involved in i-motif formation, and instead form Watson-Crick

base pairing interactions with G5 from symmetry-related molecules that promote crystal packing. Examination of the native structure suggests that the C9 bromine substitution would preclude the formation of the hybrid G-quadruplex/i-motif structure due to significant steric clashes between the bromine in Chain B and the phosphodiester backbone.

The 3′ end of the structure is a short imperfect duplex with noncanonical features. Symmetry interactions between two identical strands (Chains A and A′) form an antiparallel base pairing arrangement consisting of the brominated C9 residues at the interior, each of which is immediately flanked by a G8─A10 base pair formed through N2─N7 and N3─N6 hydrogen bonding, and finally capped by a T7─A11 Watson-Crick base pair (Figure 4.10). This duplex interacts with a second duplex that is formed from the other unique molecule (Chains B and B′) in the asymmetric unit.

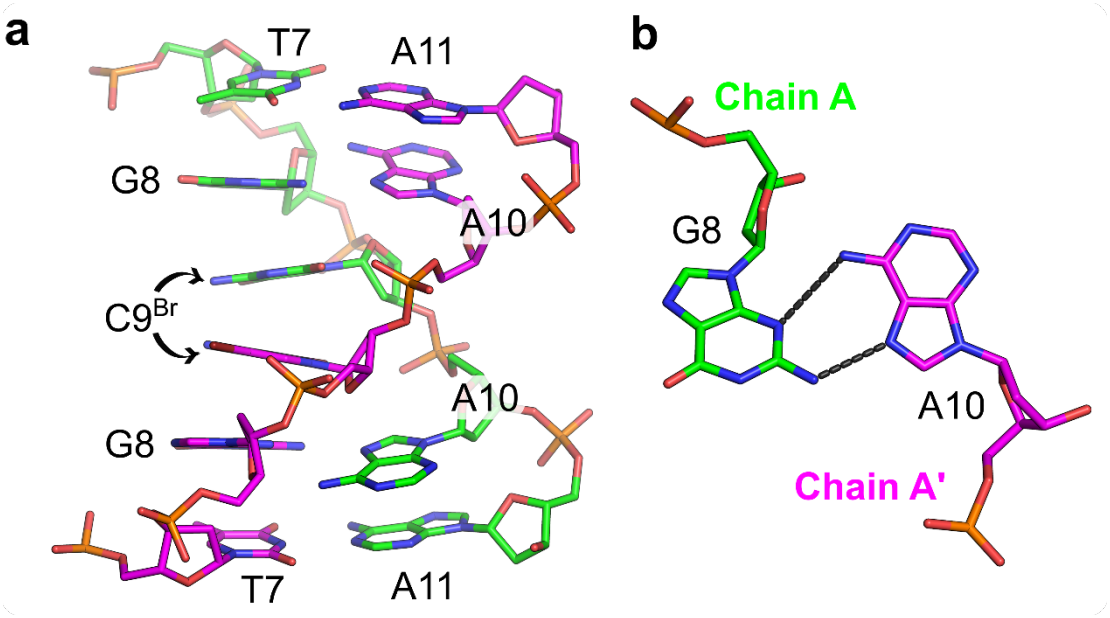

Figure 4.10. Duplex in the d(CCAGGCTGC^{Br}AA) Structure. (a) Stick representation of the 3′ antiparallel duplex formed from residues T7 through A11. **(b)** Residue G8 from Chain A (green) base pairs with A10 from Chain A′ (magenta) through the G(N2)─A(N7) and G(N3)─A(N6) hydrogen bonds.

The two distinct duplexes are held together by the G5─C9^{Br} base pair described above and by the C─G─A base triple, which is converted from the G8─A10 base pair from a single duplex (Figure 4.11).

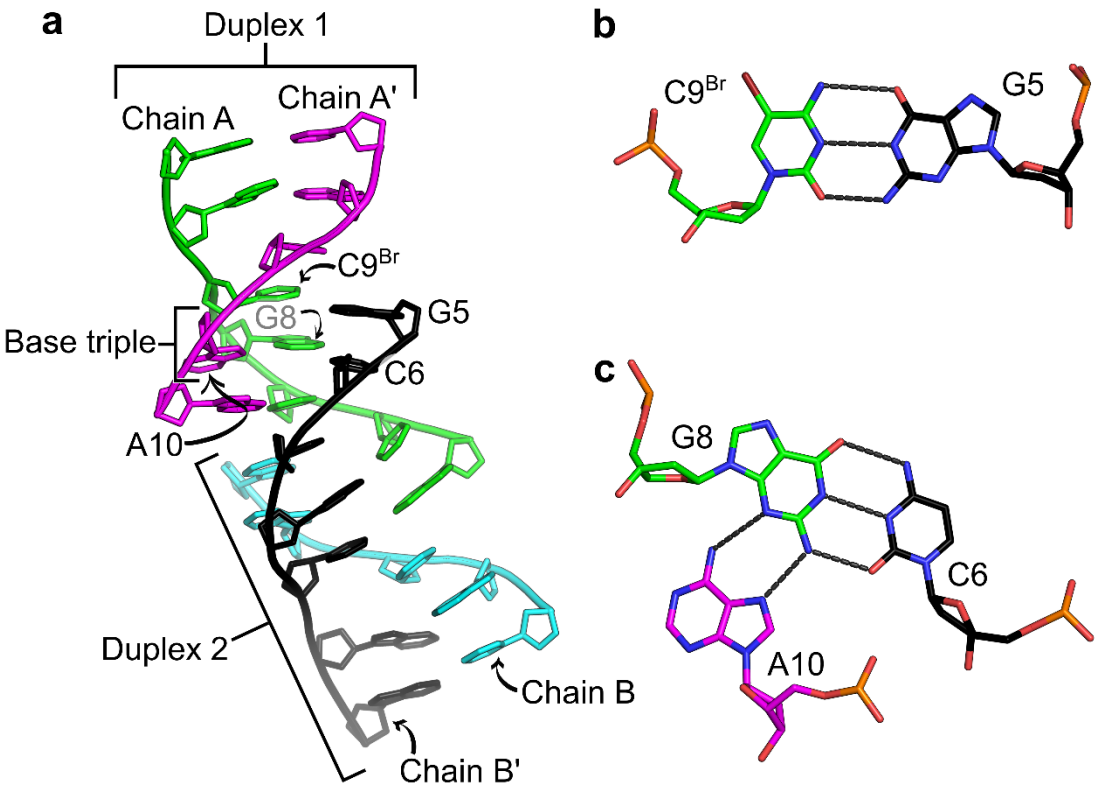

Figure 4.11. d(CCAGGC^{Br}TGCAA) Duplex Interactions. (a) Cartoon representation of the interactions between Duplex 1 (formed between Chains A and A′) and Duplex 2 (formed between Chains B and B′). **(b)** Stick representation of the G5─C9^{Br} base pair through standard Watson-Crick hydrogen bonding. (**c**) Stick representation of the C6─G8─A10 base triple. The G8─A10 base pair formed through G(N3)─A(N6) and G(N2)─A(N7) hydrogen bonding from Duplex 1 is converted into a base triple through Watson-Crick interactions with C6 from Duplex 2.

Chapter 4.2: Summary and Implications

Although previous biophysical studies characterized oligonucleotides capable of forming a parallel G4/i-motif hybrid in solution,[111] the results presented here provides the first structural snapshot of a hybrid G4/i-motif. This information provides

the beginnings of a structural paradigm for how these two distinct quadruplex motifs can coexist. Most notably, this structure suggests a requirement for spacer elements to separate the two base pairing motifs. These spacer elements serve to bridge the large differences in the interstrand backbone distances of the two motifs. This is necessitated by an exchange of the wide and narrow grooves between the individual motifs; the G-tetrad wide groove is continuous with the i-motif narrow groove and vice versa. The variable spacer regions that include the structurally integrated base triple and unpaired guanosine allow progressive changes of interstrand backbone distances to facilitate this transition.

Here, I have provided an analysis on the crystal structure of the hybrid G4/i-motif formed by d(CCAGGCTGCAA) and have highlighted the main features of this unique tetrameric assembly. In the next chapter, I will discuss the distinct characteristics of this oligonucleotide and related sequences in the solution state, which will provide insight into the stability of this hybrid structure.

Chapter 4.3: Experimental Procedures

Chapter 4.3.1. Oligonucleotide Synthesis and Purification

The 11-mer, d(CCAGGCTGCAA), the U7-Br derivative, d(CCAGGCU^{Br}GCAA), and the C9-Br derivative, d(CCAGGCTGC^{Br}AA), were synthesized using standard phosphoramidite chemistry on an Expedite 8909 Nucleic Acid Synthesize (PerSeptive Biosystems, Inc.), with reagents from Glen Research (Sterling, VA). All three oligonucleotides were purified using the Glen-Pak cartridges according to the manufacturer's protocol.

Chapter 4.3.2. Crystallization

Sitting drops of d(CCAGGCTGCAA) were set up by mixing 1 μL of 500 μM DNA solution with 2 μL of crystallization solution (30% PEG400, 20 mM barium chloride, 10 mM spermidine, and 30 mM Bis-Tris at pH 8.5). Sitting drops of the U7-Br derivative were set up by mixing 1 μL of 500 μM DNA solution with 2 μL of crystallization solution (25% PEG400, 40 mM barium chloride, 10 mM spermidine, and 30 mM Bis-Tris at pH 8.5. These drops were equilibrated against 300 μL of 5% PEG400 in the well reservoir at 22°C for 15–20 hours, followed by subsequent equilibration with 3–4 μL of glacial acetic acid added to the well reservoir. Crystals were observed 2 days after the addition of acid. Crystals were removed from the drops by nylon cryoloops and directly cryo-cooled in liquid nitrogen.

The C9-Br derivative was crystallized by mixing 3 μL of 500 μM DNA solution with 3 μL of crystallization solution (15% MPD, 120 mM calcium chloride, 20 mM lithium chloride, 8 mM spermidine, and 30 mM sodium cacodylate at pH 5.5). Crystallization was performed at 22°C and in sitting drops, which were equilibrated against 300 μL of 20% MPD in the well reservoir. Crystals were observed 2 days after plating. Crystals were removed from the drops by nylon cryoloops, dipped in 30% MPD, and cryo-cooled in liquid nitrogen.

Chapter 4.3.3. Data Collection and Structure Determination

Diffraction data for 11-mer and the U7-Br derivative were collected at the APS 24-ID-C beamline. Diffraction data for the C9-Br derivative were collected at the APS 22-BM beamline.

Data processing for the 11-mer and the U7-Br derivative was carried out in XDS[81] and Aimless.[78, 112] Diffraction data for the C9-Br derivative were indexed and integrated using iMosflm.[76] In both derivative datasets, initial phases were determined by SAD phasing, using CRANK2[113] and SHELX[97] in CCP4i2.[82] Two bromine sites were identified in each map, which enabled model building of two chains of each derivative in Coot.[80] Subsequent refinement was carried out in Refmac.[83, 84] The refined U7-Br derivative structure was used as a molecular replacement search model in Phaser[79] for the native oligonucleotide. Further refinement was carried out in Refmac and additional model building was performed in Coot. The PDB-REDO web server[98] was used to conduct *k*-fold cross-validation of R_{free} values on all three structures and to generate the final models. Final refinement statistics are shown in Table 4.1. The final models were analyzed using Pymol and x3DNA-DSSR[109] and are deposited in the PDB.

Chapter 5: d(CCAGGCTGCAA) Solution Characterization

This chapter is adapted from "Chu, B.; Zhang, D.; Paukstelis, P.J., A DNA G-quadruplex/i-motif hybrid. *Nucleic Acids Res. 47* (22): 11921 – 11930." D. Zhang acquired the nuclear magnetic resonance spectroscopy data and conducted the initial partial peak assignment.

Chapter 5.1: Motivation for Solution State Characterization

From the crystal structure described in Chapter 4, we observed the coexistence of the G4 and i-motif in the same structure forming from a single DNA sequence. As both of these motifs are known to exist in physiological conditions, there would be significant implications if the hybrid G4/i-motif structure were found to exist stably in biological settings. Because the individual motifs are associated with cancer diseases in humans, it is possible that the hybrid structure can also form in such conditions. If this is the case, we can potentially develop therapeutics to target this unique structure for more specific cancer treatments. To begin to characterize the solution behavior of the d(CCAGGCTGCAA) oligonucleotide, I probed the oligomeric state and stability of this and related sequences using several solution-based biophysical methods, including NMR, CD, and thermal denaturation via UV absorbance.

Chapter 5.2: Results and Discussion

Chapter 5.2.1. Oligomeric State Analysis via NMR

First, we conducted 1D ^{1}H-NMR, 2D-NOESY, and 2D-TOCSY experiments on the 11-mer oligonucleotide to directly assess the structure in solution. These spectra suffered from signal crowding and the appearance of multiple conformations made complete proton assignment difficult. The 1D ^{1}H-NMR profile is shown in Figure 5.1. Sequential assignment allowed identification of the nucleobase, H1′, and H2′ protons

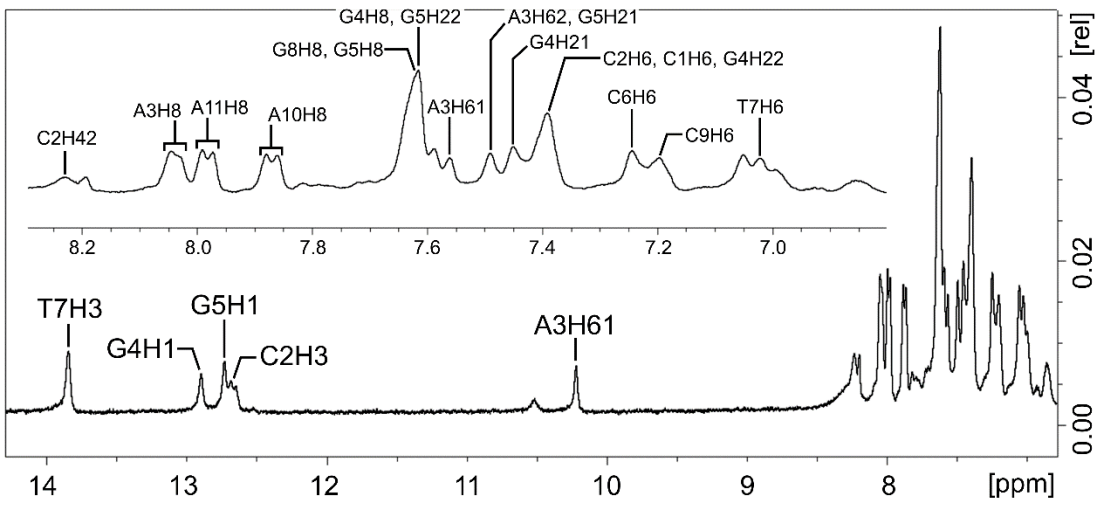

Figure 5.1. 1D ^{1}H-NMR spectrum of the 11-mer. The assigned peaks in the imino and nucleobase (inset) regions are labeled. The C2H3 and the H8 protons of the adenosine residues show up as degenerate signals. The peak at 10.5 ppm is an artifact.

in at least one conformation. Sugar-to-base connectivities were observed from C2 through G4 (Figure 5.2) and from C6 through A11 (Figure 5.3). Four imino proton signals were observed and assigned to T7H3, G4H1, G5H1, and C2H3 (Figure 5.1/Table 5.1). The C2H3 signal at 12.6 ppm is degenerate, indicating multiple

Table 5.1. Chemical Shift Values (in ppm) of ^{1}H Assignments Obtained from 2D-NOESY/TOCSY NMR Spectra of the 11-mer, d(CCAGGCTGCAA).

	Imino	Amino	H8/H6	H5/Methyl	H1′	*H2′, H2″	H3′	H4′	*H5′, H5″
C1	-		7.39	5.53	5.75	1.73, 1.66	4.34	3.89	3.48, 3.43
C2	12.66	8.23, 6.85	7.40	5.65		2.05, 1.90	4.69		
A3a	-	7.56, 7.49							
A3b	-	10.21, -	8.04	-	5.63	2.54, 2.62	4.79		
G4	12.89	7.45, 7.39	7.61	-	5.74	2.29, 2.62	4.79		
G5	12.72	7.62, 7.48	7.62	-	5.58	2.28, 2.61	4.66		
C6	-	6.54, 6.25	7.22	4.97	5.87	1.80, 2.22	4.54	4.08	
T7	13.83	-	7.03	1.41	5.52	1.65, 2.00	4.61	3.86	
G8	-		7.63	-	5.71	2.26, 2.34	4.66		
C9	-		7.19	5.58	5.66	1.49, 1.95	4.44	3.78	
A10	-		7.86	-	5.73	2.21, 2.35	4.63	3.99	3.74, 3.66
A11	-		7.98	-	5.96	2.20, 2.40	4.47	3.93	

*H2′/H2″ and H5′/H5″ protons were not stereospecifically assigned.

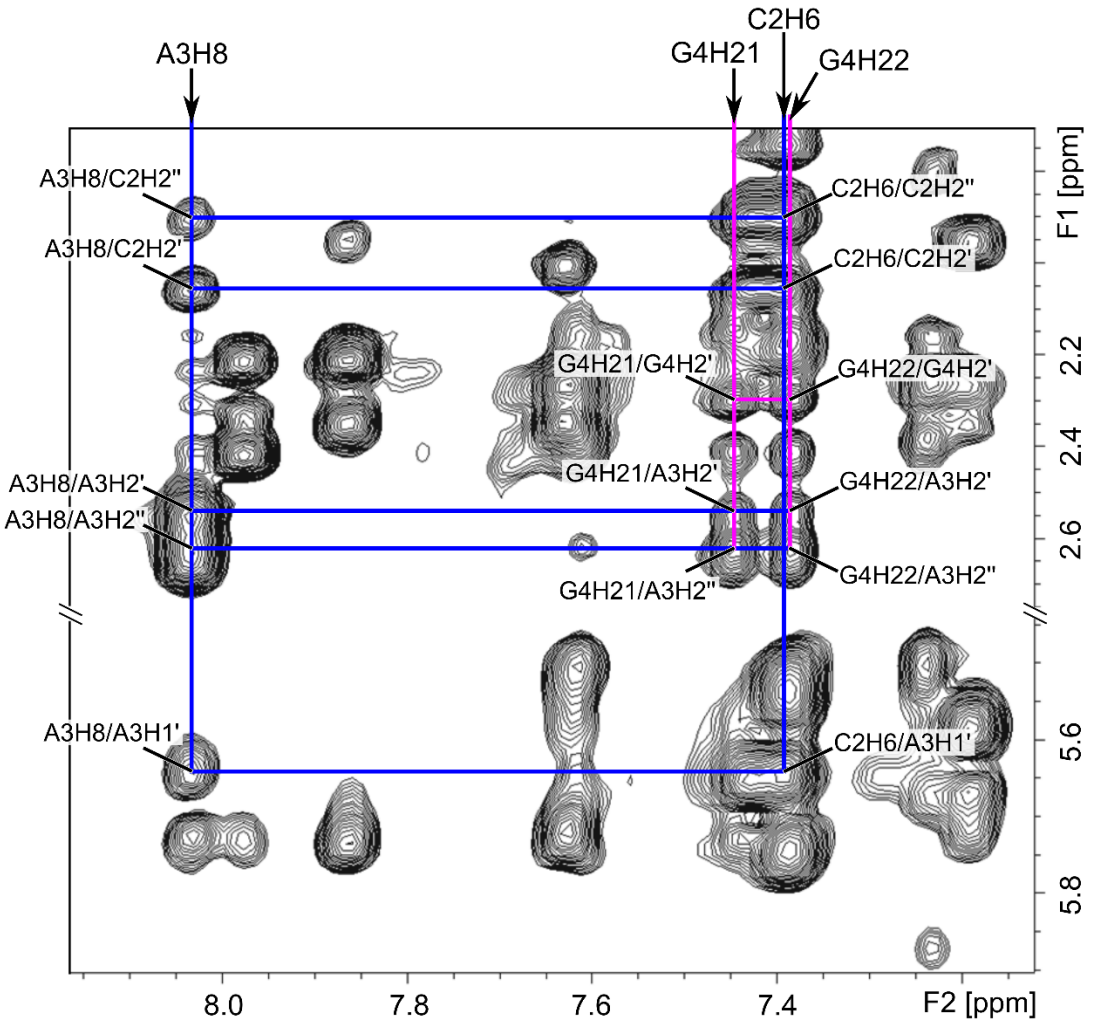

Figure 5.2. Regions of the 11-mer 2D-NOESY NMR Spectra Showing Connectivity from C2 to G4. The C2H6/A3H1′ and A3H8/A3H1′ cross-peaks demonstrate the proximity between C2 and A3, whereas the G4H21/A3H2′, G4H21/A3H2″, G4H22/A3H2′, and G4H22/A3H2″ cross-peaks confirm the connectivity from A3 to G4.

conformations that are consistent with the crystal structure. These assignments allowed identification of several key structural features.

Three key structural features were confirmed by NMR analysis. First, the C2H3 signal demonstrates protonation at this position and evidence for a C─C^+ base pair of the i-motif. This was the only CH3 resonance observed. Typically, this proton is

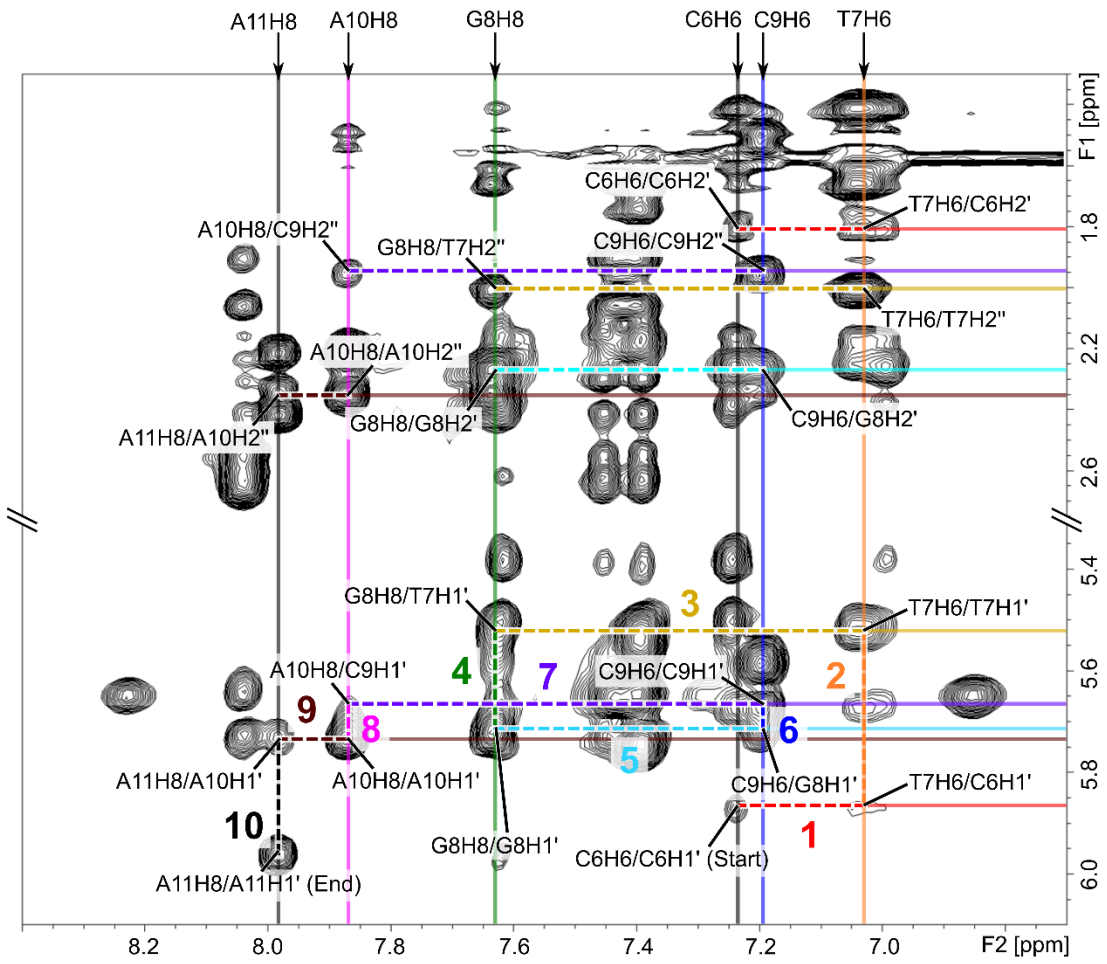

Figure 5.3. Sequential Intra-strand Connectivities for C6→T7→G8→C9→A10→A11 of the 11-mer. Cross-peaks in the non-exchangeable proton regions of the 2D-NOESY NMR spectra are labeled and indicated by intersecting lines. The sugar-to-base (H8/H6 to H1′) connectivities are colored and numbered 1–10. NOEs to H2′ or H2″ are indicated in the same color scheme and demonstrate internal consistency.

observed at chemical shift values near 15 ppm, though in this case, there was a significant upfield shift to 12.6 ppm (Figure 5.4a). Cross-peaks to C2H41, C2H42, and C2H5 confirmed the assignment (Figure 5.4a). This large perturbation may be the result of cation-π interactions, with the localized positive charge at the C2H3 position stacking with the pyrimidine ring of G8 (Figure 5.4b). The absence of C1H3 and C9H3

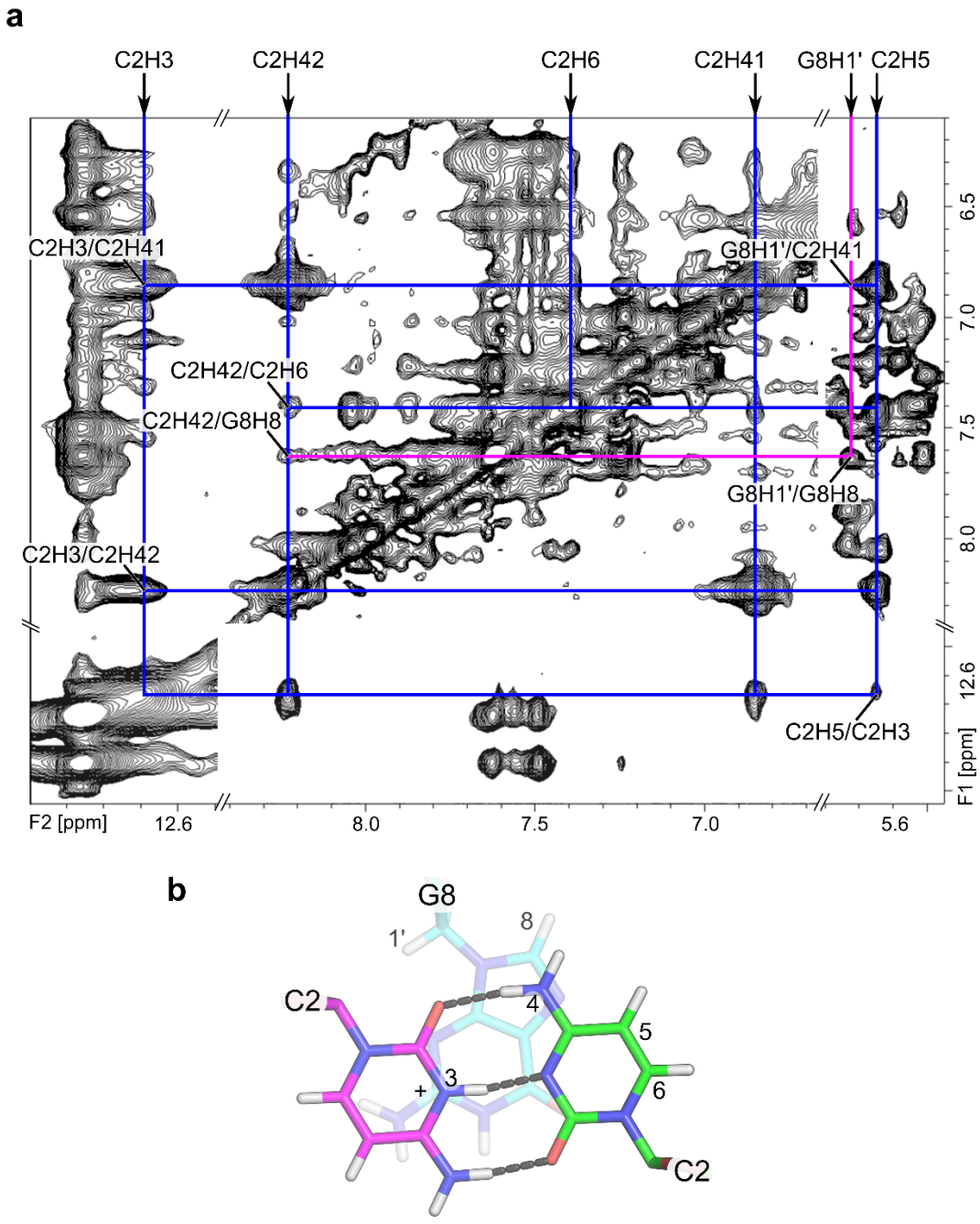

Figure 5.4. Regions of the 11-mer 2D-NOESY NMR Spectra Showing Cross-peaks Confirming the C2─C2$^+$ Base Pair. (a) The C2H3/C2H41, C2H3/C2H42, and C2H5/C2H3 cross-peaks indicate internal consistency, whereas the C2H42/G8H8 and G8H1′/G8H8 cross-peaks confirm the proximity of C2H3 and G8. **(b)** Top view of the C2─C2$^+$ base pair stacking on the unpaired G8 residue. Relevant atom positions are numbered.

signals could be due to weak or transient hydrogen bonding between the cytosines in solution, which has been previously reported in other structures consisting of multiple C─C$^+$ base pairs.[10] NOEs confirmed the proximity of C2H3 and G8, as anticipated from the crystal structure (Figure 5.4b/Table 5.1). Second, imino NOE cross-peaks confirmed the hydrogen bonding between T7 and A3 and additional NOEs between two independently assigned A3 residues indicated the formation of the A─A─T base triple (Figure 5.5a,b/Table 5.1). Third, cross-peaks between the imino protons G4H1 and G5H1 suggest hydrogen bonding between the guanosine residues, while resonances between guanosine H8 protons and neighboring guanosine imino and amino protons indicate their interaction through Watson-Crick and Hoogsteen faces (Figure 5.6a,b/Table 5.1).

Importantly, these structural features were all internally consistent with the stacking arrangement observed in the crystal structure. Resonances between the G-tetrad guanosine imino protons (G4H1, G5H1) and multiple members of the base triple (T7H3, A3H61, A3H62) confirm the stacking between the base triple and G-tetrad (Figure 5.5a). Further, the G5H8/T7H5, G4H8/T7H1′, and G4H8/T7H4′ cross-peaks confirm the arrangement of the guanosines of the G-tetrad in the *syn-anti-syn-anti* topology with respect to the A─A─T base triple (Figure 5.6a). The G8H8/T7H1′ cross-peak demonstrates the connectivity from T7 to G8 (Figure 5.3), supporting the sequential link from the base triple to the unpaired G8 residue (Figure 5.6c). Lastly, the C2H42/G8H8 resonance confirms the stacking between the G8 and the C2─C2$^+$ base pair (Figure 5.4a,b). Although these NMR data do not allow independent structure determination, they are consistent with the three major base pairing motifs and support

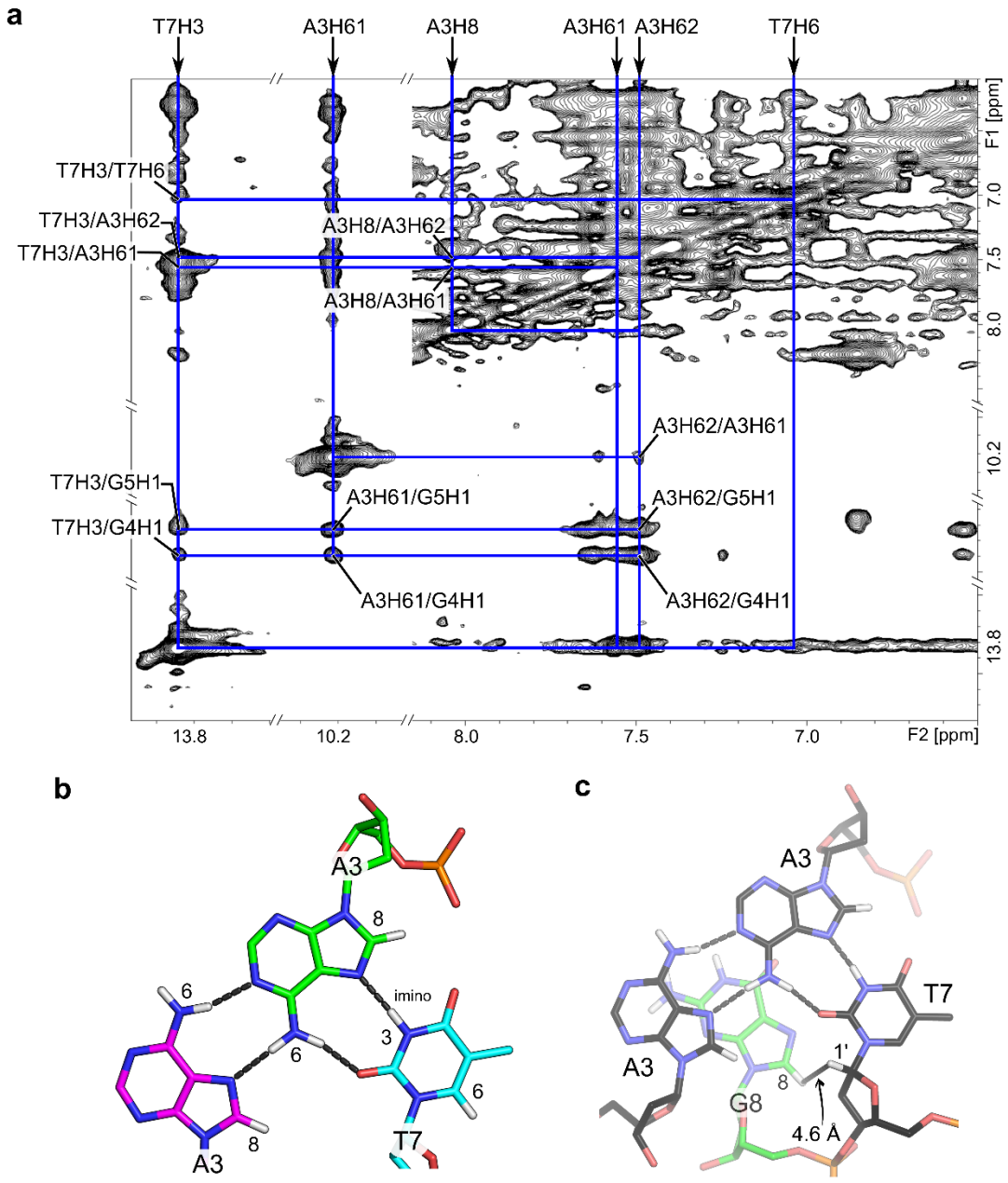

Figure 5.5. Regions of the 11-mer 2D-NOESY NMR Spectra Showing Cross-peaks Confirming the A─A─T Base Triple. (a) The T7H3/A3H61 and T7H3/A3H62 cross-peaks confirm the T7─A3 base pair and the A3H8/A3H61, A3H8/A3H62, and A3H62/A3H61 cross-peaks from two independently assigned A3 residues provide evidence for the A3─A3 base pair. NOEs between the guanosine imino protons and the T7H3, A3H61, and A3H62 confirm the stacking between the base triple and G-tetrad. **(b)** Stick representation of the A─A─T base triple. Relevant atom positions are labeled. **(c)** Top view of the A─A─T base triple (black) stacking on the unpaired G8 residue (green). The distance between T7H1′ and G8H8 is 4.6 Å.

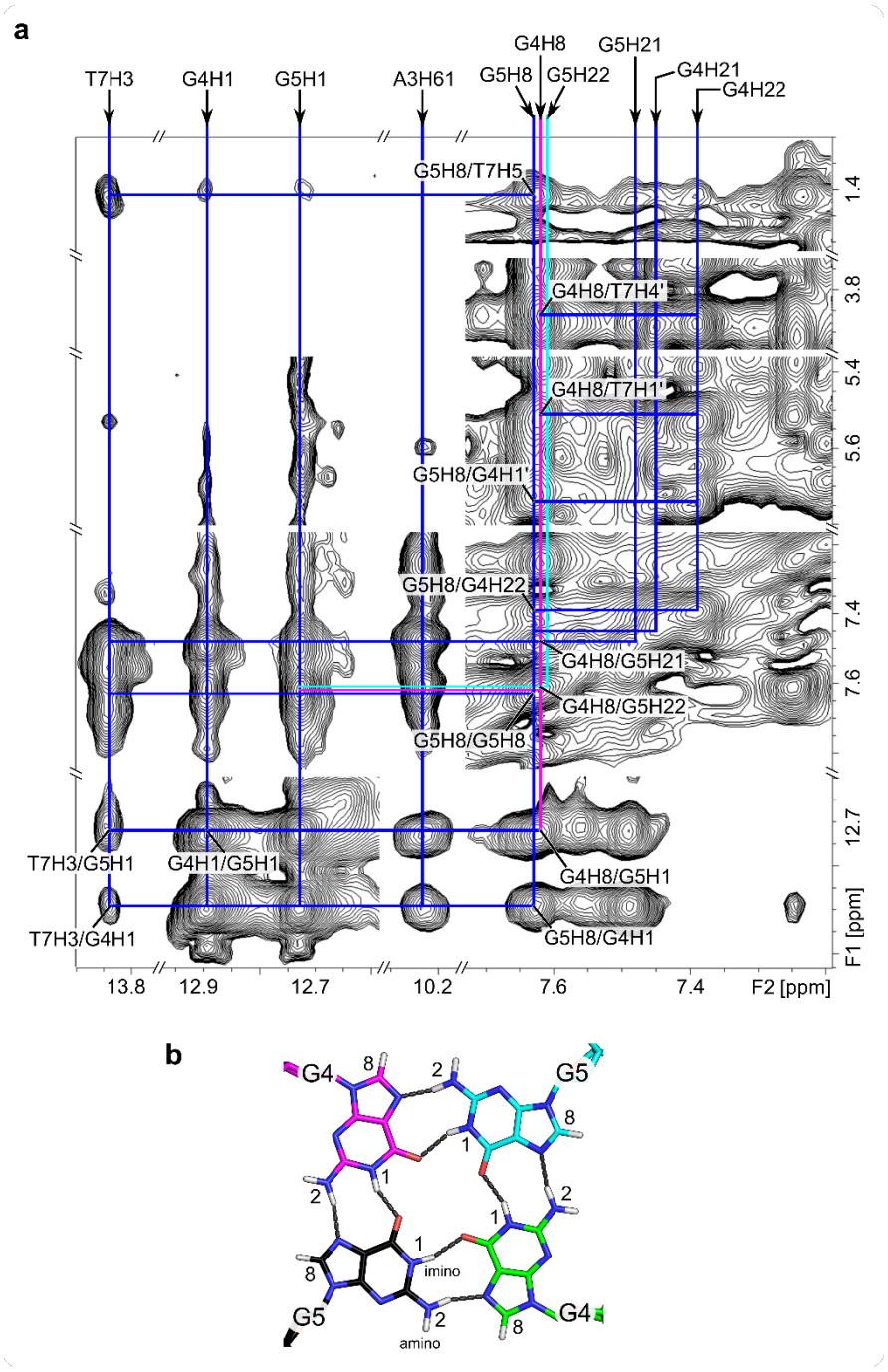

Figure 5.6. Regions of the 11-mer 2D-NOESY NMR Spectra Showing Cross-peaks Confirming the G-tetrad. (a) The G4H1/G5H1 cross-peak, in conjunction with the G4H8/G5H21, G4H8/G5H22, and G5H8/G4H22 cross-peaks, confirm hydrogen bonding between G4 and G5 through their Watson-Crick and Hoogsteen faces. The G5H8/T7H5 cross-peak, as well as resonances between G4H8 and the sugar protons of T7, confirm the arrangement of the guanosines in the G-tetrad in relation to the A─A─T base triple. **(b)** Stick representation of one G-tetrad with relevant atom positions labeled.

the sequential stacking order from the G-tetrad to the base triple to the unpaired G8 residue and finally to the C2─C2$^+$ base pair in the crystal structure.

Chapter 5.2.2. Tandem repeats alter the oligomeric solution state

The crystal structure suggested that the flexibility at the A11 position could allow tandem sequence repeats to form a dimeric quadruplex (Figure 5.7), analogous

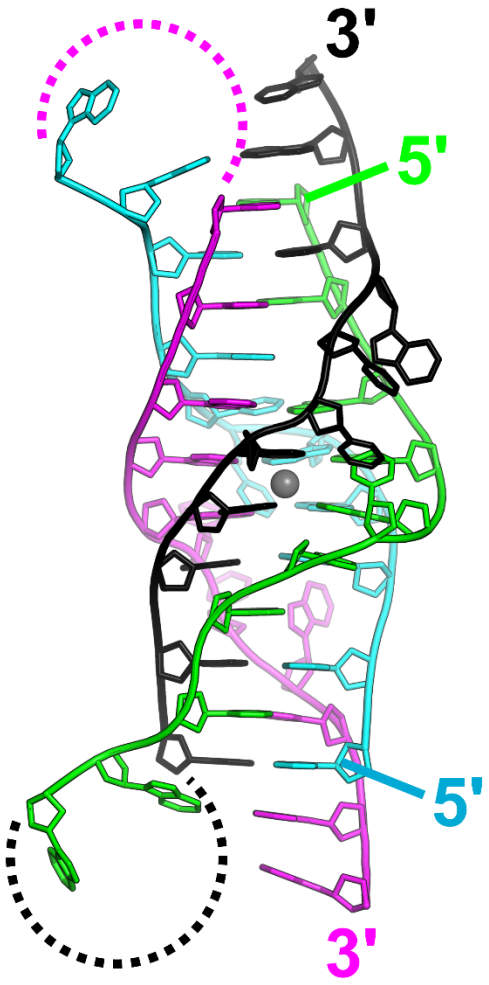

Figure 5.7. Proposed Model of the 22-mer Oligonucleotide. The barium-stabilized tetrameric hybrid quadruplex formed by the 11-mer is adapted to demonstrate the predicted dimeric assembly by the 22-mer. The cyan molecule is connected to the magenta molecule by a dotted segment to form one chain of the dimer. The green molecule is connected to the black molecule by a dotted segment to form the second chain of the dimer. The 5′ and 3′ ends of the two chains are indicated. The gray sphere represents a barium ion.

to loops in bi- and unimolecular G4- and i-motif-forming sequences. I synthesized the 22-mer, d(CCAGGCTGCAACCAGGCTGCAA), which contains one tandem repeat of the 11-mer sequence (Table 5.2), and compared it to the 11-mer by 1D ^{1}H-NMR, CD, and UV absorption spectroscopy.

Table 5.2. Variants of the d(CCAGGCTGCAA) Oligonucleotide Discussed in this Study.

	Name	d(Sequence)*
0	11-mer (Native)	CCAGGCTGCAA
1	22-mer (Tandem Repeat)	CCAGGCTGCAA**CCAGGCTGCAA**
2	C6A	CCAGG**A**TGCAA
3	C6G	CCAGG**G**TGCAA
4	C6T	CCAGG**T**TGCAA
5	T7A	CCAGGC**A**GCAA
6	T7C	CCAGGC**C**GCAA
7	T7G	CCAGGC**G**GCAA
8	G8A	CCAGGCT**A**CAA
9	G8C	CCAGGCT**C**CAA
10	G8T	CCAGGCT**T**CAA

*The variations from the native 11-mer sequence are underlined and indicated in bold.

First, I looked at the CD character of the 11-mer in conditions similar to the 40 mM $BaCl_2$ crystallization condition. CD spectra of the 11-mer oligonucleotide titrated with Ba^{2+} (up to 100 mM) showed the appearance of a positive band at ~240 nm, a strong negative band at ~255 nm, a positive band at 280 nm, and a weak negative band at ~295 nm (Figure 5.8a). While these spectral features are not identical to those of the antiparallel G4 or i-motif alone, they appear to be consistent with a hybrid structure containing both of these motifs. The antiparallel G4 shows a distinctive positive band at ~240 nm and a negative band at ~260 nm and the i-motif shows a characteristic negative band at ~260 nm and a positive band at ~290 nm.[114, 115] The observed spectra

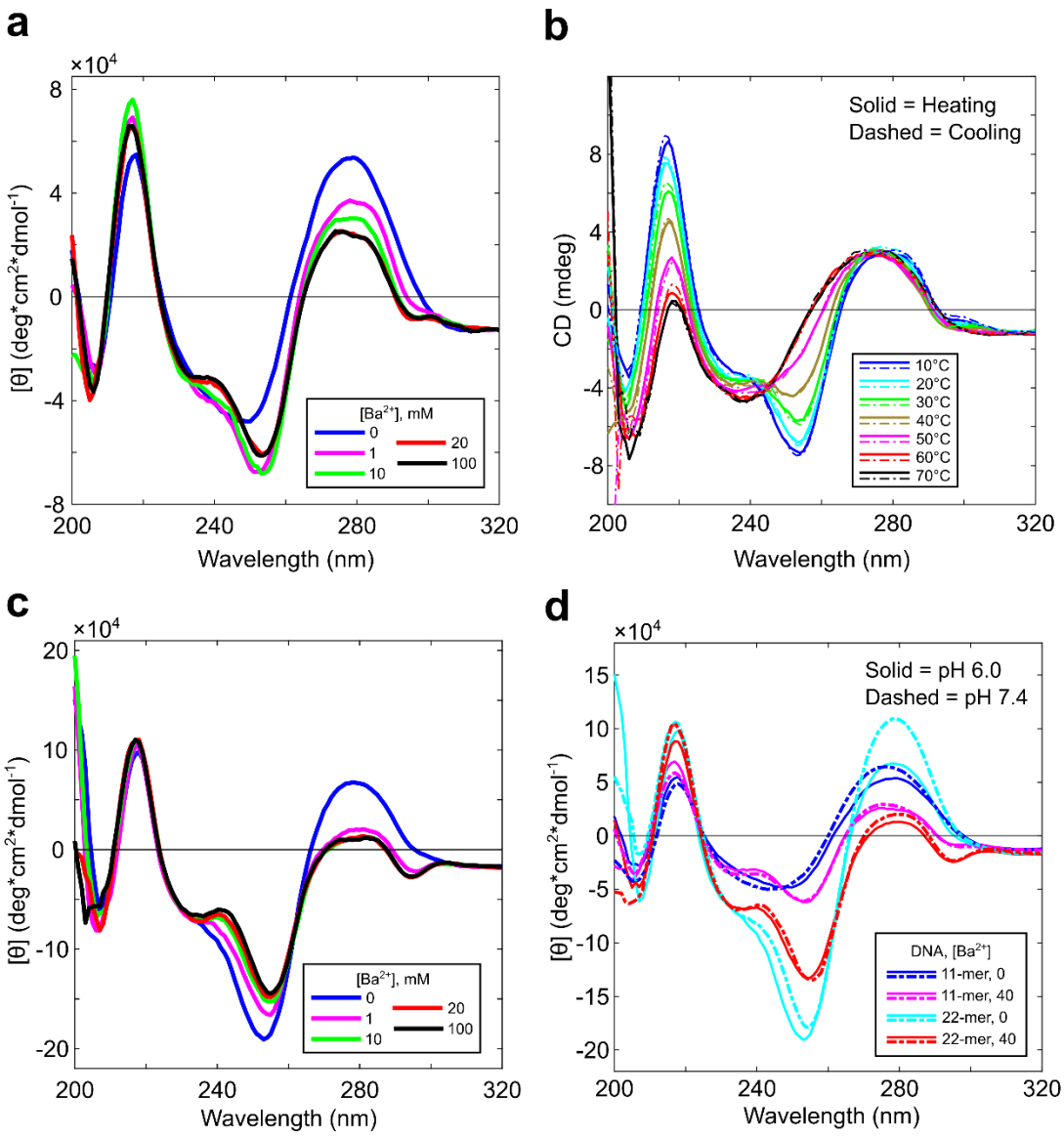

Figure 5.8. CD Spectra of the 11-mer and 22-mer in $BaCl_2$ Conditions. (a) Ba^{2+} titration up to 100 mM for the 11-mer. **(b)** Forward (solid) and reverse (dashed) CD melting curves of the 11-mer at 40 mM Ba^{2+} collected at 10°C increments between 10°C and 70°C. **(c)** Ba^{2+} titration up to 100 mM for the 22-mer. **(d)** CD spectra of the 11-mer and 22-mer in sodium cacodylate buffer alone or supplemented with 40 mM Ba^{2+} at pH 6.0 (solid) and 7.4 (dashed).

show a slight blue shift of the negative band to ~255 nm and the positive band to ~280 nm and in the presence of the weak negative band at ~295 nm. The presence of both motifs within the same structure, as well as additional interactions from the A─A─T

base triple and capping A─A base pair, likely contribute to the deviations from the characteristic spectra. CD melting analysis suggested that the observed spectral features were due to the formation of a specific structure, as the characteristic peaks disappeared with increasing temperature. Structural assembly appears to be reversible, as shown by the nearly identical forward and reverse temperature dependence spectra (Figure 5.8b).

The 22-mer had a similar CD profile with more pronounced characteristic peaks, suggesting that the tandem repeat forms the same or similar structure as that of the 11-mer (Figure 5.8c). Further, the 1D ^{1}H-NMR profile of the 22-mer showed peaks at the same chemical shift ranges in the imino region (Figure 5.9), also indicating

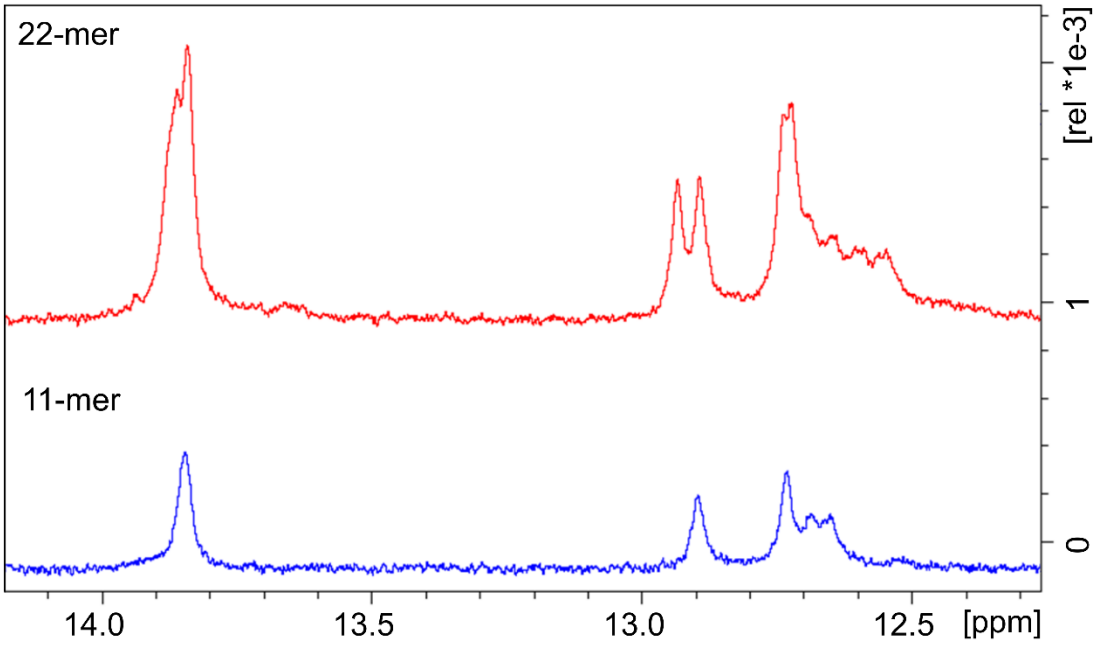

Figure 5.9. Comparison of the Imino Regions of the 1D ^{1}H-NMR Spectra Between the 11-mer and 22-mer. The imino region of the 11-mer (blue) and 22-mer (red) profiles in 30 mM sodium cacodylate buffer at pH 6.0 supplemented with 40 mM $BaCl_2$ are shown to present the peaks in the same chemical shift region and to demonstrate that the pattern for the 22-mer is more complicated. The 11-mer and 22-mer oligonucleotides were prepared at a final concentration of 500 and 300 µM, respectively.

structural similarities. As expected, the 22-mer showed a more complicated spectra with additional peaks as a result of the longer sequence, which hindered subsequent NMR analyses. Nonetheless, both the similar CD and 1D ^{1}H-NMR profiles suggest that the 22-mer, like the 11-mer, assembles into the hybrid G4/i-motif structure.

Chapter 5.2.3. pH Dependence and Thermal Stability

Interestingly, structural assembly was not pH-dependent, as the same characteristic CD profile was observed at pH 6.0 and 7.4 for both sequence lengths (Figure 5.8d). This suggests that the hybrid G4/i-motif structure can form at physiological pH, which has been observed in other i-motifs.[29, 116] To further explore the effect of pH on structure formation, I altered the pH environment of the 11-mer sample by titrating in sodium hydroxide and monitored the imino regions of the 1D ^{1}H-NMR profile. While typical i-motifs structures form in acidic conditions and have been shown to stably exist at neutral pH, they have not been observed at alkaline environments. We hypothesized that the hybrid G4/i-motif structure would not stably exist in basic conditions. However, titration up to pH 10.0 showed a 1D ^{1}H-NMR profile that is very similar to that at pH 6.0. The imino signals assigned to T7H3, G4H1, G5H1, C2H3 are present at the same relative intensity (Figure 5.10), suggesting the formation of the hybrid G4/i-motif. The noticeable differences between the two spectra are seen in the lack of degeneracy of the C2H3 peak at 12.6 ppm and in the absence of the artifact at 10.5 ppm. The manifestation of the C2H3 resonance as one signal indicates that the C2─C2^{+} base pair is likely in one conformation and that the overall sample is more uniform. These preliminary results suggest that there is no significant change in structural assembly at pH 10. However, i-motif formation in alkaline

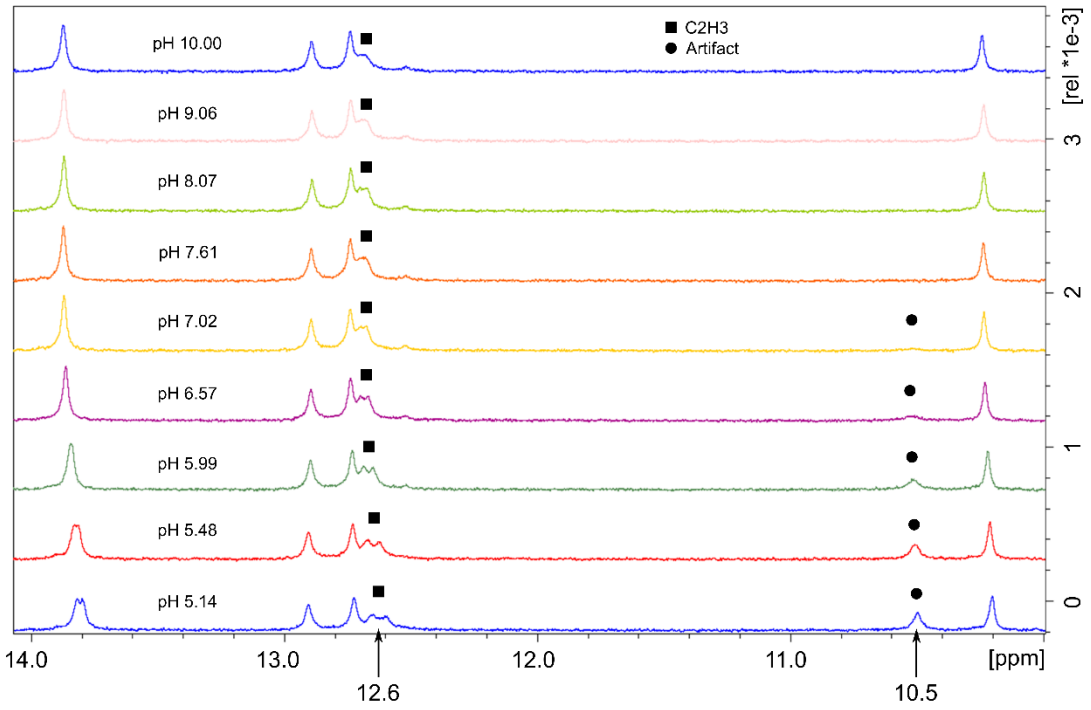

Figure 5.10. NaOH Titration of the 11-mer Monitored by 1D ^{1}H-NMR. Chemical shifts in the imino region are shown from pH 5.14 to pH 10.00. The two signals that undergo noticeable changes with increasing pH are at 12.6 ppm and 10.5 ppm. The 12.6 ppm signal, which is assigned to C2H3, is denoted by a black square. The 10.5 ppm signal is an artifact and is denoted by a black circle. The sample was initially prepared in 500 µL volume containing 300 µM DNA, 40 mM $BaCl_2$, and 30 mM sodium cacodylate buffer at pH 5.14. The pH was increased by titrating in 2 M NaOH.

conditions remains dubious, as protonation of the cytosine N3 seems very unlikely at pH values 10^6 times more basic than its typical pK_a at 4.3.[106] Additional studies will be necessary to reach more definitive conclusions.

Next, I investigated the thermal stabilities of the hybrid G4/i-motif structure formed by both the 11- and 22-mer oligonucleotides. UV melting analysis showed a dramatic difference in melting temperature (T_m) between these two assemblies: 41.7±1.3°C for the tetrameric assembly and 73.7±2.5°C for the dimeric assembly (Figure 5.11/Table 5.3). The fewer number of DNA strands in the dimeric quadruplex

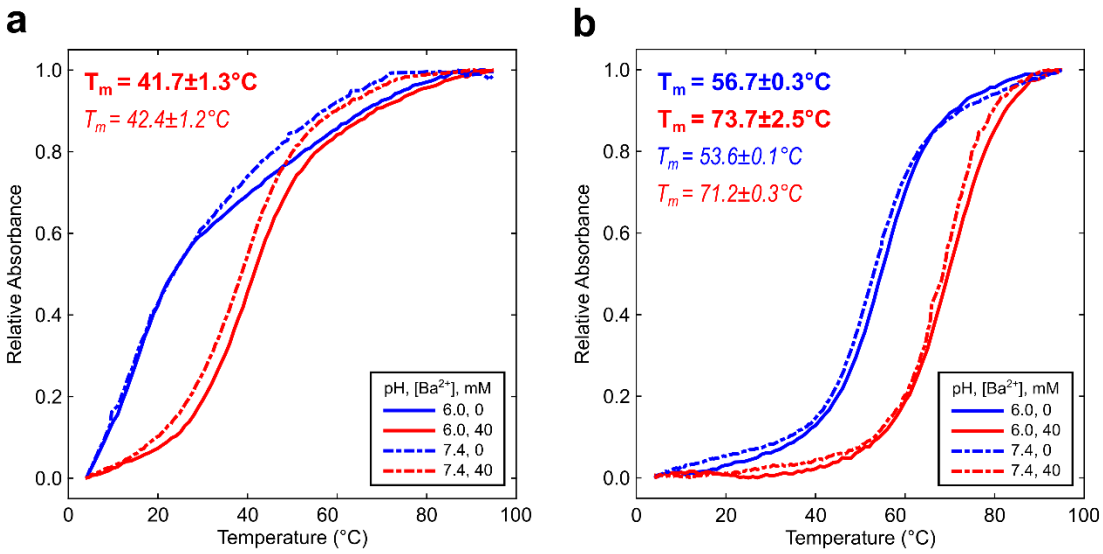

Figure 5.11. Comparison of Absorbance Spectra at Acidic and Neutral pH. (a-b) Melting curves for both pH values at 0 mM Ba^{2+} (blue) and 40 mM Ba^{2+} (red) are shown for the **(a)** 11-mer and **(b)** 22-mer in 30 mM sodium cacodylate buffer. Solid curves and T_m values in bold correspond to pH 6.0. Dashed curves and T_m values in italics correspond to pH 7.4.

Table 5.3. Melting Temperature (T_m) Values for All Reported Experiments.

Condition		T_m (°C)	
		11-mer	**22-mer**
pH 6.0	**Sodium Cacodylate**	N/A	56.7±0.3
	40 mM Ba^{2+}	41.7±1.3	73.7±2.5
	40 mM Ca^{2+}	38.1±1.4	64.4±1.0
	40 mM Sr^{2+}	41.9±1.2	70.6±1.2
	100 mM K^{+}	37.2±3.9	62.9±0.2
	100 mM Na^{+}	37.3±4.1	65.3±1.3
pH 7.4	**Sodium Cacodylate**	N/A	53.6±0.1
	40 mM Ba^{2+}	42.4±1.2	71.2±0.3

likely results in reduced end fraying, which can account for the apparent stability increase in melting experiments. Moreover, the thermal stabilities of both the tetrameric and dimeric assemblies at neutral pH were not significantly different (42.4±1.2°C and 71.2±0.3°C, respectively), as shown by the nearly superimposable melting curves (Figure 5.11/Table 5.3). Along with the CD, these results indicate the formation of

similar structural assemblies with comparable stabilities at physiological pH, suggesting the robustness of the hybrid G4/i-motif structure.

Chapter 5.2.4. Cation Dependence and Specificity

Initial conditions chosen for solution studies on the hybrid G4/i-motif were based on the observation of the Ba^{2+} coordination of the two G-tetrads in the crystal structure. However, monovalent and other divalent cations have also been shown to stabilize G4s, with monovalents organized on every G-tetrad plane and divalents between the planes. The general trend for G4 stabilization by cations has been observed in the following order: $Sr^{2+} > Ba^{2+} > Ca^{2+} > Mg^{2+} > K^{+} > Rb^{+} > Na^{+} > Li^{+} = Cs^{+}$.[117, 118] From this list, I chose to analyze both the 11- and 22-mer oligonucleotides in two divalent, Ca^{2+} and Sr^{2+}, and two monovalent, K^{+} and Na^{+}, conditions and compared each group to the Ba^{2+} results.

CD spectra of the 11-mer in Ca^{2+} and Sr^{2+} were similar to that in Ba^{2+}, showing a shoulder at ~240 nm, a negative band at ~255 nm, and a positive band at 280 nm, but lacking the negative band at ~295 nm (Figure 5.12a). Further, the shoulder observed in Ca^{2+} and Sr^{2+} is weaker than that in Ba^{2+}. Because of the weak shoulder and the absence of the negative band at ~295 nm, it is unclear if the 11-mer forms a structure identical to the hybrid G4/i-motif in Ca^{2+} or Sr^{2+} conditions. On the other hand, CD spectra of the 22-mer in Ca^{2+} and Sr^{2+} showed the characteristic ~240 nm shoulder and the negative bands at ~255 nm and ~295 nm at the same relative intensities as Ba^{2+} (Figure 5.12b). This is a strong indication that the 22-mer also assembles into the hybrid G4/i-motif structure in Ca^{2+} and Sr^{2+} conditions. UV melting analysis of the 11-mer in Ca^{2+} and Sr^{2+} showed a two-state transition curve for both conditions, with similar T_m values

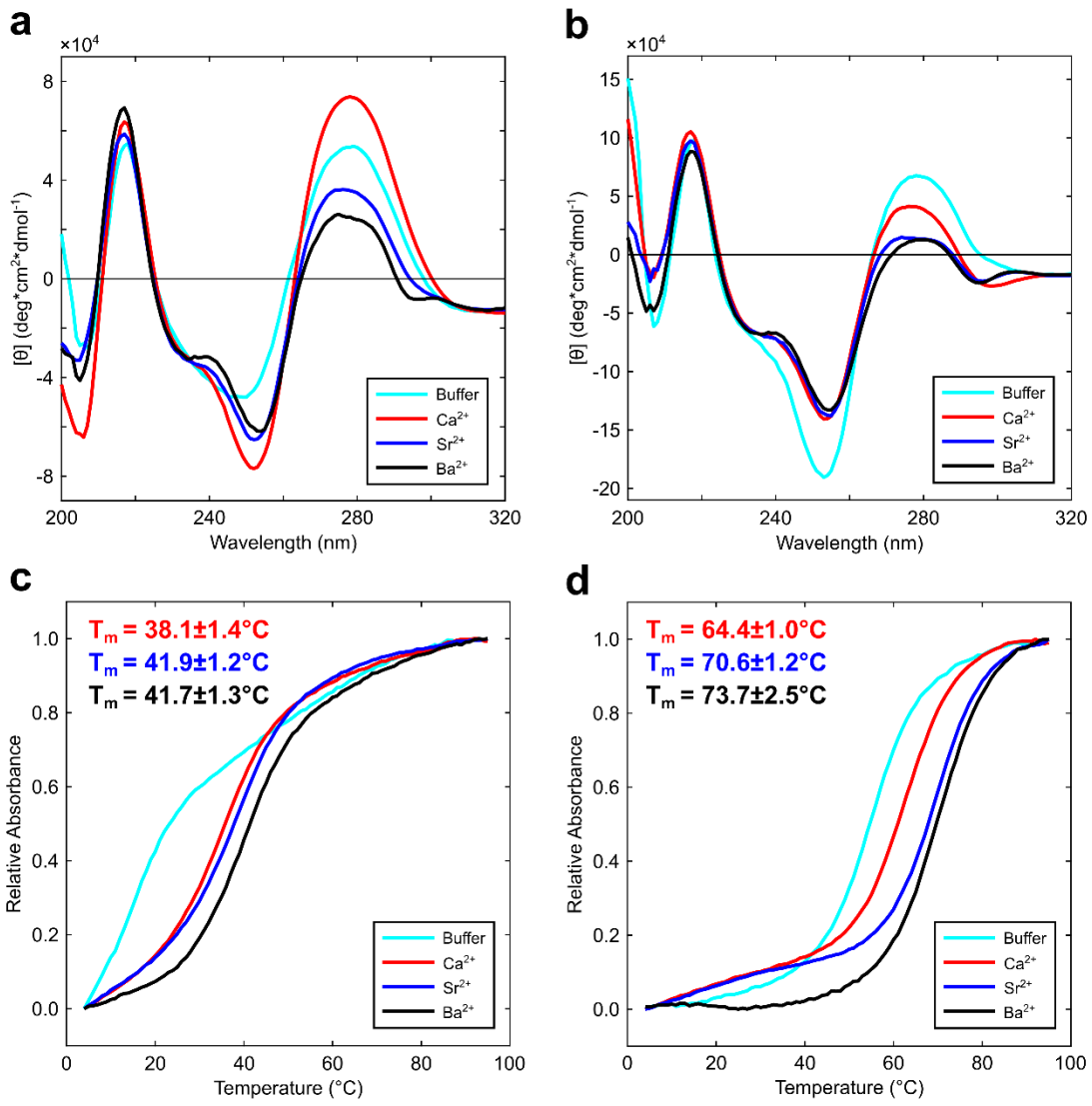

Figure 5.12. Comparison of CD and Absorbance Spectra in Divalent Cation Conditions. (a-b) CD spectra of the **(a)** 11-mer and **(b)** 22-mer are shown in 30 mM sodium cacodylate buffer at pH 6.0 alone or supplemented with 40 mM Ca^{2+}, Sr^{2+}, or Ba^{2+}. **(c-d)** Melting curves of the **(c)** 11-mer and **(d)** 22-mer are shown in the same conditions. T_m values for all two-state transition curves are indicated.

(38.1±1.4°C and 41.9±1.2°C, respectively) as that in Ba^{2+} (Figure 5.12c/Table 5.3). While it is unclear if the structure formed is the hybrid G4/i-motif, we can conclude that the assembly has comparable stability. Interestingly, the 22-mer, which appears to form the hybrid G4/i-motif in Ca^{2+} and Sr^{2+}, has decreased stability in Ca^{2+} with a T_m

of 64.4±1.0°C, but similar stability in Sr^{2+} with a T_m of 70.6±1.2°C, as compared to Ba^{2+} (Figure 5.12d/Table 5.3).

Both the tetrameric and dimeric assemblies showed a preference for Ba^{2+} over monovalent cations. The spectra for both sequence lengths differed slightly in monovalent cations (K^+ or Na^+), with respect to Ba^{2+}, showing a shoulder at ~240 nm and a negative band at ~255 nm, but lacking the ~295 nm negative band (Figure 5.13a). These bands were largely absent from the 11-mer in buffer alone, suggesting additional cations were necessary for structure formation. The 22-mer showed comparable CD profiles between 100 mM monovalent conditions and buffer only (Figure 5.13b), suggesting that the Na^+ cation from the cacodylate buffer was sufficient to induce some assembly. However, because both the 11-mer and 22-mer show a noticeably weak shoulder at ~240 nm and also lack the negative band at ~295 nm, it is not clear if the hybrid G4/i-motif is formed in monovalent conditions. Thermal denaturation experiments of the 11-mer showed no observable melting transition in conditions containing the buffer alone, whereas melting transitions were observed with additional K^+ or Na^+ (37.2±3.9°C and 37.3±4.1°C, respectively; Figure 5.13c/Table 5.3). For the 22-mer, a melting transition at 56.7±0.3°C was observed in buffer, while additional K^+ or Na^+ increased the T_m (62.9±0.2°C and 65.3±1.3°C, respectively; Figure 5.13d/Table 5.3).

From these results, we observe two different trends of the strength of cation stabilization. For the 11-mer, the order shows $Ba^{2+} \approx Sr^{2+} > Ca^{2+} > K^+ \approx Na^+$. While the melting transitions show slight differences, the CD profiles in Ca^{2+} and Sr^{2+} are comparable to those in K^+ and Na^+ (Figure 5.12a, 5.13a), indicating ambiguity in the

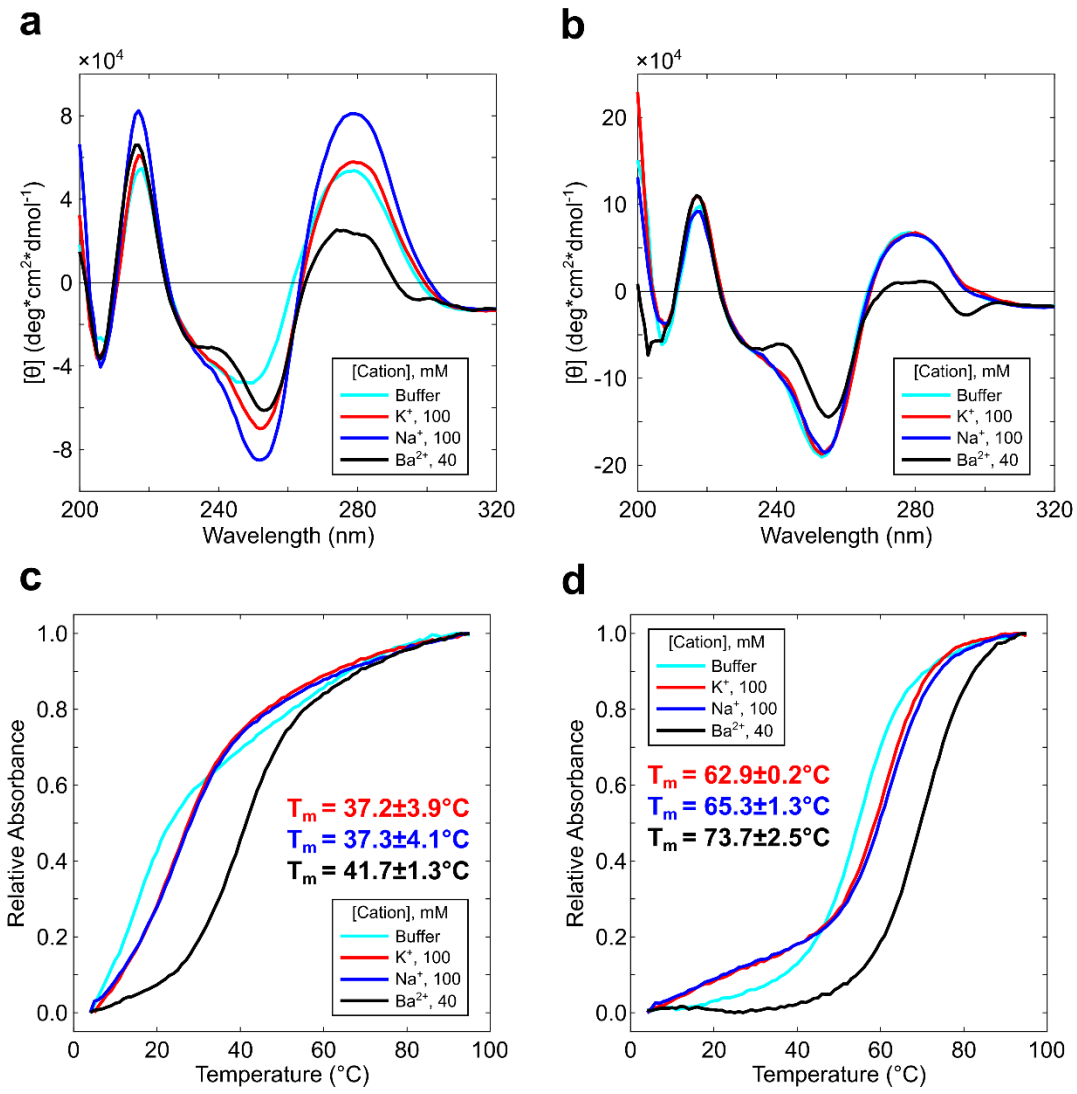

Figure 5.13. Comparison of CD and Absorbance Spectra Between Monovalent and Ba^{2+} Conditions. (a-b) CD spectra of the **(a)** 11-mer and **(b)** 22-mer are shown in 30 mM sodium cacodylate buffer at pH 6.0 alone or supplemented with 100 mM K^+, 100 mM Na^+, or 40 mM Ba^{2+}. **(c-d)** Melting curves of the **(c)** 11-mer and **(d)** 22-mer are shown in the same conditions. T_m values for all two-state transition curves are indicated.

type of structural assembly. For the 22-mer, however, CD spectra in Ca^{2+}, Sr^{2+}, and Ba^{2+} conditions clearly show pronounced characteristic peaks, indicating that divalent cations induce assembly into the hybrid G4/i-motif structure (Figure 5.12b, 5.13b). The cation stabilization order shows $Ba^{2+} > Sr^{2+} > Na^+ > Ca^{2+} > K^+$. For both the 11-mer

and 22-mer oligonucleotides, the observed T_m values in monovalent conditions were lower than those in Ba^{2+}. Even among divalents, both sequence lengths showed a preference for Ba^{2+}, suggesting that the Ba^{2+} cation plays a significant role in structural stability.

Chapter 5.2.5. Analysis of Mutations of the Variable Spacer Region

In Chapter 4.1.4, I discussed the variable spacer region, which is comprised of the C6, T7, and G8 residues and bridges the central G-tetrads and the peripheral i-motif. Based on the crystal structure, we proposed that these residues are necessary to accommodate the geometric differences between the two motifs. To explore the structural importance of these residues, I synthesized all nine possible point mutations (Table 5.2) and compared the CD profiles in Ba^{2+}.

The C6A, C6G, and C6T variants showed CD profiles lacking the characteristic peaks associated with the hybrid G4/i-motif (Figure 5.14a), suggesting the formation of a different structure. In the T7N group, the T7A and T7C variants also lacked the

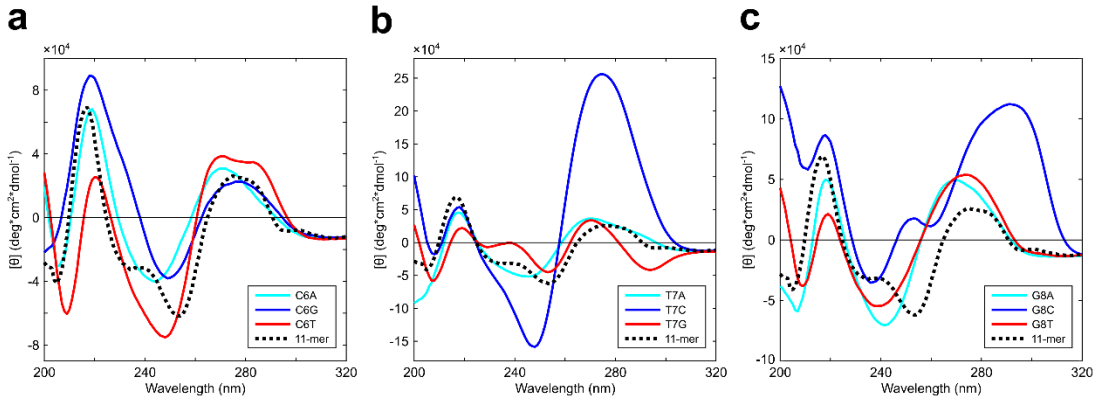

Figure 5.14. Comparison of CD spectra of d(CCAGGCTGCAA) Variants. (a-c) The profiles of **(a)** C6N **(b)** T7N **(c)** G8N are compared with the 11-mer reference spectra (dotted black). All DNA were prepared at a final concentration of 100 μM in 30 mM sodium cacodylate buffer at pH 6.0 containing 40 mM $BaCl_2$.

distinctive peaks; however, T7G displayed a characteristic profile very similar to that of the 11-mer (Figure 5.14b). This prompted me to further investigate this variant oligonucleotide by comparing the imino region of the 1D ^{1}H-NMR profile with that of the 11-mer. The results showed starkly different peak patterns between 8 and 14 ppm (Figure 5.15, red). The T7G variant profile shows additional peaks, as compared to the 11-mer. While some peaks are in the same chemical shift region as the 11-mer, many are in different ones as well. This suggests that the local chemical environment of the T7G structure is quite different from the 11-mer hybrid G4/i-motif. The differences in the NMR profiles do not preclude the possibility of formation of the hybrid G4/i-motif by the T7G oligonucleotide; the observed results merely indicate that there are significant differences in the hydrogen bonding patterns of the structure. Further studies will be necessary to elucidate these structural differences.

Lastly, the G8N group showed similar profiles in the G8A and G8T

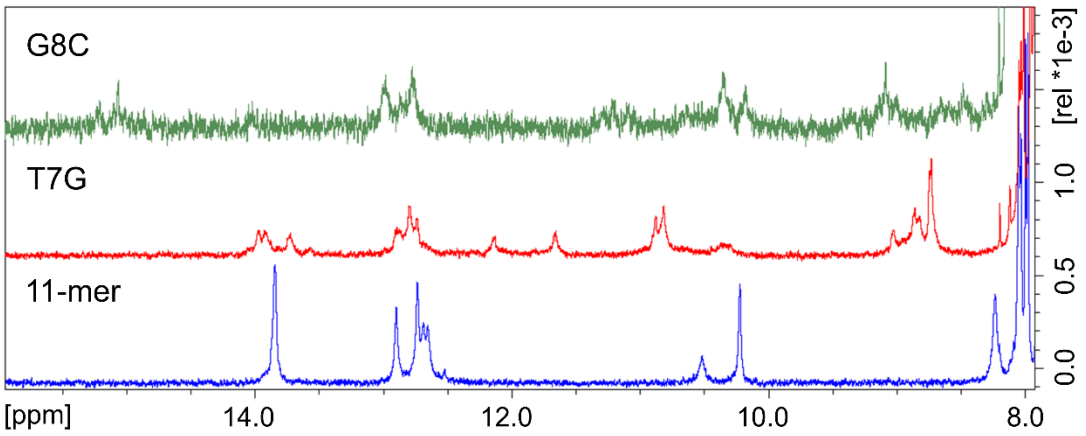

Figure 5.15. Comparison of the 1D ^{1}H-NMR Profiles Between the 11-mer and Variants. The imino regions of the T7G (red) and G8C (green) variants are juxtaposed with that of the 11-mer (blue). Both variants present additional signals and show different peak patterns in the 8.0–15.0 ppm chemical shift range. All DNA were prepared at a final concentration of 300 μM in 30 mM sodium cacodylate buffer at pH 6.0 containing 40 mM $BaCl_2$.

oligonucleotides, both of which lack the hybrid G4/i-motif characteristic peaks, whereas G8C showed a completely different profile than that of the 11-mer (Figure 5.14c). The G8C profile is more representative of an i-motif,[38] with a broad positive band at 290 nm. 1D ^{1}H-NMR analysis of G8C showed weak imino signals at 15 ppm in addition to an altered peak pattern (Figure 5.15, green), confirming the formation of a structure completely different from the 11-mer hybrid G4/i-motif described in Chapter 4. These preliminary results show that the variant oligonucleotides have structural differences from the native G4/i-motif hybrid, but further solution studies combined with crystal structure analyses will be essential to gain additional insight.

Chapter 5.3: Biological Implications

Biologically, the hybrid G4/i-motif hybrid structure hints at the potential complexity of noncanonical DNA structures that may be harbored within genomes. The demonstration that the DNA studied here forms a highly stable dimeric structure from tandem repeats suggests that longer repetitive sequences may have the ability to form complex structures, perhaps containing existing known DNA motifs. Repetitive DNA makes up >50% of the human genome,[69] with microsatellite (1–10 nt), minisatellite (10 to several hundred nt), and macrosatellite (up to thousands of nt) repeats, making up ~3%.[119] Satellite DNA is involved in a variety of biological functions and pathologies,[120] and repeat sequences have been implicated as drivers of evolution through the formation of noncanonical structures that result in genomic instability.[121] Though there are now some examples for how repetitive DNA can impact biological function, the structural basis for this is largely unknown. The discovery of new, stable noncanonical DNA structures suggests the possibility that these repeat sequences can

form complex motifs that may not be predictable from existing sequence/structure relationships.

Chapter 5.4: Experimental Procedures

Chapter 5.4.1. Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR data were acquired on a Bruker Avance III 600-MHz spectrometer equipped with a Cryo-TCI probe. 1D ^{1}H-NMR experiments were conducted using varying DNA concentrations and different buffer conditions. All 1D spectra were prepared with 7% D_2O, collected at 10°C, and analyzed in TopSpin. Specific details will be described in the figure captions. For 2D ^{1}H-NMR experiments, the 11-mer oligonucleotide was prepared at 500 μM in 30 mM sodium cacodylate buffer at pH 6.0 containing 40 mM $BaCl_2$ and 7% D_2O. The sample was lyophilized and dissolved in 100% D_2O for subsequent analysis. A combination of 2D-NOESY and 2D-TOCSY experiments were performed, in which the mixing time was set to 100 and 90 ms, respectively. For both experiments, multiple mixing times were initially tested and the ones that yielded spectra with the best resolution were selected. Both 1D and 2D experiments were conducted at 10°C. The oligonucleotide sequential assignment was conducted using the CARA program.[95]

A combination of NOESY and TOCSY spectra of the 11-mer was used to generate proton assignments. First, the characteristic cytosine H5/H6 NOE signals were identified in the 2D-TOCSY spectrum, allowing cytosines to be distinguished from the other three nucleotides. Next, identification of NOEs between nucleobase protons and sugar H1′ and H2′ protons in the 6.5–8.0 ppm region established sequential

connectivity between multiple residues. We observed more signals than expected, indicating the presence of multiple conformations. In addition, overlapping resonances in this region complicated the assignment. Nonetheless, the initial assignments were internally consistent and were verified through proton signals in the 5–6 ppm range. Finally, imino signals were assigned based on NOEs between the protons on its own ring and those from their base pairing partner. Exchangeable proton signals were confirmed from the NOESY and TOCSY spectra collected on the 100% D_2O sample.

Chapter 5.4.2. Circular Dichroism (CD) Spectroscopy

CD spectra were acquired using the Jasco J-810 spectropolarimeter fitted with a thermostated cell holder. Samples were prepared in 30 mM sodium cacodylate buffer at pH 6.0 or 7.4 containing differing concentrations of monovalent or divalent cations. Specific experimental details are provided in the figure captions. All 11-mer and 22-mer samples were prepared at final DNA concentrations of 100 and 75 μM, respectively. Variant (C6N, T7N, G8N) oligonucleotide samples were prepared at final DNA concentrations of 100 μM. All samples were equilibrated for 12–18 hours at 4°C prior to the acquisition of the spectra at room temperature. CD spectra were collected from 200 to 320 nm at a rate of 50 nm/min and with a data pitch of 1.0 nm. For melting experiments, the sample was allowed to dwell for 7 minutes at the temperature set point.

Chapter 5.4.3. Thermal Denaturation

UV melting spectra were acquired using the Cary100 Bio UV-visible spectrophotometer equipped with a 12-cell sample changer and a Peltier

heating/cooling system. The sample chamber was purged with N_2 throughout both melting and annealing data collection runs. Samples were prepared in 30 mM sodium cacodylate buffer at pH 6.0 or 7.4 containing differing concentrations of monovalent or divalent cations. Specific experimental details are provided in the figure captions. All 11-mer and 22-mer samples were prepared at final DNA concentrations of 14.4 and 7.25 μM, respectively, and were equilibrated for 15–20 hours at 4°C prior to the acquisition of the spectra. They were then were transferred to self-masking quartz cuvettes with 1.0 cm path length for UV absorbance measurements. All spectra were collected at 260 nm. An initial fast heating ramp from 4°C to 95°C at 10°C/min was done to remove any pre-formed aggregates. Data were collected every 1°C during a slow cooling ramp from 95°C to 4°C at 1°C/min and a subsequent slow heating ramp at the same temperature range and rate. Thermal melting analyses and curve fitting were conducted using MATLAB.

Chapter 6: Conclusions and Future Perspectives

Chapter 6.1: Summary

Chapter 6.1.1: The d(CGTAAGGCG) Fold-back Quadruplex

In Chapters 2 and 3, I discussed the self-assembly of the d(CGTAAGGCG) and three variant oligonucleotides into a non-G-quadruplex fold-back quadruplex through both Watson-Crick and noncanonical interactions. The tetrameric assembly encloses a central cation binding pocket and features a hexad base pairing arrangement. Solution studies revealed that structure formation is dependent on divalent cations. While the native oligonucleotide assembles in the presence of magnesium ions, the three variants demonstrate that barium and calcium ions can also be incorporated into the central cavity. Although other fold-back structures have been reported in the literature, this is the first fold-back structure that forms a tetramer and is specific for divalent cations.

Chapter 6.1.2: The d(CCAGGCTGCAA) Hybrid Quadruplex

In Chapters 4 and 5, I described the hybrid G-quadruplex/i-motif structure formed by the d(CCAGGCTGCAA) oligonucleotide. The structure features a Ba^{2+}-stabilized G-quadruplex, which is flanked on either side by a base triple formed through noncanonical interactions and an i-motif. While the G4 and i-motif have been observed individually in different structures, this is the first assembly in which both are seen simultaneously. It is somewhat surprising that these two motifs with such large geometric differences are coexisting in one structure, but analysis of this hybrid assembly reveals a spacer region that facilitates gradual changes in the interstrand

backbone distances, thus allowing for both the G4 and i-motif to coexist in one structure.

Chapter 6.2: Broader Impacts and Future Perspectives

The B-form double helix has been the icon of DNA structure since its discovery in 1953, but the identification of numerous non-duplex structures in the last 67 years has greatly expanded the field of DNA structural biology. My doctoral work further adds to the repertoire of noncanonical DNA structures. In this dissertation, I have described two types of structures with previously unobserved structural features. The elucidation of the fold-back tetramer and the hybrid G4/i-motif structures suggests that there are greater biological roles for non-duplex structures. Not only does my work add to the growing structural diversity of DNA, it also validates the structure-before-function approach for discovering new DNA structures.

There is great potential for genomes to contain sequences capable of forming noncanonical DNA structures that we do not even know about yet. But how do we go about finding new structures if we don't know what they look like? This structure-before-function method employs the screening of a library of short, randomized DNA sequences and therefore allows us to search a large amount of sequence space in an unbiased way with the goal of identifying biologically relevant non-duplex DNA structures. Based on the two types of structures identified in my doctoral research, we can envision that many more new DNA motifs can be uncovered through this innovative approach. And beyond DNA, we can even apply this method to RNA sequences and identify new types of RNA structures that may also have biological

significance, which would further expand our understanding of noncanonical nucleic acid structures.

In addition to the two families of DNA oligonucleotides described in this dissertation, I have attempted to crystallize and solve the structures of many more from the sequence library. While not all the sequences I worked with crystallized, I moved forward with the ones that developed crystalline material from the initial screening. To date, I have obtained diffraction-quality crystals and attempted to collect diffraction data for the oligonucleotides listed in Table 6.1, but have not been able to progress towards structure determination due to either poor data quality or difficulties in data processing.

Table 6.1. List of DNA Oligonucleotides that Yielded Robust Crystals.

Name	d(Sequence)	Length (nt)	Diffraction Data Status
R1-07	GAGCGAGAACAG	12	Low Resolution
R1-17	GAGGTGATTGA	11	Poor Anomalous Signal
R1-95	CGAAGGTCCCTGT	13	Low Resolution
R2-27	TAACCATCCCA	11	No Diffraction
R2-57	ACTCCGCCCTTT	12	Poor Anomalous Signal
R3-19	AGCGCCATGGCG	12	Poor Diffraction
R3-48	GGTCAGATGT	10	Poor Anomalous Signal
R3-60	TCCGCGGGAT	10	No Diffraction
R3-72	TCTTCCACCA	10	Poor Diffraction
Rui22	CGGATGGGCTA	11	No Diffraction

Although I have collected diffraction data at reasonable resolution for the R1-17, R2-57, and R3-48 sequences, the poor anomalous signal has posed a significant challenge in data processing. Without accurate identification of the heavy atom positions in the electron density maps, we have not been able to build atomic models that reflect the DNA structural assembly with high certainty. For these three DNA

families, it would be worthwhile to meticulously analyze and merge the diffraction data that we have already collected in order to obtain data with stronger anomalous signal. As for the other seven DNA families, the next immediate experiments would be focused on optimizing conditions to grow new crystals because every crystal is unique and can have varying degrees of order, leading to different diffraction patterns. Suggestions include identifying new crystallization conditions that would propagate slower crystal growth and improving cryo-cooling conditions to minimize ice formation, both of which could lead to more ordered crystals.

Ultimately, by identifying new DNA motifs that utilize noncanonical base pairing interactions, we can better understand the predictability of noncanonical motifs. This would complement our current knowledge of Watson-Crick base pairing and significantly expand the available toolkit to rationally design more precise DNA crystal structures that are targeted for specific nanotechnological applications.

References

1. Franklin, R. E.; Gosling, R. G., Molecular configuration in sodium thymonucleate. *Nature* **1953,** *171* (4356), 740-741.
2. Watson, J. D.; Crick, F. H., Genetical implications of the structure of deoxyribonucleic acid. *Nature* **1953,** *171* (4361), 964-967.
3. Watson, J. D.; Crick, F. H., Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **1953,** *171* (4356), 737-738.
4. Donohue, J., Hydrogen-bonded helical configurations of polynucleotides. *Proc. Natl. Acad. Sci. U.S.A.* **1956,** *42* (2), 60-65.
5. Hoogsteen, K., The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta. Crystl.* **1963,** *16*, 907-916.
6. Gellert, M.; Lipsett, M. N.; Davies, D. R., Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. U.S.A.* **1962,** *48* (12), 2013-2018.
7. Gehring, K.; Leroy, J. L.; Gueron, M., A tetrameric DNA structure with protonated cytosine.cytosine base pairs. *Nature* **1993,** *363* (6429), 561-565.
8. Salisbury, S. A.; Wilson, S. E.; Powell, H. R.; Kennard, O.; Lubini, P.; Sheldrick, G. M.; Escaja, N.; Alazzouzi, E.; Grandas, A.; Pedroso, E., The bi-loop, a new general four-stranded DNA motif. *Proc Natl Acad Sci USA* **1997,** *94* (11), 5515-5518.
9. Robinson, H.; van der Marel, G. A.; van Boom, J. H.; Wang, A. H., Unusual DNA conformation at low pH revealed by NMR: parallel-stranded DNA duplex with homo base pairs. *Biochemistry* **1992,** *31* (43), 10510-10517.
10. Robinson, H.; Wang, A. H., 5'-CGA sequence is a strong motif for homo base-paired parallel-stranded DNA duplex as revealed by NMR analysis. *Proc. Natl. Acad. Sci. U.S.A.* **1993,** *90* (11), 5224-5228.
11. Sen, D.; Gilbert, W., Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **1988,** *334* (6180), 364-366.
12. Williamson, J. R., Guanine quartets. *Curr. Opin. Struct. Biol.* **1993,** *3* (3), 357-362.
13. Davis, J. T., G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. *Angew. Chem. Int. Ed. Engl.* **2004,** *43* (6), 668-698.
14. Williamson, J. R., G-quartet structures in telomeric DNA. *Annu. Rev. Biophys. Biomol. Struct.* **1994,** *23*, 703-730.
15. Hud, N. V.; Plavec, J., The Role of Cations in Determining Quadruplex Structure and Stability. In *Quadruplex Nucleic Acids.*, Neidle, S.; Balasubramanian, S., Eds. The Royal Society of Chemistry: Cambridge, UK, 2006; pp 100-130.
16. Leroy, J. L.; Gueron, M., Solution structures of the i-motif tetramers of d(TCC), d(5methylCCT) and d(T5methylCC): novel NOE connections between amino protons and sugar protons. *Structure* **1995,** *3* (1), 101-120.
17. Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S., Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **2006,** *34* (19), 5402-5415.
18. Gueron, M.; Leroy, J. L., The i-motif in nucleic acids. *Curr. Opin. Struct. Biol.* **2000,** *10* (3), 326-331.

19. Todd, A. K.; Johnston, M.; Neidle, S., Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* **2005,** *33* (9), 2901-2907.
20. Huppert, J. L.; Balasubramanian, S., Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **2005,** *33* (9), 2908-2916.
21. Kocman, V.; Plavec, J., Tetrahelical structural family adopted by AGCGA-rich regulatory DNA regions. *Nat. Commun.* **2017,** *8*, 15355.
22. Mukundan, V. T.; Phan, A. T., Bulges in G-Quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.* **2013,** *135* (13), 5017-5028.
23. Heddi, B.; Martin-Pintado, N.; Serimbetov, Z.; Kari, T. M.; Phan, A. T., G-quadruplexes with (4n-1) guanines in the G-tetrad core: formation of a G-triad.water complex and implication for small-molecule binding. *Nucleic Acids Res.* **2016,** *44* (2), 910-916.
24. Jiang, H. X.; Cui, Y.; Zhao, T.; Fu, H. W.; Koirala, D.; Punnoose, J. A.; Kong, D. M.; Mao, H., Divalent cations and molecular crowding buffers stabilize G-triplex at physiologically relevant temperatures. *Sci. Rep.* **2015,** *5*, 9255.
25. Zhang, N.; Gorin, A.; Majumdar, A.; Kettani, A.; Chernichenko, N.; Skripkin, E.; Patel, D. J., V-shaped scaffold: a new architectural motif identified in an A x (G x G x G x G) pentad-containing dimeric DNA quadruplex involving stacked G(anti) G(anti) x G(anti) x G(syn) tetrads. *J. Mol. Biol.* **2001,** *311* (5), 1063-1079.
26. Lim, K. W.; Phan, A. T., Structural basis of DNA quadruplex-duplex junction formation. *Angew. Chem. Int. Ed. Engl.* **2013,** *52* (33), 8566-8569.
27. Karg, B.; Mohr, S.; Weisz, K., Duplex-Guided Refolding into Novel G-Quadruplex (3+1) Hybrid Conformations. *Angew. Chem. Int. Ed. Engl.* **2019,** *58* (32), 11068-11071.
28. 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. *Nat. Biotechnol.* **2015,** *33* (8), 877-881.
29. Wright, E. P.; Huppert, J. L.; Waller, Z. A. E., Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. *Nucleic Acids Res.* **2017,** *45* (6), 2951-2959.
30. Mergny, J. L.; Lacroix, L.; Han, X.; Leroy, J. L.; Helene, C., Intramolecular Folding of Pyrimidine Oligodeoxynucleotides into an i-DNA Motif. *J. Am. Chem. Soc.* **1995,** *117* (35), 8887-8898.
31. Benabou, S.; Avino, A.; Lyonnais, S.; Gonzalez, C.; Eritja, R.; De Juan, A.; Gargallo, R., i-motif structures in long cytosine-rich sequences found upstream of the promoter region of the SMARCA4 gene. *Biochimie* **2017,** *140*, 20-33.
32. Fujii, T.; Sugimoto, N., Loop nucleotides impact the stability of intrastrand i-motif structures at neutral pH. *Phys. Chem. Chem. Phys.* **2015,** *17* (26), 16719-16722.
33. Sunami, T.; Kondo, J.; Kobuna, T.; Hirao, I.; Watanabe, K.; Miura, K.; Takenaka, A., Crystal structure of d(GCGAAAGCT) containing a parallel-stranded duplex with homo base pairs and an anti-parallel duplex with Watson-Crick base pairs. *Nucleic Acids Res.* **2002,** *30* (23), 5253-5260.
34. Escaja, N.; Viladoms, J.; Garavis, M.; Villasante, A.; Pedroso, E.; Gonzalez, C., A minimal i-motif stabilized by minor groove G:T:G:T tetrads. *Nucleic Acids Res.* **2012,** *40* (22), 11737-11747.

35. Muser, S. E.; Paukstelis, P. J., Three-dimensional DNA crystals with pH-responsive noncanonical junctions. *J. Am. Chem. Soc.* **2012,** *134* (30), 12557-12564.
36. Tripathi, S.; Zhang, D.; Paukstelis, P. J., An intercalation-locked parallel-stranded DNA tetraplex. *Nucleic Acids Res.* **2015,** *43* (3), 1937-1944.
37. Esmaili, N.; Leroy, J. L., i-motif solution structure and dynamics of the d(AACCCC) and d(CCCCAA) tetrahymena telomeric repeats. *Nucleic Acids Res.* **2005,** *33* (1), 213-224.
38. Assi, H. A.; Harkness, R. W.; Martin-Pintado, N.; Wilds, C. J.; Campos-Olivas, R.; Mittermaier, A. K.; Gonzalez, C.; Damha, M. J., Stabilization of i-motif structures by 2'-beta-fluorination of DNA. *Nucleic Acids Res.* **2016,** *44* (11), 4998-5009.
39. Kang, C.; Berger, I.; Lockshin, C.; Ratliff, R.; Moyzis, R.; Rich, A., Stable loop in the crystal structure of the intercalated four-stranded cytosine-rich metazoan telomere. *Proc. Natl. Acad. Sci. U.S.A.* **1995,** *92* (9), 3874-3878.
40. Berger, I.; Kang, C.; Fredian, A.; Ratliff, R.; Moyzis, R.; Rich, A., Extension of the four-stranded intercalated cytosine motif by adenine.adenine base pairing in the crystal structure of d(CCCAAT). *Nat. Struct. Biol.* **1995,** *2* (5), 416-429.
41. Weil, J.; Min, T.; Yang, C.; Wang, S.; Sutherland, C.; Sinha, N.; Kang, C., Stabilization of the i-motif by intramolecular adenine-adenine-thymine base triple in the structure of d(ACCCT). *Acta. Crystallogr. D. Biol. Crystallogr.* **1999,** *55*, 422-429.
42. Mir, B.; Soles, X.; Gonzalez, C.; Escaja, N., The effect of the neutral cytidine protonated analogue pseudoisocytidine on the stability of i-motif structures. *Sci. Rep.* **2017,** *7*, 2772.
43. Day, H. A.; Pavlou, P.; Waller, Z. A., i-Motif DNA: structure, stability and targeting with ligands. *Bioorg. Med. Chem.* **2014,** *22* (16), 4407-4418.
44. Escaja, N.; Gomez-Pinto, I.; Viladoms, J.; Pedroso, E.; Gonzalez, C., The effect of loop residues in four-stranded dimeric structures stabilized by minor groove tetrads. *Org Biomol Chem* **2013,** *11* (29), 4804-4810.
45. Escaja, N.; Pedroso, E.; Rico, M.; Gonzalez, C., Dimeric solution structure of two cyclic octamers: Four-stranded DNA structures stabilized by A:T:A:T and G:C:G:C tetrads. *J Am Chem Soc* **2000,** *122* (51), 12732-12742.
46. Escaja, N.; Gomez-Pinto, I.; Pedroso, E.; Gonzalez, C., Four-stranded DNA structures can be stabilized by two different types of minor groove G:C:G:C tetrads. *J Am Chem Soc* **2007,** *129* (7), 2004-2014.
47. Gallego, J.; Chou, S. H.; Reid, B. R., Centromeric pyrimidine strands fold into an intercalated motif by forming a double hairpin with a novel T:G:G:T tetrad: solution structure of the d(TCCCGTTTCCA) dimer. *J Mol Biol* **1997,** *273* (4), 840-856.
48. Liu, H. H.; Shen, F. S.; Haruehanroengra, P.; Yao, Q. Q.; Cheng, Y. S.; Chen, Y. Q.; Yang, C.; Zhang, J.; Wu, B. X.; Luo, Q.; Cui, R. X.; Li, J. X.; Ma, J. B.; Sheng, J.; Gan, J. H., A DNA structure containing AgI-mediated G:G and C:C base pairs. *Angew Chem Int Ed Engl* **2017,** *56* (32), 9430-9434.
49. Leonard, G. A.; Zhang, S.; Peterson, M. R.; Harrop, S. J.; Helliwell, J. R.; Cruse, W. B. T.; d'Estaintot, B. L.; Kennard, O.; Brown, T.; Hunter, W. N., Self-

association of a DNA loop creates a quadruplex: crystal structure of d(GCATGCT) at 1.8 Å resolution. *Structure* **1995,** *3* (4), 335-340.
50. Thorpe, J. H.; Teixeira, S. C. M.; Gale, B. C.; Cardin, C. J., Crystal structure of the complementary quadruplex formed by d(GCATGCT) at atomic resolution. *Nucleic Acids Res* **2003,** *31* (3), 844-849.
51. Mir, B.; Serrano, I.; Buitrago, D.; Orozco, M.; Escaja, N.; Gonzalez, C., Prevalent sequences in the human genome can form mini i-motif structures at physiological pH. *J Am Chem Soc* **2017,** *139* (40), 13985-13988.
52. Rhodes, D.; Lipps, H. J., G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* **2015,** *43* (18), 8627-8637.
53. Lam, E. Y.; Beraldi, D.; Tannahill, D.; Balasubramanian, S., G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.* **2013,** *4*, 1796.
54. Biffi, G.; Tannahill, D.; McCafferty, J.; Balasubramanian, S., Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **2013,** *5* (3), 182-186.
55. Henderson, A.; Wu, Y.; Huang, Y. C.; Chavez, E. A.; Platt, J.; Johnson, F. B.; Brosh, R. M. J.; Sen, D.; Lansdorp, P. M., Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.* **2014,** *42* (2), 860-869.
56. Huppert, J. L.; Balasubramanian, S., G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **2007,** *35* (2), 406-413.
57. Brooks, T. A.; Kendrick, S.; Hurley, L., Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J.* **2010,** *277* (17), 3459-3469.
58. 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.
59. Dexheimer, T. S.; Sun, D.; Hurley, L. H., Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter. *J. Am. Chem. Soc.* **2006,** *128* (16), 5404-5415.
60. 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. *J. Am. Chem. Soc.* **2017,** *139* (25), 8522-8536.
61. 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. *J. Am. Chem. Soc.* **2009,** *131* (31), 10878-10891.
62. Onel, B.; Carver, M.; Agrawal, P.; Hurley, L. H.; Yang, D., The 3'-end region of the human PDGFR-beta core promoter nuclease hypersensitive element forms a mixture of two unique end-insertion G-quadruplexes. *Biochim. Biophys. Acta. Gen. Subj.* **2018,** *1862* (4), 846-854.
63. Piazza, A.; Serero, A.; Boule, J. B.; Legoix-Ne, P.; Lopes, J.; Nicolas, A., Stimulation of gross chromosomal rearrangements by the human CEB1 and CEB25 minisatellites in Saccharomyces cerevisiae depends on G-quadruplexes or Cdc13. *PLoS Genet.* **2012,** *8* (11), e1003033.

64. Abou Assi, H.; Garavis, M.; Gonzalez, C.; Damha, M. J., i-Motif DNA: structural features and significance to cell biology. *Nucleic Acids Res.* **2018,** *46* (16), 8038-8056.
65. Dzatko, S.; Krafcikova, M.; Hansel-Hertsch, R.; Fessl, T.; Fiala, R.; Loja, T.; Krafcik, D.; Mergny, J. L.; Foldynova-Trantirkova, S.; Trantirek, L., Evaluation of the Stability of DNA i-Motifs in the Nuclei of Living Mammalian Cells. *Angew. Chem. Int. Ed. Engl.* **2018,** *57* (8), 2165-2169.
66. 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. *Nat. Chem.* **2018,** *10* (6), 631-637.
67. Niu, K.; Zhang, X.; Deng, H.; Wu, F.; Ren, Y.; Xiang, H.; Zheng, S.; Liu, L.; Huang, L.; Zeng, B.; Li, S.; Xia, Q.; Song, Q.; Palli, S. R.; Feng, Q., BmILF and i-motif structure are involved in transcriptional regulation of BmPOUM2 in Bombyx mori. *Nucleic Acids Res.* **2018,** *46* (4), 1710-1723.
68. Balasubramanian, S.; Hurley, L. H.; Neidle, S., Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug. Discov.* **2011,** *10* (4), 261-275.
69. Saha, S.; Bridges, S.; Magbanua, Z. V.; Peterson, D. G., Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.* **2008,** *36* (7), 2284-2294.
70. Paukstelis, P. J.; Seeman, N. C., 3D DNA Crystals and Nanotechnology. *Crystals* **2016,** *6* (8), 97.
71. Geng, C.; Paukstelis, P. J., DNA crystals as vehicles for biocatalysis. *J. Am. Chem. Soc.* **2014,** *136* (22), 7817-7820.
72. Paukstelis, P. J., Three-dimensional DNA crystals as molecular sieves. *J. Am. Chem. Soc.* **2006,** *128* (21), 6794-6795.
73. Kantardjieff, K. A.; Rupp, B., Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci* **2003,** *12* (9), 1865-1871.
74. Kettani, A.; Gorin, A.; Majumdar, A.; Hermann, T.; Skripkin, E.; Zhao, H.; Jones, R.; Patel, D. J., A dimeric DNA interface stabilized by stacked A.(G.G.G.G).A hexads and coordinated monovalent cations. *J Mol Biol* **2000,** *297* (3), 627-644.
75. Ho, B. K.; Gruswitz, F., HOLLOW: generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct Biol* **2008,** *8* (49).
76. Leslie, A. G. W.; Powell, H. R., Processing diffraction data with mosflm. In *Evolving Methods for Macromolecular Crystallography.*, Read, R. J.; Sussman, J. L., Eds. NATO Science Series: Springer, Dordrecht, 2007; Vol. 245, pp 41-51.
77. Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D., Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* **2012,** *68*, 352-367.
78. Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S., Overview of the CCP4 suite and current developments. *Acta. Crystallogr. D. Biol. Crystallogr.* **2011,** *67*, 235-242.

79. McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J., Phaser crystallographic software. *J. Appl. Crystallogr.* **2007,** *40*, 658-674.
80. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K., Features and development of Coot. *Acta. Crystallogr. D. Biol. Crystallogr.* **2010,** *66*, 486-501.
81. Kabsch, W., XDS. *Acta. Crystallogr. D. Biol. Crystallogr.* **2010,** *66*, 125-132.
82. Potterton, L.; Agirre, J.; Ballard, C.; Cowtan, K.; Dodson, E.; Evans, P. R.; Jenkins, H. T.; Keegan, R.; Krissinel, E.; Stevenson, K.; Lebedev, A.; McNicholas, S. J.; Nicholls, R. A.; Noble, M.; Pannu, N. S.; Roth, C.; Sheldrick, G.; Skubak, P.; Turkenburg, J.; Uski, V.; von Delft, F.; Waterman, D.; Wilson, K.; Winn, M.; Wojdyr, M., CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta. Crystallogr. D. Struct. Biol.* **2018,** *74*, 68-84.
83. Murshudov, G. N.; Skubak, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A., REFMAC5 for the refinement of macromolecular crystal structures. *Acta. Crystallogr. D. Biol. Crystallogr.* **2011,** *67*, 355-367.
84. Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of macromolecular structures by the maximum-likelihood method. *Acta. Crystallogr. D. Biol. Crystallogr.* **1997,** *53*, 240-255.
85. Viladoms, J.; Escaja, N.; Frieden, M.; Gomez-Pinto, I.; Pedroso, E.; Gonzalez, C., Self-association of short DNA loops through minor groove C:G:G:C tetrads. *Nucleic Acids Res* **2009,** *37* (10), 3264-3275.
86. Ilavsky, J.; Jemian, P. R., Irena: tool suite for modeling and analysis of small-angle scattering. *J Appl Crystallogr* **2009,** *42*, 347-353.
87. Konarev, P. V.; Volkov, V. V.; Sokolova, A. V.; Koch, M. H. J.; Svergun, D. I., PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J Appl Crystallogr* **2003,** *36*, 1277-1282.
88. Svergun, D.; Barberato, C.; Koch, M. H. J., CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* **1995,** *28*, 768-773.
89. Meisburger, S. P.; Pabit, S. A.; Pollack, L., Determining the locations of ions and water around DNA from X-ray scattering measurements. *Biophys J* **2015,** *108* (12), 2886-2895.
90. Maizels, N., Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat Struct Mol Biol* **2006,** *13* (12), 1055-1059.
91. 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 Res* **2007,** *35* (22), 7429-7455.
92. Fry, M.; Loeb, L. A., The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. *Proc Natl Acad Sci USA* **1994,** *91* (11), 4950-4954.
93. Gacy, A. M.; Goellner, G.; Juranic, N.; Macura, S.; McMurray, C. T., Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* **1995,** *81* (4), 533-540.

94. Usdin, K., NGG-triplet repeats form similar intrastrand structures: implications for the triplet expansion diseases. *Nucleic Acids Res* **1998,** *26* (17), 4078-4085.
95. Keller, R. L. J., *The Computer Aided Resonance Assignment Tutorial.* CANTINA Verlag: Goldau, Switzerland, 2004.
96. Franke, D.; Petoukhov, M. V.; Konarev, P. V.; Panjkovich, A.; Tuukkanen, A.; Mertens, H. D. T.; Kikhney, A. G.; Hajizadeh, N. R.; Franklin, J. M.; Jeffries, C. M.; Svergun, D. I., ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr* **2017,** *50*, 1212-1225.
97. Sheldrick, G. M., Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta. Crystallogr. D. Biol. Crystallogr.* **2010,** *66*, 479-485.
98. Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A., The PDB_REDO server for macromolecular structure model optimization. *IUCrJ.* **2014,** *1(Pt 4)*, 213-220.
99. Esposito, V.; Galeone, A.; Mayol, L.; Oliviero, G.; Virgilio, A.; Randazzo, L., A topological classification of G-quadruplex structures. *Nucleos. Nucleot. Nucl.* **2007,** *26*, 1155-1159.
100. Cang, X.; Sponer, J.; Cheatham, T. E., Explaining the varied glycosidic conformational, G-tract length and sequence preferences for anti-parallel G-quadruplexes. *Nucleic Acids Res.* **2011,** *39* (10), 4499-4512.
101. Marathias, V. M.; Bolton, P. H., Structures of the potassium-saturated, 2:1, and intermediate, 1:1, forms of a quadruplex DNA. *Nucleic Acids Res.* **2000,** *28* (9), 1969-1977.
102. Mao, X.; Marky, L. A.; Gmeiner, W. H., NMR structure of the thrombin-binding DNA aptamer stabilized by Sr2+. *J. Biomol. Struct. Dyn.* **2004,** *22* (1), 25-33.
103. 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. *J. Am. Chem. Soc.* **2014,** *136* (14), 5249-5252.
104. Zhang, D.; Huang, T.; Lukeman, P. S.; Paukstelis, P. J., Crystal structure of a DNA/Ba2+ G-quadruplex containing a water-mediated C-tetrad. *Nucleic Acids Res.* **2014,** *42* (21), 13422-13429.
105. Pan, B.; Xiong, Y.; Shi, K.; Deng, J.; Sundaralingam, M., Crystal structure of an RNA purine-rich tetraplex containing adenine tetrads: implications for specific binding in RNA tetraplexes. *Structure* **2003,** *11* (7), 815-823.
106. Tang, C. L.; Alexov, E.; Pyle, A. M.; Honig, B., Calculation of pKas in RNA: on the structural origins and functional roles of protonated nucleotides. *J. Mol. Biol.* **2007,** *366* (5), 1475-1496.
107. Lim, K. W.; Alberti, P.; Guedin, A.; Lacroix, L.; Riou, J. F.; Royle, N. J.; Mergny, J. L.; Phan, A. T., Sequence variant (CTAGGG)n in the human telomere favors a G-quadruplex structure containing a G.C.G.C tetrad. *Nucleic Acids Res.* **2009,** *37* (18), 6239-6248.
108. Malliavin, T. E.; Snoussi, K.; Leroy, J. L., The NMR structure of [Xd(C2)4] investigated by molecular dynamics simulations. *Magn. Reson. Chem.* **2003,** *41*, 18-25.

109. Lu, X. J.; Bussemaker, H. J.; Olson, W. K., DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **2015,** *43* (21).
110. Phan, A. T.; Gueron, M.; Leroy, J. L., The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. *J. Mol. Biol.* **2000,** *299* (1), 123-144.
111. Zhou, J.; Amrane, S.; Korkut, D. N.; Bourdoncle, A.; He, H. Z.; Ma, D. L.; Mergny, J. L., Combination of i-motif and G-quadruplex structures within the same strand: formation and application. *Angew. Chem. Int. Ed. Engl.* **2013,** *52* (30), 7742-7746.
112. Evans, P. R.; Murshudov, G. N., How good are my data and what is the resolution? *Acta. Crystallogr. D. Biol. Crystallogr.* **2013,** *69(Pt 7)*, 1204-1214.
113. Skubak, P.; Pannu, N. S., Automatic protein structure solution from weak X-ray data. *Nat. Commun.* **2013,** *4*, 2777.
114. Kypr, J.; Kejnovska, I.; Renciuk, D.; Vorlickova, M., Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.* **2009,** *37* (6), 1713-1725.
115. Bishop, G. R.; Chaires, J. B., Characterization of DNA structures by circular dichroism. *Curr. Protoc. Nucleic Acid Chem.* **2002**, 7.11.1-7.11.8.
116. Kovanda, A.; Zalar, M.; Sket, P.; Plavec, J.; Rogelj, B., Anti-sense DNA d(GGCCCC)n expansions in C9ORF72 form i-motifs and protonated hairpins. *Sci. Rep.* **2015,** *5*, 17944.
117. Guschlbauer, W.; Chantot, J. F.; Thiele, D., Four-stranded nucleic acid structures 25 years later: from guanosine gels to telomer DNA. *J. Biomol. Struc. Dyn.* **1990,** *8* (3), 491-511.
118. Venczel, E. A.; Sen, D., Parallel and antiparallel G-DNA structures from a complex telomeric sequence. *Biochemistry* **1993,** *32* (24), 6220-6228.
119. Treangen, T. J.; Salzberg, S. L., Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **2012,** *13* (1), 36-46.
120. Bagshaw, A. T. M., Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. *Genome Biol. Evol.* **2017,** *9* (9), 2428-2443.
121. Xie, K. T.; Wang, G.; Thompson, A. C.; Wucherpfennig, J. I.; Reimchen, T. E.; MacColl, A. D. C.; Schluter, D.; Bell, M. A.; Vasquez, K. M.; Kingsley, D. M., DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* **2019,** *363* (6422), 81-84.