UNIVERSITY OF CALIFORNIA
SANTA CRUZ

HYDRATION OF r(UGGGGU) QUADRUPLEXES.
A dissertation submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in

BIOMOLECULAR ENGINEERING

by

Alastair Fyfe

June 2012

The Dissertation of Alastair Fyfe is approved:

Professor William G. Scott, Chair

Professor Kevin Karplus

Professor Eugene Switkes

Tyrus Miller
Vice Provost and Dean of Graduate Studies
Table of Contents

List of Figures v
List of Tables vii
Abstract viii
Dedication x
Acknowledgments xi

1 Introduction 1

2 Guanine Quadruplexes : General Characteristics 7
   2.1 Structural Invariants 8
       2.1.1 G-quartet Geometry 9
       2.1.2 G-quartet Cation Coordination 13
   2.2 Structural Polymorphism 14
   2.3 Frequency of G4 Sequence Elements 17
   2.4 Genomic Distribution of G4 Sequence Elements 22
   2.5 Putative Function 25
       2.5.1 Transcriptional Regulation: G4 Folds in DNA 28
       2.5.2 Translational Regulation: G4 Folds in RNA 36
       2.5.3 Replication and G4-linked Genomic Instability 41
       2.5.4 Telomere Maintenance 42
       2.5.5 mRNA Splicing 43

3 Observed UGGGGU Quadruplex Structures 44
   3.1 Methods 45
   3.2 Crystallographic Data 46
       3.2.1 Molecular Replacement 47
       3.2.2 Refinement 49
   3.3 Oligo Assemblies 51
3.4 Alternate Crystal Packings ........................................... 56
3.5 Backbone Conformation ............................................. 57
3.6 Quartet Cation Coordination ....................................... 58
3.7 Disorder ........................................................................ 64
3.8 Internal Cavity ............................................................ 68

4 Solvent Structure .......................................................... 70
4.1 What Does Bound “Water” Model? ................................. 72
4.2 Classifying Bound Waters ............................................ 75
  4.2.1 Hydration Variation by Atom Type ............................. 76
4.3 Bound Waters with Similar Association Patterns .............. 85
  4.3.1 Groove-bound Waters .............................................. 89
  4.3.2 Phosphate-bound Waters ......................................... 92
  4.3.3 Ribose O2’-bound Waters ....................................... 92
  4.3.4 Interstitial Crystal-contact Waters ............................ 94
  4.3.5 Internal Cavity Waters ............................................. 96

5 Gradient-flow Segmentation of Crystallographic Electron Density Maps100
5.1 Smoothing ................................................................. 104
5.2 Algorithms for Piecewise-linear Approximation ............... 106
  5.2.1 Input and Definitions .............................................. 107
  5.2.2 Computation of MS3 Sets ....................................... 108
  5.2.3 Computation of MS2 Sets ....................................... 108
  5.2.4 Classifying Boundary Voxels .................................. 109
  5.2.5 Output and Analysis ............................................. 109
5.3 Results ....................................................................... 110
  5.3.1 Difference Map Segmentation ................................. 116

6 Calculation of Molecular Surfaces and Derived Properties by Signed
  Distance Functions and Fast Marching Methods .................. 119
6.1 Molecular Surfaces .................................................... 120
6.2 Implicit Surfaces and Fast Marching Methods ................. 122
6.3 Computing Surface-derived Properties .......................... 126

7 Conclusions and Future Directions ................................. 131
  7.1 Future Directions ...................................................... 133

A Software ..................................................................... 136

B Distance to Surface Calculations .................................. 139

Bibliography ................................................................. 141
## List of Figures

1.1 Guanine quartet secondary structure and quartet stacking. 2

2.1 Right-handed helical nucleic acid conformations. 10
2.2 Cation coordination in quadruplex conformations. 15
2.3 Loop motifs in quadruplex structures. 17
2.4 Possible strand orientation in tetramolecular, bimolecular and uni-molecular quadruplex complexes. 18
2.5 Transient formation of G4 folds during replication and transcription. 27
2.6 Alternative G-quadruplex-forming elements in the c-MYC nuclease hypersensitivity element NHIII. 30
2.7 G-loop formation during transcription of G-rich immunoglobin class switch regions. 34
2.8 Quadruplex-forming sequences in the 5' UTR of an E. coli reporter gene. 38

3.1 Octamer assembly for UGGGGU quadruplexes. 52
3.2 Least squares superposition of chains R6A and R6B from P1B structure. 53
3.3 Helical stacking of four-stranded UGGGGU assemblies. 55
3.4 Arrangement of quadruplex columns in tetragonal (P42₁2) and orthorhombic (C22₂₁) space groups. 56
3.5 Cation coordination by Guanidine and Uridine Quartets. 61
3.6 Phased anomalous difference map for structure P1B. 62
3.7 Alternate conformation of a phosphate group in UGGGGU structure P1A. 64
3.8 Isotropic B factor distribution for assembly R6,8 from dataset P1B. 66
3.9 Internal cavity of structure P1B. 69

4.1 Distribution of surface waters for model P1A. 76
4.2 Hydration spine in the minor groove of B-form DNA. 90
4.3 Hydration spine in a DNA, parallel-stranded, G-quadruplex, PDB entry 352D. 91
4.4 Phosphate-bound waters in UGGGGU data set P1A. 93
4.5 Bound waters in contact with the ribose O2' hydroxyl group in UGGGGU data set P1A. 95
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>Bound waters involved in crystal contacts in UGGGGU data set P1A.</td>
<td>98</td>
</tr>
<tr>
<td>4.7</td>
<td>Waters in the internal cavity of UGGGGU data set P1A.</td>
<td>99</td>
</tr>
<tr>
<td>5.1</td>
<td>Decomposition of a two-dimensional scalar domain into Morse-Smale basins</td>
<td>105</td>
</tr>
<tr>
<td>5.2</td>
<td>Smoothing of the Morse-Smale decomposition of a two-dimensional domain</td>
<td>106</td>
</tr>
<tr>
<td>5.3</td>
<td>Single atom density basin.</td>
<td>111</td>
</tr>
<tr>
<td>5.4</td>
<td>Peak height vs. total density of descending basins for RNA atoms in</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>structure P1B.</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Volume characteristics of descending basins for RNA atoms in structure</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>P1B.</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>Rank order of total density in kicked omit maps for all modeled RNA</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>atoms in structure P1B.</td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>Commonly-used molecular surfaces.</td>
<td>121</td>
</tr>
<tr>
<td>6.2</td>
<td>The depth of A, B and Z-form DNA grooves as calculated by the travel</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>depth algorithm.</td>
<td></td>
</tr>
<tr>
<td>B.1</td>
<td>Solvent to surface distance calculation.</td>
<td>140</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Helical parameters of four right-handed helical conformations. . . . . . . 11
3.1 Crystallographic Data Reduction Statistics . . . . . . . . . . . . . . . . 48
3.2 Refinement Statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
3.3 RNA backbone torsion angles for structure P1B. . . . . . . . . . . . . 59
4.1 Hydration of UGCGGU Guanine Atoms . . . . . . . . . . . . . . . . . . . 78
4.2 Hydration of UGCGGU Uracil Atoms . . . . . . . . . . . . . . . . . . . 79
4.3 Hydration of UGCGGU Ribose/Phosphate Atoms . . . . . . . . . . . . . 80
4.4 Surface Atom Contacts of Modeled Waters . . . . . . . . . . . . . . . . 87
Abstract

Hydration of r(UGGGGU) quadruplexes.

by

Alastair Fyfe

For about 50 years it has been known that guanine-rich sequence can, under appropriate conditions, adopt a distinctive, four-stranded, helical fold known as a G-quadruplex. Interest in quadruplex folds has grown in recent years as evidence of their biological relevance has accumulated from both sequence analysis and function-specific assays. The folds are unusually stable and their formation appears to require close management to maintain cell health; regulatory failure correlates with genomic instability and a number of cancer phenotypes.

This thesis examines, by x-ray crystallography, the solvent structure of a previously reported tetramolecular RNA quadruplex, UGGGGU stabilized by Sr\(^{2+}\) ions. Crystal forms of the octameric assembly formed by this sequence exhibit unusually strong diffraction and anomalous signal enabling the construction of reliable models to a resolution of 0.85 Å. The solvent structure confirms hydration patterns reported for other nucleic acid helical conformations and provides support for the greater stability of RNA quadruplexes relative to DNA. Novel features detected in the octameric RNA assembly include a new crystal form and evidence of multiple conformations, among which one leading to the formation of a well-hydrated internal cavity.
Though solvent is generally acknowledged to play a fundamental role in nucleic acid structure, its characterization from diffraction data remains challenging. To assist with this task, the thesis investigates two novel additions to the crystallographic methods arsenal. The first is segmentation of electron density maps into Morse-Smale basins characterized by uniform gradient flow. The second is the use of level set Fast Marching methods to compute the full distance field defined by the molecular surface. Both techniques show promise, though additional work will be required to yield effective tools. Gradient-flow segmentation provides an unambiguous way to gather all map density associated with a modeled atom and enables the calculation of novel volumetric properties including total basin density. Distance field calculation provides a unified framework for combining molecular surface calculation with surface-related queries including pocket and cavity detection and solvent travel depth.
To Pamela,

thank you.
Acknowledgments

This is an opportunity to express my gratitude to Bill for providing a problem, bench space and resources to work on it, and endless encouragement, guidance and support. I thank my committee for the guidance and direction they have provided. If not for Kevin’s bioinformatics or Gene’s chemistry lectures I may never have developed an interest in structural biology and without Kevin’s timely reminders about looming deadlines I am quite certain this thesis would never have been written.

I would like to thank all members, past and present, of the Scott lab. In particular, Monika Martick and Michael Robertson were extraordinary teachers, patient, and generous with their time.

Pete Dunten of the Stanford Synchrotron Radiation Laboratory (SSRL) has been a continuing friend, mentor and guide to crystallographic practice.

In all human endeavors, practitioners reserve particular esteem and gratitude for those who provide their tools. For programmers, a well-designed and implemented class library is as valuable as concise and expressive notation. Kevin Cowtan’s Clipper libraries are an outstanding C++ class library for crystallographic computing; I’m indebted to Kevin for writing Clipper and for providing guidance in its use.

Attila Gyulassy, and other members of the UC Davis Institute for Data Analysis and Visualization (IDAV) provided helpful guidance in the application of their algorithms for Morse-Smale partitioning.
Chapter 1

Introduction

Nucleic acid sequences in which guanines are both abundant and distributed in accordance with a distinct but permissive pattern have a propensity to fold into a G-quadruplex motif [107] comprised of stacks of planar G-quartets (Section 2.1 and Figures 1.1, 2.3). Once formed, such structures are stabilized by a complement of electrostatic, stacking and hydrogen-bonding forces that render them unusually stable relative to other nucleic acid conformations. Though many examples of in vitro folded structures obtained from G-rich sequences are available [92, 14], the extent to which such structures occur in vivo and their physiological roles remain open questions.

Our understanding of this field remains preliminary, however the following generalizations appear broadly applicable: a growing inventory of structures confirms the stability and conformational diversity of G-quadruplex folds; sequences with the potential to adopt a quadruplex fold are abundant, though globally counterselected, and exhibit a markedly non-uniform genomic distribution suggestive of selective pres-
Figure 1.1: Guanine quartet secondary structure and quartet stacking. An intramolecular fold with two edge wise loops, one diagonal loop and anti-parallel strand orientation is shown.

This thesis examines, by x-ray crystallography, the structure of bound solvent at the surface of an RNA hexamer, UGGGGU, whose conformation adopts a guanine quadruplex fold. Though it is widely recognized that solvent is an essential component of macromolecular structure [117], its examination remains challenging. Distinguishing electron density attributable to solvent from background noise is not straightforward and the stereochemical ‘prior knowledge’, such as permissible torsion angles, bond lengths, chiral volumes and planarity constraints, that guides the interpretation of electron-density maps and their transformation into reliable atomic models is not available for ordered solvent. The macro-molecular crystal structure examined has been previously
reported [25], thus the main contribution of this thesis is a detailed examination of its solvent structure.

Residency times for water molecules near molecular surfaces as calculated from molecular dynamics and NMR studies range from $10^{-2}$ to $10^{-10}$ seconds [62]. Consideration of this result requires an interpretation of bound solvent that differs somewhat from interpretation of the macromolecular model itself. Rather than statically bonded atoms, bound solvent identifies minima in the potential of mean force near the molecular surface. Solvent is modeled at specific sites that are briefly but consistently occupied by indistinguishable, interchangeable, solvent molecules [62].

Mapping these minima helps characterize how a macromolecule participates in molecular recognition and binding, including its interaction with designed small-molecule agents. Quadruplex folds, and the sequences from which they form, have been implicated in a wide range of diseases including various cancer and premature aging disorders. Accordingly, they are being actively investigated as targets for therapeutic intervention [89] as illustrated by CX-3543, an inhibitor of rDNA synthesis currently in Phase II clinical trials [29].

The data analyzed in this work, whose resolution ranges from 0.9 to 1.5 Å, displays unusually high resolution relative to current PDB depositions, particularly for RNAs. Thus, independently of insights into quadruplex structure, these data provide an opportunity for reliable quantification of RNA structure and hydration parameters, and this analysis is therefore of more general interest.

The thesis is organized into seven chapters. The second introduces general
characteristics of G-quadruplexes, including their structural and sequence properties, along with an overview of known results on physiological function. This is an extensive and rapidly growing field, so this overview is necessarily limited. It is intended to frame the UGGGGU hexamer considered here in a broader research context. The third and fourth chapters consider, respectively, the structure of the RNA polymers and of the surrounding solvent, counterions and ordered waters, observed in the data sets collected.

Chapter 3, in addition to describing the RNA structure, summarizes the characteristics and quality indicators of the data sets collected and the novel crystal packings observed. Though solvent generally plays a supporting role in the presentation of structural results, here it is the leading character and the RNA polymers provide a largely invariant backdrop for the examination of hydration patterns. A key theme of Chapter 4 is to classify bound waters according to their location relative to properties of the RNA surface such as charge and accessibility. This chapter also discusses technical difficulties associated with identifying solvent sites, interpretation of bound waters and validation of modeled solvent.

The next two chapters, 5 and 6, discuss methods, whose application to crystallographic data is novel, developed as part of the preceding investigations. Chapter 5 is concerned with segmenting crystallographic maps into regions of uniform density gradient flow. The goal is to adapt known techniques for partitioning three-dimensional scalar density fields into Morse-Smale basins [42]. to a novel application, the segmentation of electron-density maps These methods have been developed primarily in the context of data-visualization, but are readily applicable to map interpretation. Inte-
grating density over a basin provides a robust measure of scattering at a given site; this attribute is a more sensitive discriminator than peak height or iso-contour level, the two measures currently in widespread use.

Chapter 6 discusses the use of signed distance functions, and their implementation by Fast Marching methods, to construct molecular surfaces and calculate derived properties such as pocket depth. Implicit surfaces and their implementation by narrow band Fast Marching methods have proven to be a versatile mathematical technique applicable to a wide range of applications including image processing, materials science and computer vision [108]. They have been previously shown to be effective in the construction of molecular surfaces [16]. Here this approach is extended to calculate derived properties of the surface. Development of the methods described in these two chapters was motivated by necessity. Density partitioning is intended to help characterize and identify ambiguous solvent structure and signed distance function calculations assist in classifying neighborhoods of the molecular surface such as pockets and cavities.

The thesis concludes with a summary of results obtained and possible directions for future work. A closing appendix summarizes software modules developed as part of this work.

Analysis of the r(UGGGGU) structure described in Chapters 3 and 4 was a collaborative effort. Monika Martick grew the initial crystals from previously published conditions and collected a preliminary high-resolution data set. Peter Dunten assisted with data collection, provided guidance in the use of crystallographic software and determined initial structure solutions via iterative rigid body refinement when standard
molecular replacement software proved unsuccessful. William Scott provided overall
guidance and direction and resolved countless interfering hurdles. All remaining tasks,
as well as the methods-oriented work described in Chapters 5 and 6, were undertaken
by the author.
Chapter 2

Guanine Quadruplexes : General Characteristics

Interest in quadruplex structures has grown in recent years as evidence of their biological relevance has accumulated from both sequence analysis and function-specific assays [82]. This chapter provides background information on G4 folds, reviewing their structure, the genomic frequency and distribution of sequence motifs, thermodynamic properties and putative biological roles. Though the structural sections highlight similarities and differences between the UGGGGU hexamer and related structures, most of the remainder reviews general results on G4 folds without distinguishing DNA and RNA forms.
2.1 Structural Invariants

Briefly, a quadruplex structure consists of a stack of two or more guanine quartets, optionally connected by loops. A G-quartet is a nucleic acid secondary motif formed by four planar guanine bases held symmetrically about a four-fold axis by two hydrogen bonds between the Hoogsteen face of one guanine and the Watson-Crick face of its neighbor (Figure 1.1). Stacking interactions among the planar quartets and coordination by a central monovalent or divalent cation yield guanine-quadruplex structures which combine the four-stranded, stacked quartet motif with a diverse range of folds parametrized by the number of polymer chains involved, strand polarity and the length and composition of loop nucleotides [14].

Successive G-quartet layers are formed by adjacent guanosines, hence the general criteria for sequence searches of three-quartet folds is “four or more runs of three or more guanosines”. The rise between successive layers is about 3 Å. To avoid the repulsive π effect of coaxial stacking, successive layers are offset by about 20 degrees, a displacement that provides a compromise between the repulsive and attractive forces in stacking effects. The successive displacement enforces a right-handed helical twist on the nucleic acid backbone and results in an overall right-handed helical conformation common among nucleic acid folds. As in other helical nucleic acid structures, including B and A-form DNA, quadruplex folds include a groove between adjacent strands. The floor of the groove is formed by base atoms and its opening by atoms in the solvent-exposed sugar-phosphate chain (Fig. 6.2).
2.1.1 G-quartet Geometry

A comparison of key structural parameters between A and B-form DNA and two quadruplex structures is given in Figure 2.1 and Table 2.1. Panel (a) shows a well-studied B-DNA structure, the Dickerson-Drew dodecamer 5’-CGCGAATTCCGCG-3’, and its complementary strand at 1.9 Å resolution from PDB entry 1BNA. Panel (b) displays a DNA decamer 5’-AGGGGCCCCCT-3’ which adopts an A-form duplex from data collected to 1.1 Å, PDB entry 440D. Panel (c) displays the UGGGGU RNA examined in this thesis from dataset P1B at 0.9 Å resolution. Only one of the two strands in the asymmetric unit is shown along with its symmetry rotations about the four-fold axis. Panel (d) displays another parallel, four-chain, quadruplex, the DNA counterpart of the structure examined in this thesis, TGGGGT, at 0.95 Å from PDB entry 352D.

Salient differences between these conformations are evident from visual inspection and quantified by calculation of base-pair parameters as shown in Table 2.1. In B and A-form DNA, the two strands are anti-parallel and not equidistant. As a consequence of the latter, two grooves of different width result: the major and minor grooves. However in quadruplex structures, four-fold symmetry among the parallel strands requires that the four grooves also be symmetry-equivalent. The different orientation of the bases is also reflected in larger values for the intra-strand opening parameter in the two quadruplex structures. Also, whereas buckle is pronounced among the TGGGGT guanines, the propeller twist prominent in B-form DNA is absent.

The rise between successive base pairs among the four types of right-handed
Figure 2.1: Right-handed helical nucleic acid conformations: (a) B-form DNA duplex from PDB entry 1BNA, (b) A-form DNA duplex from entry 440D (c) parallel four hexamer UGGGGU structure from dataset P1B and (d) parallel, four hexamer TGGGGT structure from PDB entry 352D. For the two quadruplex structures in panels (c) and (d), the distant, least visible strand, has been omitted for clarity. The curved path obtained from spline interpolation of phosphate backbone atoms highlights the backbone structure.
### Base Pair Conformation Parameters

<table>
<thead>
<tr>
<th></th>
<th>B-DNA</th>
<th>A-DNA</th>
<th>TGGGGT</th>
<th>UGGGGU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rise</td>
<td>3.35</td>
<td>2.65</td>
<td>2.39</td>
<td>2.61</td>
</tr>
<tr>
<td>Helical Twist</td>
<td>36.0</td>
<td>31.4</td>
<td>20.0</td>
<td>21.5</td>
</tr>
<tr>
<td>Minor Groove Width</td>
<td>4.94</td>
<td>10.14</td>
<td>2.16</td>
<td>4.27</td>
</tr>
<tr>
<td>Major Groove Width</td>
<td>11.42</td>
<td>4.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Minor Groove Depth</td>
<td>5.27</td>
<td>0.618</td>
<td>2.90</td>
<td>2.42</td>
</tr>
<tr>
<td>Major Groove Depth</td>
<td>4.45</td>
<td>11.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Opening</td>
<td>1.0</td>
<td>2.3</td>
<td>98.1</td>
<td>91.2</td>
</tr>
<tr>
<td>Buckle</td>
<td>0.3</td>
<td>-0.9</td>
<td>8.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>Propeller Twist</td>
<td>-13.7</td>
<td>-9.0</td>
<td>3.0</td>
<td>4.8</td>
</tr>
<tr>
<td>Tilt</td>
<td>-0.2</td>
<td>-0.1</td>
<td>-1.9</td>
<td>11.3</td>
</tr>
<tr>
<td>Roll</td>
<td>-0.3</td>
<td>7.4</td>
<td>-13.6</td>
<td>-2.8</td>
</tr>
</tbody>
</table>

Table 2.1: Helical parameters of four right-handed helical conformations. Structure identifiers are as given in Figure 2.1, calculations are from program Curves+ [59].
helices are comparable, however the quadruplexes exhibit narrower helical twist. Groove width for the UGGGGU quadruplexes is close to that of the minor groove in B-DNA, but considerably narrower for the TGGGGT quadruplex. Grooves in both quadruplexes are shallower than in B-DNA. These measurements also confirm the observation that the features of major and minor grooves are reversed between B and A-form DNA, with the A-form minor groove having width comparable to that of B-form major groove (Fig. 6.2).

Measurements of sugar pucker and glycosylic angle also vary across these conformations. Whereas sugar pucker in B-form DNA includes C2’endo, C1’exo, O1/O4’endo, C3’endo and C4’exo conformations, in A-DNA the C3’ endo sugar pucker appears almost exclusively. For the two parallel quadruplexes, the UGGGGU structure shows both C3’endo and C2’endo modes whereas puckers in the TGGGGT quadruplex include C2’endo, C1’ exo and O1/O4’ endo at the 3’ terminal thymine.

The glycosyl torsion angle, $\chi$, defined by the guanine O4’-C1’-N9-C4 atoms assumes an anti conformation in both quadruplexes and in A-form DNA. For TGGGGT, $\chi$ angles in the first strand range from $-135.2$ to $-106.5$ and in the first UGGGGU strand from $-165.0$ to $-121.7$. In A-form DNA, $\chi$ values range from $-172.7$ to $-145.7$. On the other hand, the B-form dodecamer exhibits a $\chi$ range from $-135.1$ to $-89.6$ and thus includes nucleotides in the high-anti range, $[-110, -60]$. 
2.1.2 G-quartet Cation Coordination

In addition to helically stacked quartets, a second defining feature of quadruplex structures is the presence of stabilizing cations in a central channel bounded by the guanine O6 and N1-H atoms. The hydrogen bonding network within each G quartet places adjacent carbonyl oxygens in close proximity, 3.16 Å in the UGGGGU P1B structure. Furthermore, though the 20° twist between successive quartets avoids direct coaxial stacking of the guanosines, pronounced buckle in the guanine planes can reduce rise. For example in the P1B structure, the O6 carbonyl oxygens of neighboring quartets can be within 2.96 Å. In G-quartet folds, the resulting accumulation of negative charge is invariably offset by a central cation.

Different monovalent and divalent cations exhibit different stabilizing effects and coordination geometries. Comparison of stabilizing effects based on melting temperatures of 8-bromoguanosine gels yielded the order $K^+ >> Rb^+ > NH_4^+ > Na^+ > Li^+$ for monovalent and $Sr^{2+} >> Ba^{2+} > Ca^{2+} > Mg^{2+}$ for divalent cations. [95]. Stabilization by different cations can have profound structural consequences as illustrated by completely different structures reported for the human telomeric sequence 5’AGGG(TTAGGG)3-3’ in $Na^+$ solution and as crystallized in the presence of $K^+$ [64, 92]. The physiological concentration of intracellular $K^+$ in mammalian cells is 139mM relative to 12mM for $Na^+$ [71, 68]. Since $K^+$ is both more abundant and exerts a stronger stabilizing effect on G4 folds, it is believed to be the more significant cation in vivo. The strong stabilizing effect of $Sr^{2+}$ is likely a significant factor in the
high-resolution diffraction observed for UGGGGU tetraplexes.

Coordination geometry varies with charge and atomic radius. For Na\(^+\), the 0.95 Å atomic radius enables coordination in the G-quartet plane, whereas for bulkier ions such as K\(^+\) with radius 1.33 Å and NH\(_4\)\(^+\) with radius 1.43 Å the cation must be coordinated between planes [95]. Placement is typically not uniform across all planes. Figure 2.2 compares placement of Sr\(^{2+}\) and Na\(^+\) cations in the UGGGGU and TGGGGT quadruplexes. In panel (a), the four Sr\(^{2+}\) cations are sandwiched between two adjacent quartet planes in a bipyramidal-antiprism geometry [25] that distorts the planarity of each pair of quartets. This arrangement, which omits cations in alternating quartets, is unique to Sr\(^{2+}\) and believed to result from electrostatic repulsion [95]. The two 3’ U-tetrads are coordinated by Na\(^+\). Panel (b) shows the ordering of Na\(^+\) cations in the TGGGGT structure. At each end of the quartet stack, cations occupy the quartet plane but gradually shift to an inter-plane coordination toward the center, enabling seven cations to coordinate the eight quartets.

2.2 Structural Polymorphism

Right-handed helical conformation and the coordination of quartet carbonyls by dehydrated cations are universal features of G4 folds. However, variation of additional parameters enables formation of a broad and diverse range of structures. Three key parameters underlying this polymorphism are the syn/anti disposition of the glycosyl bond χ angle, the parallel or anti-parallel orientation of adjoining strands, and the
Figure 2.2: Cation Coordination in Quadruplex Conformations. The eight quartet assemblies from (a) UGGGGU structure dataset P1B and (b) TGGGCT structure, PDB ID 352D; Sr$^{2+}$ cations are in magenta, Na$^+$ in teal. The backbone chain is omitted for clarity. The two four-strand assemblies comprising each octamer are shown in red and green.
form of the connecting loops in uni-molecular and bi-molecular complexes.

The number of consecutive G runs dictates the stoichiometry of a complex [92]. Tetramolecular quadruplexes can form from strands containing a single G run, as is the case in the structure we investigated in this thesis. Bi-molecular complexes generally require two distinct G-runs in each chain and uni-molecular complexes require at least four. The number of guanines within a run determines the number of stacked quartets. Stability of the fold increases with the number of G-quartets. Though there is evidence supporting stable bi-quartet structures [128], most structural and sequence studies have focused on three and four-quartet structures.

Loops are required in bi and monomolecular complexes to allow formation of four distinct strands at the corners of a quartet. Classification of loops in deposited structures yields the four forms shown in Figure 2.3: edge-wise, diagonal double-chain and V-loop. Loops are one the most intriguing aspect of G4 folds. Their length, structure and composition have significant effects on stability, molecular recognition and polymorphic potential. Unfortunately, they are not a feature of the tetramolecular structure examined in this thesis.

The orientation of the four strands in a quadruplex is another key structural parameter. For a given strand, the two adjoining neighbors can have either the same, opposite or mixed polarity yielding a parallel, anti-parallel or mixed conformation. This parameter is highly correlated with syn/anti disposition of the \( \chi \) torsion angle about the C1'-N9 bond. In parallel-stranded structures, all \( \chi \) angles typically have an anti conformation, as is the case with UGGGGU. However anti-parallel and mixed struc-
tures typically exhibit an alternating mix of *syn* and *anti* conformations. Schematic illustration of possible strand orientations in tetra, bi and uni-molecular complexes is shown in Figure 2.4.

Current PDB depositions for G4 folds give evidence of wide conformational diversity. However, a theoretical study of geometrically possible topologies enumerated a number of variations that have not yet been observed [127]. Thus it seems premature to assume the current inventory represents an exhaustive sampling of G4 structural space.

### 2.3 Frequency of G4 Sequence Elements

The structural results summarized in the previous section demonstrate that, at least under laboratory conditions, G-rich nucleic acid sequences can adopt a quadruplex fold comprised of one, two or four polymers. Before considering the putative biological
function of these folds, it is relevant to review their prevalence and distribution in genomic (DNA) and transcribed (RNA) sequences. A number of bioinformatic studies have considered this question over the past decade, starting with initial searches of human genomic data [120, 49] around 2004. The available results, though based on simplifying assumptions, provide a preliminary approximation to the frequency and distribution of potential G4-folding sequence (PG4S) elements.

A commonly-used definition for PG4S motifs is four or more runs of three or more guanines separated by one to seven loop nucleotides, a requirement that may be expressed as

$$G_{x_1}N_{L_1}G_{x_2}N_{L_2}G_{x_3}N_{L_3}G_{x_4}$$

(2.1)

where $x_i$ and $L_i$ denote the respective lengths of the G-run and loop nucleotides. In the
human genome, separate studies yielded a very similar number of 375,000 potential sites [120, 49] confirming that the motif is common. In evaluating the overall significance of quadruplex folds, it is important to assess whether this number reflects more or less than what might be expected to occur by chance. However, as with all genomic probabilistic assessments, there is the problem of constructing credible null models against which to compare the observed frequency. One approach to the problem is described below. In addition to this problem, G4 accounting involves additional difficulties specific to PG4S:

- PG4S frequently occurs in repetitive DNA including telomeric repeats, rDNA and microsatellite regions. It is not clear whether such data should be treated as short, distinct or overlapping PG4S, or as one or a few long PG4S.

- Guanines in G4 structures can appear as corners of a G-tetrad or as loop elements, thus making the boundaries of loops ambiguous.

- The PG4S motif is permissive and thus common. As specified in 2.1, it also incorporates a number of known motifs among which quadruplex formation in unlikely to play a role. These include CpG islands associated with cytosine methylation and binding sites for SP1 and other transcription factors.

- Current searches have focused on intramolecular G4s, a choice supported by low physiological concentration of nucleic acids. Though reasonable, this simplification excludes consideration of intermolecular structures formed from two or four polymers. Structural results, including the RNA considered here, support stable folds for such combinations.
– Searches have only considered loop lengths limited to 1-7 nucleotides, though there is yet little evidence to rule out longer loops. Thus current estimates may only represent a lower bound on active PG4S elements.

To assess how the results of PG4S searches compare with what might be expected to arise by chance, Huppert et al [48, 49] compared similar sequence elements in two null models. The frequency of four classes of sequence patterns were calculated, with classes defined by replacing G in (2.1) by C, A and T. C-patterns would be expected on the strand complementary to a PG4S motif whereas A and T patterns served as controls for gauging observed G/C patterns. The resulting counts across all human chromosomes were 376,000 and 3,259,000 for GC and AT patterns respectively.

To assess whether these counts differed significantly from what might be expected by chance, an initial null model modeled DNA as a simple Bernouilli stream with independent nucleotide probabilities derived from genome-wide averages. The expected frequency of GC and AT patterns predicted by this model were only 8000 and 300,000 respectively, far fewer than observed. The discrepancy may be explained by failure of the null model to account for any local effects, particularly the significant correlation in the frequency distributions of adjacent nucleotides. Inspection of diad distributions confirmed that independence of adjacent nucleotides, as assumed by the Bernoulli stream model, is not a realistic assumption for genomic data.

A more elaborate but realistic estimate of nucleotide probabilities was obtained by segmenting genomic data into windows and considering local base distributions. In
this approach, DNA was divided into windows of a given size and the diad frequency of successive pairs calculated within each window. These probabilities were then used to generate a stream of simulated DNA of the same size to which the search algorithm was applied, thus maintaining local probabilities of both nucleotides and nucleotide pairs. Different window sizes were evaluated. A window of 75-bp yielded an AT-pattern count of 3,260,000, in good agreement with the value observed in actual genomic data. However, the corresponding GC-pattern count for a 75-bp window was 514,000, significantly higher than the 376,000 PG4S motifs observed. The discrepancy between observed and expected GC patterns suggests a global effect against formation of PG4S sequences.

A different approach was used in PG4S searches by Eddy and Maizels. The “G4 Calculator” [50] program does not include loop length as an explicit constraint. Required inputs are window size, a shift offset, the minimum number of consecutive G’s in a run and the minimum number of runs. The program scans the input, successively calculating new windows by applying the shift offset, and searches each window for a motif meeting the two G-run criteria. Upon completion, the ratio of windows yielding a match to the number of windows searched is reported. Omission of an explicit loop length constraint may be detrimental as the composition and length of loops have been found to be a distinctive characteristic of G4 structures [49]. However, the ability to represent multiple G runs as a single motif may correspond more closely to physical reality. Evidence from thermodynamic studies indicates multiple intramolecular quartets can easily form and dissociate within a multi-run sequence.

In summary, current methods for finding and counting PG4S motifs confirm
they are a common occurrence in genomic DNA. Present estimates must be considered preliminary, and reliable screens for identifying intra and inter-strand G4 folds await development of more reliable methods for predicting fold formation potential from sequence. In current practice, experimental confirmation of G4 folds at putative, sequence-predicted, sites relies on application of well-established, G4-specific, biophysical assays including circular dichroism (CD) spectra, UV-melting curves [76] and expression level of reporter constructs in *E. coli* [129].

### 2.4 Genomic Distribution of G4 Sequence Elements

The discrepancy between expected and observed PG4S frequencies supports the hypothesis of global negative selection against G4-forming sequences. Additional evidence of selection for and against this motif comes from analysis of the genomic locations favored by PG4S elements. Evidence of a highly non-uniform distribution has emerged in increasing detail since initial PG4S counts were reported around 2004.

Given the approximate nature of analysis tools available for identifying PG4S motifs, results on motif location must also be considered preliminary. Nevertheless, a number of trends have been identified and confirmed by independent researchers. In roughly chronological order, key results obtained to date include the following:

- The identification of G-rich repeats with high PG4S frequency and specific biological function, including telomeres, immunoglobulin switch regions, and minisatellites is well-established and precedes more recent bioinformatic screens[70].
PG4S concentration was found to be significantly reduced in coding exons [49, 31].

The windowed Markov model described above was applied over exon sequences to compare observed and expected frequencies of the motif given by (2.1) for all four bases. T-patterns again served as a control and a ratio of 0.48 resulted for observed G to T patterns, whereas a null model simulation predicted a ratio of 0.91. Furthermore, the ratio of observed to expected G-patterns was only 0.53. These results demonstrate a clear bias against PG4S in coding regions. Under-selection can be explained by bias against formation of G4 structures in unwound DNA that would interfere with transcriptional processivity.

Among known human RefSeq genes, there is a clear association between gene function, as assessed by the Gene Ontology classification [40] and both tails of the PG4S frequency distribution. Genes that exhibit a low PG4S concentration include G-protein coupled receptors and genes associated with the ubiquitin cycle and cell division. At the other extreme, high concentration was found among transcription factors and genes with roles in development, growth factor activity and signaling. Further analysis revealed that high PG4S occurs among proto-oncogenes as classified by OMIM [2], whereas tumor-suppressors exhibit lower than median PG4S [31]. The above distinctions reach statistical significance at low P-levels. Furthermore, a similar pattern was observed when examining gene cDNAs, indicating that the distinction applies to both exons and introns.

Among human RefSeq genes, there is a strong positive correlation, 0.73, between
PG4S frequency in template and non-template strands [31]. The predicted slope is less than one (0.83), suggesting that transient G4 fold formation in a transcription bubble is less common than in a replication fork.

- The transcription start site (TSS) delineates a clear boundary in PG4S frequency among human RefSeq genes. Though strong peaks in PG4S frequency occur within 300-bp windows on each side of the TSS, more detailed analysis reveals very different characteristics. Upstream, many G-rich sites are also recognizable as sequence elements with known function. These include CpG dinucleotides and G-rich binding sites for known transcription factors including SP1, KLF, EKLF, MAZ, EGR-1 and AP-2[32, 121]. The effect of masking upstream sequence for the SP1 factor alone is nearly sufficient to eliminate the PG4S frequency peak, an observation that makes it less likely that PG4S motifs play a widespread role as promoter elements, though this conclusion has been disputed [97].

- The PG4S distribution downstream of the TSS among human RefSeq genes is not affected by masking CpG or transcription factor binding motifs. Two additional trends are observed in this region, enrichment favoring the non-template strand and the 5' end of the transcript [32]. In their analysis of this region, Eddy and Maizels separately considered the full transcript, cDNA, coding region and first two introns of all RefSeq genes. Strikingly, PG4S on the non-template strand occurs within a 100-bp window at the 5' end of the first intron among 48% of human genes and among 24% for the template strand.
In cDNA sequences, the first 5′ window includes PG4S elements in 22% of genes. Though many of these are likely to be found in 5′-untranslated (UTR) regions, this remains to be confirmed. G4 structures on the non-template strand near the start of transcripts could play a regulatory role either as DNA, as shown schematically in panel (b) of Figure 2.5, or in the nascent mRNA transcript. Structures in the 5′UTR would persist in mature mRNA and could serve in translational control.

To determine conservation of G-rich first introns (GRIN1) elements, Eddy and Maizels replicated the analysis in mouse, chicken, frog and zebrafish, finding them to be conserved through frogs (41% mouse, 42% chicken).

In summary, the highly non-uniform distribution of PG4S motifs demonstrates clear selective pressure for and against formation in specific genomic regions. Whereas a bias against formation in coding exons can be rationalized as favoring translational processivity, it is not yet clear why there appears to be a bias in favor of G4 fold formation. In particular, the functional roles, mechanisms and associated binding factors for GRIN1 elements await exploration.

2.5 Putative Function

The preceding three sections have reviewed properties of known G4 structures and the frequency and distribution of PG4S elements. A natural question is whether any of this has biological significance. Observation of quadruplex folds \textit{in vivo} is not straightforward, and confirmation of quadruplex-specific function in G-rich sequences
is more sparse, unsurprisingly, than frequency and location indications from bioinfor-
matic screens or in vitro structure models. However, considerable evidence from assays
directed at diverse cellular processes has accumulated and representative examples are
summarized in the following sections. The role of G4 folds in transcriptional and transla-
tional control, replication, telomere maintenance and alternative splicing are considered
in turn.

A few preliminary observations may help unify these diverse and disparate re-
results. For DNA, G4 formation is likely a transient event that accompanies replication
and transcription once guanidines are released from the stability of Watson-Crick (WC)
pairing in the B-form duplex. Thus, during transcription, G4 folds can be expected on
the non-template strand, and possibly in the nascent mRNA, whereas during replication
they might be expected on both lagging and leading strands as summarized schemati-
cally in Figure 2.5. In this regard, G4 folds are one of several, transient, non-B DNA
secondary structures, a class that includes Z-DNA, DNA triplexes such as H-DNA and
a variety of unusual structures such as hairpins and slipped-DNA associated with triplet
repeats [123]. For RNA, release from a restraining duplex form is not required, and G4
formation would be expected to occur more frequently. However, to date, investigation
of RNA G4 folds has lagged DNA. A notable exception are the translational repression
results described in section 2.5.2. The protocol recently developed by Wieland and Har-
tig [129] provides an effective tool for assessing, in vivo, the impact of PG4S elements
in the 5’UTR and this approach has been applied to test potential sites identified by

26
Some common underlying themes emerge from considering the diverse functional evidence summarized in the following sections:

- Unwinding of duplex DNA enables formation of G4 folds in sequences with appropriate G-rich sequence. PG4S elements in gene promoters tend to coincide with nuclease hypersensitive regions indicating that the transition is facilitated in accessible euchromatin.

- Once formed, quadruplexes tend to be relatively stable. However, they exist in a dynamic equilibrium with single-stranded DNA and duplex-DNA. This balance is influenced by a number of factors including concentration of stabilizing cations, such as Na\(^+\) and K\(^+\), proteins that stabilize or unwind G4 folds and negative supercoiling introduced by an upstream polymerase [97, 55].

- Extensive polymorphism is a common feature among competing G4 structures at a given site. This effect has been clearly documented in the multiple structures reported for the human telomeric repeat [64] and c-MYC promoter [112].
Formation of a stable G4 fold may interfere with “generic” processing of nucleotides at that site, such as replication or transcription. This appears to be the definitive aspect of G4 folds as regulatory elements as evidenced by replicative malfunction [56], transcriptional inhibition [112] and arrest [124] or telomerase inhibition [134].

Proteins that unwind G4 folds and allow processing to resume, such as DEAH helicases [70], or factors that inhibit their formation such as pot1 [135] are key components of G4 regulatory roles.

It should be stressed that applicability of the above generalizations is limited by our incomplete understanding of this emerging field. For example, while proteins responsible for unwinding G4 folds during replication have been identified, (RECQ/FANCJ), no counterpart has been found for unwinding G4 folds in mRNA 5’UTR elements that inhibit translation. Similarly, no model has yet been proposed to account for the remarkable enrichment of PG4S elements at the 5’ extremity of the first intron of human genes (GRIN).

2.5.1 Transcriptional Regulation: G4 Folds in DNA

Evidence in favor of a role for G4 folds in transcriptional regulation may be divided into examples specific to individual genes, notably  c-MYC and c-KIT, and evidence for a broader systemic interaction with the transcriptional apparatus. Each of these is described below.
2.5.1.1 Quadruplexes in Promoter Regions

The *c-MYC* gene encodes a well-known transcription factor that participates in regulation of diverse cellular functions related to cell proliferation, differentiation and apoptosis. Normally expressed at elevated levels during embryogenesis and in rapidly-dividing tissue such as epidermis, it is also a potent oncogene whose overexpression or deregulation has been associated with diverse cancer phenotypes, including small-cell lung cancer, melanoma, myeloid leukemia, breast, colon and cervical cancer [90]. Regulation of *c-MYC* and the effects it exercises on its target genes are complex, multi-faceted, networks that remain only partially understood, notwithstanding an extensive literature. Since suppression of *c-MYC* expression has been found to inhibit tumor growth via apoptosis or induced cell differentiation, its regulation is an important cancer drug target [87].

Though multiple promoters and start sites are involved in *c-MYC* transcription, two promoters, P1 and P2, account for most activity [66]. A number of DNase I hypersensitive sites have been detected in the upstream region and one of these, NHIII1, a 27-bp sequence upstream of P1, controls up to 95% of transcription and consists of a G-rich, PG4S element on the template strand [113, 112].

This element has been the subject of extensive investigations by Hurley, Patel, Yang and others [93, 112, 89]. Analysis of candidate folds is complicated by polymorphism due to five G runs two of which incorporate four guanines, one more than required for a three-tetrad structure as summarized in Figure 2.6. Comparison with related se-
Figure 2.6: Alternative G-quadruplex-forming elements in the c-MYC nuclease hypersensitivity element NHIII₁. Candidate guanines in NHIII are shown in green and two possible three-quartet quadruplex folds in red. The G₃N₁G₃N₂G₃N₁G₃ fold, with four consecutive G runs, was shown to be the dominant conformation [93].

Sequences identified two potential three-tetrad conformations, labeled G₃N₁G₃N₂G₃N₁G₃ and G₃N₁G₃N₆G₃N₁G₃ [97] and later NMR analysis confirmed formation of two stable parallel-strand conformations [93].

Since the N7 atom of guanines involved in tetrads shares a hydrogen bond with the neighboring N2, it is not available for methylation by dimethyl sulfate (DMS). By a series of DMS footprinting and mutational experiments, Hurley and coworkers established that the biologically relevant quadruplex is the “chair” conformation consisting of the four consecutive 3′ G runs.

Two additional experiments confirmed a direct link between formation of the “chair” fold and repression of the c-MYC gene [112]. A single G-A mutation incompatible with a G4 fold resulted in a 3-fold increase in transcription. Conversely, binding by TMPyP4, a small-molecule agent known to stabilize quadruplex structures and previously investigated as an inhibitor of telomerase activity, was shown to increase c-MYC transcription. This effect was substantiated in vivo by comparing two lymphoma cell-lines (Ramos and CA46) in which only the former retains the NHIII₁ element after chromosomal translocation of c-MYC. As expected, only the Ramos line exhibited
TMPyP4 sensitivity. The hypothesis of a regulatory function for this PG4S element was further strengthened by report of a stabilizing I-motif on the corresponding C-rich coding strand [113].

Though *c-MYC* is perhaps the best-studied example of transcriptional regulation by a *cis*-acting G4 fold, a number of other examples have been documented as reviewed recently by Qin and Hurley [97]. The *c-KIT* gene exhibits a similar pattern. This gene also regulates cell proliferation and is a potent oncogene; mutation have been implicated in the formation of gastrointestinal stromal tumors[38]. Unlike *c-MYC*, *c-KIT* codes for a membrane protein, a tyrosine kinase responsive to binding by the stem cell factor (SCF) ligand. Bioinformatic screens identified two potential quadruplexes in the sequence from -120 to -161 bp upstream of the TSS which had previously been shown to control core promoter activity. Both of PG4S motifs have been investigated in detail by the Balasubramanian and Patel labs [91, 38]. CD and UV analysis of the elements were consistent with a G4 fold at physiological conditions. An NMR-derived structure by the Patel lab revealed a three-quartet structure with a novel loop pattern distinguished by the ability of a lone guanine, outside of any G run, to maintain a quartet [91].

Other genes in which promoter-specific quadruplexes have been investigated include *Ki-ras, VEGF* and *HIF-α* [38]. GREGLIST, a database of potential G-regulated genes, screened via the QUADPARSER [49] program, provides a current inventory of candidates for further investigation [136].

Whether these and similar documented instances of G4-mediated transcrip-
tional control are examples of a widespread regulatory mechanism remains somewhat controversial. Sequence-based analysis by Eddy and Maizels [32] suggests that PG4S sequences in 5' regulatory regions are primarily sites of cytosine methylation [121] or binding for transcription factors, principally SP1. Qin and Hurley argue that the considerable evidence for quadruplex-mediated *cis* control near promoter sites contradicts this conclusion noting that the bioinformatic screen “appears to us largely to disregard the evidence presented in papers published by us and groups in Europe and Asia” [97]. However it seems likely that both authors are correct and that multiple, independent forces account for selective pressure for and against formation in different genomic regions. The role of PG4S elements in nuclease hypersensitive promoter regions does not exclude pressure against PG4S in coding regions or, the remarkable and still unexplained enrichment of GRIN elements [32].

2.5.1.2 G4 Folds In Transcription-coupled DNA Repair

Bioinformatic searches point towards PG4S enrichment on the non-template strand in particular classes of genes and suggest a more global effect than activity at specific promoter sites. As noted above, unwinding of the B-DNA duplex during transcription or replication can free nucleotides from canonical WC-pairs and promote formation of alternate intramolecular secondary structures including triplexes, Z-DNA and G4 DNA. For tetraplexes, this has been confirmed by the Maizels group in a G-rich region involved in immunoglobulin class switch recombination [70]. The resulting unwound duplex, labeled a “G-loop”, is shown schematically in panel (a) of Figure
2.5.1.2 and was directly visualized by electron microscopy, as reproduced in in panels (b) and (c). G-loops are characterized by a DNA-RNA hybrid on the template strand and stable G4 folds on the non-template strand. In immunoglobulin class switch repeats, the G4 folds in G loops are stabilized by binding to the mismatch repair protein MutSα which has been shown to bind tetraplexes with high affinity.

Though the Maizels group studied G-loops in the context of a specific locus, a recent review by Tornaletti [123] discusses an emerging hypothesis that G-loop like structures may occur on a wider scale and account, in part, for the observed link between genomic instability and G-rich regions. This hypothesis proposes that formation of non-B DNA secondary structures, including G4 folds, can lead to polymerase arrest that in turn triggers spurious activation of an error-prone transcription-coupled repair (TCR) pathway on undamaged DNA.

The nucleotide excision and DNA repair (NER) pathways serve an essential role in repairing a wide class of DNA lesions [84]. In conjunction with other DNA repair pathways, they maintain genomic integrity in the face of a constant barrage of chemical challenges. NER pathways remove a short oligonucleotide that includes the damaged site and direct replacement of the gap by DNA polymerase I via the undamaged strand. The distinguishing feature of this subset of repair mechanisms is the ability to detect and respond to structural changes in B-form duplex as diagnostics of DNA lesions. Two distinct NER subpaths are involved. Global genomic DNA repair (GCR) can sense and repair lesions throughout the genome, whereas transcription coupled repair (TCR) specifically couples repair to transcription.
Figure 2.7: G-loop formation during transcription of G-rich immunoglobulin class switch regions. G-loops are shown schematically in (a) and in electron micrographs in (b) and (c). Binding by MutSα, a mismatch-repair factor with high G4 affinity, is shown in (c). (image from [70])
A key trigger of TCR activity is detection of a stalled polymerase as an indicator of DNA pathology. Thus processivity of RNAPII is used as a timely and sensitive monitor of DNA damage, particularly in heavily-transcribed genes [122]. In previous work, Tornaletti and coworkers documented that formation of G4 folds may block T7 RNA polymerase and RNAPII [124]. The G4-forming sequence used in this work was the same 27 element repeat, 5′GCTGAGCTGGGGTGAGCTGA-3′, previously shown to exhibit a G-loop structure by Maizels and coworkers. Transcription of these repeats with the G-rich sequence on the non-template strand consistently triggered partial arrest in both polymerases. In RNAPolII, the arrest site was 10 bp upstream of the first repeat, in agreement with the elongation edge measured in previous analyses [124]. Arrest was enhanced by increasing $K^+$ concentration, which would be expected to stabilize a G4 fold, and by modifying the in vitro transcription assay to permit multiple concurrent rounds of transcription. In the latter, negative supercoiling downstream of an initial polymerase would be expected to facilitate G4 fold formation ahead of its successor.

This result extended considerable previous work by the Hanawalt lab documenting polymerase arrest by other non-B DNA secondary structures including Z-DNA [27] and triplet repeats [8]. Interestingly, the latter examined repeats within the same c-MYC nuclease-hypersensitive region studied by Hurley and coworkers as an example of G4-fold transcriptional regulation.

Taken together, these results suggest that formation of non-B DNA structures as a side effect of transcription-coupled DNA unwinding may block polymerase progressivity. The link between polymerase arrest and activation of TCR pathways has been
well documented [122]. Though DNA repair is essential, it provides less fidelity than mainstream replication. The higher rate can account for greater instability in regions associated with high TCR activity which would tentatively include PG4S sequences along with CAG-CTG triplet repeats and Z-DNA.

The hypothesis, while well-founded, remains to be tested. If proven, it would provide a mechanism to account for widespread linkage between G-rich sequence and genomic instability, particularly in heavily transcribed regions. The proposed mechanism closely parallels a similar effect recently documented during DNA replication as described in Section 2.5.3.

In summary, while the scope and role of G4 DNA in transcriptional regulation remain under active investigation, conclusions regarding the underlying mechanisms are still speculative. Quadruplexes in the c-MYC [112] and c-KIT [111] promoter regions have demonstrated capacity to repress transcription rates and to be amenable to ligand-induced stabilization. However, it remains unclear whether these results generalize to the many G-rich elements uncovered in promoter regions by sequence screens. Recent analysis of human genes has shown that many putative quadruplex-forming elements display features characteristic of other regulatory motifs including binding sites for the SP1 transcription factor and cytosine methylation [32, 121].

2.5.2 Translational Regulation: G4 Folds in RNA

As noted above, G4 folds might be expected to play a more prominent role in RNA than in DNA because in RNA guanines need not be sequestered within the
constraints of a canonical Watson-Crick duplex. Though the significance of quadruplex motifs in RNA has been investigated to a lesser degree than in DNA, a number of studies have documented G4 formation as a translational suppressor. Structured elements in UTR mRNA regions are a common mechanism of translational control in both prokaryotes and eukaryotes. In this respect, G4 folds seem to exercise an inhibitory role similar to that of other 5’UTR inhibitors such as riboswitches and temperature-dependent hairpins [115].

Two publications in 2007 initially documented G4-mediated translational repression of a reporter gene. One analyzed synthetic PG4S elements in e.coli [128], and the other examined a 5’UTR element of the human NRAS gene in a cell-free translation system [57]. More recently, studies have built on the results of bioinformatic screens to further explore this effect. Arora and coworkers considered the impact of a 5’UTR PG4S on expression of the human Zic-1 zinc-finger protein [3] and Balkwill et. al applied a similar approach to an estrogen receptor whose mRNA exhibits multiple isoforms [7]. These results are summarized below.

Wieland and Hartig constructed a series of G-rich sequences expected to display a decreasing propensity to adopt stable G4 folds as shown in Figure 2.8 [128].

The G3U sequence was anticipated to adopt the most stable fold because of the single linking uracil, followed by G3U2 with a two-nucleotide loop. The number of tetrads was decreased from three to two in G2U and G2U2 and further weakened by introduction of a long loop in G2LL. These sequences, which incorporate the consensus
Figure 2.8: Quadruplex-forming sequences in the 5’ UTR of an E. coli reporter gene. Guanines in the wild-type sequence are shown in green and the unmutated Shine-Dalgarno and start codon nucleotides are shown in lower case. Mutations that enable possible 3-quartet folds are shown in red and those that enable two-quartet folds are shown in magenta. The sequence compatible with the most stable quadruplex fold, G3U, with three quartets and short loops between the Shine-Dalgarno site and the start codon, exhibited a 96% reduction of expression relative to wild type [128].

Shine-Dalgarno sequence (GAAGGAG) 14-bp upstream of the start codon for a green-fluorescent reporter gene, were transformed into an E. coli strain. Reduced expression tracked anticipated structural stability with a 96% reduction relative to wild type for the G3U fold decreasing to a 20% reduction for G2LL. Though no detailed structural studies were attempted, analysis of CD spectra and melting curves confirmed the expected variation in stability. The authors also noted that mutations introduced a clear temperature dependence on gene expression. This effect was not observed in wild type but consistently correlated with the expected stability of the G4 insertions.

Similar to the membrane-bound receptor encoded by c-KIT, NRAS is a well-studied oncogene whose expression is linked to regulating cell replication and differentiation. Genomic screens using the QUADPARSER algorithm [49] identified a PG4S element in the 5’ UTR of human NRAS,

\[
5’-UGUGGGAGGGCGGGUCUGGG-3’
\]
222-bp upstream of the TSS; the element is well-conserved in orthologous genes. To investigate the element’s potential for G4-fold formation and regulatory function, Kumari et al first measured its CD spectrum and obtained positive and negative peaks at 263 nm and 241 nm respectively, a signature characteristic of parallel-stranded G4 folds [57]. In subsequent experiments, they inserted the entire 254-bp 5’ UTR upstream of a firefly luciferase gene and demonstrated that either a truncated construct lacking the PG4S element or one containing a GGG-to-AAA mutation, expected to interfere with the presumed G4 fold, enhanced translation. In both cases they observed an approximate 3.7-fold increase in translation rate with 5’ UTR constructs lacking the capacity for a G4 fold.

In a similar approach, Arora and coworkers examined a highly conserved G-rich element [3]

5-GGGUGGGGGGGCGGGGGAGGCCGGGG-3

46-bp upstream of the TSS for the human Zic-1 gene. Its gene product is a zinc-finger transcription factor involved in development of the cerebellum. Overexpression of Zic-1 is linked with childhood medulloblastoma tumors. Translation rates were again assayed with a luciferase reporter gene, but here results were obtained in vivo in a human (HeLa) cell line. Insertion of the G4 element upstream of the start codon resulted in 80% reduction in protein synthesis and the effect was undone by guanosine to adenine mutations that destabilized the G4-forming potential of the sequence.

G4-fold mediated translational control in a more complex gene structure was
investigated by Balkwill and coworkers in a recent analysis of the human estrogen α receptor ESR1 [7]. Transcription of ESR1 is known to rely on multiple, tissue-specific, promoters, yielding several mRNA variants with different 5’UTRs subject to separate translational regulation [54]. In this case, sequence analysis with QUADPARSER [49] revealed 20 PG4S elements distributed throughout the gene’s 472928 nucleotides of which 6 among promoter elements, 10 among introns, 3 among exons and 1 in the 3’UTR. The authors focused on a specific PG4S element, QS6, located in the 5’UTR exon of one of the ESR1 isoforms. After confirming the PG4S sequence was compatible with a G4 fold by CD and UV analysis, they assayed its ability to suppress translation of a luciferase reporter gene via in vitro translation in rabbit reticulocyte lysate. A six fold increase in translation efficiency was observed with a G4-inhibiting mutation relative to wild type, supporting a role for the QS6 PG4S element in translation control.

A higher throughput approach to functional analysis of translational repression is presented in recent publications by Wieland and Hartig. These include specification of a fairly rapid, two-day, protocol for assessing the stability and translational impact of a candidate sequence [129] based on their earlier approach and application of this method to systematic evaluation of natural and synthetic PG4S elements [45].

In summary, there appears to be a growing collection of evidence, albeit in synthetic systems, that supports a repressive role for G4 folds in mRNA, notably in the 5’ UTR, by interfering with effective translation. This effect has been documented in oncogenes for which downregulation may have therapeutic benefit. Thus agents that selectively stabilize G4-folds in mRNAs may provide an alternative to the ongoing
research in transcriptional control. However, it is noteworthy that at this time there is not yet any clear indication of which helicases or other factors may be involved in unwinding, or possibly promoting, G4 formation as a general mechanism of translational control.

2.5.3 Replication and G4-linked Genomic Instability

Recent work by Tijsterman, Brosh and Boulton has established a conserved pathway linking G4 fold formation during replication and genomic instability [72]. Duplex unwinding in the replication fork can permit formation of G4 folds as shown schematically in Figure 2.5. In wild type, formation of G4 folds is inhibited by specialized helicases. However, mutation or silencing of the helicases reliably triggers replication-associated deletion of G-rich regions. This work parallels G4-associated interference with transcriptional processing [123], though in this instance the effect is not polymerase arrest but deletions in the nascent strand. Furthermore, unlike transcription, proteins responsible for maintaining wild-type behavior, members of the DEAH family of helicases, have been identified.

These results build on earlier work showing that G-rich regions were deleted in a C. elegans mutant deficient for gene F33H2.1, a DEAH helicase labeled dog-1 (deletion of guanine). Deletions consistently began at the 3' end of G-tracts but terminated at random locations about 300 nucleotides downstream [17]. Later reports established that dog-1 is the C. elegans homolog of human FANCJ [132]. This gene has been associated with the disorder Fanconi anemia characterized by congenital abnormalities
and a heightened susceptibility to agents of DNA interstrand cross-linking damage. More recently, Tijsterman and coworkers confirmed that absence of dog-1 accounted for deletions of G4PS elements and furthermore that deletions occurred only in G4 DNA [56]. Deletions occurred at approximately 4% of sites per generation and were related to G-richness at the site. Brosh and coworkers established a complementary results by RNAi-induced silencing of FANCJ in HeLa cells [130]. They demonstrated FANCJ unwinds G4 DNA, including telomere sequences, and that unwinding was inhibited by addition of telomestatin, a known G4-stabilizing agent.

In summary, this evidence indicates transient formation of G4 folds in the replication fork can trigger genomic instability unless inhibited by appropriate helicases and that surveillance of PG4S elements is a conserved mechanism.

2.5.4 Telomere Maintenance

The most clearly established domain of G4-associated regulation occurs at chromosomal termini where putative quadruplex folds in the G-rich, species-specific, duplex telomeric repeats and in the terminal, single-stranded, 3’ overhang provide specific substrates for proteins involved in maintenance of chromosome integrity. Telomerase, the ribonucleoprotein complex responsible for synthesis of telomeric sequence, is inhibited by quadruplex formation in the 3’ overhang that serves as complementary substrate to the complex’s RNA moiety [134]. hPOT1, another prominent factor in maintenance of human telomeres, maintains processivity of the telomerase complex by recognizing and disrupting quadruplex formation [135, 126]. Structures for a number of
telomeric repeats have been determined along with those of associated binding proteins [81] though interpretation of the structural results has been complicated by significant, cation-dependent, polymorphism [64, 61]. Since immortality conferred by upregulation of telomerase activity is a hallmark of many tumor cell lines, synthetic stabilization of quadruplex structures in telomeric sequence is an active area of pharmaceutical research [89].

2.5.5 mRNA Splicing

In addition to translational regulation, quadruplex-forming elements in mRNA have been shown to regulate splicing of alternate mRNA products. Binding by the fragile X mental retardation protein (FMRP) to a polymorphic G-quartet element in its own mRNA can direct translation product, between two alternate FMRP isoforms, according to intra-cellular product concentration [26]. However, direct involvement of FMRP, and its G-quartet binding site, in the splicing process has not yet been demonstrated.
Chapter 3

Observed UGGGGU Quadruplex Structures

This chapter presents structural characteristics of the UGGGGU hexamer observed in the data sets we collected. The focus here is the conformation of the polymer and the arrangement of cations in the central column formed by the quadruplex fold; the structure of solvation shells is described in Chapter 4. The experiments we describe have uncovered novel properties of this structure: two new crystal packings, one in a previous previously unreported orthorhombic space group; a well-hydrated internal cavity which results from an interesting asymmetry in the stability of the terminal 3’ uridines in the octameric assembly; and evidence for multiple conformations of the phosphate backbone.

Methods used for crystallization, data collection and refinement along with quality indicators of the data sets used are summarized in the initial section. Subsequent
sections describe the structural units that result from assembly of two, four and eight oligos, alternate crystal packings, conformation of the backbone chain, coordination of metal cations by G and U quartets and the distribution of B-factors and other aspects of disorder.

Both DNA, d(TGGGGT), and RNA, r(UGGGGU), forms of the hexamer examined in this thesis have been reported previously [25, 94, 58]. The combination of a stable quadruplex fold with no bridging loops and low solvent content yields a stable structure which exhibits remarkably strong diffraction for a nucleic acid.

3.1 Methods

We purchased UGGGGU RNA oligos from Dharmacon and removed 2’-protecting groups by incubation in 100mM acetic acid adjusted to pH 3.8 with TEMED per the manufacturer’s instructions [119]. Crystals grew in hanging-drop vapor diffusion plates using RNA concentrations of 8-10 mg/mL and previously reported conditions [25]: 40 mM sodium cacodylate buffer (pH 7.0), 20 mM magnesium chloride, 12.0 mM spermine chloride, 80 mM lithium chloride, 40 mM strontium chloride, 20 mM calcium chloride, 10% (vol/vol) methyl-2,4-pentanediol (MPD), and 35% MPD in the reservoir. Usable crystals grew within one to two weeks. Variation in the neighborhood of these conditions did not appear to affect crystal quality consistently. However, high mosaicity and irregular growth yielding multiple lattices were a common problem and required screening multiple crystals to obtain high-quality data sets. Growth of irregular crystals
was significantly improved by a simple macroseeding protocol [9].

3.2 Crystallographic Data

The thirty three data sets used for our analysis were collected on beam-lines 9-1, 9-2, 11-1 and 7-2 at the Stanford Synchrotron Radiation Laboratory (SSRL) from seventeen single crystals selected from a much greater pool of candidates. We observed three distinct crystal packings with high-resolution limits ranging from 1.4 to 0.85 Å. For several crystals, multiple data sets were collected, either to optimize anomalous differences or to improve data statistics via high and low resolution passes.

For data reduction, we relied on programs in the ccp4 [20], scalepack [88] and labelit [105] suites. Determining the correct spacegroup for these data required a non-standard iterative search as the results of the autoindexing modules of mosflm [20] and HKL2000 [88] did not lead to a solution. For the tetragonal packing, detection of the correct symmetry elements is complicated by the near equality of the $a$, $b$, and $c$, dimensions of the unit cell, a condition that nearly meets the lattice restrictions imposed by a cubic crystal system.

The data reduction search strategy was divided into three stages. An initial attempt to assess symmetry, mosaicity and resolution was first made by running labelit [105] on randomly selected pairs of images with approximately 90° separation, thus sampling estimates of these parameters over the full data set. To confirm the labelit solution for each pair, the CCP4 program pointless was also run on the mtz file produced
as part of labelit’s search. In the next stage, parameters for the most commonly reported, consensus, space group solution were used for post-refinement and wedge integration of two contiguous ranges of images starting at each pair. Pointless was again run on the mtz files resulting from the wedge integration and the resulting space group probabilities were accumulated. The most probable space group and the cell parameters yielding the lowest weighted positional error from the wedge integration were used for the last stage of processing. In this step, a number of integration and scaling runs with different parameters were applied to the full data set to find optimal settings. Data reduction summary statistics are shown in 3.1 for each packing.

3.2.1 Molecular Replacement

The structures were solved by molecular replacement starting from a previously published UGGGGU model [25]. However, initial, naive use of the phaser and molrep software did not lead to a solution. Though the tetragonal packing observed here shares the $P4_212$ space group of the previously reported UGGGGU structure [25], the length of the $c$ unit cell axis, 37.4 Å for data set P1B, is only half the length of that for PDB entry 1JG8, 74.1 Å. This observation, coupled with the four-fold symmetry expected for a quadruplex, motivated Peter Dunten to apply rigid body refinement via phenix.refine using only half of the 1J8G octamer. This search was iterated over different angles about the four-fold axis and yielded a successful initial model [30]. Similarly, consideration of the limited packing options available for placing the octameric unit within the larger orthorhombic unit cell, as illustrated in Figure 3.4, motivated a
Crystallographic Data Reduction Statistics for Data Sets in Space Group $P4_2_12$ (P1x) and $C22_1$ (P2x).

<table>
<thead>
<tr>
<th>Id</th>
<th>Resolution</th>
<th>$R_{pim}$</th>
<th>Total</th>
<th>Unique</th>
<th>Completeness</th>
<th>Multiplicity</th>
<th>$I/\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1A</td>
<td>36.67(0.89)</td>
<td>0.018(0.161)</td>
<td>138943(3594)</td>
<td>18942(1927)</td>
<td>95.0(68.4)</td>
<td>7.3(1.9)</td>
<td>20.7(3.5)</td>
</tr>
<tr>
<td>P1B</td>
<td>26.16(0.90)</td>
<td>0.034(0.301)</td>
<td>388930(8988)</td>
<td>17571(1296)</td>
<td>91.5(49.1)</td>
<td>22.1(6.9)</td>
<td>17.3(3.4)</td>
</tr>
<tr>
<td>P1C</td>
<td>37.01(1.20)</td>
<td>0.024(0.217)</td>
<td>200414(28506)</td>
<td>8310(1182)</td>
<td>100.0(100.0)</td>
<td>24.1(24.1)</td>
<td>17.6(3.8)</td>
</tr>
<tr>
<td>P1D</td>
<td>26.16(0.85)</td>
<td>0.041(0.523)</td>
<td>234881(28080)</td>
<td>22995(3291)</td>
<td>99.8(100.0)</td>
<td>10.2(8.5)</td>
<td>14.0(3.4)</td>
</tr>
<tr>
<td>P1E</td>
<td>36.99(1.27)</td>
<td>0.020(0.257)</td>
<td>61058(2114)</td>
<td>6417(601)</td>
<td>92.3(63.3)</td>
<td>9.5(3.5)</td>
<td>16.0(2.4)</td>
</tr>
<tr>
<td>P1F</td>
<td>37.05(1.02)</td>
<td>0.010(0.057)</td>
<td>427960(11310)</td>
<td>12797(1230)</td>
<td>95.0(66.5)</td>
<td>33.4(9.2)</td>
<td>57.4(9.3)</td>
</tr>
<tr>
<td>P1G</td>
<td>37.07(1.03)</td>
<td>0.028(0.136)</td>
<td>138375(5040)</td>
<td>12244(1255)</td>
<td>93.3(67.5)</td>
<td>11.3(4.0)</td>
<td>21.6(4.4)</td>
</tr>
<tr>
<td>P2A</td>
<td>47.77(1.25)</td>
<td>0.026(0.142)</td>
<td>160983(9848)</td>
<td>26360(2908)</td>
<td>95.4(74.1)</td>
<td>6.1(3.4)</td>
<td>17.3(5.0)</td>
</tr>
<tr>
<td>P2B</td>
<td>47.81(0.99)</td>
<td>0.019(0.111)</td>
<td>616625(18264)</td>
<td>50165(4647)</td>
<td>92.0(59.6)</td>
<td>12.3(3.9)</td>
<td>24.2(5.4)</td>
</tr>
<tr>
<td>P2C</td>
<td>31.90(0.97)</td>
<td>0.026(0.763)</td>
<td>640869(24660)</td>
<td>56088(6371)</td>
<td>96.4(77.3)</td>
<td>11.4(3.9)</td>
<td>18.5(1.1)</td>
</tr>
<tr>
<td>P2D</td>
<td>47.77(1.40)</td>
<td>0.029(0.344)</td>
<td>120145(8817)</td>
<td>19329(2424)</td>
<td>97.7(86.5)</td>
<td>6.2(3.6)</td>
<td>16.7(2.1)</td>
</tr>
</tbody>
</table>

Table 3.1: Crystallographic Data Reduction Statistics. Standard measures and quality indicators for the data sets used in our analysis. Definitions of resolution, $R_{pim}$ and $I/\sigma$ are available in most texts on crystallographic analysis, for example, Chapter 8 of [102].
solution in the $C_{2221}$ space group. We also relied on anomalous difference maps, as shown in Figure 3.6, for independent experimental confirmation of the models.

### 3.2.2 Refinement

Once approximate placement of the RNA chains was completed, refinement focused on identifying multiple conformations, alternate locations of the Na$^{+}$ cations at the 3’ end of the octameric unit and on identifying solvent composition and location.

The *refmac* [80], *shelxl* [109] and *phenix.refine* [1] programs were used for refinement. Rigid body refinement search with *phenix.refine* provided initial placement of the RNA chains and the program’s simulated annealing feature was used to place alternate conformations. The ability of *shelxl* to specify atom-specific settings for hydrogen placement and anisotropic B-factors was useful in limiting those parameters to better-ordered regions. Placement of hydrogen atoms yielded a definite drop in $R_{\text{free}}$ values, however not all possible hydrogens were modeled. Hydrogens on waters, hydroxyl groups, disordered atoms and atoms in alternate conformations were not included. Similarly, anisotropic B factors were not refined for alternate conformations or disordered residues. Statistics from three representative refinements are shown in Table 3.2.

Waters were placed by means of a novel approach based on partitioning density difference maps into Morse-Smale complexes as described in Chapter 5. Briefly, a preliminary model was built and waters and ions first placed via the coot [36] “find-waters” command. All waters were then removed and the resulting difference map was partitioned into basins bounded by zero-flux surfaces. Basins enclosing an acceptable
total density with a peak within plausible hydrogen bonding distance to non-hydrogen atoms of the existing working model were selected as candidate waters. Those yielding improvements in $R_{work}$ and $R_{free}$ were retained for further rounds of refinement. The procedure was iterated outward from the working model to uncover waters in secondary hydration shells.

### Refinement Statistics

<table>
<thead>
<tr>
<th></th>
<th>P1A</th>
<th>P1B</th>
<th>P2C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cell dimensions, Å</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a$</td>
<td>36.636</td>
<td>36.613</td>
<td>36.667</td>
</tr>
<tr>
<td>$b$</td>
<td>36.636</td>
<td>36.613</td>
<td>54.210</td>
</tr>
<tr>
<td>$c$</td>
<td>37.034</td>
<td>37.380</td>
<td>95.693</td>
</tr>
<tr>
<td>Non-hydrogen atoms refined</td>
<td>331</td>
<td>339</td>
<td>1296</td>
</tr>
<tr>
<td>Solvent waters</td>
<td>54</td>
<td>68</td>
<td>247</td>
</tr>
<tr>
<td>Correlation coefficient FO-FC(free)</td>
<td>0.981</td>
<td>0.971</td>
<td>0.981</td>
</tr>
<tr>
<td>Resolution range, Å</td>
<td>37.03-0.89</td>
<td>37.38-0.90</td>
<td>47.85 - 0.97</td>
</tr>
<tr>
<td>Completeness for range</td>
<td>94.83</td>
<td>91.37</td>
<td>95.93</td>
</tr>
<tr>
<td>Overall B-factor, Å$^2$</td>
<td>11.961</td>
<td>11.877</td>
<td>13.2</td>
</tr>
<tr>
<td>$R_{work}$</td>
<td>0.12302</td>
<td>0.10366</td>
<td>0.12578</td>
</tr>
<tr>
<td>$R_{free}$</td>
<td>0.13291</td>
<td>0.12188</td>
<td>0.13926</td>
</tr>
</tbody>
</table>

Table 3.2: Refinement Statistics. Quality indicators for our refinement of three representative data sets. Definitions of $R$, map correlation coefficients and overall B factor statistics are available in [102], Chapter 12, and similar texts.
3.3 Oligo Assemblies

For all unit cell arrangements we observed, the overall crystal packing can be decomposed into columns comprised of coaxially stacked oblong octamers. Each octamer consists of eight polymer chains wrapped about a central axis as summarized in Fig. 3.1. The individual RNA oligos that comprise each chain are wrapped about the four-fold axis with a right-handed helical twist as shown in panel (a).

The four guanidine residues and the 3’ uridine point towards the axis and the orbit resulting from a symmetry rotation about the four-fold yields the stacked G-quartets characteristic of a G4 fold as shown in Figure 3.1(b). The G-quartets formed by four hydrogen-bonded guanidine bases are approximately planar and all nucleotides adopt an anti conformation about the glycosylic bond with $\chi$ values around $-144^\circ$, near the center of the anti range.

Cations that stabilize the stacked G-quartets are positioned along the four-fold symmetry axis. In these structures, Sr$^{2+}$ cations are invariably intercalated between two G-quartets, as shown in magenta in Figure 3.1(b), and Na$^+$ cations coordinate the 3’ U tetrads, as shown in green in Figure 3.5(b). As summarized in Chapter two, other reported G4 structures exhibit considerable variability in this regard.

The 5’ and 3’ terminal uridines adopt very different conformations. The vector corresponding to the C1’-N1 bond of the 3’ uridine points towards the central axis but diagonally away from it in the 5’ uridine.

In all structures we examined, the four-stranded assemblies are arranged in
Figure 3.1: Octamer Assembly. (a) A single strand corresponding to chain R6A from dataset P1B. The four central guanidine residues are shown in yellow and the uridines in blue. Two Sr$^{2+}$ ions are shown on the four-fold axis, the 3’ uridine is at top. (b) The orbit resulting from rotation about the four-fold and (c) as viewed along the symmetry axis. (d) Strand R6A and R6B are related by a non-crystallographic two-fold axis. The strands superpose with an RMS distance of 0.140Å over the first five 5’ residues.
pairs of opposite strand polarity related by a two-fold axis as shown in Figure 3.1(d). Conformation of the two strands is in close agreement for the first 5 residues in 5’ to 3’ order; the 136 atoms involved can be superposed with a $rms$ deviation of 0.141. However, the 3’ uridines of the two strands exhibit very different conformations as shown in Figure 3.2. The two conformations are characterized by large differences in the backbone $\alpha(P - O_{5'})$ and $\delta(C_4' - C_3')$ torsion angles, 166.3° versus 71.9° and 139.6° versus 90.5° as shown in rows U1006 and U2006 of Table 3.3.

![Figure 3.2](image)

Figure 3.2: *Least squares superposition of chains R6A and R6B from P1B structure using atom positions from the first five residues in 5'-3' order*

The structural consequence of this difference in the corresponding four-fold orbits is that the terminal 3’ uridine tetrad tilts towards the guanine quartets in one
orbit but away from it in the other. To distinguish these conformations, we label the two strands as “R6A” and “R6B” and the corresponding quartet assemblies as “R6A_4” and “R6B_4”. Throughout, residues are numbered in 5’ to 3’ order with R6A starting at 1001 and R6B at 2001. In most figures, strand R6A is shown below R6B. Relative to the G quartets, the 3’ U-tetrad is concave in R6A_4 but convex in R6B_4.

Simultaneous application of four-fold symmetry to the two strands yields a pair of coaxial, interlocked quadruplex folds with head-to-head, 5’-5’ polarity as shown in panel (a) of Figure 3.3. Assembly of this octameric unit, designated by the label “R6_8”, involves insertion of the spalled-out 5’ uridines into the grooves created by the helical twist of the R6A_4 and R6B_4 folds. Remarkably, this insertion retains the relative twist of successive G-quartets. Thus, the eight stacked quartets of the R6_8 unit maintain an approximate 20° rotation as emphasized in panels (b) and (c) of Figure 3.3.

The oblong, octameric, R6_8 unit, which loosely resembles a beehive, terminates at the 3’ U-quartets of the R6A_4 and R6B_4 components. Column packing results from coaxial stacking of octamers and aligns the convex 3’ U-quartet of R6B_4 within the concave 3’ U-quartet of the R6A_4 complex of the adjacent octamer as shown, along with the coordinating Na⁺ cations, in panel (b) of Figure 3.5. Thus, the repeating pattern of polymer strands within each columns is R6A(3’-5’),R6B(5’-3’),R6A(3’-5’),...
Figure 3.3: Helical stacking of four-stranded assemblies. (a) Unit $R6B_4$ is shown on top in gray and the $O6$ atoms for one strand of each unit are shown in red. Each quartet is offset by about $20^\circ$ in a right-handed helical twist. (b) Arrangement of the $O6$ atoms of the eight quartets about the central cation column. $Sr^{2+}$ cations are shown in magenta. (c) Two strands of $O6$ atoms viewed down the $Sr^{2+}$ axis, the total rotation spans approximately $180^\circ$. 
3.4 Alternate Crystal Packings

We observed crystal packings in two space groups, tetragonal $P4_2\overline{1}2$ and orthorhombic $C22\overline{1}1$, as well as two distinct unit cells dimensions among among the tetragonal data sets. Different packings exhibit different solvent volumes, crystal contacts and division of labor between crystallographic and non-crystallographic symmetry operators. A structurally prominent feature of the different packings is the arrangement of the columns formed by stacked R6$_8$ units as shown in Figure 3.4.

Figure 3.4: Arrangement of quadruplex columns in tetragonal ($P4_2\overline{1}2$) and orthorhombic ($C22\overline{1}1$) space groups. Stacked R6$_8$ assemblies are seen with the central cation axis directed towards the reader. The asymmetric unit is shown in lighter colors and solvent molecules are shown in magenta.

Whereas in $P4_2\overline{1}2$ the nearest neighbors of each column are arranged in a
square, in $C_{222_1}$ the arrangement is a hexagon. The latter reduces gaps between R6,8 units yielding a lower solvent content and a greater incidence of crystallographically observable of bound solvent.

A second prominent difference between the two packings is that the R6B form of the hexamer only occurs in the tetragonal setting. In the orthorhombic space group, only the conformation with concave 3’ U-quartet occurs. In consequence, coaxial stacking of the R6,8 assemblies differs between the two space groups.

### 3.5 Backbone Conformation

The conformation of the nucleic acid backbone is specified by six torsion angles labeled $\alpha$ to $\zeta$ in sequential order from the $P-O_{5'}$ bond to $O_{3'}-P$ along the backbone. The six angles allow for much greater flexibility than the $(\phi, \psi)$ pair characteristic of a protein chain. Table 3.3 shows the values torsion angles computed with the Curves+ program [59] for structure P1B. For comparison, representative values for A and B-form DNA calculated from the 440D and 1BNAPD entries shown in Figure 2.1 are also included.

Overall, the torsion angles show considerable variation across nucleotides and agree with neither the A nor B-form duplexes, though they seem in closer agreement with the former. As expected from close superposition, the two strands are in good agreement with one another with the exception of the $\alpha$ and $\delta$ angles of the 3’ uridines (U1006 and U2006), a difference which reflects the convex-concave form of the corresponding
R6A₄ and R6B₄ assemblies.

Recent work has shown that the RNA backbone is rotameric notwithstanding the greater flexibility that results from six torsion angles as compared to the $\phi$, $\psi$ and $\omega$ angles that define the protein backbone [100, 79]. Analysis of reliable nucleic acid structures showed that adjoining torsion angles are correlated and occupy favored regions of conformational space, as has long been known for proteins. Our analysis of conformation angles from structures P1A and P1B via the Richardson lab dangle and suitename programs (suitename version 0.3.070628) [100] indicated that ten of the twelve residues could be assigned to known clusters or “suites”. The two outliers were torsion angles corresponding to the backbone between the two central guanines. Backbone stress in this area might be due to deviation of the guanine bases away from one another and towards coordinating Sr²⁺ cations. Further investigation is needed to confirm why these angles are flagged as outliers. In light of the strong experimental support for this part of the structure, including anomalous differences free from model bias, it seems likely that the discrepancy is due to omissions in the rotamer conformation database.

### 3.6 Quartet Cation Coordination

As discussed above, the close packing of bases characteristic of quartet structures places carbonyl oxygens, O₆ in G-quartets and O₄ in U-quartets, in close proximity. The repulsion between partial negative charges residing on neighboring O atoms
<table>
<thead>
<tr>
<th></th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
<th>ε</th>
<th>ζ</th>
<th>χ</th>
<th>P</th>
<th>A</th>
<th>Pucker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(P - O₅)</td>
<td>(O₅ - C₅')</td>
<td>(C₅' - C₄')</td>
<td>(C₄' - O₃')</td>
<td>(O₃' - P)</td>
<td>(C₃' - N₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R6A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1001</td>
<td>——</td>
<td>——</td>
<td>54.0</td>
<td>81.0</td>
<td>-158.7</td>
<td>-92.3</td>
<td>-143.3</td>
<td>10.5</td>
<td>41.7</td>
<td>C3'endo</td>
</tr>
<tr>
<td>G1002</td>
<td>-98.4</td>
<td>68.4</td>
<td>174.7</td>
<td>80.1</td>
<td>-152.1</td>
<td>-70.6</td>
<td>-165.0</td>
<td>8.3</td>
<td>39.7</td>
<td>C3'endo</td>
</tr>
<tr>
<td>G1003</td>
<td>-67.6</td>
<td>-172.2</td>
<td>58.5</td>
<td>150.6</td>
<td>-170.5</td>
<td>-119.2</td>
<td>-121.7</td>
<td>166.6</td>
<td>42.8</td>
<td>C2'endo</td>
</tr>
<tr>
<td>G1004</td>
<td>-76.0</td>
<td>177.4</td>
<td>42.5</td>
<td>77.0</td>
<td>-151.1</td>
<td>-73.1</td>
<td>-154.7</td>
<td>16.7</td>
<td>42.8</td>
<td>C3'endo</td>
</tr>
<tr>
<td>G1005</td>
<td>-70.4</td>
<td>177.5</td>
<td>53.3</td>
<td>78.4</td>
<td>-157.1</td>
<td>67.9</td>
<td>-146.0</td>
<td>14.6</td>
<td>42.0</td>
<td>C3'endo</td>
</tr>
<tr>
<td>U1006</td>
<td>166.3</td>
<td>141.4</td>
<td>46.2</td>
<td>139.6</td>
<td>——</td>
<td>——</td>
<td>-135.3</td>
<td>149.4</td>
<td>41.7</td>
<td>C2'endo</td>
</tr>
<tr>
<td>R6B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U2001</td>
<td>——</td>
<td>——</td>
<td>53.4</td>
<td>78.3</td>
<td>-157.6</td>
<td>-90.7</td>
<td>-147.2</td>
<td>10.6</td>
<td>42.9</td>
<td>C3'endo</td>
</tr>
<tr>
<td>G2002</td>
<td>-99.6</td>
<td>67.3</td>
<td>172.9</td>
<td>84.1</td>
<td>-153.1</td>
<td>-71.2</td>
<td>-170.5</td>
<td>6.1</td>
<td>39.5</td>
<td>C3'endo</td>
</tr>
<tr>
<td>G2003</td>
<td>-70.2</td>
<td>-168.6</td>
<td>62.4</td>
<td>152.4</td>
<td>-172.0</td>
<td>-117.1</td>
<td>-116.3</td>
<td>170.3</td>
<td>38.6</td>
<td>C2'endo</td>
</tr>
<tr>
<td>G2004</td>
<td>-78.4</td>
<td>176.7</td>
<td>47.3</td>
<td>75.2</td>
<td>-155.2</td>
<td>-70.6</td>
<td>-154.6</td>
<td>18.8</td>
<td>43.5</td>
<td>C3'endo</td>
</tr>
<tr>
<td>G2005</td>
<td>-68.4</td>
<td>175.0</td>
<td>48.2</td>
<td>75.3</td>
<td>-163.4</td>
<td>43.6</td>
<td>-149.5</td>
<td>15.4</td>
<td>43.4</td>
<td>C3'endo</td>
</tr>
<tr>
<td>U2006</td>
<td>71.9</td>
<td>176.2</td>
<td>56.3</td>
<td>90.5</td>
<td>——</td>
<td>——</td>
<td>-161.4</td>
<td>20.6</td>
<td>43.0</td>
<td>C3'endo</td>
</tr>
<tr>
<td>A DNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-52</td>
<td>175</td>
<td>42</td>
<td>79</td>
<td>-148</td>
<td>-75</td>
<td>-157</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B DNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-30</td>
<td>136</td>
<td>31</td>
<td>143</td>
<td>-141</td>
<td>-161</td>
<td>-98</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: RNA backbone torsion angles for structure P1B. Representative examples of A and B form duplexes are calculated from PDB entries 440D and 1BNA respectively. Calculations are from program Curves+ [59].
is screened by coordination of dehydrated cations in the central channel formed by the stacked quartets. The coordination patterns we observed in the UGGGGU structure exhibit three distinct coordination patterns, one defined by the two central guanine quartets and two by the U-quartets formed at the 3’ of the R6A_4 and R6B_4 3’ assemblies. As described previously, the 5’ uridines rotate away from the central channel and thus do not participate in either quartets or cation coordination.

The coordination geometry of the guanine quartets is invariant across all structures examined in this work and is summarized in panel (a) of Figure 3.5. The central cation is always Sr\(^{2+}\), centrally positioned between eight O6 carbonyl oxygens of adjacent G-quartets in a bipyramidal-antiprism geometry. The Sr-O6 distance, about 2.60 Å, is shorter than O6-O6 distances to atoms in either the same or adjacent quartets. The guanines in each quartet are tilted away from a common plane and towards the Sr\(^{2+}\) cation; a cation occurs only between every other quartet. This alternating distribution, which allows a separation of 6.43 Å between adjacent cations appears to be unique to Sr\(^{2+}\), and accommodates their electrostatic repulsion [95]. As evident from the distribution of B factors summarized in Figure 3.8, cation coordination by G quartets yields a stable, well-ordered, part of the overall structure.

Assessment of the observed Sr\(^{2+}\) coordination by the bond valence method [13] yields values in good agreement with the positive charge. For example, for Sr 3002 in structure P1B with Sr-O6 distances of 2.61 and 2.62 Å to the enclosing G-quartets, the
Figure 3.5: Cation coordination by Guanidine and Uridine Quartets. (a) Coordination of a Sr$^{2+}$ cation, shown in green, between the eight guanosine O6 atoms of two G-quartets, shown in magenta. (b) Coordination of two alternate positions of an Na$^+$ cation, shown in green, by eight uracil O4 atoms shown in cyan and magenta. The two U-quartets comprise the 3’ end of separate R6,8 assemblies. Coordinates are from the refinement of data set P1B.

calculated value is 2.08. Here valency is calculated as

$$V_i = \sum_{\text{bonds}_{i,j}} \frac{d_0 - d_{ij}}{b}$$

where the bond-valence parameter $d_0$ for (Sr$^{2+}$, O) is 2.118 [12], $b$ is 0.37 [13] and distances include all the eight symmetry-related O6 atoms.

Further confirmation for placement of modeled Sr$^{2+}$ cations is provided by a phased anomalous difference map as shown in Figure 3.6. The map was computed by phaser [1] using its log-likelihood-gradient mode with phases from the full model and unmerged $F(\pm)$ values. Clear peaks in the anomalous signal are seen for all modeled Sr$^{2+}$ cations, shown in magenta, but for neither of the alternate locations of the central Na$^+$ cation, shown in red. Additional peaks occur at some, but not all, Mg$^{2+}$ ions, shown in green, and at a few backbone phosphorus atoms.

In contrast to the stability of the SR-coordinated G-quartets, the two 3’
Figure 3.6: Phased anomalous difference map. The R6_8 assembly from structure P1B is shown in gray and a phased anomalous difference map contoured at 3σ in blue. Modeled Sr^{2+}, Mg^{2+}, and Na^+ ions are shown in magenta, green and red respectively.
uridines exhibit considerably greater variability. As with the G-quartets, the U-quartets in both R6A\textsubscript{4} and R6B\textsubscript{4} are oriented away from a plane defined by the four ribose C1’ atoms. However in R6B\textsubscript{4} this displacement is towards the adjacent guanine quartet whereas in the R6A\textsubscript{4} the displacement points away from the guanines. Here the monovalent coordinating cation is Na\textsuperscript{+}, in two alternate locations, only one of which is near the U-quartet plane.

This arrangement can be clarified by examining the coordination geometry in Figure 3.5(b). The two terminal 3’ U-quartets provide the contacts that link adjacent R6\textsubscript{8} assemblies. The carbonyl O4 atoms of the R6B\textsubscript{4} U-quartet, colored in magenta, define the vertices of square of length 3.4 Å. The central Na\textsuperscript{+} is located approximately in the plane of this square at a distance of 2.41 Å from the carbonyl O4s. The O4 atoms of the R6A\textsubscript{4} U-quartet, colored cyan, define a similar square of side 3.43 Å. However, the coordinating central Na\textsuperscript{+} is placed considerably further above this plane, nearly half way between the two U-quartets with a 2.87 Å distance to the R6A\textsubscript{4} O4 and 3.03 Å to the R6B\textsubscript{4} O4 atoms. Refinement supports modeling the two Na\textsuperscript{+} positions as alternate locations of the same cation, with approximately equal occupancy, rather than as distinct cations.

Bond-valence calculations in this case provide little support for the modeled geometry. Using a bond-valence parameter, \(d_0\), of 1.8 for (Na\textsuperscript{+}, O) \cite{bond-valence}, \(b\) of 0.37 as above, distances of 2.87 and 3.03 for the out-of-plane location and 2.41 for the in-plane location and occupancies of 0.60 and 0.40, yields valences of 0.22 and 0.30 respectively. Since in this case it is quite likely that the coordinating cation is in fact Na\textsuperscript{+}, the discrepancy
underscores limitations of the bond-valence method in identifying \( \text{Na}^+ \) ions, a point recently emphasized by George Sheldrick. [110, 78].

3.7 Disorder

![Figure 3.7: Alternate conformation of a phosphate group in UGGGU structure P1A. Isocontours of a \( \sigma_A \)-weighted 2Fo-Fc density map contoured at 2.0\( \sigma \) are shown in blue, conformation A is shown in magenta and B in cyan.](image)

The magnitude of the electron density calculated at a given map site is mathematically linked to contributions from scattering detected throughout reciprocal space. Scattering is affected by a number of factors. Some are related to the geometry of data collection, such as the crystal to detector distance and associated scattering angle.
range. Other factors depend on properties of the crystal, such as the rigidity of the macromolecule and the regularity of the crystal lattice. Atomic displacements about a mean position attenuate scattering. The attenuation is expressed mathematically by a term known as the temperature-factor, and this term is a component of the total scattering modeled for each atom: [102]

\[ f^{R_{\text{iso}}} = f_{\text{iso}}^{\theta} T_{\text{s}} \]

where

\[ T_{\text{s}} = \exp \left( -\frac{B_{\text{iso}}|S|^2}{4} \right) \].

A complete definition of the terms used to model atomic scattering is available in most crystallography texts, for example [102], section 6.1.

The relatively low disorder evident in the R6_8 structure is consistent with its tight packing, low solvent content and high resolution. Panel (a) of Figure 3.8 shows the R6_8 assembly for data set P1B with non-hydrogen atoms colored by refined B-factor value. A histogram of the B-factor values is shown in panel (b).

As expected, the central core is more well-ordered and the molecular surface less so. The tightly-bound, Sr^{2+}-coordinated, G-quartets exhibit the least disorder. Starting from the 3’ end of R6B_4 towards R6A_4, median B-factor values for the eleven-atom base moiety of the eight G quartets are 3.53, 3.17, 3.64, 3.08, 3.15, 3.29, 3.55 and 4.65 Å^2, in close agreement throughout the structure with the possible exception of residue 1005 at the 3’ of R6A_4.
P1B B-factor Distribution

Figure 3.8: Isotropic B factor distribution for assembly R6_8 from dataset P1B. (a) Structure with atoms colored by B-factor value and (b) histogram of B-factor values with corresponding color codes.
Backbone atoms, with an overall median B-factor of 5.35 Å², are more disordered than bases with a median of 3.63 Å². Among backbone atoms, those protruding further into solvent are more disordered, for example the 5’ uracils of R6A₄ and the phosphate-group oxygens relative the phosphorus atom.

Once again, the two 3’ uridine tetrads show quite different profiles. The tetrad at the 3’ end of R6B₄ is the most disordered residue with a median of 9.25 Å² whereas the B-factors of the uridine tetrad at the 3’ end of R6A₄, with a median of 5.35 Å², are close to those of the adjacent G-quartet.

Another instance of disorder evident in the UGGGGU structures is an alternate conformation for the phosphate group that connects uridine 1006 to guanidine 1005 in chain R6B as shown in Figure 3.7. A similar alternate conformation was detected in the TGGGGT structure deposited as PDB entry 352D [94], but not in the previously published UGGGGU structure [25]. Uracil 1006 is the least ordered base in the R6₈ structure. However there was little evidence to support extending the alternate conformation past the phosphate group. As evident in Figure 3.7, there is clear support for the alternate locations of the phosphorus atoms, however density for the corresponding \( O1P \) and \( O2P \) oxygens only becomes visible at about the 1.5σ level and displacement of the \( O3' \) and \( O5' \) oxygens appears minimal.
3.8 Internal Cavity

An interesting and novel feature of the UGGGGU conformation we detected in data sets with a tetragonal crystal packing is an internal, solvent accessible, cavity formed at the 3’ end of strand R6_B as shown in Figure 3.9. The “roof” of the cavity, the region proximal to the exterior, is formed by the 3’ U quartet and the floor by the adjacent G quartet. The source of this feature is a novel conformation for the 3’ uridine of strand R6_B as shown in panel (a). A multiple superposition of two chains from structure P1B and four chains from the previously published UGGGGU structure, PDB entry 1JG8, is shown. Common conformations for the 3’ uridine are colored in red. The two outliers are chain 1JG8 chain ‘B’, colored gold, and the novel R6_B uridine conformation uncovered in our data sets, colored green. Multiple superposition was calculated with the program THESEUS [118].

Though voids occur in other UGGGGU conformations, including the orthorhombic packing reported here and the previously published tetragonal packing [25], they are not internal to the nucleotide assembly but form between coaxially stacked R6,S units.

The cavity is accessible to bulk solvent by an opening at the center of the 3’ U-quartet. The central axis of the cavity is occupied by Na⁺ cations coordinated by the uracil quartets as detailed in Figure 2.2. When a solvent-accessible surface is applied, as shown in panel (b) of Figure 3.9, the volume of the cavity shrinks significantly but does not vanish indicating there is room to support movement of bound waters. The structure of the waters detected within the cavity is discussed in Section 4.3.5.
Figure 3.9: Internal cavity of UGGGGU dataset P1B. (a) Six chain superposition using the first five residues in 5’-3’ order calculated with THESEUS [118]. All four chains from PDB entry 1JG8 and both chains from structure P1B are included. The 3’ uridines are colored for emphasis: 1JG8, chain B, in gold, P1B, chain R6_B in green, others in red. (b) View of the molecular surface emphasizing the cavity. The view is clipped so as to expose the internal cavity formed by the 3’ uridines of strand R6B, apparent at the top of the image. Visible fragments of the 3’ uridines are highlighted in green.
Chapter 4

Solvent Structure

This chapter reports the detailed structure of solvent in the UGGGGU data sets we collected. Solvent plays a key role in the structure of nucleic acids [4, 117]. Its detailed analysis can contribute insight into folding thermodynamics [116], enzyme mechanism [125, 73], and molecular recognition, including small-molecule binding crucial for drug design [83, 99]. Previous analyses of nucleic acids solvent structure have revealed a number of general patterns including charge-shielding about the sugar-phosphate backbone and placement of waters with longer occupancy times in inner channels and cavities such as the “spine of hydration” reported in the minor-groove of B-form DNA [28, 52].

The results we present are consistent with these patterns. In addition, our analysis gives additional insight into structural features specific to quadruplexes: strand-bridging waters hydrogen-bonded to the O2’ hydroxyl ribose contribute to the greater stability of RNA over DNA quadruplexes; and the pattern of solvent in the 3’ internal cavity discussed in the previous chapter is consistent with the previously reported
structure of the quadruplex-bound telomerase-inhibiting drug daunomycin [18].

Different levels of hydration can play a significant role in determining nucleotide conformations. Perhaps the best known example is the preference of B and A-form DNA folds for high and low humidity conditions respectively [10]. This effect has also been shown to be important in less common helical conformations, including triplexes and quadruplexes. In G-quadruplex structures, analyses under varying solvent composition and molecular crowding conditions have documented the fundamental role of solvent in folding and stability [77, 114, 85]. Whereas addition of co-solutes destabilizes the DNA B-form duplex, it stabilized quadruplex folding as measured by melting temperature experiments [114]. Thus the solvent observed in our data sets is likely to play a significant role in stabilizing the quadruplex folds. We collected all data sets used for this analysis at cryogenic temperature. The resulting decrease in thermal motion can be expected to have increased the number of detectable hydration sites.

Modeling of bound or bulk solvent differs in fundamental ways from modeling of polymer atoms in macromolecular structures; these differences are summarized in the initial section. Subsequent sections summarize the relative hydration of different atom types and the characteristics of different classes of bound waters.

The distribution of bound solvent about a molecular surface does not lend itself to concise summary. Tools such as ribbon diagrams and skeleton traces that help tame the complexity of macromolecular structure are not available for summarizing solvent placement and distribution. Thus much of this chapter is taken up with detailed descriptions of solvent placement and distance distribution viewed from two perspectives:
the hydration of surface-accessible atoms and the contact points of individual waters. Taken together, these summaries confirm generally accepted observations about solvent placement: bound waters satisfy unfulfilled hydrogen bonding potential and shield partial charge accumulation; grooves, pockets and internal cavities are more hydrated than smoother surface areas and water contacts between multimer assemblies can help stabilize the crystal lattice. However, the main value of this analysis is not only to support these generalizations but to provide insight into the specific hydration properties of the UGGGGU quadruplex structure, a task not previously undertaken.

4.1 What Does Bound “Water” Model?

X-ray crystallography can contribute much to an understanding of the composition and distribution of solvent near accessible macromolecular surfaces. However, as emphasized in a seminal review by Levitt and Park [62], the technique provides qualitatively different insights for solvent than for macromolecular polymers. These differences are partly due to technical issues, but, more importantly, to a fundamental difference in what is being observed.

Crystallographic analysis constructs an atom-based model of electron density congruent with the experimentally-derived measurement of diffraction intensity amplitudes. These data reflect electron density in a unit cell averaged over all cells in a crystal and over the time interval required for collecting a complete set of diffraction images, typically on the order of hours. The derived model assumes an essentially
static arrangement of atoms in the unit cell. In sharp contrast, NMR and molecular
dynamics simulations indicate that water molecules in the vicinity of macromolecular
structures are very mobile with residence times that range from 50ps for surface waters
in the neighborhood of non-polar atoms to $10^{-2}$s for waters buried in pockets in the
interior[62].

Notwithstanding this rapid exchange of equivalent molecules, bound waters
are routinely modeled in the hydration shells of macromolecules and their incorpora-
tion yields significant improvements in the agreement between the model and diffraction
data as measured by R-factor statistics [15, 51]. The implication of this apparent con-
tradiction is that bound waters in the derived model represent not specific oxygen atoms
statically resident at particular locations in the unit cell, but minima in the potential of
mean force in the vicinity of the macromolecular surface. We expect solvent to occupy
all free space surrounding a macromolecule, however a site identified as the location
of a bound water molecule is distinguished by a local minimum in the force potential
relative to a neighboring, unoccupied sites. At both sites, water molecules will be in
rapid exchange with bulk solvent. However, only at the former will exchanging waters
reoccupy a site with so little variation in position as to enable detection of the associated
electron density by diffraction.

Ideally, crystallographic models should incorporate rapid solvent exchange and
represent solvent as a probability distribution rather than as a collection of static point
scatterers. The unexpectedly high R and R-free values obtained in macromolecular re-
finement, relative to values expected from models for which discrepancies are limited
to errors in amplitude measurement, have been attributed in part to inadequate solvation models [15, 51]. However, at present, the only available tools for representing a probabilistically-smeared distribution of preferred binding sites in hydration shells are simple correction factors for the volume occupied by bulk solvent and variation in the partial occupancy and crystallographic B factor parameters of individually-modeled scattering sites.

A second, more technical, consideration in crystallographically-derived solvent models is difficulty in verifying the identity of scattering sites. Water and ions are modeled at peaks that meet empirically-derived assignment conventions including charge compatibility, coordination geometry, distance and hydrogen bonding plausibility. For example, a typical criterion for interpreting a candidate peak as a water molecule is a peak height of at least $3\sigma$ in the $\sigma_A$-weighted $2m\text{FO}-\text{DFc}$ the map, a distance between 2.4 and 4.1 Å of a previously modeled atom, and an occupancy of 1 (unless placed on a special position)[5]. While helpful in distinguishing spurious density peaks from credible sites for bound solvent, these criteria are much weaker restraints that the distance, angle, planarity, and rotameric “prior information” available for modeling nucleic acid bases and protein residues.

Furthermore, even for sites that meet the above criteria, atom type assignment remains a challenge. Simulation studies indicate that sodium ions can occupy hydration sites traditionally modeled as waters in the B-DNA hydration spine [133]. Distinguishing water molecules from isoelectronic Na$^+$ and Mg$^{2+}$ cations is difficult. Occasionally, the identity of metal cations can be confirmed by anomalous difference maps. For the
data we used in this study, $\text{Sr}^{2+}$ cations can be reliably detected as peaks in phased anomalous difference maps as shown in Figure 3.6. However this is not true for lighter elements, for example even P atoms only appear sporadically. For MAD data sets with strong dispersive signal, density maps can be constructed that are free of experimental bias and of quality sufficient to identify solvent [15, 67]. This approach provides strong validation of conventionally modeled bound water sites, however it does not address the limitations of atom type assignment.

### 4.2 Classifying Bound Waters

On first inspection, detectable solvent reported alongside a macromolecular model can resemble a featureless collection of water molecules randomly distributed over the molecular surface, as shown in Figure 4.1 However, with more detailed analysis, distinct patterns emerge that can help distinguish different classes of bound waters and give insight into which regions of the surface yield binding sites specific enough to permit crystallographically-detectable waters.

The principal tools for detailed analysis are the surface atom contacts list of each modeled water molecule and the list of solvent molecules within a given distance of each surface atom. Bounds for the distance range are typically set at 2.2 Å and 3.5 Å, distances that correspond to experimentally-determined limits of coordination and hydrogen bonding.

Thus, to identify patterns of hydration, it is helpful to consider associations
from two perspectives: what parts of the macromolecule does an individual water make contact with, and which waters does each constituent atom contact. In general, this will be a many-to-many relation. The following two subsections review these associations. The first considers how hydration varies across different atom types. Section 4.3 classifies waters on the basis of the surface atoms they contact. We identify subsets of bound waters with similar association patterns and compare their properties.

### 4.2.1 Hydration Variation by Atom Type

Tables 4.1-4.3 detail hydration of atoms in the P1A model; separate tables describe the guanine and uracil bases and the sugar-phosphate backbone. For each atom type, the contacts between instances of that type and water molecules are listed.
Individual contacts are identified by their residue and water molecule identifiers. Summaries for each atom type show the total occurrences, the fraction of atoms with solvent contact and the average distance, with standard error.

The list is ordered by average hydration for that water type, that is the ratio of total water to atom occurrences. A cutoff of 3.54 Å was used in screening for water contacts. Only contacts to hydrogens included in the model are included, thus no hydroxyl hydrogens are listed.

Inspection of the table reveals that among guanines, the most hydrated atoms are those facing the outside of the G-quartet: the N2 and N3 nitrogens along with the N2H2 hydrogen, and to a lesser extent, the C2 carbon. Most other guanine atoms exhibit either no hydration or minimal hydration. These include the O6 carbonyl group in contact with the dehydrated cation in the central channel, the atoms involved in G-quartet hydrogen bonding, N7, N2H21 and N1, as well as atoms buried by other parts of the nucleotide and not exposed to solvent including N9 and C8, along with those unavailable for hydrogen bonding such as C4 and C5. The distinction is particularly sharp for the two hydrogen atoms bonded to the N2 nitrogen: no water molecules contact HN21, whereas HN22 is the most hydrated guanine atom.
Table 4.1: Hydration of UGGGU Guanine Atoms

<table>
<thead>
<tr>
<th>Atom Name</th>
<th>Contacts: (Residue Id., Water Id., Distance)</th>
<th>Hydration</th>
<th>Average Distance (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>N9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HN21</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C5</td>
<td>(2005,5002,3.4)</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>O6</td>
<td>(2005,5002,3.1)</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>(2005,5002,3.1)</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>H8</td>
<td>(1003,5022,3.5)</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>HN1</td>
<td>(2005,5002,3.5)</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>(2005,5042,3.4)</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>N2</td>
<td>(1005,5014,3.2)(2004,5063,3.4)(2005,5028,3.0)</td>
<td>3/8(0.38)</td>
<td>3.2(0.12)</td>
</tr>
<tr>
<td>N3</td>
<td>(1004,5022,2.8)(1005,5014,3.1)(1005,5023,3.5)(2004,5026,2.9)</td>
<td>6/8(0.75)</td>
<td>3.1(0.11)</td>
</tr>
<tr>
<td>HN22</td>
<td>(1004,5016,2.7)(1004,5022,2.9)(1005,5014,2.4)(2004,5026,3.0)</td>
<td>7/8(0.88)</td>
<td>2.7(0.15)</td>
</tr>
<tr>
<td></td>
<td>(2004,5063,2.5)(2005,5007,3.1)(2005,5028,2.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atom Name</td>
<td>Contacts: (Residue Id., Water Id., Distance)</td>
<td>Hydration</td>
<td>Average Distance (s.e.)</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------------------</td>
<td>-----------</td>
<td>------------------------</td>
</tr>
<tr>
<td>C4</td>
<td>(1001,5061,3.0)(2006,5006,3.2)</td>
<td>2/4(0.50)</td>
<td>3.1(0.10)</td>
</tr>
<tr>
<td>C2</td>
<td>(2006,5006,2.6)(2006,5017,2.6)</td>
<td>2/4(0.50)</td>
<td>2.6(0.00)</td>
</tr>
<tr>
<td>N3</td>
<td>(2006,5006,2.8)(2006,5017,3.4)</td>
<td>2/4(0.50)</td>
<td>3.1(0.30)</td>
</tr>
<tr>
<td>O2</td>
<td>(1006,5023,2.8)(2006,5006,3.2)(2006,5017,2.2)</td>
<td>3/4(0.75)</td>
<td>2.7(0.29)</td>
</tr>
<tr>
<td>C6</td>
<td>(2001,5020,3.5)(2006,5006,3.1)(2006,5040,2.6)</td>
<td>3/4(0.75)</td>
<td>3.1(0.26)</td>
</tr>
<tr>
<td>N1</td>
<td>(2006,5006,2.7)(2006,5017,3.3)(2006,5040,2.7)</td>
<td>3/4(0.75)</td>
<td>2.9(0.20)</td>
</tr>
<tr>
<td>C5</td>
<td>(1001,5019,3.5)(1001,5061,3.4)(2006,5006,3.3)(2006,5040,3.4)</td>
<td>4/4(1.00)</td>
<td>3.4(0.04)</td>
</tr>
<tr>
<td>H3</td>
<td>(1001,5026,3.5)(2001,5022,3.5)(2006,5006,3.4)(2006,5017,3.5)</td>
<td>4/4(1.00)</td>
<td>3.5(0.03)</td>
</tr>
<tr>
<td>H6</td>
<td>(1001,5019,3.2)(1001,5019,3.3)(2001,5020,2.9)(2006,5040,2.5)</td>
<td>5/4(1.25)</td>
<td>3.1(0.16)</td>
</tr>
<tr>
<td>O4</td>
<td>(1001,5001,2.7)(1001,5061,2.4)(2001,5015,3.4)(2001,5032,2.9)</td>
<td>6/4(1.50)</td>
<td>2.8(0.16)</td>
</tr>
<tr>
<td>H5</td>
<td>(1001,5019,2.8)(1001,5019,3.2)(1001,5061,3.2)(2001,5005,3.3)</td>
<td>7/4(1.75)</td>
<td>3.1(0.09)</td>
</tr>
</tbody>
</table>
Table 4.3: Hydration of UGGGGU Ribose/Phosphate Atoms

<table>
<thead>
<tr>
<th>Atom Name</th>
<th>Contacts: (Residue Id., Water Id., Distance)</th>
<th>Hydration</th>
<th>Average Distance (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3</td>
<td>(1006,5012,3.2)(2006,5040,2.4)</td>
<td>2/12(0.17)</td>
<td>2.8(0.40)</td>
</tr>
<tr>
<td>P</td>
<td>(2002,5061,3.5)(2004,5051,3.4)</td>
<td>2/11(0.18)</td>
<td>3.5(0.05)</td>
</tr>
<tr>
<td>C5</td>
<td>(1001,5021,3.5)(1001,5046,3.5)(1003,5037,3.5)(2006,5044,3.2)</td>
<td>4/12(0.33)</td>
<td>3.4(0.08)</td>
</tr>
<tr>
<td>C4</td>
<td>(1003,5031,3.5)(1006,5012,3.3)(2003,5039,3.5)(2005,5009,3.5)(2006,5044,3.0)</td>
<td>5/12(0.42)</td>
<td>3.4(0.10)</td>
</tr>
<tr>
<td>O5</td>
<td>(1001,5019,3.3)(1001,5019,3.4)(1001,5021,2.7)(1001,5046,2.8)(2001,5020,2.7)(2004,5028,3.4)</td>
<td>6/13(0.46)</td>
<td>3.0(0.14)</td>
</tr>
<tr>
<td>O4</td>
<td>(1002,5039,3.2)(1003,5031,2.9)(2002,5031,3.2)(2003,5039,2.9)(2005,5026,3.4)(2006,5040,3.1)</td>
<td>6/12(0.50)</td>
<td>3.1(0.08)</td>
</tr>
<tr>
<td>C1</td>
<td>(1002,5039,3.3)(1004,5022,3.4)(1005,5016,3.4)(2002,5031,3.2)(2004,5026,3.4)(2005,5063,3.2)(2006,5040,2.7)</td>
<td>7/12(0.58)</td>
<td>3.2(0.09)</td>
</tr>
<tr>
<td>O3</td>
<td>(1004,5008,2.9)(1006,5010,2.5)(1006,5012,2.7)(1006,5029,3.4)(1006,5030,3.0)(2002,5001,3.3)(2002,5013,3.0)(2004,5061,3.4)(2006,5028,3.2)(2006,5040,3.3)</td>
<td>10/13(0.77)</td>
<td>3.1(0.10)</td>
</tr>
<tr>
<td>H3</td>
<td>(1001,5046,3.3)(1003,5048,3.2)(1003,5055,2.7)(1006,5012,3.1)(1006,5029,2.8)(2001,5020,3.0)(2003,5025,3.2)(2004,5028,3.1)(2006,5040,1.9)(2006,5041,3.1)</td>
<td>10/12(0.83)</td>
<td>2.9(0.13)</td>
</tr>
</tbody>
</table>

Continued on next page...
Table 4.3 – Continued

<table>
<thead>
<tr>
<th>Atom Name</th>
<th>Contacts: (Residue Id., Water Id., Distance)</th>
<th>Hydration Average Distance (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>(1002,5039,2.6)(1004,5022,2.8)(1005,5014,3.5)(1005,5016,2.6)</td>
<td>11/12(0.92) 2.9(0.12)</td>
</tr>
<tr>
<td></td>
<td>(1005,5023,3.5)(1006,5032,2.8)(2002,5031,2.6)(2004,5026,2.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2005,5007,3.2)(2005,5063,2.3)(2006,5044,3.4)</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>(1002,5031,3.3)(1003,5016,3.4)(1004,5022,3.3)(1005,5023,3.4)</td>
<td>14/12(1.17) 3.3(0.07)</td>
</tr>
<tr>
<td></td>
<td>(2006,5035,3.2)(2006,5040,2.4)</td>
<td></td>
</tr>
<tr>
<td>H51</td>
<td>(1001,5021,3.1)(1002,5015,2.9)(1003,5003,3.1)(1003,5031,3.0)</td>
<td>16/12(1.33) 3.1(0.05)</td>
</tr>
<tr>
<td></td>
<td>(1005,5008,2.9)(1006,5044,3.5)(2002,5001,3.1)(2002,5061,3.2)</td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>(1001,5004,3.4)(1001,5039,2.8)(1003,5003,3.0)(1003,5031,3.4)</td>
<td>17/12(1.42) 3.0(0.09)</td>
</tr>
<tr>
<td></td>
<td>(1003,5037,3.0)(1004,5005,3.4)(1004,5008,3.3)(1006,5012,2.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2006,5044,2.1)</td>
<td></td>
</tr>
<tr>
<td>H2</td>
<td>(1002,5031,3.0)(1003,5016,3.0)(1003,5022,3.2)(1004,5022,3.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1005,5023,3.1)(2002,5039,2.9)(2003,5025,3.5)(2003,5026,3.4)</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page...
Table 4.3 – Continued

<table>
<thead>
<tr>
<th>Atom Name</th>
<th>Contacts: (Residue Id., Water Id., Distance)</th>
<th>Hydration</th>
<th>Average Distance (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H52</td>
<td>(2005,5017,2.6)(2006,5006,2.6)(2006,5007,2.5)(2006,5017,2.8) (2006,5035,3.2)(2006,5040,2.0)</td>
<td>18/12(1.50)</td>
<td>3.0(0.09)</td>
</tr>
<tr>
<td></td>
<td>(1001,5004,3.2)(1001,5046,3.1)(1003,5037,2.9)(1003,5055,3.2) (1004,5005,2.8)(1004,5012,2.8)(1004,5029,3.3)(1005,5062,3.3) (1005,5062,3.5)(1006,5005,2.9)(1006,5011,3.0)(1006,5012,3.3) (2001,5003,3.0)(2004,5019,3.2)(2004,5021,3.0)(2004,5051,3.2) (2005,5020,3.0)(2006,5020,3.3)(2006,5044,3.3)</td>
<td>19/12(1.58)</td>
<td>3.1(0.04)</td>
</tr>
<tr>
<td>O2P</td>
<td>(1002,5027,3.2)(1002,5033,3.1)(1002,5046,3.0)(1003,5022,2.7) (1003,5032,3.5)(1004,5014,2.9)(1004,5016,3.5)(1004,5024,3.0) (1004,5048,2.7)(1005,5024,2.7)(1005,5029,3.5)(1006,5044,2.6) (1006,5005,3.3)(2002,5020,3.3)(2002,5027,3.3)(2003,5026,2.7) (2004,5028,2.9)(2004,5047,2.7)(2004,5063,3.0)(2005,5028,2.8) (2006,5012,2.6)</td>
<td>21/11(1.91)</td>
<td>3.0(0.07)</td>
</tr>
</tbody>
</table>

Continued on next page...
Table 4.3 – Continued

<table>
<thead>
<tr>
<th>Atom Name</th>
<th>Contacts: (Residue Id., Water Id., Distance)</th>
<th>Hydration</th>
<th>Average Distance (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1P</td>
<td>(1002,5015,3.0)(1002,5054,2.9)(1003,5032,2.9)(1003,5055,3.2)</td>
<td>25/11(2.27)</td>
<td>3.0(0.05)</td>
</tr>
<tr>
<td></td>
<td>(1004,5010,3.3)(1004,5012,2.7)(1004,5038,3.0)(1005,5008,3.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1005,5029,3.0)(1005,5029,3.2)(1005,5062,3.0)(1005,5062,3.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1006,5011,2.9)(1006,5044,3.0)(2002,5001,2.9)(2002,5009,2.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2006,5017,2.6)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overall, uracil atoms are more hydrated than those of guanine, a difference that is consistent with the different role of the two bases in the UGGGGU structure, though different patterns are evident for the 5' and 3' residues. For both strands, the 5' uracil is external to the structure and well-exposed to solvent. Thus the H5 hydrogen and O4 carbonyl of the 5' residues, 1001 and 2001, are the most hydrated uracil atoms. The uracil quartet structures formed at the 3' end of the R6A and R6B strands have different conformations, but in both cases are coordinated by Na$^+$ rather than Sr$^{2+}$ cations, enabling closer water contacts. The two conformations result in sharply different hydration patterns for the 3' uracils. The 3' R6B strand uracil, residue 2006, participates in formation of the interior cavity discussed in Section 3.8, and most of its atoms, C2, C4, C6, O2,N1, C5 as well as the H3 and H6 hydrogens, participate in contacts to waters enclosed by the cavity. This list only excludes the O4 and H5 atoms facing the Na$^+$ cation. In contrast, the U quartet formed at the 3' end of strand R6A includes a single water contact, by water 5023, at atom O2 that is directed towards the external surface.

Atoms along the ribose-phosphate backbone are generally closer to the surface and thus significantly more hydrated than base atoms. The O1P and O2P oxygens of the phosphate group show high hydration values; only one instance of this atom type, one of the alternate conformations at residue 1006, does not exhibit a water contact at both O1P and O2P atoms. However, the distribution of water contacts is not symmetric. Whereas for O2P five residues are limited to a single water contact, 2006, 2005, 2003 and both conformations of residue 1006, for O1P all residues but 2006 are involved in two
or more contacts. The greater solvent exposure of O1P enables formation of multiple hydrogen bonds.

Another heavily-hydrated atom group is the ribose $O2'$ hydroxyl. All twelve instances of this atom type exhibit at least one water contact. As discussed below, this characteristic may account for the greater stability of RNA over DNA quadruplexes. Interestingly, there is little variation among the atom-water distances of the most hydrated ribose atoms: the three most hydrated atoms show the same 3.9 Å average distance with standard errors of 0.07, 0.07 and 0.05.

### 4.3 Bound Waters with Similar Association Patterns

To further explore the distribution of ordered solvent, it is helpful to group waters into similar classes on the basis of their contacts list. Such classes are not disjoint as individual waters typically associate with multiple atom types. Some classes are based on similarity of location rather than of atom type association and partitioning can be somewhat subjective for marginal cases. Nevertheless, distinct patterns of bound waters are apparent as detailed below. The six classes we considered are:

**Groove:** Waters that make up the spine of hydration occupying the grooves formed by the quadruplex right-handed helical twist.

**Phosphate-Shielding:** Waters bound to phosphate-group oxygens, $O1P$, $O2P$, $O3'$ and $O5'$, along the polymer backbone.

**Ribose-hydroxyl:** Waters bound to the ribose $O2'$ or its associated hydrogen.

85
**Interstitial Crystal Contacts:** Waters that appear to be principally involved in facilitating contacts among assemblies in neighboring unit cells.

**Cavity Bound:** Waters that occupy the large, solvent-accessible cavity formed by the quartet of 3’ uridines of strand R6_B.

**Cation-Coordinating:** Waters that play a role in coordinating the Na\(^+\), Ca\(^{2+}\) or Mg\(^{2+}\) cations detected in the solvent.

The subsets are not disjoint, nor do they include all modeled waters. For example, many groove-bound waters also include phosphate oxygens among their contacts and various additional classifications could be considered, such as waters that make contacts to other ribose atoms, including C4’ and C5’, or to solvent cations.

Table 4.4 summarizes contacts of all water molecules modeled in data set P1A. For each water, the number of contacts to different classes of surface atoms is shown, along with the average distance and standard error. The distance to the molecular surface is shown in the final column.
Table 4.4: Surface Atom Contacts of Modeled Waters

<table>
<thead>
<tr>
<th>Residue</th>
<th>B Factor</th>
<th>Polymer</th>
<th>Water</th>
<th>Cation</th>
<th>Total</th>
<th>Surface Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5017</td>
<td>14.13</td>
<td>11 3.01(0.14)</td>
<td>2 2.75(0.45)</td>
<td>0</td>
<td>13 2.97(0.13)</td>
<td>0.12</td>
</tr>
<tr>
<td>5040</td>
<td>17.46</td>
<td>11 2.64(0.15)</td>
<td>3 2.67(0.42)</td>
<td>0</td>
<td>14 2.64(0.14)</td>
<td>0.22</td>
</tr>
<tr>
<td>5039</td>
<td>11.69</td>
<td>13 3.09(0.08)</td>
<td>1 3.20</td>
<td>0</td>
<td>14 3.10(0.08)</td>
<td>0.25</td>
</tr>
<tr>
<td>5063</td>
<td>46.43</td>
<td>9 2.97(0.13)</td>
<td>3 2.80(0.17)</td>
<td>0</td>
<td>12 2.93(0.11)</td>
<td>0.61</td>
</tr>
<tr>
<td>5031</td>
<td>10.59</td>
<td>13 3.09(0.08)</td>
<td>1 3.10</td>
<td>0</td>
<td>14 3.09(0.08)</td>
<td>0.74</td>
</tr>
<tr>
<td>5026</td>
<td>8.44</td>
<td>11 3.13(0.09)</td>
<td>2 2.90(0.20)</td>
<td>0</td>
<td>13 3.09(0.08)</td>
<td>0.96</td>
</tr>
<tr>
<td>5007</td>
<td>27.12</td>
<td>9 3.07(0.08)</td>
<td>3 2.77(0.30)</td>
<td>0</td>
<td>12 2.99(0.09)</td>
<td>0.98</td>
</tr>
<tr>
<td>5006</td>
<td>17.01</td>
<td>11 3.05(0.10)</td>
<td>4 2.38(0.05)</td>
<td>0</td>
<td>15 2.87(0.11)</td>
<td>1.02</td>
</tr>
<tr>
<td>5028</td>
<td>11.78</td>
<td>7 2.91(0.17)</td>
<td>2 2.90(0.00)</td>
<td>0</td>
<td>9 2.91(0.13)</td>
<td>1.03</td>
</tr>
<tr>
<td>5022</td>
<td>11.75</td>
<td>11 3.12(0.09)</td>
<td>1 2.70</td>
<td>0</td>
<td>12 3.08(0.09)</td>
<td>1.07</td>
</tr>
<tr>
<td>5061</td>
<td>19.83</td>
<td>11 3.09(0.10)</td>
<td>1 3.00</td>
<td>1 2.30</td>
<td>13 3.02(0.10)</td>
<td>1.08</td>
</tr>
<tr>
<td>5035</td>
<td>18.17</td>
<td>3 2.93(0.27)</td>
<td>2 2.55(0.35)</td>
<td>0</td>
<td>5 2.78(0.21)</td>
<td>1.11</td>
</tr>
<tr>
<td>5023</td>
<td>14.11</td>
<td>8 3.19(0.11)</td>
<td>2 3.05(0.15)</td>
<td>1 2.70</td>
<td>11 3.12(0.10)</td>
<td>1.24</td>
</tr>
<tr>
<td>5016</td>
<td>8.00</td>
<td>7 3.09(0.14)</td>
<td>2 2.90(0.20)</td>
<td>0</td>
<td>9 3.04(0.11)</td>
<td>1.26</td>
</tr>
<tr>
<td>5012</td>
<td>10.61</td>
<td>9 2.93(0.10)</td>
<td>2 3.00(0.20)</td>
<td>0</td>
<td>11 2.95(0.08)</td>
<td>1.37</td>
</tr>
<tr>
<td>5043</td>
<td>11.14</td>
<td>1 2.40</td>
<td>3 2.67(0.33)</td>
<td>0</td>
<td>4 2.60(0.24)</td>
<td>1.38</td>
</tr>
<tr>
<td>5001</td>
<td>11.71</td>
<td>7 3.20(0.12)</td>
<td>2 3.20(0.20)</td>
<td>1 2.20</td>
<td>10 3.10(0.13)</td>
<td>1.40</td>
</tr>
<tr>
<td>5044</td>
<td>37.71</td>
<td>10 2.96(0.13)</td>
<td>1 3.10</td>
<td>0</td>
<td>11 2.97(0.12)</td>
<td>1.41</td>
</tr>
<tr>
<td>5021</td>
<td>21.83</td>
<td>6 3.08(0.12)</td>
<td>1 2.50</td>
<td>0</td>
<td>7 3.00(0.13)</td>
<td>1.43</td>
</tr>
<tr>
<td>5024</td>
<td>12.67</td>
<td>3 3.03(0.20)</td>
<td>2 2.65(0.25)</td>
<td>1 2.40</td>
<td>6 2.80(0.16)</td>
<td>1.45</td>
</tr>
<tr>
<td>5032</td>
<td>13.10</td>
<td>3 3.10(0.20)</td>
<td>2 2.95(0.45)</td>
<td>0</td>
<td>5 3.04(0.18)</td>
<td>1.45</td>
</tr>
<tr>
<td>5051</td>
<td>56.41</td>
<td>3 3.10(0.21)</td>
<td>1 2.40</td>
<td>0</td>
<td>4 2.93(0.23)</td>
<td>1.46</td>
</tr>
<tr>
<td>5014</td>
<td>13.95</td>
<td>6 3.05(0.15)</td>
<td>3 3.07(0.09)</td>
<td>1 2.70</td>
<td>10 3.02(0.10)</td>
<td>1.49</td>
</tr>
<tr>
<td>5048</td>
<td>46.92</td>
<td>2 2.95(0.25)</td>
<td>1 2.60</td>
<td>0</td>
<td>3 2.83(0.19)</td>
<td>1.49</td>
</tr>
<tr>
<td>5046</td>
<td>33.95</td>
<td>5 3.14(0.12)</td>
<td>3 3.20(0.10)</td>
<td>0</td>
<td>8 3.16(0.08)</td>
<td>1.53</td>
</tr>
<tr>
<td>5054</td>
<td>38.42</td>
<td>1 2.90</td>
<td>3 3.13(0.12)</td>
<td>0</td>
<td>4 3.08(0.10)</td>
<td>1.54</td>
</tr>
<tr>
<td>5002</td>
<td>13.71</td>
<td>4 3.27(0.10)</td>
<td>5 2.84(0.26)</td>
<td>8 2.80(0.15)</td>
<td>17 2.92(0.11)</td>
<td>1.57</td>
</tr>
<tr>
<td>5009</td>
<td>18.43</td>
<td>4 3.03(0.20)</td>
<td>3 3.07(0.18)</td>
<td>0</td>
<td>7 3.04(0.13)</td>
<td>1.58</td>
</tr>
</tbody>
</table>

Continued on next page...
<table>
<thead>
<tr>
<th>Residue</th>
<th>B Factor</th>
<th>Contacts</th>
<th>Surface Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Polymer</td>
<td>Water</td>
</tr>
<tr>
<td>5013</td>
<td>14.77</td>
<td>4 3.08(0.07)</td>
<td>1 2.80</td>
</tr>
<tr>
<td>5008</td>
<td>17.41</td>
<td>5 3.14(0.12)</td>
<td>3 3.37(0.07)</td>
</tr>
<tr>
<td>5042</td>
<td>15.91</td>
<td>1 3.40</td>
<td>5 2.42(0.13)</td>
</tr>
<tr>
<td>5020</td>
<td>13.03</td>
<td>10 3.15(0.10)</td>
<td>2 3.00(0.30)</td>
</tr>
<tr>
<td>5062</td>
<td>30.80</td>
<td>4 3.25(0.10)</td>
<td>2 1.90(1.60)</td>
</tr>
<tr>
<td>5015</td>
<td>29.85</td>
<td>3 3.10(0.15)</td>
<td>2 2.95(0.25)</td>
</tr>
<tr>
<td>5011</td>
<td>22.96</td>
<td>4 2.92(0.09)</td>
<td>4 3.03(0.12)</td>
</tr>
<tr>
<td>5019</td>
<td>28.54</td>
<td>8 3.24(0.07)</td>
<td>3 2.80(0.15)</td>
</tr>
<tr>
<td>5037</td>
<td>25.13</td>
<td>3 3.13(0.19)</td>
<td>2 2.70(0.10)</td>
</tr>
<tr>
<td>5038</td>
<td>46.47</td>
<td>1 3.00</td>
<td>2 2.80(0.20)</td>
</tr>
<tr>
<td>5003</td>
<td>13.48</td>
<td>4 3.08(0.05)</td>
<td>3 2.97(0.09)</td>
</tr>
<tr>
<td>5050</td>
<td>35.47</td>
<td>2 3.00(0.00)</td>
<td>2 2.60(0.40)</td>
</tr>
<tr>
<td>5030</td>
<td>19.46</td>
<td>1 3.00</td>
<td>3 2.93(0.03)</td>
</tr>
<tr>
<td>5055</td>
<td>37.59</td>
<td>3 3.03(0.17)</td>
<td>2 3.00(0.40)</td>
</tr>
<tr>
<td>5027</td>
<td>13.11</td>
<td>3 3.23(0.03)</td>
<td>3 2.93(0.23)</td>
</tr>
<tr>
<td>5041</td>
<td>17.08</td>
<td>2 3.25(0.15)</td>
<td>4 2.67(0.28)</td>
</tr>
<tr>
<td>5029</td>
<td>16.23</td>
<td>6 3.20(0.11)</td>
<td>3 3.17(0.03)</td>
</tr>
<tr>
<td>5033</td>
<td>17.68</td>
<td>2 3.15(0.05)</td>
<td>4 3.00(0.23)</td>
</tr>
<tr>
<td>5005</td>
<td>16.07</td>
<td>5 3.14(0.12)</td>
<td>4 2.97(0.14)</td>
</tr>
<tr>
<td>5010</td>
<td>59.96</td>
<td>2 2.90(0.40)</td>
<td>3 2.83(0.23)</td>
</tr>
<tr>
<td>5025</td>
<td>21.62</td>
<td>2 3.35(0.15)</td>
<td>5 2.92(0.20)</td>
</tr>
<tr>
<td>5004</td>
<td>14.35</td>
<td>4 3.18(0.09)</td>
<td>1 3.20</td>
</tr>
<tr>
<td>5047</td>
<td>28.88</td>
<td>1 2.70</td>
<td>2 2.70(0.10)</td>
</tr>
</tbody>
</table>

88
4.3.1 Groove-bound Waters

The structure of waters bound within the grooves formed between nucleic acid strands wrapped in helical conformations has received considerable attention in hydration studies. Multiple reports have confirmed the presence of a network of well-ordered waters with a specific hydrogen-bonding geometry termed the spine of hydration in the minor groove of B-form DNA as well in other helix-induced grooves [10].

In B-form DNA, \( \pi \)-stacking between rungs of the ladder-like helix excludes water between planar stacked bases. Furthermore, Watson-Crick hydrogen bonding excludes base atoms on the Watson-Crick faces from participating in bonds with water, though atoms on the edges, including the pyrimidine O2 and purine O6, may participate in bifurcated hydrogen bonds. Thus candidates for water-base hydrogen-bonding sites include the atoms on the sides or edge of the Watson-Crick face, specifically the purine N3, N2 and pyrimidine O2 atoms on the minor-groove side and, on the major-groove side, the guanine N7 and O6 atoms, adenine N7 and N6-H, thymine O4 and cytosine N4-H [117]. It follows that the pattern of contacts among the hydration spine waters is sequence dependent only on the major-groove side [52]. Waters that make contact to bases at these sites comprise the spine’s primary hydration layer. Additional waters that form bridging contacts among the primary layer constitute a secondary hydration layer.

This pattern is illustrated in Figure 4.2 in the context of the 1.8Å-resolution structure of Dickerson-Drew dodecamer d(CGCGAATTCGCG), PDB entry 1EHV. An
idealized outline of the minor-groove hydration pattern is shown schematically in panel (a) and a detailed map of the hydration sites observed in PDB entry 1EHV [52] in panel (b).

Figure 4.2: Hydration spine in the minor groove of B-form DNA. (a) Idealized schematic of sequence-independent contacts in B-form DNA minor-groove hydration, image from [22] (b) Map of observed hydration contacts, PDB entry 1EHV, image from [52].

These general themes, with water molecules forming hydrogen bonds to base atoms not engaged in essential structural base-pairing, are repeated in other right-handed helices, including A and Z-form helices as well as TGGGGT and UGGGGU quadruplexes. In G4 structures, essential structural hydrogen-bond pairs occur between the O6 and N1 atoms and between N2 and N7 of adjacent guanines within each quartet.
Figure 4.3 illustrates the spine of hydration in the TGGGGT quadruplex. The floor of the groove is defined by the guanine N3 and N2-H atoms of one G-quartet and the C8-H atom of the neighboring guanine of an adjacent quartet [23], as depicted in panel (b). Here the contribution to the molecular contact surface from the C8 atom is colored green and the contributions from the N3 and N2 atoms are colored red and magenta respectively. A representative row of four waters, residues 527, 284, 194 and 210, which make contacts near the floor of the groove are shown in panel (a).

Figure 4.3: Hydration spine in a DNA, parallel-stranded, G-quadruplex, PDB entry 352D [94]. (a) Waters 527, 284, 194 and 210 are shown. Atoms that define the floor of the groove near water 284, the N7 and and C8 atoms of guanine residue 72, chain H and the N3 atom of guanine residue 43, chain E, are shown in magenta. Multiple conformation of the phosphate linkage at residue 72 is shown in upper left. (b) Detail surrounding water 527. Contributions to the molecular surface of the groove floor by the N3 and N2 atoms of residue 43, chain E, and the C8 atom of residue 72 chain H are shown in red, magenta and green respectively.
4.3.2 Phosphate-bound Waters

Waters that include any of the phosphate group atoms, $O_1P$, $O_2P$, $O_3'$, $O_5'$, or $P$, among their contacts list, were classified in the subset of phosphate-bound waters. The distribution of this subset over the molecular surface is shown in panel (a) of Figure 4.4 and a detailed view in the vicinity of residue 1004 is shown in panel (b).

Inspection reveals that all phosphate groups are hydrated, as expected given the role of waters in shielding the net negative charge of the phosphate group [10]. Oxygens that protrude further from the surface, principally $O_1P$, exhibit more frequent water contacts. Though most contacts are one-to-one, phosphate-bound waters can bridge two oxygens, within the same or between adjacent phosphate groups, and as many as three waters can make contact with a given oxygen. The chain-terminating $O_3'$ and $O_5'$ oxygens are more shielded and have less water contact with the exception of the 5’ oxygen of chain R6A, residue 1001, which protrudes into solvent to make contact with three waters.

4.3.3 Ribose O2’-bound Waters

The ribose $O_2'$ hydroxyl is another well-hydrated group in the UGGGGU structure; nearly all instances of this group show at least one contact. This finding is consistent with previous analyses of RNA hydration [10, 117]. Figure 4.5 shows the overall pattern of hydration in panel (a) and a detail of a three-strand bridging network near residues 1002 and 2002 in panel (b). The color scheme is that described for Figure 4.4.
Figure 4.4: Phosphate-bound waters in UGGGGU data set P1A. (a) The distribution of waters in the asymmetric unit in contact with the ribose-phosphate backbone is shown over the symmetry-expanded molecular surface. Waters are shown in red, phosphate groups in magenta. Black lines show contacts within hydrogen-bonding distance. (b) A detailed view showing multiple waters in the hydration sphere of solvent-exposed oxygens and bridging contacts across and within phosphate groups. Phosphorus atoms are shown in magenta and O1P, O2P, O3′, O5′ oxygens are shown in yellow, purple, orange and green respectively.
Hydrogen bonding associated with the 2’-OH group has been suggested to account for greater rigidity of RNA A-duplexes [10]. Experimental data indicates RNA quadruplexes are more stable than DNA [75] and cross-strand hydrogen bonding networks such as that shown in panel (b) are likely to contribute to this effect.

### 4.3.4 Interstitial Crystal-contact Waters

One of the functions of water-mediated hydrogen bonds in the environment of a crystal is to stabilize packing contacts, enabling the formation of a regular lattice. Such interactions are generally considered to be specific to a given crystal form and thus to have less applicability to the behavior of the molecule in its native environment. To analyze waters involved in packing interactions, the contacts list can be used to separate waters for which all water-nucleotide contacts fall within the same biological assembly from those which span two or more biological assemblies, taking all symmetry operations into account. In the UGGGGU data set, the “biological assembly” is taken to be the R6₈ construct.

This class of bound waters is shown in panel (a) of Figure 4.6. A wheat-colored R6₈ unit is shown in the center surrounded by three of its ten enclosing neighbors colored in cyan. The axial polarity of the central unit is opposite to that of its neighbors. A partial symmetry expansion of the waters in the asymmetric unit is also shown. This expansion is limited to rotation about the four-fold axis necessary to replicate the entire R6₈ surface. Waters for which all polymer atoms in the contacts list lie in the central, wheat-colored R6₈ unit are colored orange, those whose endpoints all lie in or among
Figure 4.5: Bound waters in contact with the ribose O2′ hydroxyl group in UGGGGU data set P1A. Colors are as described in Figure 4.4, ribose O2′ oxygens are shown in magenta. (a) Overall water distribution. (b) Detail near residues 1001 and 2002 illustrating cross-strand stabilization of the quadruplex by bridging waters in contact with O2′ group.
the cyan-colored neighbors are colored blue. Finally, waters whose contacts list includes polymer-atoms within both the central unit and one of its neighbors are colored magenta. A line of interstitial waters is apparent along the backbone terminating at the 3’ of the R6_A strands.

Relatively few waters fall into this class. For data set P1A they are water molecules 5005, 5009, 5011, 5012, 5019, 5020, 5029 and 5044. Crystal packing appears to depend only partly on the polymer-to-polymer bridges provided by these waters. As illustrated in panel (b), the packing between adjacent R6_8 units is very close, and bridges between waters restricted to a given R6_8 unit play an equally prominent role.

4.3.5 Internal Cavity Waters

As discussed in Section 3.8, a novel feature we observed in the tetragonal packing, is an internal cavity formed by the 3’ uridines of strand R6_B. This cavity is solvent accessible by an opening around the central cation channel and is well-hydrated as shown in Figure 4.7. Four waters occupy the asymmetric unit and four-fold symmetry fills the entire volume above the 3’ G quartet.

Binding sites for these waters likely correlate with binding of a daunomycin trimer to a DNA TGGGGT quadruplex, the first structural report of a quadruplex-drug complex[18]. In this structure the thymines rotated away from the central cation channel, into the solvent, and thus did not participate in shielding the terminal G quartets. Instead, the quartet stack was terminated by a trimer of three daunomycin molecules engaged in weak π-π bonding with the outer quartets as shown in panels (c)
and (d) of Figure 4.7 [18].
Figure 4.6: Bound waters involved in crystal contacts in UGGGGU data set P1A. Waters whose contacts list spans multiple R6,8 units are shown in magenta. Waters colored in orange and blue make contacts with the wheat-colored and cyan-colored R6,8 units respectively. (b) Detail illustrating the bridging role of interstitial waters.
Figure 4.7: Waters in the internal cavity of UGGGGU data set P1A. (a) Waters in the asymmetric unit (b) Cut-away view looking down on the cavity from the 3’ terminal uridines; symmetry-expansion about the four-fold axis is shown. (c) and (d) Binding of a daunomycin trimer at the 5’ terminals of DNA quadruplex TGGGGT, PDB entry 1O0k. Images from [18].
Chapter 5

Gradient-flow Segmentation of

Crystallographic Electron Density Maps

In this chapter we discuss a novel approach to extracting and assessing the features of crystallographic maps. The method is based on partitioning the map’s three-dimensional scalar field into disjoint basins. Basin boundaries are obtained from zero-flux surfaces defined by the gradient of the map’s scalar variable. Neither the mathematics underlying this method nor the algorithms used to implement it are novel. Gradient-flow partitions are a well-established technique in mathematical analysis and differential topology that provide the foundation for Morse theory [74]. Efficient algorithms for the computation of Morse-Smale subdivision on piecewise-linear manifolds have been devised and implemented by Edelsbrunner and co-workers for two dimensions [35] and by Gyulassy and co-workers for the more difficult three-dimensional case [43, 44, 42]. However adaption of these algorithms and techniques for application in
crystallographic computation and analysis is novel.

The atomic models produced by crystallographic methods result from the interpretation of features in crystallographic maps. Such maps approximate the distribution of a scalar quantity, usually a function of the observed and computed estimates of the electron density, \( \rho \), over a domain in \( \mathbb{R}^3 \) that encompasses the molecular model. Typically, this domain is the asymmetric unit, which, by application of symmetry operations and lattice translation can cover \( \mathbb{R}^3 \). Though the underlying physical model for the distribution of \( \rho \) assumes a continuous function, the data available for analysis are limited to samples on a rectilinear grid. The spacing of this grid is conventionally chosen to be two to four times smaller than \( d_{\min} = \frac{\lambda}{2 \sin \theta_{\max}} \) where \( \theta_{\max} \) indicates the resolution limit of a given data set [102]. The Nyquist-Shannon sampling theorem ensures that sampling at this rate allows reconstruction of the density without introduction of aliasing errors.

A variety of different types of maps are employed in structure determination, refinement and validation. However, their analysis shares common characteristics. The principal features of interest are map maxima and minima, though identification of contiguous, flat and featureless areas, such as those associated with bulk solvent, is also of value. Interpretation of maps is hampered by significant levels of error. Error arises from a variety of sources, including experimental measurement, crystal packing imperfections, reliance on phases estimated from incomplete or faulty models, and ripples from truncation errors in Fourier summation. There is no definitive model for resolving the contributions to overall error from individual sources.

The principal tools currently available for examination of maps are graphical
representations of iso-value surfaces, usually in units of the map standard deviation, and peak lists, simple enumerations of local critical points. The work described in this chapter investigates the applicability of a novel approach based on segmentation of a map into regions of uniform gradient flow. The gradient of the electron density defines a vector field and thus, as with all vector fields, implicitly defines a collection of disjoint flow lines or integral curves which originate and terminate at local critical points. The crystallographic asymmetric unit may thus be partitioned into disjoint regions defined by flow lines that terminate at a common maximum. The motivating intuition underlying this approach is that volumetric properties, such as the basin size and shape associated with a feature, may contribute information and discriminating power that can help guide map interpretation. The methods are meant to complement, not replace traditional inspection of iso-value surfaces or critical points. The algorithm we used here is adapted from one developed by Gyulassy and coworkers [44] which in turn builds on prior work. Our application of these techniques to the interpretation of electron density maps is novel and in need of further investigation but seems to yield promising results.

The algorithms and terminology underlying decomposition of a scalar function on $\mathbb{R}^3$ into regions of uniform gradient flow are based on concepts that originate in a setting where $\rho$ is a well-behaved, differentiable and continuous function; these concepts are then transferred to equivalent approximations on the grid-sampled data available for analysis. The gradient of the scalar field sampled by the map, $\nabla \rho$, defines a vector field over $\mathbb{R}^3$. A curve $l(s), \mathbb{R} \to \mathbb{R}^3$ defined such that the tangent at each point in the
curve coincides with the gradient vector at the point,

\[
\frac{\partial}{\partial s} l(s) = \nabla \rho(l(s))
\]  

is an integral line \cite{138}. Gradient vectors are null at critical points of \( \rho \), thus an integral line describes a path that originates and terminates at a critical point of \( \rho \).

From their definition, two integral lines are disjoint or identical, except possibly at their endpoints, and the union of points in all integral lines covers the domain of \( \rho \).

It follows that sets of integral lines that share a common origin or terminus may be defined, and that such sets may be used to partition \( \mathbb{R}^3 \). The set of points associated with integral lines that terminate at a maximum critical point \( p \) define the descending or stable manifold of \( p \). Similarly integral lines that originate from a minimum define the unstable or ascending manifold of \( p \) \cite{138}.

To support construction of gradient-flow based basins, an additional restriction on \( \rho \) requires that the stable and unstable manifolds intersect transversally \cite{74}. Morse-Smale manifolds are then defined as the intersection of the ascending and descending manifolds. This construction is represented on a simple two-dimensional domain in Fig. 5.1. Critical points are shown as circles with filled, empty and crossed centers for minima, maxima and saddle points respectively. Panel (a) shows the stable manifold obtained by all integral lines terminating at a given maximum, panel (b) shows the unstable manifolds formed from all integral lines originating from a given minimum and panel (c) illustrates the Morse-Smale complexes obtained from the transversal intersection of the preceding manifolds.
An important feature of Morse-Smale basins is their association with critical points. In two dimensions, each basin is quadrangular and includes two saddles, along with one maximum and one minimum. In three dimensions, the basin structure is more complex, as two types of saddle critical points, 1-saddles and 2-saddles, must be accounted for. Here, basins are hexahedra, coincidentally labeled *crystals* in some publications [43], whose faces are quadrilaterals. The regular critical point structure of the quadrangles and hexahedra that comprise Morse-Smale decomposition may be used to guide topologically-guided smoothing, as described below.

### 5.1 Smoothing

As noted previously, crystallographic maps are often noisy, an attribute that complicates their interpretation. Once a decomposition of the domain into Morse-Smale basins is available, it may be used to guide smoothing of the scalar function by cancellations of pairs of critical points. Progressive simplification of the Morse-Smale partition as a method of smoothing two-dimensional piecewise linear manifolds was introduced by Edelsbrunner, Harer and colleagues [35] and extended to three dimensions by Gyulassy and co-workers [43, 44]. The method is illustrated schematically in Figure 5.2.

Panel (a) illustrates a fragment of a two-dimensional domain subdivided into six Morse-Smale basins prior to smoothing. In applying a local smoothing operation, the circled maxima-saddle pair are removed and the integral lines which flowed into the
Figure 5.1: Decomposition of a two-dimensional scalar domain into Morse-Smale basins. Image from [138]. Maxima, minima and saddle points are labeled as in Fig. 5.2.
Figure 5.2: Smoothing of the Morse-Smale decomposition of a two-dimensional domain by critical point cancellation and basin coalescing. Image from [42].

deleted points are rearranged to flow to the adjacent maximum as shown in Panel (b). 
The result is a new decomposition into four Morse-Smale basins. Importantly, changes 
to the Morse-Smale complex are restricted to the local domain. All critical points not 
canceled remain as critical points in the new complex. Effectively, the descending man-
ifold of the higher peak now includes the descending manifold of the smaller, canceled 
peak. Several alternative approaches to smoothing scalar fields are available as summa-
rized by Gyulassy et al [43]. Among these, the approach based on simplification of the 
Morse-Smale complex is noteworthy because of the fine-grained control it provides and 
the preservation of topological structure.

5.2 Algorithms for Piecewise-linear Approximation

Algorithms for the computation of Morse-Smale complexes on sampled data 
have, for the most part, originated in the data visualization literature. For the two-
dimensional case, Edelsbrunner, Harer and Zomorodian proposed an algorithm based 
on piecewise-linear approximation [35] that was later improved by Bremer et al [11]. A 
three-dimensional extension of this approach has been refined in successive publications
by Gyulassy and coworkers [43, 44]. Henkelman and coworkers introduced a similar, but independently-developed algorithm for computing ascending and descending basins in the context of charge density analysis [47]. However, this algorithm, which has subsequently been refined [104], does not compute the full Morse-Smale complex.

The algorithm we use here is adapted from the one introduced by Gyulassy et al. [44], though a different approach is used to compute the two-dimensional surfaces separating basins. A description of the algorithm for descending basin decomposition follows, the ascending case is obtained by reversing the direction of comparisons.

5.2.1 Input and Definitions.

Input consists of calculated values of $\rho$ over the usual FFT-derived crystallographic grid. No attempt to incorporate crystallographic symmetry is currently included, though this is an expected enhancement. The map grid, usually restricted to the asymmetric unit, though larger regions may be analyzed, is assumed to belong to space group P1 by previous symmetry expansion if necessary, and is wrapped by a layer of boundary voxels. This simplification is commonly used in computing density statistics over maps. Boundary voxels are assigned infinitely large or small values depending on whether ascending or descending manifolds are being computed.

The algorithm’s output consists of a partitioning of the grid voxels into disjoint subsets, classified as MS3 or MS2 sets. A grid voxel belongs to a MS3 set if it is interior to a three-dimensional basin and to a MS2 set if it is part of a two-dimensional surface that separates MS3 basins.
For each grid point $p$, a key data structure in the computation is $\text{Link}(p)$, the set of neighboring points and the related sets $\text{Link}^-(p)$ and $\text{Link}^+(p)$, the subsets of $\text{Link}(p)$ whose values are smaller or larger than $p$. This structure approximates a spherical neighborhood of $p$ and is implemented as the 27-point subgrid centered at $p$.

A subset $S$ of $\text{Link}(p)$ is considered connected if for each $s \in S$ at least one grid point in the six-point neighborhood of $s$ is also in $S$.

Initially, data values are sorted and uniqueness simulated by using the array indices to break ties at equal values. Three passes over the data are applied.

### 5.2.2 Computation of MS3 Sets.

The set of maxima is found as the set of voxels for which $\text{Link}^-(p)$ is empty; these define the seed points for each MS3. Data values are examined in sorted order and tested for membership as interior to an MS3. If $\text{Link}^+(p)$ is connected and all elements of $\text{Link}^+(p)$ are classified as interior to a given MS3, then $p$ is also considered interior to that basin. This is the same approach introduced by Gyulassy et al [44].

### 5.2.3 Computation of MS2 Sets.

Once the collection of MS3 sets is computed, a second pass screens the set of points not classified as interior. All points which are not labeled interior but whose link includes an interior point are classified as boundary points. For each boundary point, the algorithm tracks the set of MS3 complexes it is incident on and, for each MS3, the set of boundary points incident to it. The list of unique MS3 incidences, across
all boundary points, serves to label each MS2 surface instance. For each such surface, membership is obtained by computing the set intersection of the incident boundary points of all MS3 sets defining the surface.

5.2.4 Classifying Boundary Voxels.

Some points may remain that did not meet the above criteria for membership in MS3 or MS2 sets. Such points typically occur near the grid boundary. In the last stage of the algorithm, these points are classified as boundary points and, where this adjustment leads to a boundary point with no MS3 point in its link, former boundary points are reclassified as interior. Effectively this last pass expands the boundaries of MS3 to ensure all grid points are classified as interior or surface.

5.2.5 Output and Analysis.

The results of the decomposition are written to a file for later processing and inspection. Currently this file is about eight times the size of the original map, though no attempt has been made to optimize storage. Examples of typical post-processing queries include computing basin properties, such as volume and total charge and tracking the basin membership of modeled atoms. For applications that involve difference maps, including water finding, distances from the basin maximum to neighboring atoms are computed. Total charge for interior voxels is computed by averaging $\rho$ values at the voxel vertices. For boundary voxels, charge is fractionally apportioned to the incident basins.
5.3 Results

The result of the algorithm described in the preceding section is a partition of the points of the FFT-derived grid that encode the crystallographic map being examined. Each point is assigned to one of a collection of disjoint subsets that correspond to three-dimensional (MS3) or two-dimensional (MS2) elements in the ascending (or descending) segmentation of the map. Each MS3 element is associated with a distinct maximum (or minimum) and each MS2 element bounds a unique combination of MS3 elements.

It is helpful to visualize the results of this construction to understand how it differs from conventional tools for inspecting and analyzing map features. The bounding surface of points in a given MS3 basin can be built by means of grid-based polygonal surface construction algorithms such as Marching Cubes [69]. The Marching Cubes algorithm is widely used for constructing iso-contours of scalar values in $\mathbb{R}^3$ and is easily adapted to any situation where grid points can be classified as inside or outside a surface. The algorithm we used is a refinement of the original Lorensen and Cline Marching Cubes algorithm by Thomas Lewiner and co-workers [63], based on a proposal by E. V. Chernayev, that addresses some ambiguities of the original version.

Some results of partitioning a $\sigma_A$-weighted $2mF_o - DF_c$ map from a refinement of data set P1B are shown below. The model associated with this map includes 352 non-hydrogen atoms whereas the map partition yields 1616 basins. In a high-resolution, well-modeled, map each atom can be assigned to a distinct MS3 basin. In such cases,
the density enclosed by the basin can be unambiguously assigned to the atom. However, this is not generally true, and the grid point associated with an atom’s coordinates may be classified as a MS2 boundary point or multiple modeled atoms may map to the same MS3 basin.

Figure 5.3: Single atom density basin. (a) Isovalue contours of $\sigma_A$-weighted $2mF_o-DF_c$ density near atom O3' of guanine residue 1003 from structure P1b at 0.5 $\sigma$ (magenta) and 2.0 $\sigma$ (green). (b) The surface of the corresponding MS3 basin is shown in blue, slightly clipped to reveal the contours within the interior.

Panel (b) of Figure 5.3 displays the surface of the MS3 basin that corresponds to atom O3' from guanine residue 1003, a well-ordered oxygen with a B-factor of 4.56 Å$^2$. The basin, shown in blue, can be compared with isovalue contours, commonly used for feature extraction. Two contours of the map are shown in panel (a), 2.0 $\sigma$ in green.
and 0.5 $\sigma$ in magenta. The surface of the MS3 basin has been clipped to expose the enclosed contours. Some distinctive differences are apparent:

- Unlike isovalue contours, basin surfaces enclose a single peak. This property also applies to isovalue contours taken at sufficiently small values of $\sigma$ but does not hold generally. For example, panel (a) shows that in this well-ordered region of the map, the magenta-colored, 2.0 $\sigma$, contours generally enclose distinct atoms, though slight overlap is visible at the P-O2P boundary. By comparison, the green-colored, 0.5 $\sigma$, contour yields a single surface which encloses all the atoms in this region of the model. The latter situation is typical of noisier, lower-resolution, maps. Thus basins, unlike isovalue contours, can be unambiguously associated with individual atoms provided each atom is associated with a distinct local maximum.

- Typically, the volume of the map enclosed by an atom’s basin is considerably larger than the volume enclosed by the surface of the smallest isovalue contour uniquely associated with the atom, a property that follows from the defining gradient-flow characteristic of basins. As one travels outward from an atom’s density peak, values of $\rho$ decrease. As this path approaches a neighboring modeled atom, an abrupt boundary will separate the two basins. However, if the path points towards bulk solvent, the distance to the next local maximum will be greater and all points along this path will be assigned to the original basin. Thus basins generally expand towards unmodeled regions of lower density.

- Whereas all points on the surface of an isovalue contour, by definition, have the
same value of $\rho$, density values on the surface of a MS3 basin vary widely. The highest values occur near points on the vector joining covalently-bonded atoms and lower values where the basin bounds unmodeled bulk solvent.

The scattergrams shown in Figure 5.5-5.3 provide additional insight into the properties of basins. The 192 points in the scattergrams correspond to basins associated with modeled non-hydrogen RNA atoms. Atoms modeled in alternate conformations were excluded as were basins for which the Link of any interior point included the grid boundary. The latter restriction is intended to avoid any artifacts that might result from the current implementation of boundary conditions. Grid points are colored by element type. For each selected basin, the number of voxels was determined, along with two measures of basin density. Total density sums $\rho$ over all interior points along with a fractional contribution from all boundary points. Peak density simply sums the density in the 27-point sub-grid centered at the basin’s maximum.

Inspection of these plots reveals a number of interesting features. While peak density provides reasonably good discrimination between the scattering power of different element types, total basin density fares poorly with the exception of the readily recognizable phosphorus atoms. This effect is likely due to the incorporation of non-specific density as basins expand towards regions of unmodeled bulk solvent. For example, the basin with the lowest value of total density, -139.01, is associated with atom C4’ of guanine residue 2005. Though peak density for this atom is in agreement with that of other carbon atoms, as evident from Figure , the size of the basin is the second largest of any carbon atom, as shown in Figure.
Figure 5.4: Peak height vs. total density of descending basins for RNA atoms in structure P1B. For each atom, the height of the basin’s maximum density and the total basin density are plotted. Colors indicate the atom type, as shown in the legend.
Inspection of Figure 5.5 also reveals that the basins for phosphorus and nitrogen atoms are considerably more compact than for oxygens or carbons. This effect can be anticipated from the packing of the surrounding oxygens in the phosphate group, but is less readily rationalized for nitrogen.

Figure 5.5: Volume vs total density for descending basins for RNA atoms in structure P1B. For each atom, the basin volume, as measured by voxel count, and the total basin density are plotted.
5.3.1 Difference Map Segmentation

We described the application of volumetric segmentation of difference maps to detect bound water sites in Section 3.2.2. Though this approach was useful in determining the solvent model, its application was limited by the absence of reliable criteria for distinguishing density basins attributable to unmodeled solvent from noise. In other words, in this application, there is no “gold standard” against which to measure or compare performance of the method.

To investigate the value of segmentation in a more controlled, albeit artificial, environment, a series of difference maps for the known part of the model were constructed as follows. Each RNA atom in turn was removed from the final refined model for structure P1B and $\sigma_A$-weighted $mFo - DFc$ “kicked” omit maps [96] computed and segmented. The basin corresponding to the missing atom in the omit map was determined and all basins in the segmented map were ranked, in decreasing order, by magnitude of total enclosed difference density.

If a questionable part of a model is genuine, it is expected to surface in an omit map calculated with the part in question removed. Accordingly, the basin containing the omitted atom would be expected to be prominent in the difference map and thus exhibit a low rank. The use of “kicked” omit maps, averaged over approximately 50 randomly shaken models, is expected to compensate for the residual model bias remaining in the map due to the use of phases calculated before the questionable part was removed [96]. Hydrogen atoms, those in multiple conformations and those occupying basins
intersecting the map border were excluded, leaving 171 atoms. Each atom in turn was removed, the incomplete model refined, weighted difference maps calculated and segmented, and the rank of the basin enclosing the missing atom determined.

The results are shown in the histogram in panel (a) of Figure 5.6 and, as a cumulative histogram, in panel (b). From panel (a) it is apparent that the basin that covers the region of the difference map spanning the missing atom is prominent in the difference map: for 77 of the 141 atoms it has the highest rank. Furthermore, as shown in panel (b), relaxing the criterion to require the atom in question to appear among the top ten most prominent basins, correctly classified about 90% of the modeled RNA atoms as legitimate parts of the model. Additional work will be required to further tune this criterion, however, the volumetric segmentation of difference maps appears to provide a promising way of identifying prominent features.
Figure 5.6: Rank order of total density in kicked omit maps for all modeled RNA atoms in structure P1B.
Chapter 6

Calculation of Molecular Surfaces and Derived Properties by Signed Distance Functions and *Fast Marching* Methods

The atomic model fit to experimental density incorporates all covalent bonds in the macromolecule under investigation. Weak bonding interactions between atoms exclude any approach closer than a given distance, usually approximated by the sum of the respective van der Waals radii, though minor deviations in this limit may occur because of hydrogen bond or stacking interactions. Since interactions among macromolecules and between macromolecules and solvent result from weak bonding forces, it is useful to augment atomic models by van der Waals radii so as to examine how interactions at permissible distances correlate with properties such as steric fit, contact area, electrostatic charge and hydrophobicity. The principal tool for analyses of this
This chapter details the use of a relatively novel approach to calculating molecular surfaces and their properties by use of a mathematical technique, known as *Level Set Methods*. The distinguishing characteristic of these methods is characterization of the surface of interest as the level set of an advancing front and its computation as the solution to a partial differential equation modeling front expansion. The first two sections describe commonly computed molecular surfaces and introduce the level set methods and their implementation via the fast marching algorithm. The third section details how this approach is used to identify molecular surface properties and to correlate solvent sites with given surface properties.

### 6.1 Molecular Surfaces

Underlying support for the construction of molecular surfaces rests on calculation of the Lennard-Jones potential, the interaction energy of approaching neutral atoms or molecules as a function of their separation. This energy rises as $r^{12}$ for distances closer than the sum of van der Waals radii, eliminating the possibility of closer interaction in the absence of covalent or ionic bonding [41]. Three types of surfaces are commonly used in the analysis of surface properties as shown in Fig. 6.1. The van der Waals surface, shown in panel (a) is obtained by simply replacing the coordinates of each modeled atom by a sphere whose radius is the van der Waals distance of the corresponding element. The solvent accessible surface, shown in red in panel (b), is
characterized as the surface traced by the center of a ball of given radius, the probe, as it is rolled over the van der Waals surface. Lastly, the molecular surface is shown as the union of green and magenta sections in panel (c). This surface is obtained via the same rolling ball approach used in the definition of the solvent accessible surface but is comprised of two components. The contact surface, shown in green, results from the union of probe circumference points that contact any point in the van der Waals surface. The re-entrant surface, shown in magenta, spans crevices in the van der Waals surface from which solvent is excluded.

Figure 6.1: Commonly-used molecular surfaces: (a) van der Waals (b) solvent accessible and (c) solvent-excluded surface (image from [131]). The van der Waals surface, shown in blue, is the union of spheres with fixed, element-specific, radii centered at atomic coordinates. The solvent accessible surface, shown in red, is defined by tracing the center of a sphere of fixed radius, the probe, over the van der Waals surface. The solvent excluded surface, shown in green and magenta, is defined as a union of points on the probe surface in contact with the van der Waals surface. In a distance field formulation, the van der Waals and solvent-excluded surfaces result from an outward-directed front of uniform velocity and the solvent-excluded surface from an inward-directed front.

A number of different approaches for computing molecular surfaces have been proposed since their introduction by Lee and Richards in 1971 [60]. Key contributions include an algorithm for exact calculation of surfaces based on analytic geometry by Connolly [21], an approach based on alphah-shapes [65] and a widely-used package based on [103].
Among the various approaches for calculating molecular surfaces, an algorithm in widespread use and related to the distance-field methods discussed below proceeds as follows [101]. The molecule of interest is placed in a bounding box of suitable dimensions which is partitioned by a regular, orthogonal grid whose spacing depends on the desired surface quality. Each vertex in the grid is examined and marked as external if it is further than the sum of van der Waals radius and probe radius from any atom in the molecule. Other vertices are marked internal. Interface voxels, grid voxels intersected by the molecular surface are identified as those for which not all bounding vertices are internal or external. Finally, some version of the Marching Cubes algorithm, well known in computer-graphics for iso-surface construction, is used to obtain a triangulation of the molecular surface. This approach is implemented in the GRASP package for electrostatic calculation and also used for initial surface approximation in the LSMS implementation [16].

### 6.2 Implicit Surfaces and Fast Marching Methods

As evident from the preceding definitions, molecular surfaces can be viewed as an advancing front that starts from the van der Waals surface and expands in the direction of the normal vector to the surface. Casting the problem in this way gives access to an extensive inventory of Level Set Methods for tracking evolving interfaces. These methods characterize an interface as the level set of an implicit function $\phi(\vec{x}) = 0$ defined on the domain of interest [86, 108]. Thus, by introducing an appropriate implicit
function, the domain can be partitioned into an interior, defined by $\phi(\vec{x}) < 0$, an exterior where $\phi(\vec{x}) > 0$, and the interface of interest. Since the interface is dynamic, $\phi(\vec{x}(t), t)$, also depends on the time variable $t$ where $\vec{x}(t)$ tracks the path traced by a point on the interface as the front evolves. To monitor the change over time, the relation $\phi(\vec{x}(t), t) = 0$ is differentiated with respect to $t$ which, by application of the chain rule, yields

$$\frac{\partial \phi(\vec{x}(t), t)}{\partial t} + \nabla \phi(\vec{x}(t), t) \cdot \vec{x}'(t) = 0 \quad (6.1)$$

The equation can be further manipulated by using the property that the gradient of a function is normal to the isocontours defined by its level sets. Thus if $\vec{x}_0$ is a point on the interface at time 0, $\nabla \phi$ evaluated at $\vec{x}_0$ yields a normal vector which points in the direction of front evolution

$$\vec{n} = \frac{\nabla \phi}{|\nabla \phi|}$$

Also, the velocity of the evolving front at a point $\vec{x}(t)$ is defined as

$$F = \vec{x}'(t) \cdot \vec{n}$$

Combining these relations into 6.1 yields a partial differential equation (PDE) in $t$

$$\phi_t + F|\nabla \phi| = 0 \quad (6.2)$$

with initial condition $\phi(\vec{x}(t), t) = 0$. Formulating evolution of the interface in this way does nothing directly to solve or simplify the problem. In fact the introduction of the auxiliary implicit function $\phi$ introduces additional notation and complexity. However,
by expressing the problem as a PDE, computation methods for approximating solutions to PDEs on a sampled grid can be used to track the interface.

An alternative expression of 6.2 as a boundary value rather than an initial problem is possible if the velocity $F$ is a positive or negative constant. In this case, if $T(\vec{x})$ is the arrival time of the front at point $\vec{x}$, the problem is described by the Eikonal equation [108]

$$||\nabla T||F = 1$$

(6.3)

Level Set Methods provide a general framework for tracking evolving fronts, including fronts which may propagate forwards or backwards with different velocity at each point. In general, velocity may depend on a combination of global or local properties, such as curvature. However applications to molecular surfaces only need consider a simplified subset in which $\phi$ is the signed distance function from the van der Waals surface. If $S$ if the connected solid enclosed by the van der Waals surface and $\partial S$ its boundary, the signed distance function $d_S(\vec{x})$ is defined as [53]

$$d_S(\vec{x}) = sgn(\vec{x}) \ min_{x_I \in \partial S} ||\vec{x} - \vec{x_I}||$$

(6.4)

where

$$sgn(\vec{x}) = \begin{cases} 
-1 & \text{if } \vec{x} \in S \\
1 & \text{otherwise} 
\end{cases}$$

(6.5)

The signed distance function imposes a simplifying restriction on the velocity of the
Sethian [108] shows that a discretized, first-order, approximation to the squared magnitude of the gradient of $T(\vec{x})$ can be obtained by solving a quadratic equation in the distance values of the grid points in a six-point neighborhood of $\vec{x}$:

\[
||\nabla T||^2 = \begin{cases} 
\max(V_A - V_B, V_A - V_C, 0)^2 + \\
\max(V_A - V_D, V_A - V_E, 0)^2 + \\
\max(V_A - V_F, V_A - V_G, 0)^2 
\end{cases}
\]

(6.7)

where $V_A$ is the unknown point and indices B-E denote neighboring grid points at $(-1,0,0)$, $(1,0,0)$, $(0,0,1)$, $(0,0,-1)$, $(0,1,0)$ and $(0,-1,0)$. [6]. Since for distance functions the gradient of the arrival function is fixed, the above relation can be inverted to obtain the distance value at an unknown point grid point if distance values for its neighbors are available.

The Fast Marching Method (FMM) uses the above relations to iteratively compute distance function values over the entire grid by iteratively propagating distance values from the given surface. The algorithm, whose approach is similar to Dijkstra’s shortest path graph algorithm [53], separates grid points into three sets, known or frozen points, narrow band points and unknown points. Initially, the set of frozen points consists of the grid points closest to the van der Waals surface. Initial distance values for frozen grid points can be computed as distance to the nearest point on the corresponding triangular mesh. Narrow band points are immediate neighbors of frozen points for which no distance value is known. The algorithm computes distances for all narrow band grid points and among these selects the point with the smallest distance. That
point is frozen, distances are computed for the new narrow band and execution proceeds to the next iteration.

To improve efficiency a binary heap is used to maintain narrow band grid points in order of increasing distance. A higher accuracy version of the algorithm improves distance estimates by modifying 6.7 to use second-order approximations to the derivative of $T(\vec{x})$. This modification slightly increases the execution time but yields significantly better distance estimates.

### 6.3 Computing Surface-derived Properties

The use of Fast Marching methods to calculate molecular surfaces was initially reported by Can [16] and has since been investigated by other authors [137, 46]. As noted earlier, a number of algorithms are available for molecular surface calculation and level set methods provide an interesting but not clearly superior alternative. Motivation for their use rests on their ability to not only construct molecular surfaces with given probe diameter, but the entire distance field defined by the van der Waals surface, that is, values of $d_S(\vec{x})$ as given by 6.4, where $S$ is the van der Waals surface, for all $\vec{x}$ in the bounding domain in which $S$ is embedded.

Distance fields are a rich data object capable of supporting a variety of queries that arise in exploring solvent structure. Representative examples include:

**Solvent to Surface Distance.** Calculating the set atoms, including symmetry equivalents, which make contact with a given solvent molecule does not require surface
calculations and the contacts list is typically used in evaluating the plausibility and role of a given solvent molecule. However, a more direct measurement of overall proximity is the distance to the van der Waals surface. Though this can be obtained by calculating the minimal point-to-triangle distance for a given triangulation of the surface, this approach requires searching all triangles for each query. Alternatively, once a distance field is calculated, the value is immediately available by interpolating the grid points nearest to the solvent molecule.

**Bound Solvent to Bulk Solvent Distance.** A related query is the distance of a distinct solvent molecule to the amorphous surrounding layer modeled as bulk solvent. This distance measures how far a water molecule, for example, would need to travel from surrounding bulk solvent to the specific location, often in a pocket, cavity or tunnel, at which which observable density was detected. Coleman and Sharp introduced the term *Travel Depth* to describe this distance along with an algorithm for its computation [19]. The results of their algorithm as applied to the calculation of groove depth for B, A and Z-form DNA structures is shown in Fig. 6.2. To quantify this distance it is first necessary to compute the boundary of the bulk solvent region, essentially the molecular surface that results from a probe of infinite radius. Coleman and Sharp approximate this by a triangulation of the convex hull, a construction for which a number of efficient algorithms are known [24]. However, once the enclosing surface is available, the travel distance can alternatively be obtained by using the FMM to calculate a distance
The depth of A, B and Z-form DNA grooves as calculated by the travel depth
algorithm (image from [19]).

Figure 6.2:

field which progresses inward and then interpolating grid points as above. To
prevent propagating the inward-advancing front into the interior of $S$, the step of
the Fast Marching algorithm which selects the grid point with minimal distance
from among those in the narrow band must be modified to return an arbitrarily
large value for grid points interior to $S$.

**Solvent Pockets and Cavities.** The implicit functions introduced for level-set based
methods provide an effective means of defining sub-regions of interest by construct-
ing suitable expressions. Osher reviews the main elements of these expressions [86].
For example given two surfaces $S_1$ and $S_2$ defined by implicit functions $\phi_1(\vec{x})$ and
$\phi_2(\vec{x})$, the function $\phi(\vec{x}) = \min(\phi_1(\vec{x}), \phi_2(\vec{x}))$ provides an implicit function for the
union of the interior of $S_1$ and $S_2$ whereas setting $\phi(\vec{x}) = max(\phi_1(\vec{x}), \phi_2(\vec{x}))$ provides an implicit function for their intersection. Changing the sign of the implicit function $\phi(\vec{x}) = -\phi_1(\vec{x})$ yields a function which characterizes the complement of $S_1$.

These operations can be combined. If $S_1$ is the surface of a macromolecule and $S_2$ is the surface of an interior cavity, then $-\phi_2(\vec{x})$ is a signed distance function for the entire region outside the cavity and $\phi(\vec{x}) = max(\phi_1(\vec{x}), -\phi_2(\vec{x}))$ is a signed distance function for the interior of the macromolecule excluding the cavity.

This approach was exploited by Zhang and Bajaj [137] for calculating pockets of protein surfaces. They introduced the pocket function $\phi_P$ defined by

$$\phi_P(\vec{x}) = \min(d_S(\vec{x}), d_T(\vec{x}) - t)$$

(6.8)

where $d_S$ and $d_T$ are distance functions from the van der Waals and limiting, bulk solvent, surface. Rather than representing the limiting surface as a convex hull, their algorithm expands the van der Waals surface outward to an arbitrary distance $d_S(\vec{x}) = t$ to yield an approximately spherical enclosing surface and the value of $t$ is applied in 6.8. The points at which the inward-marching algorithm does not reach $S$ correspond to the mouths of internal regions. The number of mouths in contact with a given region, $m$, may be used to classify it as an internal cavity ($m = 0$), a pocket ($m = 1$), a simple tunnel ($m = 2$) or a multi-entrance tunnels ($m >= 3$).

**Surface Curvature.** The features of a molecular surface play an important role in
determining hydration as reviewed by Levitt and Park [62]. Water in narrow, deep, crevices will be subject to weaker hydrogen-bonding attraction from bulk solvent than water in broader surface depressions. Calculating surface curvature near a pocket’s mouth can help distinguish pockets with an abrupt opening to the surface. Here again level set methods provide useful computational aids [86]. For an implicit function $\phi$ which is a signed distance function, the property $|\nabla \phi| = 1$ implies the curvature $\kappa$ is available as

$$\kappa(\vec{x}) = \phi_{xx}(\vec{x}) + \phi_{yy}(\vec{x}) + \phi_{zz}(\vec{x})$$

(6.9)

where the second partial derivatives may also be approximated by finite differences over a grid, for example

$$\frac{\partial^2 \phi}{\partial x^2} \approx \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2}$$

(6.10)

Thus finite differences similar to those used in estimating the gradient of $T$ in the Eikonal equation 6.3 may also be used to estimate the curvature of the front.

In summary, the calculation of a signed distance field over a grid of suitable resolution can provide a flexible computational framework for investigating various aspects of solvent distribution in the vicinity of the macromolecular surface. For such applications, an effective way to compute the signed distance field is via a specific variant of level set methods known as Fast Marching Method (FMM) algorithms. A variety of competing algorithms for computing signed distance fields are available [53]. The FMM approach appears to provide a reasonable mix of accuracy, speed and ease of implementation and to be well suited for molecular surfaces.
In closing, this section summarizes the results we obtained in this thesis and outlines directions for additional research. The field of high-resolution RNA crystallography is small. The structures reported in Chapter 3 are of general interest because they enable examination of detail usually not visible at lower resolution. Furthermore, we uncovered a number of novel features of the UGGGGU hexamer including:

- A crystal packing in an orthorhombic space group, $C221$, not previously reported for this or similar quadruplex structures. An interesting feature of this packing is the presence of a complete, four-stranded, assembly within the asymmetric unit.

- Different conformations for the 3’ uridine residues of the two chains in the tetragonal space group. One of these, labeled $R6_B$, has not been previously reported and accounts for the formation of a well-hydrated internal cavity. The cavity separates the 3’ terminal U quartet from its adjacent G quartet, leaving the latter unshielded. Thus, the waters detected in the interior of the cavity likely serve
a similar shielding function to molecules of the anti-cancer drug daunomycin reported in complex with the DNA quadruplex TGGGGT [18].

- A novel crystal packing with smaller lattice in the previously reported $P42_12$ space group. The spermine molecule which bridged two R6-8 units and was a likely stabilizing force in the lattice of the previously reported structure [25] was not detected, though included in the crystallization conditions.

- A well-supported alternate conformation in the phosphate backbone of the 3' residue of one of the polymer chains.

Our examination of solvent structure in Chapter 4, has reinforced general themes previously reported for nucleic acid hydration including a concentration of bound waters along the negatively-charged phosphate backbone and along the floor of grooves formed by the helical conformation. In addition, some quadruplex and RNA-specific features were detected in the solvent distribution. These include cross-chain stabilization by bridging waters in contact with the RNA O2' group and the absence of any waters within regions of the quadruplex central channel occupied by quartet-stabilizing cations.

As described in Chapters 5 and 6, we adapted and extended two existing computational methods, volumetric segmentation based on Morse-Smale basins and distance field calculation via level set computation, for application to crystallographic map-interpretation and solvent structural analysis respectively. The computation of a signed distance field from the van der Waals surface provides a versatile data structure
that supports a variety of queries including:

- construction of solvent-accessible surfaces with arbitrary probe radius;
- detection of internal regions such as pockets, cavities and tunnels;
- the distance of modeled solvent molecules from the molecular surface or from bulk solvent.

The signed distance field can be efficiently calculated with FMM algorithms. An additional benefit of this approach is calculation of the curvature of the molecular surface by reuse of terms used to approximate the gradient of the implicit function used to compute the distance field.

Partitioning of density into basins bounded by zero-flux surfaces by algorithms for computing Morse-Smale complexes is a novel and promising addition to the crystallographer's toolbox. A three-dimensional basin carries more information about features of the density distribution recorded in a crystallographic map than two-dimensional isovalue surfaces or one dimensional peak and pit lists. Furthermore, the simplification of noisy maps by persistence-guided critical point cancellation provides a useful and controllable way of smoothing maps so as to expose their principal features.

7.1 Future Directions

With the possible exception of the structure of the RNA hexamer, all the above results are incomplete. Thus the main contribution of the investigations we described
has been to make apparent the need for further research. Some directions for additional work include the following.

Much has been learned about G4 folds since the initial screens of genomic data were applied approximately eight years ago. The protocols recently introduced by Wieland and Hartig [129] provide a relatively easy way to assess G4 fold formation and stability \textit{in vivo} for putative PG4S elements. These protocols supplement traditional biophysical methods such as CD spectroscopy and UV melting curves. The additional information now available should enable more discriminating screens for PG4S elements, possibly along the lines of HMMER [34].

The significance of the PG4S enrichment found in specific genomic regions, most strikingly the occurrence of GRIN1 elements in 48% of human genes, remains unclear. A potentially valuable tool in this exploration is the identification of protein domains which bind G4 folds, however this area remains largely unexplored. A number of associations have been well characterized, including proteins involved in telomere function and DEAH helicases [70]. However, to date, only one review on the general properties of G4 binding proteins has been completed [39] and no structures of protein-quadruplex complexes have been deposited.

Identification of solvent by X-ray crystallography is imprecise for reasons previously discussed. Alternative methods, including molecular dynamics and NMR can provide valuable complementary information on residence times and solvent identity. Application of such methods to the structures reported here would improve understanding of its solvent structure.
The applicability of basin partitioning methods in identifying solvent sites has been demonstrated in Chapter 4. Nevertheless, a number of crucial questions regarding these methods remain to be explored before they can be assumed to be generally-useful tools for map analysis. Key issues are excluding regions where density fluctuations do not exceed background error and assessing the comparative merits of the two competing approaches to approximating a continuous distribution over the sampled grid, piecewise linear and simplicial complexes.
Appendix A

Software

We developed a number of software components as part of this thesis. Sources are available in the svn repository at http://diablo.ucsc.edu/svn/basins. C++ components rely significantly on Kevin Cowtan’s clipper and Eugene Krissel’s MMDB libraries [36]. In most cases, we implemented core functionality in C++ and bound it to a Python interpreter to facilitate scripting. These include:

**dr**

_A Python-based package for automating data reduction and preliminary analysis. Development of this package was motivated by difficulty in determining the correct space group assignments as described in Section 3.2. Starting from an XML-encoded description of the image set, the following processing steps are applied in sequence to produce a scaled reflection file suitable for further analysis._

The processing is divided into three stages. An initial attempt to assess the symmetry, mosaicity and resolution is made by running _labelit_ [105], on randomly selected pairs of images which sample the full data set. To confirm _labelit_’s solution
for each pair, *pointless* [37] is also run on the mtz file produced as part of labelit’s search. In the next stage, parameters for the most commonly reported, consensus, space group solution are used for post-refinement and wedge integration of two contiguous ranges of images starting at the first image of each pair. Pointless is again run on the mtz files resulting from the wedge integration and the resulting space group probabilities are cumulated. The most probable space group and the cell parameters yielding the lowest weighted positional error from the wedge integration are used for the third stage of processing. Here a number of integration and scaling runs with different parameters are applied to the full data set to find optimal settings.

**asu** A C++ program that moves the contents of an arbitrary pdb file into the asymmetric unit (asu). To reduce computation, basin decomposition of maps is usually limited to maps spanning the asu, thus it is useful to also refer pdb coordinates to this volume.

**chsolv** A C++ program that assembles a list of candidate water sites from a given basin decomposition of a difference $F_o - F_c$ map. Basins of acceptable size located a reasonable distance from previously modeled atoms are considered candidate water locations for the next round of refinement. The peak height threshold and minimum-maximum distance range to a modeled atom are supplied as input.

**dgrid** A collection of Python and C++ modules that implement the basin segmentation algorithm described in Chapter 5. The decomposition of the input map is
written as a state file that records all grid points in each basin. Though the state file is large, this separation allows separate post processing and analysis of the decomposition.

**bstat** A collection of C++ and R [98] modules for the analysis of volumetric segmentation of maps.

**p2i** A collection of C++ and Python modules for automating iterative refinement and the generation of shelxl [109] input files.
Appendix B

Distance to Surface Calculations

To test the level-set based distance field calculation discussed in Chapter 6, we calculated distances of solvent atoms to the molecular surface as follows. The program EDTSURF [131] was used to calculate a triangulated molecular surface in PLY format. The subset of triangles in the neighborhood of the solvent atom of interest was obtained from a bounding box and the triangle with minimum distance to the surface was identified, as shown in Figure B.1.

To obtain the minimal point to triangle distance, we implemented an algorithm by Schneider and Eberly [106]. For a given atom \( \vec{a} \) and triangle \( \vec{v}_1, \vec{v}_2, \vec{v}_3 \), this algorithm computes the nearest point in the triangle, \( \vec{p} \). To determine whether the atom is interior to the surface, the surface normal at one of the vertices, \( \vec{v}_1 \), was projected onto the vector from \( \vec{a} \) to \( \vec{p} \). Since the EDTSURF program by convention uses surface normals pointing towards the surface exterior, the sign of the resulting projection determines whether \( \vec{a} \) occurs on the interior or exterior of the triangle.
Figure B.1: Solvent to surface distance calculation. The triangle closest to an internal water is highlighted in red. Surface triangulation was calculated with the program EDTSURF [131].
Bibliography


[29] D Drygin, A Siddiqui-Jain, S O’Brien, M Schwaebe, A Lin, J Bliesath, C B Ho,


[44] A Gyulassy, V Natarajan, V Pascucci, and B Hamann. Efficient computation of


[59] R Lavery, M Moakher, J H Maddocks, D Petkeviciute, and K Zakrzewska. Con-


[67] L Liu, M L Quillin, and B W Matthews. Use of experimental crystallographic
phases to examine the hydration of polar and nonpolar cavities in t4 lysozyme.


[98] R Development Core Team. *R: A Language and Environment for Statistical Com-


[136] R Zhang, Y Lin, and C T Zhang. Greglist: a database listing potential g-
