Crystallographic and Biophysical Studies
of Telomeric RNA and DNA
G-quadruplex-Drug Complexes

A thesis submitted for the degree of
Doctor of Philosophy of the University of London

Gavin William Collie

Cancer Research UK Biomolecular Structure Group
The School of Pharmacy,
University of London

November 2011
This thesis describes research conducted in the School of Pharmacy, University of London between October 2008 and September 2011 under the supervision of Dr. Gary Parkinson. I certify that the research described is original and that any parts of the work that have been conducted by collaboration are clearly indicated. I also certify that I have written all the text herein and have clearly indicated by suitable citation any part of this dissertation that has already appeared in publication.

Signature

Date 22/01/2012
Dedicated to the memory of

Jenny Collie
ACKNOWLEDGEMENTS

Thanks to my supervisors Dr. Gary Parkinson and Prof. Stephen Neidle, for supporting me throughout my project and for agreeing to hire me in the first place all those years ago. My time in the Biomolecular Structure Group wouldn’t have been the same (or enjoyable) without the many fantastic people who work there and I cannot thank them enough. Caterina, thank you for trying to teach me Italian for four years. Alan, you’ve helped me too many times for me to list, so I will just say, thank you. Nancy, thank you for making the lab a more enjoyable place to be and for being a dangerous amount of fun on a night out. Jarmila, for emotional and nutritional support, I am sincerely grateful. Aaron, I have to thank you for generally being awesome. I am also incredibly grateful to Katja Galesa, for teaching me more than should have been possible about molecular biology, and doing so with unbelievable patience. I also have to thank Irene Dougherty, for helping me an uncountable number of times in the past few years. And for giving me a coffee machine. I would also like to thank Rossella Promontorio, for being an excellent student and making the final few months of my PhD surprisingly enjoyable.

For my research trips to Liege, I am sincerely thankful to Valerie Gabelica and Frederic Rosu, for their expertise and guidance, as well as for all they did to make me feel welcome in Belgium. Thanks to COST action MP0802 for kindly funding these research trips. I am extremely grateful to Oxford Diffraction (Agilent) and the Diamond Light Source synchrotron, for making their data collection facilities available. Thanks also to the School of Pharmacy, for funding my PhD.

For unquestionably contributing to the maintenance of my sanity in different ways, I am hugely thankful to everyone at The Mill Hill Brazilian Jiu Jitsu Academy and my brother. Finally, I am immensely thankful to my parents and family, for always supporting me, and to my girlfriend Sara, without whom I doubt I’d have managed to write this thesis.
ABSTRACT

The terminal regions of eukaryotic chromosomes – the telomeres – play a crucial role maintaining genomic integrity and in the regulation of cellular senescence and apoptosis. These regions – composed of a complex array of proteins and nucleic acid structures – are found to be dysregulated in a high percentage of cancers, and as a result, several components of mammalian telomeres are being actively investigated as potential anti-cancer drug targets. The work in this thesis has focused on a group of unusual nucleic acid structures which form from telomeric DNA and RNA molecules. These structures – termed G-quadruplexes – are of high relevance to anti-cancer drug design, as the stabilisation of such structures with small molecule compounds provides a therapeutic route to inhibiting a key oncogenic telomeric protein – telomerase.

The overall aim of the work reported here has been to provide high resolution structural data – primarily in the form of X-ray crystal structures – concerning telomeric DNA and RNA G-quadruplexes, with a focus on the application of such structures to telomere-directed anti-cancer drug design. Several high resolution crystal structures of G-quadruplexes formed from human telomeric sequences were solved, providing atomic-scale details of RNA folding and G-quadruplex-drug interactions. The structures presented include the first crystallographic descriptions of native and ligand-bound G-quadruplexes formed from human telomeric RNA, as well as several human telomeric DNA G-quadruplex-ligand complexes, including the first complexes determined involving the much-studied human 22-mer sequence, d(AGGG[TTAGGG]3).

Additional biophysical data is also presented, complementing the crystallographic data by providing details of telomeric G-quadruplex higher-order structure formation, G-quadruplex stability and ligand-binding data.

The high-resolution crystallographic structural data, combined with the biophysical findings, provide a comprehensive picture of telomeric G-quadruplex folding, ligand-recognition and multimerisation, and reveal important differences between telomeric DNA and RNA G-quadruplexes of relevance to the future design and development of telomeric G-quadruplex-directed therapeutic compounds.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>4</td>
</tr>
<tr>
<td>Abstract</td>
<td>5</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>6</td>
</tr>
<tr>
<td>List of Figures and Tables</td>
<td>12</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>16</td>
</tr>
<tr>
<td><strong>Chapter 1 - Introduction</strong></td>
<td>17</td>
</tr>
<tr>
<td>1.1 Proteins vs Nucleic Acids</td>
<td>18</td>
</tr>
<tr>
<td>1.2 Structural Diversity of Nucleic Acids</td>
<td>19</td>
</tr>
<tr>
<td>1.2.1 Duplex DNA</td>
<td>19</td>
</tr>
<tr>
<td>1.2.2 Triplex DNA</td>
<td>20</td>
</tr>
<tr>
<td>1.2.3 Structural RNA</td>
<td>20</td>
</tr>
<tr>
<td>1.2.4 G-quadruplex nucleic acid</td>
<td>21</td>
</tr>
<tr>
<td>1.3 Human Telomeres as a Source of G-rich DNA</td>
<td>26</td>
</tr>
<tr>
<td>1.3.1 Telomere maintenance and telomerase</td>
<td>28</td>
</tr>
<tr>
<td>1.3.2 The role of shelterin</td>
<td>29</td>
</tr>
<tr>
<td>1.3.3 Telomeres and cancer</td>
<td>31</td>
</tr>
<tr>
<td>1.3.4 G-quadruplexes formed from human telomeric DNA</td>
<td>32</td>
</tr>
<tr>
<td>1.3.4.1 High resolution structures of human telomeric DNA</td>
<td>33</td>
</tr>
<tr>
<td>G-quadruplexes</td>
<td>37</td>
</tr>
<tr>
<td>1.3.4.2 Structural diversity of telomeric DNA G-quadruplexes</td>
<td>37</td>
</tr>
<tr>
<td>1.3.4.3 Evidence for the existence of (telomeric) G-quadruplexes</td>
<td>38</td>
</tr>
<tr>
<td>in vivo</td>
<td>38</td>
</tr>
<tr>
<td>1.3.4.4 Telomeric DNA G-quadruplexes as anti cancer drug targets</td>
<td>39</td>
</tr>
<tr>
<td>1.3.4.5 Structural data for telomeric DNA G-quadruplex-ligand complexes</td>
<td>42</td>
</tr>
<tr>
<td>1.4 Other Sources of G-rich DNA in the Genome</td>
<td>47</td>
</tr>
<tr>
<td>1.4.1 G-quadruplex formation in promoter regions</td>
<td>47</td>
</tr>
<tr>
<td>1.4.2 Additional roles of DNA G-quadruplexes in the genome</td>
<td>48</td>
</tr>
</tbody>
</table>
## 1.5 Human Telomeres as a Source of G-rich RNA

1.5.1 G-quadruplex RNA in the genome

1.5.2 Telomeric RNA (TERRA/TelRNA) – A source of G-rich RNA

1.5.3 TERRA regulation and protein interactions

1.5.4 G-quadruplex structures formed from TERRA

1.5.5 High resolution structural studies of TERRA G-quadruplexes

1.5.6 Structural monomorphism of TERRA G-quadruplexes

1.5.7 Higher-order structure formation within TERRA sequences

1.5.8 Biological role(s) of TERRA G-quadruplexes?

1.5.9 TERRA G-quadruplexes as drug targets?

## 1.6 Other sources of G-rich RNA in the genome

1.6.1 G-quadruplexes in mRNA untranslated regions

1.6.1.1 Discovery and principles

1.6.1.2 mRNA 5' UTR G-quadruplexes as drug targets

1.6.2 G-quadruplexes in fragile X mental retardation syndrome

1.6.3 Additional roles of RNA G-quadruplexes in the genome

## 1.7 X-ray crystallography as a tool in G-quadruplex research

## 1.8 Overall aims, focus of thesis and summary of results

## CHAPTER 2 – X-RAY CRYSTALLOGRAPHIC STUDIES OF A NATIVE TELOMERIC RNA G-QUADRUPLEX

2.1 Background

2.1.1 RNA G-quadruplexes in the genome

2.1.2 Current high-resolution structural data available for RNA G-quadruplexes

2.1.3 Crystallographic structural data available for telomeric DNA G-quadruplexes

2.2 Aims

2.3 Methods and materials

2.3.1 Sample preparation

2.3.2 Crystallisation

2.3.3 Data collection and processing

2.3.4 Structure determination and refinement

2.3.5 Validation and deposition in the Protein Databank

2.3.6 Structure analysis and visualisation software

2.4 Results and discussion

2.4.1 Crystal structure of a telomeric RNA G-quadruplex: overall topology and general features
2.4.2 Propeller-loop structure ................................................................. 83  
2.4.3 Overall comparison to equivalent NMR and DNA structures ............. 85  
2.4.4 Sugar puckering and backbone torsion angles .................................... 87  
2.4.5 C2' hydroxyl group interactions and water structure ......................... 90

2.5 General discussion and conclusions ...................................................... 93  
2.5.1 Comparison of RNA and DNA G-quadruplex structures ..................... 93  
2.5.2 Implications for drug design ............................................................ 94

CHAPTER 3 – X-RAY CRYSTALLOGRAPHIC STUDIES OF A TELOMERIC RNA  
G-QUADRUPLEX ACRIDINE COMPLEX ............................................... 96

3.1 Background ........................................................................................... 97  
3.1.1 RNA G-quadruplexes as drug targets ................................................. 97  
3.1.2 Telomeric DNA G-quadruplexes as drug targets ................................. 98  
3.1.3 Acridines as telomeric G-quadruplex binding compounds ................... 99  
3.1.4 High-resolution G-quadruplex-acridine structures ...............................100  
3.1.5 Di-substituted triazole acridines: biophysics and biology .....................102

3.2 Aims ........................................................................................................ 104

3.3 Methods and materials ......................................................................... 105  
3.3.1 Sample preparation ............................................................................105  
3.3.2 Crystallisation ......................................................................................105  
3.3.3 Data collection and processing ...........................................................106  
3.3.4 Structure determination and refinement ..............................................107  
3.3.5 Validation and deposition in the protein databank .............................108

3.4 Results and discussion ..........................................................................109  
3.4.1 Crystal structure of a telomeric RNA G-quadruplex-acridine complex:  
overall topology and general features ....................................................109  
3.4.2 RNA quadruplex-acridine binding: overall mode of interaction ........111  
3.4.2.1 RNA G-quadruplex-acridine binding: π-π stacking .........................112  
3.4.2.2 RNA G-quadruplex-acridine binding: direct hydrogen bonds ..........114  
3.4.3 Propeller-loop rearrangement ............................................................114  
3.4.4 Propeller loop stabilisation: hydrogen bonding and water contacts ........118

3.5 General discussion and conclusions ..........................................................120  
3.5.1 Unprecedented RNA behaviour .........................................................120  
3.5.2 Implications for drug design ...............................................................122
CHAPTER 4 - X-RAY CRYSTALLOGRAPHY OF A PARTIAL TELOMERIC DNA G-QUADRUPLEX-ACRIDINE COMPLEX

4.1 Background ............................................................................................................. 125
4.2 Aims ........................................................................................................................ 126
4.3 Methods and materials .......................................................................................... 127
  4.3.1 Sample preparation ................................................................................. 127
  4.3.2 Crystallisation .......................................................................................... 127
  4.3.3 Data collection and processing .............................................................. 128
  4.3.4 Structure determination and refinement .............................................. 129
  4.3.5 Validation and deposition in the protein databank .............................. 130
4.4 Results and discussion ............................................................................................ 132
  4.4.1 Partial crystal structure of a telomeric DNA quadruplex-acridine complex:
    overall topology and general features ................................................ 132
  4.4.2 Quality of the electron density and resulting DNA-complex model .... 134
  4.4.3 Quality of the electron density: considerations .......... 135
  4.4.4 DNA G-quadruplex-acridine binding ....................................................... 136
4.5 General discussion and conclusions ..................................................................... 140
  4.5.1 Ligand binding mode: comparison to DNA-acridine complexes .......... 140
  4.5.2 Implications for drug design ................................................................... 143
    4.5.2.1 Implications for telomeric DNA G-quadruplex targeting .... 143
    4.5.2.2 Implications for telomeric RNA G-quadruplex targeting .... 144

CHAPTER 5 - X-RAY CRYSTALLOGRAPHIC STUDIES OF INTRAMOLECULAR TELOMERIC DNA G-QUADRUPLEX-NAPHTHALENE DIIMIDE COMPLEXES

5.1 Background ............................................................................................................. 147
  5.1.1 Modified naphthalene diimides as potent telomeric DNA G-quadruplex
    binding compounds .............................................................................. 147
  5.1.2 Crystal structures of naphthalene diimide
    G-quadruplex-complexes ..................................................................... 149
  5.1.3 A recent subset of modified naphthalene diimide compounds .......... 151
5.2 Aims ........................................................................................................................ 153
5.3 Methods and materials .......................................................................................... 154
  5.3.1 Sample preparation ................................................................................. 154
  5.3.2 Crystallisation .......................................................................................... 154
  5.3.3 Data collection and processing .............................................................. 155
  5.3.4 Structure determination and refinement .............................................. 156
  5.3.5 Validation and deposition in the protein databank .............................. 157
5.4 Results and discussion

5.4.1 Overview of results

5.4.2 Crystal structures of two telomeric DNA G-quadruplex-ligand complexes: overall topology and general features

5.4.3 Quality of electron density and resulting models

5.4.4 Naphthalene diimide-G-quadruplex binding mode – general

5.4.5 Telomeric DNA G-quadruplex-BMSG-SH3 complex

5.4.5.1 G-quartet stacking and symmetry stabilisation of BMSG-SH3

5.4.5.2 BMSG-SH3-groove interactions within the asymmetric unit

5.4.6 Telomeric DNA G-quadruplex-BMSG-SH4 complex

5.4.6.1 G-quartet stacking and symmetry stabilisation of BMSG-SH4 binding

5.4.6.2 BMSG-SH4-groove interactions within the asymmetric unit

5.4.7 Comparison of BMSG-SH3 and BMSG-SH4 binding modes

5.4.8 Diffraction data for the BMSG-SH5-quadruplex complex

5.5 General discussion and conclusions

CHAPTER 6 - BIOPHYSICAL STUDIES OF G-QUADRUPLEXES FORMED FROM HUMAN TELOMERIC RNA AND DNA

6.1 Background

6.1.1 Tertiary structure of human telomeric G-quadruplexes

6.1.1.1 Models of telomeric DNA G-quadruplex tertiary structure

6.1.1.2 Models of telomeric RNA G-quadruplex tertiary structure

6.1.2 Human telomeric G-quadruplex-ligand interactions

6.1.3 Practical considerations

6.1.3.1 Electrospray ionisation mass spectrometry – utility and validation

6.1.3.2 Circular dichroism – utility and limitations

6.1.3.3 Polyacrylamide gel electrophoresis – utility and limitations

6.2 Aims

6.3 Methods and materials

6.3.1 Electrospray ionisation mass spectrometry

6.3.1.1 Sample preparation

6.3.1.2 Mass spectrometer parameters and data collection

6.3.2 Circular dichroism

6.3.3 Polyacrylamide gel electrophoresis

6.3.4 Additional
# LIST OF FIGURES AND TABLES

## CHAPTER 1 - INTRODUCTION

| Figure 1.1 | Fundamentals of duplex and G-quadruplex nucleic acids | 22 |
| Figure 1.2 | Structural aspects of G-quadruplexes | 25 |
| Figure 1.3 | Proteins and G-quadruplexes at human telomeres | 27 |
| Figure 1.4 | Representative structures of human telomeric G-quadruplexes | 36 |
| Figure 1.5 | Human telomeric G-quadruplex binding compounds | 42 |
| Figure 1.6 | NMR-derived telomeric G-quadruplex ligand complex structures | 43 |
| Figure 1.7 | Human telomeric G-quadruplex-ligand complexes solved by X-ray crystallography | 45 |
| Figure 1.8 | Summary of the main cellular roles in which RNA G-quadruplexes have been implicated | 51 |
| Figure 1.9 | Origin of telomeric RNA (TERRA/telRNA) | 52 |
| Figure 1.10 | NMR structure of a G-quadruplex formed from telomeric RNA/TERRA | 55 |
| Figure 1.11 | Telomeric G-quadruplex-drug binding | 59 |
| Figure 1.12 | G-quadruplex formation in 5' untranslated regions of mRNA | 61 |

| Table 1.1 | Summary of current structural models of telomeric G-quadruplex-ligand complexes | 46 |
| Table 1.2 | Outline of thesis | 69 |

## CHAPTER 2 - X-RAY CRYSTALLOGRAPHIC STUDIES OF A NATIVE TELOMERIC RNA G-QUADRUPLEX

| Figure 2.1 | RNA G-quadruplex crystals formed from the sequence, r(UBrAGGGUUAGGGU) | 76 |
| Figure 2.2 | X-ray diffraction from a native RNA G-quadruplex crystal | 76 |
| Figure 2.3 | Overall folding and topology of the telomeric RNA G-quadruplex crystal structure | 81 |
| Figure 2.4 | Two asymmetric units associating through UAUA tetrads | 82 |
| Figure 2.5 | Electron density for the RNA G-quadruplex | 82 |
| Figure 2.6 | Crystal packing interactions: implications for loop structure | 84 |
| Figure 2.7 | Hydrogen bonding within the loop region | 85 |
| Figure 2.8 | Structural overlays of telomeric G-quadruplex structures | 86 |
| Figure 2.9 | Comparison of backbone torsion angles of RNA and DNA G-quadruplex crystal structures | 89 |
| Figure 2.10 | Influence of the C2' hydroxyl group on sugar pucker conformation | 90 |
| Figure 2.11 | Hydrogen bonding and water structure within DNA and RNA G-quadruplex groove regions | 91 |
| Figure 2.12 | Correlation between sugar pucker and C2' hydroxyl group interaction | 92 |
| Figure 2.13 | C2'-OH group interactions: G-quartet and backbone stabilisation | 92 |
Table 2.1. Data collection and refinement statistics for the telomeric RNA quadruplex crystal structure ..............................................................79

Table 2.2. Distribution of sugar puckering for three related G-quadruplex structures ............................................................................................ 88

CHAPTER 3 - X-RAY CRYSTALLOGRAPHIC STUDIES OF A TELOMERIC RNA G-QUADRUPLEX ACRIDINE COMPLEX

Figure 3.1 Chemical structure of the acridine core .................................................................99
Figure 3.2 DNA G-quadruplex acridine crystal structures ..................................................102
Figure 3.3 Triazole-acridine series of G-quadruplex binding compounds .........................103
Figure 3.4 A single crystal of the RNA G-quadruplex-acridine complex .............................106
Figure 3.5 X-ray diffraction pattern of the RNA-quadruplex-acridine complex .................107
Figure 3.6 Overall topology and structure of the RNA quadruplex acridine complex .........110
Figure 3.7 Electron density for the RNA-acridine complex ..................................................111
Figure 3.8 Crystal packing interactions ...............................................................................113
Figure 3.9 Purine octet binding surface of the RNA G-quadruplex .......................................113
Figure 3.10 Acridine-RNA binding .....................................................................................115
Figure 3.11 From intermolecular quadruplex to intramolecular quadruplex ......................116
Figure 3.12 Loop remodelling of acridine complex compared to native RNA ..................117
Figure 3.13 Stabilisation of the octet adenine residues .......................................................119
Figure 3.14 G-quadruplex groove inaccessibility ...............................................................122

Table 3.1 Data collection and refinement statistics for the RNA G-quadruplex-acridine complex .........................................................................108
Table 3.2 Comparison of ribose puckering of native and acridine-bound RNA G-quadruplexes ........................................................................118

CHAPTER 4 - X-RAY CRYSTALLOGRAPHY OF A PARTIAL TELOMERIC DNA G-QUADRUPLEX-ACRIDINE COMPLEX

Figure 4.1 X-ray diffraction of DNA-acridine complex crystals .........................................129
Figure 4.2 Overall folding and features of the DNA G-quadruplex-acridine complex .........133
Figure 4.3 Electron density maps for the DNA-acridine complex .......................................135
Figure 4.4 Crystal packing and ligand binding ....................................................................139
Figure 4.5 Comparison of acridine-DNA quadruplex binding modes ................................142

Table 4.1 Data collection and refinement statistics for the DNA G-quadruplex-acridine complex .......................................................................131
Table 4.2 FRET-melting data for acridine-based G-quadruplex binding compounds tested at 1 μM (°C) ....................................................................141
CHAPTER 5 - X-RAY CRYSTALLOGRAPHIC STUDIES OF INTRAMOLECULAR TELOMERIC DNA G-QUADRUPLEX-NAPHTHALENE DIIMIDE COMPLEXES

Figure 5.1 Naphthalene diimide core................................................................. 148
Figure 5.2 Representative members from a series of tetra-substituted naphthalene diimide compounds................................................................. 149
Figure 5.3 DNA G-quadruplex-naphthalene diimide crystal structures........... 151
Figure 5.4 Methyl-piperazine modified ND compounds................................. 152
Figure 5.5 DNA G-quadruplex-naphthalene diimide complex crystals............. 155
Figure 5.6 X-ray diffraction patterns of the DNA-quadruplex-naphthalene complex crystals................................................................. 158
Figure 5.7 Overall folding and features of complexes formed between a human telomeric G-quadruplex and the ND compounds BMSG-SH3 and BMSG-SH4................................................................. 162
Figure 5.8 Electron density for the BMSG-SH3-G-quadruplex complex............ 164
Figure 5.9 Electron density for the BMSG-SH4-G-quadruplex complex............ 165
Figure 5.10 General mode of DNA 22-mer – naphthalene diimide interaction.. 166
Figure 5.11 Symmetry-related base stacking interactions: BMSG-SH3............. 167
Figure 5.12 Hydrogen bonding contacts between BMSG-SH3 and the DNA G-Quadruplex........................................................................................................ 168
Figure 5.13 BMSG-SH3-phosphate electrostatic interactions.......................... 170
Figure 5.14 Hydrogen bonding contacts between BMSG-SH4 and the DNA G-Quadruplex........................................................................................................ 172
Figure 5.15 Symmetry-related base stacking interactions: BMSG-SH4............ 173
Figure 5.16 Interactions between BMSG-SH4 sidechains and quadruplex groove components................................................................. 174
Figure 5.17 Ligand B-factor distribution.......................................................... 176

Table 5.1 Data collection and refinement statistics for three telomeric DNA-ligand complexes................................................................. 159
Table 5.2 Summary of crystallographic results for the three ND compounds used in these studies................................................................. 160

CHAPTER 6 - BIOPHYSICAL STUDIES OF G-QUADRUPLEXES FORMED FROM HUMAN TELOMERIC RNA AND DNA

Figure 6.1 Models of telomeric DNA G-quadruplex tertiary structure............ 184
Figure 6.2 Proposed model of TERRA G-quadruplex higher-order structure.... 185
Figure 6.3 Identification of G-quadruplex formation in mass spectra: example........................................................................................................ 188
Figure 6.4 ESI-mass spectra: Strand stoichiometry and G-quadruplex detection for telomeric RNA and DNA sequences................................................................. 196
Figure 6.5 Stability of monomeric G-quadruplex species................................ 197
Figure 6.6 Stability of dimeric G-quadruplex species..................................... 199
Figure 6.7 Circular dichroism spectra for annealed telomeric RNA and DNA sequences................................................................. 201
Figure 6.8  Native polyacrylamide gels of telomeric RNA and DNA in various monovalent cations. ................................................................. 203
Figure 6.9  Partially denaturing PAGE of RNA22 sequence. .................. 205
Figure 6.10 Details of naphthalene diimide compounds used in this study. 205
Figure 6.11 ESI-MS analysis of telomeric RNA and DNA ligand interactions. 207
Figure 6.12 Crystallographic data for a telomeric RNA G-quadruplex-FC4ND-10 complex. ................................................................. 213

Table 6.1  Oligonucleotide sequences used for ESI-MS, CD and PAGE analysis. ................................................................. 192
Table 6.2  Details of small-molecule ligands used for ESI-MS analysis. 192
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>Angstrom ($= 10^{-10}$ metres)</td>
</tr>
<tr>
<td>A</td>
<td>Adenine nucleotide</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CD</td>
<td>Circular Dichroism</td>
</tr>
<tr>
<td>C-rich</td>
<td>Cytosine rich</td>
</tr>
<tr>
<td>d</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>ESI-MS</td>
<td>Electrospray Ionisation Mass Spectrometry</td>
</tr>
<tr>
<td>FRET</td>
<td>Fluorescence Resonance Energy Transfer</td>
</tr>
<tr>
<td>G</td>
<td>Guanine nucleotide</td>
</tr>
<tr>
<td>G-quadruplex</td>
<td>Guanine-quadruplexes</td>
</tr>
<tr>
<td>G-quartet</td>
<td>Guanine-quartet</td>
</tr>
<tr>
<td>G-rich</td>
<td>Guanine-rich</td>
</tr>
<tr>
<td>HPLC</td>
<td>High Performance Liquid Chromatography</td>
</tr>
<tr>
<td>hTER</td>
<td>Human Telomerase RNA (also ‘hTR’ and ‘hTERC’)</td>
</tr>
<tr>
<td>hTERC</td>
<td>Human Telomerase RNA Component (also ‘hTR’ and ‘hTER’)</td>
</tr>
<tr>
<td>hTERT</td>
<td>Human Telomerase Reverse Transcriptase</td>
</tr>
<tr>
<td>hTR</td>
<td>Human Telomerase RNA (also ‘hTER’ and ‘hTERC’)</td>
</tr>
<tr>
<td>IC$_{50}$</td>
<td>Inhibitory Concentration: concentration of a compound required to inhibit a biological process by 50%</td>
</tr>
<tr>
<td>kbp</td>
<td>Kilobase pair</td>
</tr>
<tr>
<td>MALDI-ToF</td>
<td>Matrix Assisted Laser Desorption Ionisation Time of Flight</td>
</tr>
<tr>
<td>MPD</td>
<td>2-methyl-2,4-pentanediol</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>Mass Spectrometry</td>
</tr>
<tr>
<td>ND</td>
<td>Naphthalene Diimide</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>PAGE</td>
<td>Polyacrylamide Gel Electrophoresis</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Databank</td>
</tr>
<tr>
<td>PEG</td>
<td>Polyethylene glycol</td>
</tr>
<tr>
<td>POT1</td>
<td>Protection of telomeres 1 (shelterin protein)</td>
</tr>
<tr>
<td>r</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Square Deviation</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>T</td>
<td>Thymine nucleotide</td>
</tr>
<tr>
<td>TelRNA</td>
<td>Telomeric RNA (also ‘TERRA’)</td>
</tr>
<tr>
<td>TERRA</td>
<td>Telomeric Repeat-Containing RNA (also ‘telRNA’)</td>
</tr>
<tr>
<td>T$_m$</td>
<td>Melting temperature</td>
</tr>
<tr>
<td>TRAP</td>
<td>Telomere Repeat Amplification Protocol</td>
</tr>
<tr>
<td>TRF1</td>
<td>Telomeric Repeat Binding Factor 1 (shelterin protein)</td>
</tr>
<tr>
<td>TRF2</td>
<td>Telomeric Repeat Binding Factor 2 (shelterin protein)</td>
</tr>
<tr>
<td>U</td>
<td>Uracil nucleotide</td>
</tr>
<tr>
<td>U$_{br}$</td>
<td>5-bromo-uracil nucleotide</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region (of mRNA)</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
</tbody>
</table>
1.1 PROTEINS vs NUCLEIC ACIDS

The fundamental and ubiquitous role of proteins in biological processes has been fully appreciated for many years now, illustrated by the fact that most human pathologies have a protein-related basis, and consequently the majority of pharmaceutical drugs currently available are directed towards specific protein-targets. While proteins do indeed satisfy the majority of a cell's catalytic, structural and regulatory needs, the contribution of folded nucleic acid structures towards these processes is gaining appreciation. While the main view of nucleic acid structure tends to be the double-helical duplex form of DNA — discovered by Watson, Crick, Franklin and Gosling\textsuperscript{1-4}, and subsequently proved to be the storage format of human genetic information — there is in fact a quite incredible range of three-dimensional shapes and motifs available to DNA and RNA, which are involved in almost as wide a range of cellular processes as proteins. The existence and roles of nucleic acid structures are still being discovered, perhaps surprising considering that the full mapping of the human genome was completed over seven years ago\textsuperscript{5}. This may in part be due to the difficulty in predicting three-dimensional shape from a nucleic acid sequence (consider, for example, the RNA pseudo-knot motif) — a technique which has received considerable attention in the field of protein structure and function\textsuperscript{6}. Considering the difficulty in understanding and predicting the final three-dimensional shape of a given DNA or RNA molecule based on its sequence alone, high-quality experimental data is exceedingly valuable in providing information and answers regarding nucleic acid folding and structure. As with protein-based research, the two main techniques for determining atomic-scale details of macromolecules — NMR spectroscopy and X-ray crystallography — have contributed significantly to the understanding of nucleic acid structure and function\textsuperscript{7-11}. The work described in this thesis has been aimed at building upon the relatively limited amount of high-resolution structural data available for nucleic acids, using primarily X-ray crystallography, and focusing on an unusual form of nucleic acid — the G-quadruplex — a motif of emerging importance in a range of cellular process and pathologies, including gene expression, telomere regulation and crucially, cancer biology.
1.2 STRUCTURAL DIVERSITY OF NUCLEIC ACIDS

1.2.1 DUPLEX DNA

As mentioned above, probably the most familiar structural form of nucleic acid is the double-helical Watson-Crick duplex form of DNA. While several structural forms of duplex DNA exist—depending upon hydration levels and, to some degree, the experimental technique used—the key features of duplex DNA are shared between the forms, these being the formation of a two-stranded structure with the nucleic acid strands arranging in an anti-parallel orientation, and held together by hydrogen bonding interactions between A:T and G:C base pairs (figure 1.1a). This results in a double helical structure with non-equivalent grooves—major and minor—both of which have been targeted extensively with small-molecule ligands, with the aim of disrupting the expression of specific genes at the DNA level. The biologically relevant forms of duplex DNA include A-DNA, B-DNA and Z-DNA—with B-DNA being the form determined by Watson and Crick, and the main form thought to exist in cells. B-DNA is 20 Å wide, with a height of 3.38 Å per base-pair step (i.e. the distance between two stacked bases), and a right handed helical twist of 36.7°, which makes a complete turn about its axis every 10.5 base pairs. A-DNA is essentially a dehydrated form of DNA, being observed primarily in crystallographic experiments. A-DNA, again, is right-handed, but with reduced helical twist and base-pair rise dimensions compared to B-DNA (32.7° and 2.54 Å, respectively). Due to the reduced helical twist of A-DNA, a complete helical turn involves 11 base-pairs for this form of DNA. The third main biologically relevant form of duplex DNA, Z-DNA, differs more drastically from B-DNA in that the double-helix adopts a left-handed twist. Although several Z-DNA binding proteins have been identified, the biological role of this form of DNA is yet to be fully determined.

Duplex DNA, being the format utilised by most higher-order organisms for the storage of genetic information, represents arguably one of the most fundamentally important biological structures ever determined. X-ray diffraction played no small part in the initial determination (and subsequent refinement) of the duplex model for DNA structure. Initial X-ray fibre diffraction data were later enhanced with single-
crystal x-ray diffraction data\textsuperscript{18}, which was followed by a number of studies into duplex-ligand complexes involving therapeutically promising groove-binding and intercalating agents\textsuperscript{12,19,20}.

1.2.2 TRIPLEX DNA

The ability of three separate nucleic acid strands to associate stably into a ‘triplex’ helical structure was first shown in 1976 through fibre diffraction techniques\textsuperscript{21}. It was subsequently shown that, in certain sequence contexts, a third nucleic acid strand can hydrogen bond to a duplex entity, forming a stable triple-helical structure. The third strand binds to the duplex through the Hoogsteen base hydrogen bonds, and can associate in either a parallel or anti-parallel orientation, depending on the sequence (and therefore available hydrogen bonding contacts)\textsuperscript{15}. Although triplexes have not been implicated convincingly in cellular processes, they have shown potential for use as gene-specific therapeutic agents\textsuperscript{22,23}. The principal of such gene regulation – termed the ‘anti-sense approach’ – involves the design of an oligonucleotide sequence with triple-base complementarity for a target sequence – the hybridisation of such a sequence to the target duplex sequence (e.g. in a promoter or gene-expression control sequence) would then be expected to disrupt (or modulate) gene expression\textsuperscript{15}. Despite the significant potential of triplex-based therapeutics, two key obstacles which limit progress include the sequence-specific nature of triplex formation, and the sensitivity of administered oligonucleotides to nuclease degradation.

1.2.3 STRUCTURAL RNA

In contrast to the relatively limited number of structural forms available to deoxyribonucleic acid, a huge number of structural motifs and folds have been identified and characterised for ribonucleic acid molecules. The structural diversity of RNA is linked to the wider range of cellular activities to which RNA molecules contribute, compared to DNA. Some of the key folded RNA structures include transfer RNAs\textsuperscript{24,25} (the first large RNA structure solved by X-ray diffraction), ribosomal RNA\textsuperscript{26,27}, ribonucleases\textsuperscript{28}, ribozymes\textsuperscript{29,30} and riboswitches\textsuperscript{31}. As with proteins, many large RNA
structures are assembled using a toolbox of discrete motifs and domain structures. For RNA, such motifs and domain folds include hairpins, duplexes, triplexes, bulges, junctions, tetra-loops, ribose-zippers and pseudoknots. Importantly, many folded RNA structures possess catalytic activity, making RNA an intermediate between DNA and proteins in terms of shared function as well as an intermediate in the transfer of information (i.e. as the mediator between DNA and proteins, as stated in the central dogma of molecular biology: DNA > RNA > protein). The range of structures and functions available to RNA molecules and sequences highlights the importance of three-dimensional nucleic acid structures in the cell and the need to understand the folding of nucleic acid molecules at atomic scale resolution.

1.2.4 G-QUADRUPLEX NUCLEIC ACID

While the majority of cellular DNA is expected to exist in the double-helical duplex form composed of stacked Watson-Crick hydrogen bonded base pairs, there exists an alternative hydrogen bonding arrangement available to both DNA and RNA – the G-quadruplex – a four-stranded structural form of nucleic acid which is based upon the guanine-quartet motif (termed G-quartet or G-tetrad) (figure 1.1b). The G-quartet is a square-planar arrangement of four guanines, held together through hydrogen bonding interactions utilising the Watson-Crick and Hoogsteen faces of the guanine bases. The existence of G-quartets was discovered in 1962 by Gellert and coworkers, who used fiber diffraction experiments to show that the previously observed gel formation of concentrated solutions of guanylic acid was the result of strong guanine-guanine base associations in the form of G-quartet formation. G-quadruplexes are formed from the stacking of two or more G-quartets, resulting in highly stable four-stranded helical structures stabilised by intra-G-quartet hydrogen bonding and π-π stacking interactions. Additionally, monovalent cations play an integral role in G-quadruplex stabilisation; they can coordinate to the O6 oxygens of the guanine bases, typically conferring significant stabilisation to the structure. The ionic radius of the coordinating cation has a significant influence on its G-quadruplex stabilising effects, with the accepted order of stabilisation (from high to low) being $K^+ > Na^+ > Rb^+ > NH_4^+ > Cs^+ > Li^+$.
**Figure 1.1 Fundamentals of duplex and G-quadruplex nucleic acids.**

(a) Duplex DNA is formed from two strands, arranged as an anti-parallel right-handed helix, held together through a combination of Watson-Crick base-pairing hydrogen bonding and π-π base stacking forces. Duplex DNA has two non-equivalent grooves, in which coordinated Mg$^{2+}$ ions are often observed. (b) G-quadruplexes can be formed from one, two or four guanine-rich nucleic acid strands, which can arrange to form a variety of parallel, anti-parallel or mixed (parallel/anti-parallel) right-handed helical structures, based upon π-π stacked G-quartets. G-quartets are held together through base-base hydrogen bonding, involving the Watson-Crick and Hoogsteen faces of the bases. G-quadruplex groove types can be classed as either narrow, medium or wide. Monovalent cations play a pivotal role in G-quadruplex stabilisation via G-quartet coordination (a coordinating potassium ion is shown as a purple sphere). A table summarising the key differences between duplex and G-quadruplex nucleic acids is shown at the bottom. PDB files 3OEI and 2J7Z were used to generate the cartoon representations shown in (a) and (b) respectively.
In contrast to duplex DNA, the strand orientation within G-quadruplexes may be parallel, anti-parallel, or a mix of parallel and anti-parallel, which will depend upon sequence context, ionic conditions, cation identity and, of emerging importance, the presence of molecular crowding conditions (discussed further below 1.3.4.3). Strand orientation is intimately linked to the torsion angle of the glycosidic bond, which covalently links the guanine base to the deoxyribose or ribose sugar of the nucleic acid backbone. The glycosidic torsion angles (χ) for the guanines of an all-parallel G-quadruplex are typically in the anti range (-120° > χ > 180°), with an anti-parallel G-quadruplex topology requiring the adjacent bases to flip to the syn glycosidic conformation (0° < χ < 90°), in order to allow appropriate Hoogsteen hydrogen bond formation. The distribution of glycosidic angles of the guanine bases (and therefore strand orientation) will affect G-quadruplex groove widths, which are different from the two grooves of duplex DNA (i.e. major and minor), in that G-quadruplexes typically involve four grooves of varying width – defined as narrow, medium or wide.

Although G-quadruplexes are described as ‘four-stranded’ structures, they can in fact be single-, double- or four-stranded structures, depending on the sequences involved (figure 1.2). For example, a sequence such as d(TGGGGT) associates to form a stable four-stranded G-quadruplex with four stacked G-quartets, whereas a sequence such as d(AGGGTTAGGGTTAGGGTTAGGG) folds as a single-stranded (i.e. unimolecular) G-quadruplex, with three stacked G-quartets (underlined) linked through TTA trinucleotide sequences (figure 1.2c and d). Nucleotide stretches which link the G-quartets are termed ‘loops’, even if such a ‘loop’ is a single nucleotide in length. There are three types of loops seen in G-quadruplex structures:

**Lateral** – the simplest loop arrangement, linking two adjacent strands, resulting in a relative anti-parallel strand orientation (figure 1.2d).

**Diagonal** – a loop linking two strands across a G-quartet face. The resulting strands will have a relative anti-parallel orientation (figure 1.2d).

**Propeller** – a loop linking guanines from opposite faces of a G-quadruplexes.

This loop maintains a parallel strand orientation (figure 1.2d).
The fundamental requirement for G-quadruplex formation is the presence of a sufficient number of guanine residues in a given sequence to allow the formation of at least two G-quartets. Such guanine-rich sequences with the potential to form these highly stable G-quadruplex structures have been located throughout the human genome\textsuperscript{41,42}, and more recently, have also been identified within the human transcriptome\textsuperscript{43,44}. One of the key areas of the human genome in which G-quadruplexes have been implicated are the terminal regions of the chromosomes – the telomeres – which are intimately involved in cellular apoptosis and senescence regulation, and consequently, are of high importance to cancer biology. Telomeric regions contain both G-rich RNA and G-rich DNA sequences with the potential to form stable G-quadruplex structures – the work described in this thesis has been aimed at providing structural data for these G-quadruplexes, with an emphasis on the importance of such structures to cancer research and their potential as drug-targets.
a) Four strands  
(tetramolecular)

b) Two strands  
(bimolecular)

c) Single strand  
(unimolecular/intramolecular)

d) Loop types

Figure 1.2 Structural aspects of G-quadruplexes.  
a-c) G-quadruplexes may be formed from the association of one, two or four guanine-rich nucleic acid strands. Each guanine base is represented by a coloured cuboid (coloured separately for each strand), with the backbones and backbone direction (from 5’ to 3’) indicated by the black arrows.  
d) Three main types of linking loops are observed in naturally occurring G-quadruplex forming sequences: lateral (linking two adjacent strands), diagonal (linking two strands across a G-quartet face) and propeller (linking guanines from opposite faces of a G-quadruplex). The type and arrangement of the loops ultimately defines the final G-quadruplex topology (i.e. the strand orientations).
1.3 HUMAN TELOMERES AS A SOURCE OF G-RICH DNA

Telomeres are the terminal regions of eukaryotic chromosomes and are composed of an intricate and carefully regulated nucleoprotein complex. The nucleic acid component of the telomere comprises thousands of non-coding tandem hexameric repeats of the sequence, d(TTAGGG)\textsuperscript{45}. These G-rich repeats exist as double-stranded DNA for approximately 10-12 kilobasepairs, with a 3\textquotesingle terminal single-stranded overhang of between 100 and 200 nucleotides\textsuperscript{46}. The DNA of the terminal overhang — being single-stranded and G-rich — is able to form stable G-quadruplex structures, which have been studied extensively for over 10 years now (discussed in full below, section 1.3.4) (figure 1.3). The protein component of the telomere consists primarily of a six-membered protein complex termed ‘shelterin’, which plays a major role in protecting and regulating telomeric regions\textsuperscript{47} (figure 1.3a). Shelterin, in complex with single- and double-stranded TTAGGG-repeated telomeric DNA, serves to protect the ends of eukaryotic/human chromosomes from:

1) Unwanted degradation;
2) Unwanted double-strand break repair processes, and;
3) Chromosomal end-to-end fusions.

Although perhaps not immediately apparent, these three processes are linked closely with two key cellular states: cellular senescence and genome instability. Consequently, telomere biology is of fundamental importance to two areas of pathological research, namely; cancer and aging. The role of telomere regulation in cancer biology is one of the main themes of the work described in this thesis — specifically, the role of G-quadruplexes formed from telomeric nucleic acid in cancer biology. However, before the significance of telomeric G-quadruplex structures in cancer research is considered, an overview of telomere regulation will first be given.
INTRODUCTION – CHAPTER 1

a) Telomere length regulation
2) Telomere capping (protection)
3) DNA damage response

b) Stabilisation of G-quadruplex with drug molecule

Figure 1.3 Proteins and G-quadruplexes at human telomeres.
a) Figure showing the localisation and interactions of the six shelterin proteins and telomerase at the telomere. Shelterin includes proteins which are able to specifically recognise both single-stranded and double-stranded telomeric repeat DNA. Shelterin largely performs a protective role at the telomere. Telomerase is responsible for actively adding d(TTAGGG) repeats to the 3' single-stranded overhang, thereby compensating for the loss of DNA following each round of replication. b) Model for G-quadruplex formation within the G-rich single-stranded overhang of human telomeres. G-quadruplex formation (blue cuboids) would be expected to disrupt telomerase function by sequestering its template (i.e. single-stranded telomeric DNA) and well as by displacing components of the Shelterin complex, such as POT1. Additional stabilisation of telomeric G-quadruplexes with small-molecule therapeutic compounds (yellow) is a viable route to artificially modulating telomerase activity, as a means to selectively target cancer cells.
1.3.1 TELOMERE MAINTENANCE AND TELOMERASE

The existence of extended, non-coding repeated G-rich sequences at the ends of linear chromosomes was first reported by Blackburn and Gall in 1978, who showed the linear chromosomes of *Tetrahymena thermophila* to be composed of tandem d(GGGGTT) repeats\(^48\). Following from this, the existence of repeated G-rich sequences were identified in a number of organisms, including humans\(^45\) - showing repeated G-rich motifs at telomeres to be a phenomenon conserved throughout a diverse range of eukaryotic organisms.

Functional investigations into the telomeres of normal human somatic cells revealed these regions to shorten by around 100 base pairs (bps) following each round of replication, eventually resulting in cellular senescence or apoptosis when a critically short telomere length was reached\(^49,50\). Telomere shortening is a result of the inability of DNA polymerase to fully replicate the terminal hundred basepairs of the chromosome (termed 'the end replication problem'), a phenomenon which was proposed to be a 'mitotic clock' mechanism, by which telomere length regulates the age of a cell\(^51,52\). This mechanism involves the cell identifying critically short telomeres as regions of genomic instability, consequently triggering DNA damage response pathways, leading ultimately to cellular senescence (i.e. no further replication) or apoptosis (i.e. programmed cell death)\(^53\).

Importantly, it was discovered that the telomeres of stem cells - which possess extended replicative ability - shortened at a far reduced rate compared to somatic cells\(^50\). Details of the mechanism behind such telomere-maintenance were provided by seminal work performed by Greider and Blackburn, who discovered a reverse-transcriptase enzyme in ciliates – telomerase – to be responsible for the maintenance of telomere length\(^54\). Human telomerase was identified soon after, comprising a heterodimeric enzyme composed of a catalytic protein component, hTERT, and a structural RNA component, hTR (which contains the template sequence for telomeric DNA repeat addition)\(^55\). The structure and function of telomerase has since been studied extensively, and is now known to function by adhering to single-stranded telomeric DNA and actively adding TTAGGG repeats to the 3' ends of mammalian chromosomes, using the intrinsic CAAUCCCAAUC motif of the RNA subunit (hTR)\(^49,56\).
(figure 1.3a). It is through this mechanism of telomere extension that telomerase is responsible for the extended replicative potential of stem cells, as well as for the immortalised phenotype developed by many cancers\textsuperscript{56}. Indeed, around 85\% of human cancers tested were shown to have re-activated telomerase activity\textsuperscript{57}, making this enzyme a highly attractive target for anti-cancer drug-design.

### 1.3.2 THE ROLE OF SHELTERIN

The shelterin complex consists of six unique adapter proteins which bind and protect telomeric DNA by forming a protective molecular ‘cap’. This cap ensures the telomeric DNA does not trigger a DNA damage response pathway, as well as preventing chromosomal end-to-end fusions and exonucleolytic degradation\textsuperscript{58,59}. Shelterin is also intimately involved in the regulation of telomerase activity, primarily by controlling access of telomerase to the telomere (which is the substrate of telomerase). The shelterin proteins are able to specifically recognise and bind single-stranded telomeric repeat DNA, double-stranded-repeat DNA and telomerase (figure 1.3a). The shelterin proteins include:

1) **TRF1** (telomeric repeat binding factor 1): TRF1 specifically interacts with double-stranded telomeric repeat DNA as a homodimer, and is considered to be a negative regulator of telomere length. TRF1 is anchored to the shelterin complex though interactions with TIN2 (see below).

2) **TRF2** (telomeric repeat binding factor 2): a homologue of TRF1, TRF2 also binds double-stranded telomeric repeat DNA as a homodimer and is involved in regulation of telomere length. Interactions with other shelterin proteins include TIN2 and RAP1 (see below).

3) **TIN2** (TRF-interacting nuclear factor 2): TIN2 lacks DNA-binding abilities but rather serves as a key anchor in the shelterin complex – as well as binding TRF1 and 2, TIN2 also interacts with TPP1. TIN2 is considered a negative regulator of telomere length.
4) **RAP1** (TRF2-interacting protein): lacking DNA binding activity, RAP1 regulates telomere length through TRF2 binding.

5) **POT1** (protection of telomeres-1): POT1 binds selectively to single-stranded telomeric repeat DNA and is anchored to the shelterin complex through the TPP1 protein. POT1 is intimately involved in the control of telomere length (both positively and negatively) and has been shown to resolve/disrupt telomeric G-quadruplex structures.

6) **TPP1** (also called PTOP – POT1 and TIN2 organising protein): TPP1 anchors POT1 to the shelterin complex. TPP1 controls the recruitment of POT1 to the telomeres, and has been shown to interact directly with telomerase.

A considerable volume of work has been reported concerning the shelterin complex and its role in telomere biology and cancer progression, a significant portion of which has involved mouse-knockout studies, which have highlighted the protective role of the shelterin proteins at the telomere. Additional data suggests shelterin protects telomeres through the stabilisation of the ‘T-loop’ structure – an arrangement detected by electron microscopy which involves the ‘folding back’ and intercalation of the single-stranded 3’ overhang into the sub-terminal duplex G-rich region of the telomere. Based on the available data, it appears that telomere-length in telomerase-active cells (i.e. cancer and stem cells) is regulated via a feedback mechanism, whereby long telomeres result in increased shelterin recruitment, which blocks further telomerase-driven telomere extension. Conversely, shorter telomeres are associated with reduced shelterin loading, thereby increasing telomerase-driven telomere extension. The role of shelterin in telomerase-inactive cells (i.e. somatic cells) appears to be limited to the protection of genomic integrity, such as the prevention of chromosomal end-to-end fusions and double-strand break repair processes. In telomerase inactive cells, telomere length eventually shortens such that the association of shelterin is prevented, and hence, apoptotic pathways are triggered as a result of genomic instability.
Although many of the interactions and effects of the shelterin proteins are known, the precise role of each protein remains to be fully understood. This is almost certainly due to the complexity of telomere biology, for example, the functions of shelterin proteins (and other regulatory proteins) can switch (i.e. from negative regulators to positive regulators of telomere length) depending on the status of the telomere\(^\text{61,62}\). Additionally, a vast number of additional signalling, adapter and enzymatic proteins are involved in controlling telomere status (for example PINX1, Apollo, Tankyrase-1, Ku70/80\(^\text{63-65}\)), complicating the signalling networks further. It is quite clear, therefore, that telomere regulation is a complex process, and further research will no doubt shed more light on the full role(s) of each telomere-related protein. Perhaps the factor of shelterin function most pertinent to the focus of this thesis is the observation that POT1 is able to resolve (i.e. ‘unwind’) telomeric G-quadruplex structures\(^\text{66}\), providing evidence for a biologically relevant role for such G-quadruplex structures.

### 1.3.3 Telomeres and Cancer

Cancer refers to a group of diseases characterised by the unrestricted and excessive proliferation of cells. In order for a cancer to develop, the tumorigenic cells must bypass the cellular mechanisms which have evolved to prevent uncontrolled cell growth – these include the evasion of both cell death (apoptotic) signals and cellular senescence signals. The circumvention of cellular senescence (and ultimately cell death) relies largely upon mutations in the p53 and pRb (retinoblastoma protein) tumour suppression pathways\(^\text{56}\). Loss of tumour suppression signals, however, is not sufficient for the immortalisation of cancerous cells – cellular immortalisation requires the reactivation of a telomere-stabilisation mechanism. This is because, as mentioned, the inability of normal cells to fully replicate the terminal regions of their linear chromosomes results in successive telomere shortening, eventually leading to programmed cell death when a critically short length is reached. There are two primary mechanisms by which cancerous cells can avoid cell death as a consequence of telomere-depletion, which involve either 1) the re-activation of telomerase, or 2) the activation of ALT (alternative lengthening of telomeres) pathways\(^\text{67}\). Telomerase re-
activation appears to be the primary method by which cancer cells avoid telomere shortening – with around 85% of human cancers displaying upregulated telomerase activity\(^5\). Because of the high incidence of telomerase activity in cancer cells (along with low-to-zero activity in normal somatic cells), telomerase is considered to be a highly attractive anti-cancer drug target. Many such anti-telomerase strategies exist\(^6\), which include the design of therapeutic compounds to selectively bind and inhibit: 1) the protein component of telomerase, hTERT; 2) the RNA component of telomerase, hTR; 3) telomerase regulatory proteins (such as POT1, TPP1, TRF1/2) and 4) the substrate of telomerase – telomeric DNA. The final strategy is of key importance to the work reported here, and will be described in detail below. Briefly – the principle of telomeric-DNA targeting (for telomerase inhibition) is based upon the stabilisation of G-quadruplex structures formed from the single-stranded 3' overhang of telomeres with small molecule compounds, with the aim of disrupting telomerase binding/activity, or indeed the binding or activity of any telomere-regulatory protein (such as POT1) (figure 1.3b). Before considering the methods, results and success of such a strategy, the drug-targets themselves – the telomeric G-quadruplexes – will first be considered.

### 1.3.4 G-QUADRUPLEXES FORMED FROM HUMAN TELOMERIC DNA

G-quadruplexes – four-stranded structures based upon stacked G-quartet motifs – can be formed from G-rich sequences composed of four runs of 3-5 guanines interspersed with connecting ‘loop’ sequences between (typically) 1-8 bases, i.e. \(G_xL_yG_xL_yG_x\) – where \(x = 3-5\) guanines and \(y = 1-8\) loop nucleotides. The human telomeric repeat sequence very comfortably satisfies these parameters – the minimal repeat-sequence, \(d(TTAGGG)\), when repeated four times (i.e. \([TTAGGG]_4\)) results in a stable G-quadruplex structure composed of three-stacked G-quartets, with three connecting loops composed of TTA trinucleotides, positioned external to the G-quartet core (figures 1.2c and 1.3b). Although four-repeats of the human telomeric sequence (e.g. \([TTAGGG]_4\)) would appear to allow a rather limited number of structural arrangements, there are in fact around ten distinct topological arrangements reported for sequences containing four telomeric repeats (figure 1.4). While the G-quartets (which can be

---

32
considered the 'core' of any G-quadruplex structure) remain consistent between the reported human telomeric structures, the strand orientation (i.e. parallel/anti-parallel) and loop arrangements are highly variable.

1.3.4.1 HIGH-RESOLUTION STRUCTURES OF HUMAN TELOMERIC DNA

G-QUADRUPLEXES

The first structural description of a G-quadruplex formed from human telomeric DNA was reported by Patel and co-workers in 1993, who determined a high-resolution NMR structure for the sequence, d(AGGG[TTAGGG])₃ (figure 1.4a, PDB entry 143D). [Note: all subsequent PDB entry codes are indicated in bold italics.] This structure, solved in sodium ionic conditions, was shown to fold as an intramolecular (i.e. unimolecular) G-quadruplex, with all four strands running anti-parallel, connected by three 'edge-wise' TTA loops (sometimes referred to as a 'basket type' quadruplex). This somewhat globular structure was rather different to the subsequent crystallographic structural determination of the same sequence, which revealed an all-parallel stranded G-quadruplex, with three propeller-type loops (figure 1.4d, 1KF1). An additional bimolecular structure was also solved by crystallography at the same time: the sequence d(TAGGGTTAGGGT) was also shown to fold as a parallel stranded bimolecular G-quadruplex with propeller-type loops (1KBP). The crystal structures caused somewhat of a controversy, as all biophysical (as well as NMR) data at the time pointed to an anti-parallel arrangement for human telomeric G-quadruplex structures. The main cause of the structural differences was ascribed to the differences in coordinating cation used: the NMR structure was solved in the presence of sodium, whereas the crystal structure was determined in potassium salt conditions. This suggested a mechanism of cation-induced topology shifting of telomeric G-quadruplexes, with understandable connotations for telomere biology. The importance of cation identity – while still important – appears to be overshadowed by recent investigations into the role of molecular crowding (and pseudo-molecular crowding) on G-quadruplex topology (discussed below).

Following from the crystal structure determinations, subsequent NMR structures solved in potassium conditions began to reveal the complex structural
diversity of telomeric G-quadruplexes. Two 'hybrid' structures were reported for two difference 26mer sequences: 'hybrid-1', formed from the sequence, d(AAAGGG[TTAGGG]_{3}AA)^{73} (2HY9) and 'hybrid-2', formed from the sequence, d([TTAGGG]_{4}TT)^{74} (2KPZ). Both structures display a mix of parallel and anti-parallel strand associations, with two edge-type loops and one propeller-type loop (figure 1.4c). In addition to the hybrid structures, a number of '3+1'-type G-quadruplexes were determined by NMR (in potassium conditions) for a variety of human telomeric sequences^{75–77}. These '3+1' structures – termed so due to the common feature of having a single-strand in an opposite orientation – displayed a diverse arrangement of edge-type and propeller-type loops. A representative member of this class is shown below, displaying a propeller-edge-edge loop arrangement^{77} (figure 1.4b, 2GKU). Other examples of '3+1' G-quadruplex structures include: 2JSL, 2JSM, 2JSK and 2JSQ^{75} and 2E4I^{76}. Subsequent to the '3+1' structure determinations, an additional structural arrangement was solved (by NMR again) involving a basket-type arrangement – but with just two-G-quartets forming the central core^{78} (2KF7 and 2KF8).

The NMR work has been pivotal in highlighting the significant structural polymorphism of human telomeric DNA G-quadruplexes. When combined with X-ray crystallographic data, the following classes of unimolecular human telomeric G-quadruplex structures can be defined:

1) All parallel-stranded (1KF1, 1K8P, 2LD8)^{40,79};

2) All anti-parallel-stranded (143D)^{39};

3) A mix of 3 parallel to 1 anti-parallel strands ('3+1' and 'hybrids') (2GKU, 2HY9, 2JSL, 2JSM, 2JSK, 2JSQ and 2E4I)^{73–77}, and;

4) All anti-parallel-stranded with a central core of two G-quartets (2KF7 and 2KF8)^{78}.

It is quite apparent, therefore, that there is a high degree of polymorphism intrinsic to human telomeric G-quadruplex structures, dependent upon cation type, the method
of analysis (i.e. NMR or crystallography) as well as the sequence studied – particularly the flanking 5' and 3' terminal residues, which can influence G-quadruplex structure significantly (as seen in the two 'hybrid' structures). As well as polymorphism in topology (i.e. loop-type and strand orientation), close inspection of the many structures reveals a high degree of diversity in individual base arrangement.
Figure 1.4 Representative structures of human telomeric G-quadruplexes.
Structures of intramolecular G-quadruplexes formed from human telomeric repeat sequences, solved by NMR (a-c) and X-ray crystallography (d). Cartoon representations are shown on the left, with schematic representations and additional information shown on the right. Each box in the schematics represents three stacked G-quartet layers. Cartoons are coloured according to the Nucleic Acid Database colour scheme, which is used throughout this thesis (unless stated otherwise): guanine = green, adenine = red, thymine = blue, potassium = purple (sphere), backbone = grey.
1.3.4.2 STRUCTURAL DIVERSITY OF TELOMERIC DNA G-QUADRUPLEXES

Based on the crystallographic and NMR structures, there is a clear structural flexibility available to G-quadruplexes formed from telomeric repeat sequences. Understanding structural polymorphism is critical to the understanding of a molecule's biological function, however, it is arguably more important to understand structural polymorphism if the molecule in question is a potential a drug target, as is the case with telomeric G-quadruplexes. So what are the causes of such structural diversity in telomeric DNA structures? As mentioned, it was initially thought that cation-type was the determining factor in defining G-quadruplex topology, however, it now appears that DNA concentration and molecular crowding effects are playing significant roles. Cation identity is undeniably important in influencing G-quadruplex stability, and to some degree topology, however, recent and invaluable work by Renciuk and co-workers has shown DNA concentration to be absolutely critical in defining topology. In this work, a single DNA sequence d(GGG[TTAGGG]₃) was shown to form anti-parallel, parallel and 'mixed' (i.e. '3+1') G-quadruplex structures depending upon DNA concentration alone — independent of ionic conditions (i.e. sodium or potassium). Additional work into the effects of molecular-crowding-mimicking agents on G-quadruplex topology has shown the use of low molecular weight polyethylene glycols (PEGs) to favour (and even induce) the parallel topology for the telomeric G-quadruplex. However, the situation is further complicated by the observation that PEGs physically interact with G-quadruplex DNA, thereby voiding PEGs as true molecular crowding agents. A particularly interesting recent finding has been the NMR structural determination of the d(AGGG[TTAGGG]₃) sequence in molecular crowding conditions (mimicked by the use of high percentages of PEG200), which revealed a parallel-stranded structure highly similar to the crystal structure reported for the same sequence almost 10 years previously.

It remains unclear exactly which conformation of the human telomeric G-quadruplex is most prevalent in the cell (if any indeed are present) — it is likely that all reported forms (i.e. hybrids, all-parallel and all-anti-parallel) can exist in cellular conditions. Considering the importance of molecular crowding and quadruplex concentration, a significant step forward may be the accurate determination of the
concentration of telomeric DNA G-quadruplexes in the cell. This of course leads to the question – do G-quadruplexes formed from human telomeric DNA (or any G-rich sequence) actually exist in the cell?

1.3.4.3 EVIDENCE FOR THE EXISTENCE OF (TELOMERIC) G-QUADRUPLEXES IN VIVO

Considering the vast amount of research that has been performed on G-quadruplex DNA – including G-quadruplexes formed from telomeric repeat DNA – it is perhaps surprising that conclusive evidence demonstrating the existence of G-quadruplex structures in vivo is lacking. This may be because it is technically challenging to prove the folding of a specified nucleic acid sequence in a cellular setting.

The identification of protein binding partners is generally good evidence for the in vivo existence of a specific structure, and many such G-quadruplex binding proteins have indeed been identified, albeit not necessarily all for telomeric G-quadruplex DNA\textsuperscript{81}. More importantly, a functional protein-DNA G-quadruplex interaction was discovered by Soldatenkov and co-workers, who showed the catalytic activity of hPARP (poly ADP-ribose polymerase) to be significantly increased upon binding to human telomeric G-quadruplexes\textsuperscript{82}. Strong evidence for the in vivo existence of G-quadruplex structures is also provided by Schaffitzel and co-workers, who have successfully used telomeric G-quadruplex specific antibodies to show the presence and localisation of such structures in ciliates\textsuperscript{83,84}.

Evidence for the existence of non-telomeric G-quadruplexes has also been provided. Bioinformatic analysis of the human genome has shown a strong bias for potential G-quadruplex forming sequences to occur in gene regulatory regions (such as promoter regions)\textsuperscript{85,86}. Further work has identified a similar trend in the genomes of lower organisms\textsuperscript{87}, suggesting an evolutionary bias for such sequences (and therefore for such structures) (see section 1.4.1 below). Experimental evidence in support of the in vivo existence of promoter region G-quadruplexes has also been provided, for example, antibodies raised against human promoter region G-quadruplex structures by phage display have very convincingly shown the presence of such structures in vivo\textsuperscript{88}.

These findings, coupled to the high stability and propensity of telomeric and non-telomeric G-quadruplexes to form in vitro, suggest that it is highly likely that such
non-canonical DNA structures are able to form in vivo. However, conclusive proof of the existence of G-quadruplex structures in vivo would represent a considerable advance for the field.

1.3.4.4 TELOMERIC DNA G-QUADRUPLEXES AS ANTI-CANCER DRUG TARGETS

Regardless of whether one is convinced of the in vivo existence of G-quadruplex structures within the 3' terminal regions of human chromosomes, for almost ten years now, researchers have nevertheless applied considerable effort to designing small molecule compounds to selectively bind and stabilise such G-quadruplex structures (see figure 1.3b). As mentioned above, the purpose of such research is to inhibit telomerase activity through the sequestration of its substrate (i.e. single-stranded telomeric DNA) with the aim of selectively inhibiting cancer cell growth. Telomerase is a particularly attractive anti-cancer drug target for two main reasons:

1) Telomerase activity in somatic cells is low-to-zero, making this protein a potentially cancer-cell specific drug target.

2) Telomerase re-activation (or an alternative mechanism to reinstate telomere extension) is critical for cell immortalization. For this reason, telomerase could almost be considered the miracle-target of cancer research, as it is an almost ubiquitous requirement for tumourigenesis.

As with any drug discovery program, the main challenges in designing effective human telomeric DNA G-quadruplex binding small-molecule compounds include:

1) achieving target selectivity (i.e. minimising off-target interactions) and;

2) achieving target affinity (i.e. strong ligand-target binding)

Drug target selectivity in this scenario includes duplex vs G-quadruplex selectivity, as well as selectivity for a nucleic acid type (i.e. RNA vs DNA selectivity) and selectivity for
INTRODUCTION - CHAPTER 1

a G-quadruplex type (e.g. telomeric G-quadruplexes vs promoter region G-quadruplexes). While the majority of telomeric G-quadruplex-directed small molecule compounds do indeed possess excellent duplex vs G-quadruplex selectivity, designing compounds with the ability to discriminate between G-quadruplex type (e.g. telomeric vs promoter) remains more challenging.

The first evidence that a small molecule compound could inhibit telomerase activity through G-quadruplex stabilisation was demonstrated by Sun and co-workers in 1997, who showed that a G-quadruplex binding anthraquinone-based ligand was able to reduce human telomerase activity in a telomere-length-measurement assay (TRAP - telomere repeat amplification protocol\(^2\). Since then, a vast number of telomeric DNA G-quadruplex targeted ligands have been designed and characterized\(^{68,90,91}\). Few, however, have been tested in cell lines, and even fewer still have been tested \textit{in vivo} (e.g. in tumour xenograft models). Small molecule ligands designed to bind selectively and with high affinity to telomeric G-quadruplex structures typically share three common features:

1) a large aromatic core (to maximise \(\pi-\pi\) stacking interactions with the G-quartet surface);

2) a positive charge (to utilize the negative phosphate groups of the DNA backbone), and;

3) diverse sidechain functionality (to maximise the opportunities to introduce specific functional groups to enhance interactions with quadruplex grooves/loops/individual bases).

Telomeric G-quadruplex interacting compounds are often rationally designed for optimal binding to telomeric G-quadruplex structures\(^{93-95}\). This method uses a combination of high-resolution native and drug-complex structures (i.e. NMR and X-ray structures) and computational molecular modelling techniques\(^{96,97}\). Typically, synthesized compounds are subsequently assessed for G-quadruplex binding using a variety of biophysical techniques, such as fluorescence resonance energy transfer
(FRET), surface plasmon resonance (SPR), circular dichroism (CD) and electrospray ionisation mass spectrometry (ESI-MS).

Screening compounds in appropriate cell lines is a vital step in evaluating ligand suitability; as well as determining the ability of a compound to kill cancerous cells, it is equally important to ensure toxicity to normal healthy cells is acceptably low. Three examples of telomeric DNA quadruplex binding ligands which have performed well in cell-based assays and in vivo tests include BRACO-19\textsuperscript{98}, RHPS\textsuperscript{4,100} and telomestatin\textsuperscript{101} (figure 1.5). In cellulo cytotoxicity assays on a human uterus carcinoma cell line showed BRACO-19 to effectively inhibit telomerase activity (within 24 hours), followed by induction of cellular senescence (after 15 days). Additionally, in vivo mouse xenograft studies showed BRACO-19 to reduce tumour growth by 96% (compared to controls), with evidence provided that the drug was acting through a telomerase-displacement mechanism (as intended)\textsuperscript{98}. RHPS\textsuperscript{4} performed similarly to BRACO-19, inhibiting cancer cell growth both in cellulo and in tumour xenograft studies, with evidence supporting a telomerase-displacement mechanism of action (i.e. loss of telomerase activity)\textsuperscript{99,100}. Telomestatin, differing from BRACO-19 and RHPS\textsuperscript{4} in being a natural product, nevertheless produced similar results – telomestatin effectively reduced tumour sizes in xenograft models, with reduced telomerase activity detected in tumour samples\textsuperscript{101}.

Despite the activity of these compounds, as yet, none have progressed to clinical trials. Indeed, of the many compounds which have been designed and synthesised to bind telomeric G-quadruplex DNA, none have yet reached clinical trials, probably due to the challenges of designing and evaluating truly specific and selective compounds. (N.B. the only G-quadruplex binding compound to reach clinical trials is quarfloxin/CX-3543 – strictly a non-telomeric G-quadruplex binding agent\textsuperscript{102}.) Avoiding off-target binding is a challenge to any drug discovery programme, perhaps even more so in the field of telomeric G-quadruplex drug design, considering the huge number of potential G-quadruplex forming sequences present within the human genome (> 300, 000\textsuperscript{31,42}), as well as the recent discovery that the well-studied telomeric repeat sequence, d(TTAGGG), can in fact be found in several locations of the genome outside of the telomeric regions\textsuperscript{103}.
While NMR has been responsible for uncovering the structural diversity of G-quadruplexes formed from human telomeric repeat DNA, it is crystallography which has arguably been the workhorse in providing structural data for human telomeric G-quadruplex-ligand complexes (table 1.1, page 46).

Several high-resolution NMR G-quadruplex-complex structures exist, however, these models all involve short sequences, such as d(TGGGGT), resulting in tetramolecular G-quadruplex topologies which lack the important intricate loop arrangements of the folded intramolecular telomeric G-quadruplexes. The first such NMR structure was solved in 2003 by Searle and co-workers, who revealed the binding of the pentacyclic acridinium compound, RHPS4, to the four stranded G-quadruplex formed from the sequence d(TTAGGGT)\textsuperscript{104} (figure 1.6a, 1NZM). This highly potent G-quadruplex binding compound was shown to stack effectively over both of the available external G-quartets, resulting in a 2:1 ligand:quadruplex stoichiometry.
Recently reported structures solved by Randazzo and co-workers have shown the exclusive interaction of Distamycin-A (and a related analogue) with the groove regions of a G-quadruplex formed from the model sequence, d(TGGGGT)\textsuperscript{105,106} (figure 1.6b, \texttt{2KVY} and \texttt{2JT7}). Distamycin-A displays an unusual mode of G-quadruplex-interaction in that two ligand-pairs localise to two separate quadruplex grooves, with no interaction at all between the ligands and the G-quartets.

\begin{itemize}
\item[a)] \textbf{RHPs4 – d(TTAGGGT) complex}
\texttt{PDB Id: 1NZM}
\item[b)] \textbf{Distamycin-A – d(TGGGGT) complex}
\texttt{PDB Id: 2KVY}
\end{itemize}

\textbf{Figure 1.6 NMR-derived telomeric G-quadruplex ligand complex structures.}
a) Structure of a parallel-stranded tetramolecular G-quadruplex formed from the sequence, d(TTAGGGT), in complex with the pentacyclic acridinium compound, RHPs4 (yellow spheres). b) Structure of a parallel-stranded tetramolecular G-quadruplex formed from the sequence, d(TGGGGT), in complex with an analogue of distamycin-A (yellow and purple spheres). Four distamycin-A molecules bind this G-quadruplex as dimers. DNA is shown as grey cartoon, with potassiums shown as purple spheres.

The first G-quadruplexes studied by X-ray crystallography were tetraplexes formed from the model sequence, d(TGGGGT), and bimolecular quadruplexes formed from the \textit{Oxytricha nova} telomeric repeat sequence, d(GGGGTTTTGGGG)\textsuperscript{38,107–109}. As well as providing early details of native quadruplex folding, these sequences were also used for the first quadruplex ligand complex structures\textsuperscript{110,111}. The first structural determination of a human telomeric G-quadruplex-ligand complex (which included TTA loop sequences) was not provided until 2007, when the crystal structure of a complex formed between the tetra-substituted porphyrin, TmPyP4, and a bimolecular
G-quadruplex formed from the human telomeric sequence, d(TAGGGTTAGGG), was
determined\textsuperscript{112} (figure 1.7a, 2HRI). It was expected that the significant aromatic
moieties of TMPyP4 would interact with a G-quadruplex by \( \pi-\pi \) stacking over an
external G-quartet, however, this crystal structure revealed the ligand to be interacting
exclusively with loop and terminal adenine and thymine residues. Following from this,
the successful crystal structure determination of a complex formed between the
acridine, BRACO-19, and a bimolecular quadruplex formed from the human telomeric
sequence, d(TAGGGTTAGGGT), revealed an interesting and (arguably) more
biologically relevant mode of ligand binding\textsuperscript{113} (figure 3.2b [chapter 3, page 102],
3CES). BRACO-19 was seen to stack effectively over the 3' G-quartet, with the
sidechains interacting through hydrogen bonds and water bridges with the DNA. This
structure also provided information concerning the interaction of ligands with multiple
G-quadruplex subunits, as crystal packing contacts showed the ligand to be stabilised
from ‘below’ by an additional G-quadruplex. The remaining four crystal structures
involving human telomeric G-quadruplex-ligand complexes all involve tetra-substituted
naphthalene diimide compounds\textsuperscript{97,114} (figure 1.7b, 3CDM and 3CCO. N.B. – two of
these structures are yet to be released from the Protein Databank). These compounds
– which have large aromatic cores modified with four flexible sidechains – all exhibited
an unexpected mode of ligand binding to telomeric G-quadruplexes, in that, although
these ligands did indeed bind to external G-quartets (as hoped), there were also
extensive binding sites created exclusively from TTA loop residues. These naphthalene
diimide complexes also include the only ligand complexes involving intramolecular
human telomeric G-quadruplexes – two complex structures were solved using the
sequence, d(TAGGG[TTAGGG]\textsubscript{3}). The intramolecular G-quadruplex is arguably the most
relevant topology to study concerning human telomeric G-quadruplexes, as this
represents the expected single-repeat G-quadruplex unit of single-stranded telomeric
DNA.
a) TmPyP4 – d(TAGGGTTAGGG) complex
PDB Id: 2HRI

b) Naphthalene diimide – d(TAGGG[TTAGGG]₃) complex
PDB Id: 3CDM

Figure 1.7 Human telomeric G-quadruplex-ligand complexes solved by X-ray crystallography.
a) Complex formed between a bimolecular G-quadruplex formed from two strands of the sequence, d(TAGGGTTAGGG), and the porphyrin molecule, TmPyP4 (yellow spheres). Both bound ligand molecules are interacting exclusively with loop and terminal residues, in contrast to the expected mode of ligand binding (i.e. G-quartet stacking). b) Complex formed between an intramolecular G-quadruplex formed from the sequence, d(TAGGG[TTAGGG]₃), and a tetra-substituted naphthalene diimide compound. Although this compound binds to G-quartet surfaces as expected (yellow spheres), it also binds to surfaces generated exclusively from loop residues (purple spheres). Chemical structures of the compounds are shown next to both cartoon representations. DNA is coloured grey, with the ligands shown as yellow or purple spheres and potassium ions shown as large purple spheres.

The crystal structures, along with NMR data, have aided in the rational design of human telomeric G-quadruplex interacting compounds, and have highlighted an unexpected diversity in quadruplex-ligand binding modes. For example, telomeric-quadruplex targeted ligands were initially considered to interact primarily through π-stacking interactions with the G-quartet, however, crystal structures have repeatedly illustrated a major role of loop residues in ligand binding. As with all
INTRODUCTION – CHAPTER 1

crystallographically determined structures, such structures should be considered with caution, as often it is not clear whether the drug-quadruplex interactions seen in the crystal structures are the result of crystal packing interactions within the lattice, or whether they are representative of potential in vitro/in vivo interactions. The main caveat applicable to the NMR derived G-quadruplex-complexes is that these structures, so far, have involved tetramolecular G-quadruplexes (often using non-wild-type sequences) and hence are lacking the important TTA-loops of bi- and unimolecular G-quadruplexes. More generally, it is unfortunate that quadruplex-ligand complexes are so challenging to study by NMR, as the solution-phase structures determined by NMR spectroscopy would be exceptionally useful in determining the validity of crystallographically determined drug complexes.

Table 1.1 Summary of current structural models of telomeric G-quadruplex-ligand complexes.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Species</th>
<th>Method</th>
<th>Strands</th>
<th>Topology</th>
<th>Ligand</th>
<th>PDB Id</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>d(TTAGGGT)</td>
<td>Human/ model</td>
<td>NMR</td>
<td>4</td>
<td>Parallel</td>
<td>RHPS4</td>
<td>1NZM</td>
<td>Gavathiotis et al 2003^°^</td>
</tr>
<tr>
<td>d(TGGGGT)</td>
<td>model</td>
<td>NMR</td>
<td>4</td>
<td>Parallel</td>
<td>Distamycin-A</td>
<td>2JT7</td>
<td>Martino et al 2007^°^</td>
</tr>
<tr>
<td>d(TGGGGT)</td>
<td>model</td>
<td>NMR</td>
<td>4</td>
<td>Parallel</td>
<td>Distamycin-A (analogue)</td>
<td>2KVV</td>
<td>Consconati et al 2010^°^</td>
</tr>
<tr>
<td>d(TGGGGT)</td>
<td>model</td>
<td>X-ray</td>
<td>4</td>
<td>Parallel</td>
<td>Daunomycin</td>
<td>100K</td>
<td>Clark 2003^°^</td>
</tr>
<tr>
<td>d(GGGGTGGGGG)</td>
<td>O. nova</td>
<td>X-ray</td>
<td>2</td>
<td>Anti-para</td>
<td>Acridine</td>
<td>1L1H</td>
<td>Haider 2003^°^</td>
</tr>
<tr>
<td>d(GGGGTGGGGG)</td>
<td>O. nova</td>
<td>X-ray</td>
<td>2</td>
<td>Anti-para</td>
<td>Acridines</td>
<td>3EM2^*</td>
<td>Campbell et al 2009^°^</td>
</tr>
<tr>
<td>d(GGGGTGGGGG)</td>
<td>O. nova</td>
<td>X-ray</td>
<td>2</td>
<td>Anti-para</td>
<td>Acridines</td>
<td>3NYP, 3NZ7</td>
<td>Campbell et al 2011^°^</td>
</tr>
<tr>
<td>d(TAGGGTTAGGG)</td>
<td>Human</td>
<td>X-ray</td>
<td>2</td>
<td>Parallel</td>
<td>Porphyrin</td>
<td>2HRI</td>
<td>Parkinson et al 2007^°^</td>
</tr>
<tr>
<td>d(TAGGGTTAGGG)</td>
<td>Human</td>
<td>X-ray</td>
<td>2</td>
<td>Parallel</td>
<td>BRACO-19</td>
<td>3CE5</td>
<td>Campbell et al 2008^°^</td>
</tr>
<tr>
<td>d(TAGGGTTAGGG)</td>
<td>Human</td>
<td>X-ray</td>
<td>1</td>
<td>Parallel</td>
<td>Naphthalene diimide</td>
<td>3CDM</td>
<td>Parkinson et al 2008^°^</td>
</tr>
<tr>
<td>d(TAGGGTTAGGG)</td>
<td>Human</td>
<td>X-ray</td>
<td>2</td>
<td>Parallel</td>
<td>&quot;</td>
<td>3CCO</td>
<td>&quot;</td>
</tr>
<tr>
<td>d(TAGGGTTAGGG)</td>
<td>Human</td>
<td>X-ray</td>
<td>1</td>
<td>Parallel</td>
<td>&quot;</td>
<td>tbr^*</td>
<td>Haider et al 2011^°^</td>
</tr>
<tr>
<td>d(TAGGGTTAGGG)</td>
<td>Human</td>
<td>X-ray</td>
<td>2</td>
<td>Parallel</td>
<td>&quot;</td>
<td>tbr^*</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

* A total of 7 structures are reported in this work, PDB Ids ; 3EM2, 3EQW, 3EUI, 3ERU, 3ES0, 3ET8, 3EUM. # to be released. Structures of human sequences are highlighted in grey
1.4 OTHER SOURCES OF G-RICH DNA IN THE GENOME

While the focus of this thesis is on G-quadruplex structures formed from human telomeric DNA (and RNA), it should be noted that putative G-quadruplex forming sequences have been discovered throughout the human genome, which impacts upon the relevance and significance of telomeric DNA (and RNA) G-quadruplexes. For this reason, the key extra-telomeric sources of G-quadruplex forming DNA sequences will be considered below.

1.4.1 G-QUADRUPLEX FORMATION IN PROMOTER REGIONS

Bioinformatic analysis of the human genome in 2005-6 revealed for the first time the significant prevalence of potential G-quadruplex forming sequences within our DNA, with over 300,000 sequences identified as being capable of forming G-quadruplex structures\textsuperscript{41,42}. In-depth analysis revealed a highly biased distribution of these potential G-quadruplex forming sequences in gene promoter regions (with a distinct absence in gene-coding regions), suggesting such motifs to be regulating gene expression at the sequence level. The proposed mechanism of G-quadruplex-based regulation of gene-expression involves the formation of a G-quadruplex structure upstream of the transcription start site, which blocks the progress of RNA polymerase, thereby inhibiting gene expression.

The existence of such potential G-quadruplex forming sequences within gene promoter regions offers the possibility to control gene expression at the transcription level, through stabilisation of such structures with small-molecule compounds. Early work on the \textit{c-MYC} oncogene promoter region demonstrated that the formation of a G-quadruplex structure upstream of the transcription start site could indeed reduce gene expression\textsuperscript{117}. Importantly, this work further showed that addition of a known G-quadruplex stabilising compound – the porphyrin, TmPyP4 – could further inhibit gene expression. Following from this, potential G-quadruplex forming sequences located within the promoter regions of a number of genes (primarily onocogenes) have been studied as potential drug targets, including \textit{c-KIT}\textsuperscript{118-120}, \textit{BCLz}\textsuperscript{121-123}, \textit{HIF1α}\textsuperscript{124} and \textit{VEGF}\textsuperscript{125,126} (as well as further work on \textit{c-MYC}\textsuperscript{118,127}). As with telomeric quadruplex drug
design, promoter-region quadruplex directed compounds have also been designed based on high-resolution structures – so far solved exclusively by NMR\textsuperscript{128}. In a manner similar to telomeric quadruplex targeted ligands, the majority of reported promoter-directed quadruplex binding agents are capable of effectively selecting between duplex and quadruplex DNA, however, intra-quadruplex specificity remains a challenge.

### 1.4.2 ADDITIONAL ROLES OF DNA G-QUADRUPLEXES IN THE GENOME

While the majority of DNA G-quadruplex research to date has focused on G-quadruplex structures formed from telomeric and promoter region sequences – mainly due to the relevance of such structures to cancer and pathological research – there exist a number of additional regions of the genome in which DNA G-quadruplexes have been implicated. These include:

1) **Immunoglobulin switch regions**

G-rich regions have been identified in the DNA recombination processes of immunoglobulin class switching (which involves heavy chain rearrangements). G-quadruplex formation within these G-rich regions has been implicated in the switching process\textsuperscript{129,130}.

2) **Minisatellites**

Minisatellites are tracts of between 10 and 60 basepairs of repeated short DNA sequences. G-quadruplex formation of the triplet repeat, d(CGG), has been shown to play a possible role in the pathogenesis of the fragile X mental retardation syndrome\textsuperscript{131–133}.

3) **Ribosomal DNA and helicases**

A conserved family of helicases have shown the unique ability to resolve (i.e. unwind) G-quadruplex structures\textsuperscript{134,135}. The localisation of G-rich helicase target-sequences has implicated such G-quadruplex structures to be involved in
the rDNA transcription process, as well as in a number of genomic-instability linked diseases (e.g. Bloom's syndrome and Werner's syndrome)\textsuperscript{134,135}. 
1.5 HUMAN TELOMERES AS A SOURCE OF G-RICH RNA

1.5.1 G-QUADRUPLEX RNA IN THE GENOME

RNA G-quadruplex structures have long been known to exist within genomes, and have been implicated in a diverse array of biological functions and processes (figure 1.8). Early work identified a role for RNA G-quadruplex structures in viral genomes, such as in HIV genome dimerisation and in translational recoding within herpes simplex viruses\(^\text{136,137}\). Selective binding of a viral protein to RNA quadruplex structures was also reported\(^\text{138}\), and single-celled organisms were found to utilise RNA G-quadruplex structures, for example, in the control of slippage transcription in bacteria\(^\text{139}\). Significantly, four stranded RNA structures were also found to play a role in several important process within higher organisms, including mRNA turnover in mice\(^\text{140}\), post-transcriptional control of IGF-2 (insulin-like growth factor 2) expression in rats and humans\(^\text{141}\) and in translational control of FGF-2 (fibroblast growth factor 2) expression in humans\(^\text{142}\). Work within the last 10 years has built on these areas, as well as identifying a number of additional roles for quadruplex RNA in biological settings. Two areas of RNA G quadruplex research in particular have received considerable attention within the past few years: 1) the recently discovered G-rich telomeric repeat RNAs (termed 'TERRA')\(^\text{143,144}\) and 2) G-rich sequences located within 5' untranslated regions (UTRs) of mammalian mRNA sequences\(^\text{43}\). TERRA and 5' UTR mRNA quadruplexes are both of potentially high relevance to cancer biology, and consequently, a considerable portion of research effort has focused on therapeutic implications and targeting of such structures. One of the main themes of this thesis is the formation of G-quadruplex structures formed from telomeric-repeat containing RNA (TERRA). The current understanding and work for such structures will be considered below, followed by a brief consideration of non-TERRA G-quadruplex structures in the genome.
1.5.2 TELOMERIC RNA (TERRA/TelRNA) — A SOURCE OF G-RICH RNA

As described above (section 1.3), the terminal regions of mammalian chromosomes exist as tandem repeats of the sequence, d(TTAGGG). For many years this region was thought to be transcriptionally silent, however, it was shown in 2007 that the C-rich strand of the telomere is in fact actively transcribed, resulting in variable length RNA sequences composed of r(UUAGGG) repeats\(^{143,144}\). These transcripts, termed TERRA or telRNA, are transcribed from sub-telomeric regions by RNA polymerase II, and were found to be between 0.1 and 9 kilobases in length (figure 1.9). The shelterin protein, TRF-2, was identified as a possible transcriptional regulator of TERRA levels, as the binding of TRF-2 to RNA polymerase II led to an increase in TERRA production\(^{144}\). Evidence for TERRA having a functional role in the cell was also provided, as TERRA was shown to be involved in chromatin remodelling and, importantly, was shown to inhibit
Telomerase activity in vitro. TERRA has now been identified in humans, mice, zebrafish, and yeast, such apparent evolutionary conservation suggesting an important role for TERRA within cells.

**Figure 1.9** Origin of telomeric RNA (TERRA/telRNA).
The C-rich strand of human telomeric DNA is transcribed by RNA polymerase II, resulting in RNA transcripts up to 9000 nucleotides in length composed of tandem (UUAGGG) repeats (termed TERRA or telRNA, coloured red). These transcripts have been shown to play a role in chromatin remodelling, telomerase activity and in the regulation of telomere length.

### 1.5.3 TERRA REGULATION AND PROTEIN INTERACTIONS

Following the discovery of TERRA, proteins able to bind to these sequences were soon identified. The shelterin proteins, TRF1 and TRF2, plus elements of the origin recognition complex (ORC) were all shown to be capable of binding directly to TERRA, such interactions suggesting a role in heterochromatin structure at the telomere.
Two independent TERRA binding sites were identified within telomerase, one within the constitutive RNA component (hTR/hTERC) and interestingly, another within the protein component itself (hTERT)\textsuperscript{148}. This work further showed that the addition of TERRA sequences to telomerase protein \textit{in vitro} resulted in a significant reduction in telomerase activity (as shown by Shoeftner and Blasco\textsuperscript{144}). While the TERRA-binding site within the hTR subunit is expected to comprise the telomeric-repeat template region (i.e. AAUCCC), the TERRA-binding site within the hTERT subunit is not yet characterised. The interaction of TERRA with telomerase, and its inhibitory effect, are aspects of TERRA function which could be of key importance to cancer biology (and therefore of subsequent importance to anti-cancer drug design).

More recently, Lopez de Silanes and co-workers used a MALDI-ToF-based assay to identify a vast array of TERRA binding proteins, many of which were in the hnRNP family of RNA binding proteins (heteronuclear ribonucleoproteins)\textsuperscript{149}. In contrast to identifying TERRA-binding proteins, Cech and co-workers described an example of a protein with specific mechanisms to avoid telomeric RNA binding\textsuperscript{150}. Critical residues within the single-stranded telomeric repeat binding protein, POT1, were shown to be responsible for POT1's ability to bind specifically to single-stranded telomeric DNA rather than TERRA. The details of TERRA regulation and localisation are yet to be fully unravelled, but so far several proteins have been identified as regulators of TERRA levels, such as TRF-2 (mentioned above) and the MLL (mixed lineage in leukaemia) transcription factor\textsuperscript{151}. It is likely that many more important proteins involved in TERRA regulation and function will be discovered, which will hopefully lead to an understanding of the biological significance of TERRA.

\textbf{1.5.4 G-QUADRUPLEX STRUCTURES FORMED FROM TERRA}

As described above, the G-rich single-stranded 3' ends of chromosomes are able to fold into stable four-stranded G-quadruplex structures (section 1.3.4). Considering the similarity of G-rich telomeric RNA sequences to their G-rich DNA counterparts, it is perhaps not surprising that shortly after their discovery, TERRA transcripts were shown to be able to form stable G-quadruplex structures. This was first demonstrated by Xu and co-workers, who used 1D NMR and circular dichroism to show that the sequences
r(UAGGGU), r(UAGGGUUAGGGU) and r([UUAGGG]₆) formed stable tetramolecular, bimolecular and unimolecular parallel G-quadruplexes, respectively⁵²-⁵⁴. The tetramolecular quadruplex, r(UAGGGU), was subsequently shown to be stabilised by the formation of a 3' U-tetrad⁵⁵. Interestingly, and in contrast to the equivalent DNA sequences, the RNA quadruplexes studied were seen to form parallel structures in both Na⁺ and K⁺ ionic conditions.

Additionally, Xu and co-workers demonstrated the presence of DNA-RNA intermolecular hybrid quadruplexes in vitro⁵⁶,⁵⁷. Such hybrids have the potential to form between TERRA sequences and the single-stranded G-rich DNA overhang of the telomere. It is possible that such hybrid formation could play a role in the regulation of telomere maintenance, via the disruption of the telomerase complex⁵⁸.

1.5.5 HIGH-RESOLUTION STRUCTURAL STUDIES OF TERRA G-QUADRUPLEXES

The first atomic-level insight into the structure of telomeric RNA G-quadruplexes was provided by Martadinata and Phan, who used NMR spectroscopy to determine the structure of a bimolecular quadruplex formed from the sequence r(UAGGGUUAGGGU)⁵⁹ (figure 1.10). The structure involves a parallel stranded G-quadruplex (with RNA bases all in the anti conformation) with one chain-reversal propeller type loop per strand. To date, however, there have been no structural determinations for unimolecular TERRA G-quadruplexes, in contrast to the considerable number of unimolecular structures reported for human telomeric DNA G-quadruplex forming sequences. Possible reasons for the lack of high-resolution structural data for unimolecular TERRA G-quadruplexes will be considered in chapter 6.

1.5.6 STRUCTURAL MONOMORPHISM OF TERRA G-QUADRUPLEXES

A prevalent theme within telomeric RNA G-quadruplex research is that the topology of these structures appears to be far less dependent on the environmental conditions when compared to equivalent DNA sequences. For example, the sequence r(AGGG[UUAGGG]₃) was shown to form a predominantly parallel G-quadruplex structure (as measured by CD) regardless of molecular crowding or ionic conditions⁶₀.
The equivalent DNA sequence, however, formed parallel, anti-parallel or mixed (i.e. '3+1') G-quadruplex topologies depending upon the monovalent cation type (i.e. sodium or potassium) and the presence of a molecular crowding agent (PEG200). This work illustrates the apparent rigidity in RNA G-quadruplex structural topology – indeed, all telomeric (and non-telomeric) RNA G-quadruplexes studied to date have displayed a parallel topology. This contrasts quite significantly with the array of reported folds for DNA sequences (see section 1.3.4.1 above). It is not clear whether this is because RNA G-quadruplexes are indeed restricted to a narrow range of topologies, or simply that potential structural variations are yet to be discovered. As with studies into DNA G-quadruplexes, it is likely that NMR will prove to be the most successful tool in the search for variant RNA G-quadruplex structures.

**Bimolecular TERRA G-quadruplex**

PDB id: 2KBP | Sequence: r(UAGGGUAGGGU) | parallel | K⁺ cation

![Bimolecular TERRA G-quadruplex diagram](image)

**Figure 1.10 NMR structure of a G-quadruplex formed from telomeric RNA/TERRA.**
The TERRA sequence, r(UAGGGUAGGGU), was shown by NMR to fold as a parallel stranded bimolecular G-quadruplex, with two propeller-type loops. The structure is shown as a cartoon (a) and schematic (b). The cartoon is coloured according to the NDB colour scheme: guanine = green, adenine = red, uracil = cyan, backbone = grey.

**1.5.7 HIGHER-ORDER STRUCTURE FORMATION WITHIN TERRA SEQUENCES**

TERRA sequences can be up to 9000 bases in length, and as such, a single TERRA transcript has the potential to accommodate hundreds of discrete G-quadruplex structures. Therefore, in addition to the study of how individual TERRA G-quadruplexes form...
fold, it is important to understand any possible 'higher-order' (i.e. tertiary) structure, (i.e. the mechanism by which multiple quadruplex subunits interact with one another), as this could have implications concerning the possible biological function of TERRA G-quadruplexes. For example, the involvement of G-quadruplex structures in mRNA processing has been reported\(^{140}\), and it is possible that G-quadruplex formation within TERRA sequences could be involved in a similar process. Understanding TERRA quadruplex higher order structure is also important if protein binding partners are to be identified, as it would be expected that the structural motif presented by unfolded TERRA would be very different to that presented by TERRA folded into compact quadruplex arrangements.

Investigations into this area by Randall and Griffith using electron microscopic techniques, demonstrated the existence of potassium-dependent, highly compact structures within TERRA transcripts of approximately 600 nucleotides in length\(^{163}\). Enzymatic digestion of these transcripts followed by polyacrylamide gel electrophoretic (PAGE) analysis revealed a PAGE pattern of bands differing from each other in size by around 24 nucleotides. This very convincingly suggests quadruplexes form within TERRA transcripts in a beads-on-string pattern, with UUA loops interspersing quadruplex units. Gel shift studies performed by Martadinata and Phan identified a propensity for telomeric RNA sequences to multimerise – analysis of 10-mer and 9-mer RNAs \(r\{GGGUUAGGGU\}\) and \(r\{GGGUUAGGG\}\) by native PAGE showed the presence of dimers and higher-order structures, respectively\(^{159}\). Unimolecular TERRA G-quadruplexes have also shown a strong tendency to multimerise in PAGE experiments\(^{164}\). The multimerisation of RNA G-quadruplex structures will be discussed further in chapter 6.

1.5.8 BIOLOGICAL ROLE(S) OF TERRA G-QUADRUPLEXES?

The identification of specific protein-binding partners for a molecule of interest is usually strong evidence that such a molecule has a biologically relevant role. However, as yet, no specific TERRA G-quadruplex interacting proteins have been discovered. Several proteins able to bind TERRA sequences (folded or not) have been identified however (section 1.5.3), and it will be important to determine whether these proteins
are interacting with folded or unfolded TERRA sequences. The identification of specific TERRA G-quadruplex binding proteins would be a significant step in determining whether TERRA G-quadruplexes have a biologically relevant purpose. Past work on RNA quadruplexes in yeast have shown that specific RNA–quadruplex-protein recognition is certainly possible.

It was observed that G-quadruplex formation within long RNA transcripts was able to confer RNase resistance. It is reasonable to suggest that G-quadruplex formation within TERRA transcripts could play a role in regulating the cellular levels of these molecules. If G-quadruplex formation within TERRA molecules proves to be a mechanism of regulating RNA levels, then it will be important to determine the precise RNase(s) involved in regulation, as early work on mRNA processing showed quadruplex elements within mouse mRNA to promote exoribonuclease degradation, rather protect against it.

An additional potential role for TERRA G-quadruplex structures in cells comes from the finding that DNA-RNA hybrid G-quadruplex structures are able to form from telomeric sequences, suggesting that telomeric RNA may be involved in telomere regulation through a mechanism involving sequestration of the substrate of telomerase (i.e. telomeric DNA).

The finding that the crucial shelterin protein, POT1, has apparently evolved a mechanism to selectively avoid TERRA adds to already strong evidence that these transcripts may play an important role at the telomere. Whether G-quadruplex structures have biologically relevant roles in vivo is yet to be determined, but recent evidence of RNA G-quadruplex formation within living cells provided by Xu and co-workers adds support to the notion that such structures may indeed be biologically functional. In this work, short TERRA oligos labelled with light-activated fluorophores were transfected into HeLa (cervical cancer) cells, allowing G-quadruplex formation to be positively detected by fluorescence measurements. Further evidence for the existence of RNA (and DNA) G-quadruplex structures in cells has been provided recently by Muller and co-workers, who used a biotin affinity tagged G-quadruplex-selective small-molecule probe to successfully isolate G-quadruplexes from human cells.
1.5.9 TERRA G-QUADRUPLEXES AS DRUG TARGETS?

With the discovery of telomeric RNA, the G-quadruplex structures formed from these transcripts now need to be considered in the context of telomere-focused drug design. As telomeric DNA quadruplex structures have been targeted for anti-cancer drug design for over 10 years, it is important to understand how these ligands could interact with the potentially very similar RNA structures, (but which possess very different biology) in order to minimise off-target effects (figure 1.11). For example, it has been shown that TERRA can inhibit telomerase activity in vitro – primarily by hybridising to (and thereby sequestering) the template region of the RNA subunit of telomerase (hTR)^144,148. One would expect that TERRA molecules would need to be unfolded (i.e. not folded into a G-quadruplex structure) to allow this hybridisation (and thereby inhibition) to occur. Stabilisation of telomeric RNA G-quadruplexes with small molecule compounds might therefore increase telomerase activity, by reducing the amount of unfolded TERRA available for binding to and inhibiting telomerase. As most telomeric DNA G-quadruplex binding compounds have been designed without concern for DNA vs RNA selectivity, it is possible that many of these compounds will indeed bind TERRA G-quadruplexes with reasonably high affinity. There is currently limited ligand binding data available concerning telomeric RNA interacting compounds^169. Further research – both structural and biophysical – will be needed to determine the significance (if any) of TERRA G-quadruplex stabilisation by telomeric G-quadruplex-directed small molecule compounds. At the very least, it is the concern of any drug discovery programme to understand drug-target binding and minimise off-target binding as much as practically possible.

Although as yet, TERRA G-quadruplexes are not validated drug targets, considering the key role TERRA has been shown to play in telomere biology and telomerase regulation, there is a high likelihood that TERRA G-quadruplexes will be validated as drug targets in the near future. In addition, the development of a TERRA G-quadruplex-selective compound would also be a valuable tool as a molecular probe, in order to investigate the possible role of TERRA G-quadruplexes at the telomere.
**Figure 1.11 Telomeric G-quadruplex-drug binding.**
a) The stabilisation of telomeric DNA G-quadruplex structures with small-molecule therapeutic compounds has been shown to inhibit telomerase activity and induce senescence in cancer cells. b) However, the effects of the binding of these same drug compounds to the potentially similar TERRA G-quadruplexes are not known – this is an aspect of telomere-based research that warrants further investigation. DNA guanines are shown as blue cuboids in (a), with RNA guanines shown as red cuboids in (b).
1.6 OTHER SOURCES OF G-RICH RNA IN THE GENOME

1.6.1 G-QUADRUPLEXES IN mRNA UNTRANSLATED REGIONS

1.6.1.1 DISCOVERY AND PRINCIPLES

As described in section 1.4.1, bioinformatic studies into potential G-quadruplex forming sequences within the human genome revealed an enrichment of such sequences within promoter regions, suggesting that G-quadruplexes might play a role in gene expression. Subsequent analysis of the human transcriptome revealed a similar trend, with potential G-quadruplex forming sequences identified in a number of mRNA 5' untranslated regions (UTRs). The ability of an mRNA 5' UTR G-quadruplex forming sequence to inhibit translation was first shown by Kumari and co-workers, who used a dual luciferase reporter assay to demonstrate the ability of a G-rich sequence located within the 5' UTR of the NRAS proto-oncogene to down regulate protein expression. The mechanism of gene regulation in this situation is proposed to involve G-quadruplex formation within a 5' UTR physically blocking the progress of the ribosome complex, thereby inhibiting translation (figure 1.12).

The work of Balasubramanian's group identified potential G-quadruplex forming sequences in the 5' UTRs of a number of genes and oncogenes, which have subsequently been experimentally confirmed to be able to regulate gene expression. These genes include those encoding the apoptosis regulator BCL2, the shelterin protein TRF2, and the oestrogen receptor α gene. In addition to the human transcriptome, bioinformatic analysis of a plant transcriptome (Arabidopsis thaliana) revealed a similar prevalence and distribution of potential G-quadruplex forming sequences.

1.6.1.2 mRNA 5' UTR G-QUADRUPLEXES AS DRUG TARGETS

The discovery of a G-quadruplex motif within the 5' UTR of the NRAS oncogene capable of inhibiting translation opened up a new avenue of G-quadruplex directed drug design. The principle of regulating gene expression through stabilisation of
mRNA UTR G-quadruplex structures with small molecule compounds is similar to that of DNA promoter region G-quadruplex drug design, in the former case, disruption of the ribosome complex being the aim (figure 1.12b). An attractive aspect of mRNA 5’ UTR G-quadruplex targeting is that, with no complementary strand, G-quadruplex forming sequences within the mRNA are not subject to competition with duplex formation (in contrast to the situation with DNA promoter sequences), and hence may be more available to form G-quadruplex structures (figure 1.12). As with promoter region G-quadruplex targeting, the majority of 5’ UTR mRNA G-quadruplex studies have focused on oncogenic genes. 5’ UTR G-quadruplex sequences which have been successfully stabilised by synthetic compounds (resulting in a reduction in gene expression) include the NRAS proto-oncogene\textsuperscript{174}, the TRF2 gene\textsuperscript{171} and the aurora A protein kinase gene\textsuperscript{175}.

**Figure 1.12** G-quadruplex formation in 5’ untranslated regions of mRNA.

a) Model mRNA system, including a potential G-quadruplex forming sequence located between the ribosome binding site and the translation start site (i.e. in the untranslated region, UTR). If such a sequence does not form a G-quadruplex, the ribosome complex is not hindered by secondary structure formation and is able to translate the gene. b) Alternatively, G-quadruplex formation blocks the progress of the ribosome complex, inhibiting gene expression. Stabilisation of such structures with small-molecule compounds is expected to allow artificial control of gene expression (e.g. in an anticancer therapeutic scenario).
1.6.2 G-QUADRUPLEXES IN FRAGILE X MENTAL RETARDATION SYNDROME

RNA G-quadruplex structures have been linked to two similar forms of intellectual disability: Fragile X mental retardation syndrome (FXS) and fragile X E-associated mental retardation (FRAXE), both of which result from the aberrant expression of two specific RNA G-quadruplex binding proteins, FMRP and FMRP2, respectively (FXS mental retardation protein 1 and 2). FMRP was shown to bind G-rich RNA targets in the 5' untranslated regions of several genes (PP2A\(^{176}\) and MAPB1\(^{177}\)), resulting in a reduction in translation, predicted to be via disruption of ribosome function. FMRP was also shown to bind to G-quadruplex structures located within the coding region of the FMR1 gene itself\(^{178}\), and within the mRNA of the APP gene (amyloid precursor protein\(^{179}\)), thereby regulating translational activity. RNA G-quadruplex targets of FMRP were also identified within the 3' UTR regions of a number of other genes, however, the role of these structures is yet to be confirmed\(^{177}\). The G-quadruplex forming sequence within the coding region of FMR1 was also shown to be a potential exon splice enhancer, with both FMRP and FMRP2 shown to bind to this structure, consequently affecting mRNA splicing patterns\(^{180,181}\).

G-quadruplex structures located within the mRNAs targeted by the FMRP and FMRP2 proteins represent an unexplored area of mRNA quadruplex ligand-design. By stabilisation of FMRP RNA G-quadruplex-binding-targets with selective drugs, it could be possible to mimic the effects of FMRP-G-quadruplex binding and thereby ameliorate the effects of low FMRP levels in FXS patients. This is yet to be attempted, however, the successful targeting of oncogenic 5' UTR mRNA G-quadruplexes with small molecule compounds\(^{171,174}\), suggests such a strategy may be effective in FXS treatment.

1.6.3 ADDITIONAL ROLES OF RNA G-QUADRUPLEXES IN THE GENOME

In addition to G-quadruplex formation within TERRA transcripts, 5' UTR mRNAs and FXS-related genes, recent findings have been reported concerning the role of RNA G-quadruplex structures in a range of additional cellular processes. For example, interesting recent findings by Wanrooij and co-workers have outlined a role for RNA G-
quadruplex structures in mitochondrial transcription termination\textsuperscript{182}. RNA G-quadruplex formation has also been discovered within the RNA component of telomerase (hTR/hTERC)\textsuperscript{183} – suggesting that RNA G-quadruplex structures could exist as discrete structural motifs within much larger RNA structures, in a manner similar to pseudoknot or hairpin motifs. Indeed, the RNA helicase protein, DHX36 (also G4R1, “G4-resolvase 1”), was recently shown to interact with this G-quadruplex, affecting hTR accumulation at telomeres and thereby affecting telomerase activity\textsuperscript{184}.

Progress has also been made on the role of RNA G-quadruplexes in viral genomes. For example, Shen and co-workers demonstrated that strand transfer in HIV-1 retroviral genomes is promoted by RNA G-quadruplex formation\textsuperscript{185}. A role for RNA quadruplexes was also identified in Epstein-Barr virus nuclear antigen function\textsuperscript{186}.

A significant body of evidence implicates RNA G-quadruplex structures in a wide variety of biological processes – and in a diverse range of organisms – strongly suggesting such structures to have a far more prevalent functional presence than previously thought. Structural data for RNA G-quadruplexes is yet to catch up with that of DNA G-quadruplexes, however, considering the potential significance of TERRA and mRNA G-quadruplexes in cancer research, it is likely that high-quality structural data for such RNA structures will soon become available. Indeed, providing such data has been a focus of the work described in this thesis, with X-ray crystallography being the primary structural technique used.
1.7 X-RAY CRYSTALLOGRAPHY AS A TOOL IN G-QUADRUPLEX RESEARCH

The ability to directly visualise molecules at an atomic-scale level of detail is exceedingly informative, whether studying the folding of a molecule for basic research purposes or investigating the interactions of a drug molecule with a macromolecular target as part of a drug discovery programme. There are two main methods for obtaining such data – NMR spectroscopy and X-ray crystallography – both of which have been utilised extensively in G-quadruplex research. While both methods provide high quality structural data (and indeed are highly complementary), X-ray crystallography has been the main method used in the work described here, and as such, the limitations and capabilities of this technique will be briefly considered.

Macromolecular structure determination by X-ray crystallography involves the exposure of a single, well-ordered crystal of a macromolecule to high-intensity X-rays, followed by measurement of the intensities of the diffracted X-rays. Each diffraction intensity measurement is the sum of constructive and destructive wave interference of X-rays scattered from all atoms within a plane of the crystal lattice. Specifically, it is the electrons of the target molecule which scatter the X-rays, therefore, the intensities of the diffracted X-rays are directly related to the electron density distribution of the crystal lattice. A Fourier transform can be used to calculate the electron density distribution using the diffracted X-ray intensity data plus information concerning the phase angles of the X-rays (the phase angle being the position of a wave’s maximum height relative to an arbitrary point). It is into this electron density map that the atomic model is fitted – the end product of a structure determined by X-ray crystallography is therefore an atomic model which best fits the electron density (which is derived from the experimental data).

Importantly, the experimental data is not the only information used to determine the final atomic model, as well-established parameters for macromolecular geometry (e.g. for bond lengths, bond angles and torsion angles) are incorporated into the model fitting process – this is termed ‘refinement’. Refinement parameters are included for two main reasons: 1) to ensure macromolecular geometry remains within reasonable limits, and 2) to increase the observation-to-parameter ratio. Observation-to-parameter ratio refers to the number of data points (i.e. observations) to which we
can refine the parameters of our final atomic model (the model parameters being 4n+1, where n = number of atoms). Without the inclusion of refinement parameters (i.e. geometric restraints and constraints), the 4n+1 model parameters can only be refined using the diffraction intensity observations. The inclusion of geometric restraints/constraints can therefore be considered as the addition of further experimental observations, significantly increasing the accuracy of the final atomic model. Refinement procedures are described further in methods section 2.3.4.

The resolution of the diffraction data will determine the level of detail and information provided by a given crystal structure. Diffraction data in the region of 3.5 – 4 Å will allow backbones and strand associations to be determined; data between 3 and 3.5 Å will allow nucleic acid bases to be modelled with reasonable confidence; resolution better than 3 Å will allow most components of a macromolecule to be resolved, including (concerning nucleic acid) nucleotide bases, ribose rings, backbone atoms, hydroxyl groups, ordered water molecules, ordered ions and bound ligand molecules (e.g. natural ligands or drug molecules). Ultimately therefore, high-quality crystal structures allow atom-scale details of macromolecular inter- and intra-molecular interactions.

As with all experimental methods, there are some important limitations to crystallographically determined structures, which have implications concerning the utility of the final atomic model. The first step in the crystallographic structure determination process – crystallisation of the target macromolecule – is technically one of the most challenging steps, mainly due to the difficulty in forcing molecules which have evolved in an aqueous environment (i.e. the cytoplasm) to self-associate into a well-ordered, ‘insoluble’ crystals. As well as being technically challenging, the crystalline nature of crystallographically determined structures often draws criticism for generating ‘static’ and/or ‘dehydrated’ structures, which may not be physiologically representative. This is fair criticism, but the following points need to be considered:

1) Macromolecular crystals and (and the models derived thereof) are not dehydrated, but in fact typically have solvent contents of between 20 and 70 %. This differs significantly to the ‘typical’ view
of a crystal – such as an inorganic salt crystal – which will generally be highly rigid and contain minimal solvent.

2) Macromolecular crystals are not necessarily ‘static’ – a measure of regional and atomic flexibility of a crystal structure is provided by the B-factor (also termed ‘temperature factor’) – a crystallographic parameter which defines atomic mobility (i.e. thermal motion) within the crystal lattice (measured in Å$^2$).

3) If an equivalent NMR-derived structure exists for a crystal structure, then this can provide excellent corroboration for a crystallographic model. Often NMR-derived and crystallographically determined atomic structures correlate well, albeit usually providing an alternative structural perspective.

Ultimately however, crystal structures do generate relatively rigid models (particularly compared to high-resolution NMR-derived models), as regions of high flexibility (such as loop regions) will not scatter X-rays sufficiently, and hence will not be resolved in electron density maps. Therefore, as with any experimental technique, it is beneficial to validate crystallographic data with additional high- and low-resolution biophysical techniques, such as (ideally) NMR spectroscopy, mass spectrometry or dynamic light scattering.

The ability to study the interactions of drug molecules with macromolecular targets is a particularly important use of X-ray crystallography, as this allows the rational design and improvement of therapeutic compounds. There are, however, two main limitations to this application. The first is that, for a component (such as a drug molecule) to be visualised in electron density, it must occupy an equivalent position within the crystal lattice in approximately 50 % of the repeated units (e.g. the macromolecules). This therefore requires a drug molecule with high affinity as well as high specificity for a binding site, if it is to be visualised in the final model (i.e. in electron density). A second issue with drug binding (and crystal structures in general) is that often macromolecules can form potentially biologically unexpected contacts with
neighbouring molecules in the crystal lattice. Indeed, these contacts – termed ‘crystal packing contacts’ – are absolutely required for crystal growth, however, they can often generate new surfaces and binding sites which may not exist in a cellular environment.

These limitations do not necessarily restrict the use or utility of crystallography, rather, consideration of such limitations ensures that the final model (i.e. the crystal structure) is not over-analysed or misinterpreted. Indeed, X-ray crystallography has been used to great success in G-quadruplex research, providing atomic-scale details of G-quadruplexes formed from model and wild-type sequences\(^{38,107-111}\), followed by the first high-resolution insight into the interactions of drug molecules with G-quadruplex structures formed from human telomeric sequences\(^{112-114}\) (see table 1.1, above). Such structures have formed the basis of drug discovery programmes, aimed at generating effective telomere-directed anti-cancer drugs.
The overall aim of the work described herein has been to generate high-quality structural data for G-quadruplex structures formed from human telomeric RNA and DNA – with a focus on the more recently discovered telomeric RNA, for which there is currently very limited structural data. Considering the success of X-ray crystallography in DNA G-quadruplex research, this has been the main method used in the work reported here for the generation of high-resolution structural data. In addition to X-ray crystallographic data, biophysical methods were also employed, in order to provide lower-resolution yet complementary data concerning the behaviour and folding of human telomeric G-quadruplex structures.

While the research reported here encompasses both basic research and drug discovery, a key focus has been the importance and potential application of human telomeric G-quadruplex structures to anti-cancer drug design. The structural data produced in this work consequently reflects this focus, which includes crystal structures of: a native G-quadruplex formed from human telomeric RNA (chapter 2); an RNA G-quadruplex-drug complex (solved using the same RNA sequence) (chapter 3); a drug-complex for the equivalent DNA G-quadruplex (chapter 4); and subsequently, two high-resolution intramolecular DNA G-quadruplex-drug complexes (chapter 5). The final results presented here – comprising a series of biophysical studies – are similarly balanced, consisting of basic research into telomeric RNA G-quadruplex tertiary structure (or ‘higher-order’ assemblies) followed by telomeric RNA and DNA G-quadruplex ligand-binding studies (chapter 6). An outline of the following work is given in table 1.2, below.
### Table 1.2 Outline of thesis.

<table>
<thead>
<tr>
<th>Method</th>
<th>Chapter</th>
<th>Nucleic Acid Sequence</th>
<th>Type</th>
<th>Ligand</th>
<th>Resolution/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Introduction)</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>X-ray crystallography</strong></td>
<td>2</td>
<td>r(U₉₆AGGGUUAGGGGU) RNA</td>
<td>Native RNA</td>
<td>2.2 Å</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>r(UAGGGUUAGGGGU) RNA</td>
<td>Acridine* complex</td>
<td>2.6 Å</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>d(TAGGGTTAGGGT) DNA</td>
<td>Acridine* complex</td>
<td>3.2 Å</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>d(AGG[GTTAGGG]₃) DNA</td>
<td>ND* complex</td>
<td>2.3 Å</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>(TTAGGG)n (UUAGGG)n n = 2, 4, and 8 DNA + RNA ND⁵ series, BRACO-19, native studies</td>
<td></td>
<td>CD, PAGE, ESI-MS⁶</td>
<td></td>
</tr>
<tr>
<td>(Conclusions)</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

a) acridine = di-substituted triazole acridine, FD-121; b) ND = naphthalene diimide, BMSG-SH3; c) ND = naphthalene diimide, BMSG-SH4; d) ND = naphthalene diimide series; e) CD = circular dichroism; PAGE = polyacrylamide gel electrophoresis; ESI-MS = electrospray ionisation mass spectrometry.
CHAPTER 2
X-RAY CRYSTALLOGRAPHIC STUDIES OF A NATIVE TELOMERIC RNA G-QUADRUPLEX
2.1 BACKGROUND

2.1.1 RNA G-QUADRUPLEXES IN THE GENOME

RNA sequences capable of forming G-quadruplex structures have been identified throughout the human genome (as well as in the genomes of lower organisms) but there is, as yet, very limited high-resolution structural data available for these potential (or proved) structures. This is surprising, as many of the cellular processes and roles in which RNA G-quadruplex structures are thought to be involved are strongly linked to human pathologies – particularly cancer – although many other potential diseases are implicated. The key regions of the human genome and cellular roles in which RNA quadruplex structures have been implicated includes:

1) 5' and 3' untranslated regions of mRNA transcripts\(^{43,44}\);
2) the coding regions of FXS mRNA\(^{177,178}\);
3) mRNA processing involving mXRP1 exoribonuclease\(^{140}\);
4) IGF-II regulatory regions\(^{141}\);
5) mitochondrial transcription termination\(^{182}\);
6) the RNA component of telomerase (hTR)\(^{183,184}\); and
7) telomeric repeat RNA (TERRA)\(^{143,144}\).

Although all of the above mentioned putative-G-quadruplex-forming RNA sequences are of potential biological importance, the research described in the following chapters is focused on G-quadruplexes formed from human telomeric repeat RNA (TERRA). As outlined in section 1.5.2, human telomeres were recently shown to be actively transcribed – a finding that contradicted the current understanding – with the resulting transcripts composed of hundreds of tandem r(UUAGGG) repeats\(^{143,144}\). These RNA molecules – termed TERRA or telRNA – are emerging as important components of telomere biology, and have been found to be involved in several key telomere-related processes, such as regulation of telomere length, regulation of telomerase activity and chromatin remodelling\(^{143,144,148}\). Shortly after the identification of TERRA, these G-rich RNAs were shown to form stable G-quadruplex structures,
analogous to the equivalent G-rich telomeric DNA sequences\textsuperscript{152,153,161}. The focus of the following chapters has been to provide structural data for TERRA G-quadruplexes, in order to increase the current understanding of these newly discovered RNA molecules. Such structural data could inform on RNA folding and stability, and aid in the identification of protein binding partners. Additionally, it was hoped that such data could form the basis for the rational design of telomeric RNA-directed therapeutic agents. To this end, a high-resolution crystal structure of a G-quadruplex formed from human telomeric RNA (TERRA) was determined, the utility and possible importance of which will be described and discussed in this chapter.

2.1.2 CURRENT HIGH-RESOLUTION STRUCTURAL DATA AVAILABLE FOR RNA G-QUADRUPLEXES

The utility of structural data falls into two categories: 1) basic research (what does this molecule do and why?); and 2) drug discovery (can we exploit and control the properties of this structure with therapeutic compounds?). With the relatively recent discovery of TERRA – and TERRA G-quadruplex structures – there is a need to provide structural data for these molecules to satisfy both categories of research (i.e. basic research and drug discovery). High-resolution structural data currently available for telomeric RNA G-quadruplexes is limited to a single NMR structure, involving a bimolecular G-quadruplex formed from the human repeat sequence, r(UAGGGUUAGGGU)\textsuperscript{159} (figure 1.10, page 55). This structure provided the first atomic-scale details of TERRA G-quadruplex folding, and gave insight into the sugar puckering and geometry of these RNA structures.

This is the extent to which human TERRA G-quadruplexes have been studied with high-resolution techniques, however, a series of ultra-high resolution crystal structures of RNA tetraplexes are available, providing exquisite details of RNA tetraplex folding\textsuperscript{187–192}. This series of tetraplex crystal structures, solved by Pan and co-workers, involves six different tetraplexes formed from various related hexamer ribo-oligonucleotide sequences (typically based around a ‘UGAGGU’ motif). These structures allow the direct visualisation of individual hydrogen bonding contacts, sugar pucker geometry and cation coordination. Interestingly, a wide variety of cations ($\text{Na}^{+}$,}
K⁺, Ba²⁺, Sr²⁺) were identified in various binding sites within these tetraplex structures — including Ba²⁺ dication coordination within a G-quartet

2.1.3 CRYSTALLOGRAPHIC STRUCTURAL DATA AVAILABLE FOR TELOMERIC DNA G-
QUADRUPLEXES

The use of crystallography to study telomeric DNA G-quadruplex structures dates back almost 20 years, with the first crystal structure being reported in 1992 of a bimolecular G-quadruplex formed from the O. nova telomeric repeat sequence, d(GGGTTTTG)¹⁰⁷ (which was subsequently retracted). Following from this, many tetrameric G-quadruplexes formed from ‘model’ sequences (such as, d[TGGGT]) have been studied³⁸,¹⁰⁸, as well as a determination of a validated O. nova G-quadruplex¹⁰⁹. Both the tetraplexes and the O. nova G-quadruplex have been studied in complexes with small molecule compounds (discussed in chapter 3), however, the G-quadruplex crystal structures of most relevance here are those solved by Parkinson and co-workers in 2002, involving G-quadruplexes formed from human telomeric repeat DNA. In this work, two crystal structures were solved — one describing a bimolecular quadruplex formed from the human repeat sequence d(TAGGGTTAGGGT), and the second structure describing an intramolecular quadruplex formed from the sequence d(AGGG[TAGGG])₃. Both of these G-quadruplexes were shown irrefutably to be parallel stranded — a finding that contrasted starkly to the anti-parallel arrangement of the current NMR-derived structure. Regardless of the topology controversy, these human quadruplex crystal structures were of sufficiently high quality to form the basis for the rational design of human telomeric G-quadruplex directed drugs. While many subsequent crystal structures of human telomeric G-quadruplex-ligand complexes have been solved⁹⁷,¹¹²-¹¹⁴, these native structures proved to be highly useful for initial drug design, as well as in providing a potential model for telomeric DNA G-quadruplex ‘higher-order’ structure.
2.2 AIMS

The crystallographic work performed by others on model RNA tetraplexes and human telomeric DNA highlights the utility of crystallography in basic structural research. A good-quality crystal structure (i.e. better than 3 Å resolution) can yield a quite incredible amount of information, ranging from gross overall structural features (such as RNA folding, topology, base-orientation, and sugar pucker conformation), to finer structural details (such as water structure, intramolecular hydrogen bonding and ion-binding). For this reason, crystallographic studies were undertaken in order to provide further details of G-quadruplex structures formed from human telomeric RNA (TERRA). The main questions set out to be answered were:

1) What are the specific intramolecular contacts which hold TERRA G-quadruplex structures together?
2) How do the C2'–OH groups influence the quadruplex?
3) Is the crystal structure in agreement with the NMR structure?
4) How does the RNA G-quadruplex differ from the equivalent DNA G-quadruplex and why?
5) Can these differences be exploited for drug-design?
2.3 METHODS AND MATERIALS

2.3.1 SAMPLE PREPARATION

The RNA sequence used in these studies, r(U_GGGUAGGGU), was purchased from Eurogentec (Belgium) (HPLC purified) and used without further purification (U_G corresponds to 5-Bromo-uracil). A stock solution was prepared by dissolving the lyophilized RNA in RNase-free water to a final single-strand concentration of 3 mM. Prior to annealing, the RNA was further diluted to 1.5 mM single-strand concentration and buffer exchanged into 50 mM potassium cacodylate (pH 6.5) and 30 mM potassium chloride. An annealing protocol was used in order to ensure G-quadruplex formation, the procedure consisted of heating the RNA sample to 90°C for 5 minutes (in order to disrupt any unspecified secondary structure) followed by slow cooling to room temperature (overnight) to allow correct G-quadruplex folding to take place. The annealing protocol used here has been repeated exactly for all G-quadruplex samples described in this thesis (unless stated otherwise).

2.3.2 CRYSTALLISATION

Crystals of the RNA G-quadruplex were grown using the hanging drop vapour diffusion method, in Qiagen screw-cap 24-well plates. Crystals were grown in hanging drops composed of 1 μL of the annealed RNA sample mixed with 1 μL of the crystallisation solution (composed of 15% 2-methyl-2,4-pentanediol [MPD], 150 mM sodium chloride, 50 mM sodium cacodylate [pH 6.5] and 5 mM spermine). This drop was equilibrated against a well containing 15% MPD. Crystals appeared overnight and were left to grow for 1 week before data collection. Final crystal dimensions were approximately 50μm x 50μm x 100μm (figure 2.1).

2.3.3 DATA COLLECTION AND PROCESSING

X-ray diffraction data were collected on a single RNA crystal flash-frozen in a nitrogen stream (105 K), on an Oxford Diffraction Xcalibur NovaT X-ray machine. This closed-
tube home-source generates standard Cu Kα x-rays of wavelength 1.542 Å. 110 frames of data were collected in 1.0° slices, with an exposure time for each frame of 5 minutes. Diffraction data were collected to 2.2 Å resolution (figure 2.2), and were processed (indexed, integrated and merged) using CrysalisPRO (Oxford Diffraction) and scaled using Scala193 (from the CCP4 suite194). The crystals belong to space group P3121, with unit cell dimensions of a = b = 57.58 Å, c = 38.38 Å, α = β = 90°, γ = 120°. Data collection statistics are shown in table 2.1.

Figure 2.1 RNA G-quadruplex crystals formed from the sequence, r(U_BrAGGGUUAGGGU). RNA G-quadruplex crystals grown in hanging drops using MPD as a precipitant. The precipitate visible here is a result of non-specific RNA aggregation induced by the spermine additive present in the crystallisation conditions. Although precipitate is generally unfavourable in crystallisation experiments, in this case, the presence of this polyamine additive was critical for crystal growth and did not seem to reduce crystal quality. These crystals are approximately 100 μm in the longest dimension.

Figure 2.2 X-ray diffraction from a native RNA G-quadruplex crystal. Diffraction image recorded on a CCD detector (Titan, Oxford Diffraction) produced from a 5 minute exposure of an RNA crystal to CuKα X-rays. The spots are sharp and well resolved and produced good-quality electron density maps.
2.3.4 STRUCTURE DETERMINATION AND REFINEMENT

Although the RNA oligonucleotide used in these studies contained a Bromo-uracil modification, it was not necessary to utilize the anomalous scattering potential of the bromine to obtain phase information, as this structure solved easily by molecular replacement. The molecular replacement program Phaser (from CCP4) was used to obtain initial phase angle estimates\textsuperscript{195}, using the native human telomeric DNA G-quadruplex structure (PDB Id: \textbf{1K8P}) as a search model\textsuperscript{40}. This search model was modified to exclude all non-guanine components, i.e. only the three stacked G-quartets were used as a search model. Initial $2F_o - F_c$ maps showed clear electron density for the G-quartets and potassium ions, as well as residual density for the loop residues and bromines (seen in $F_o - F_c$ maps also).

Model building and structure refinement were performed using Coot\textsuperscript{196} and Refmac5\textsuperscript{197}, using the maximum-likelihood restrained refinement method. This refinement method involves adjustment of the atomic model (i.e. adjustment of atomic coordinates, B-factors and translation-libration-screw [TLS] parameters) to best fit the experimental data (i.e. the observed structure factor amplitudes generated from single-crystal X-ray diffraction). ‘Restrained’ refinement refers to the inclusion of known geometric parameters in the refinement process, in order to: 1) ensure the geometry (e.g. bond lengths, bond angles, torsion angles and aromatic planarity) of the final model is in agreement with accepted measurements, and 2) to increase the observation-to-parameter ratio.

Refinement was monitored by use of the ‘residual’ ($R_{\text{work}}$) and ‘free residual’ ($R_{\text{free}}$) factors – these methods provide measures of how well the refined model correlates to the experimental data\textsuperscript{198}. For the $R_{\text{free}}$ calculations, 5% of the observed structure factor amplitude data were set aside before the refinement process and used to provide an independent assessment of model accuracy. As well as atomic coordinate positions, individual isotropic B-factors were also refined throughout the process, with translation, libration, screw (TLS) motion determination parameters included towards the end of the refinement (each RNA strand was divided into two TLS groups)\textsuperscript{199,200}. B-factors provide a measure of atomic mobility within the crystal lattice, while TLS parameters allow regional mobility (such as that of loop regions) to be
accounted for during the refinement process. The refinement protocol described here was used for all structures reported in this thesis, unless stated otherwise.

The final model was refined using data between 11.69 Å and 2.2 Å, with final R and R_free values of 21.6% and 23.1% respectively. The final overall (isotropic) B-factor for the model (including solvent molecules) is 24.26 Å². Additional refinement statistics are shown in table 2.1.

### 2.3.5 VALIDATION AND DEPOSITION IN THE PROTEIN DATABANK

The final atomic model was validated using the RCSB web server (www.deposit.rcsb.org) and has no bond lengths or bond angles greater than 6 times the root mean square deviation (RMSD) from standard dictionary values. The RMSDs of bond lengths and bond angles is 0.013 Å and 0.859° respectively.

The atomic coordinates and structure factors have been deposited in the Protein Databank with accession code 3IBK (deposition date: 16/7/09). This structure was first reported in the Nucleic Acids Research journal in 2010 (see Appendix A).

### 2.3.6 STRUCTURE ANALYSIS AND VISUALISATION SOFTWARE

All structures reported in this thesis have been analysed using 3DNA and PyMol. All figures have been prepared using PyMol.
Table 2.1 *Data collection and refinement statistics for the telomeric RNA G-quadruplex crystal structure.*

<table>
<thead>
<tr>
<th><strong>Data collection</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>r(UbrAGGGUUAGGGU)</td>
</tr>
<tr>
<td>Space group</td>
<td>P3$_1$21</td>
</tr>
<tr>
<td>Unit cell dimensions:</td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>57.58, 57.58, 38.38</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.00, 90.00, 120.00</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>49.87 - 2.20 (2.25 - 2.20)</td>
</tr>
<tr>
<td>R$_{int}$ (%) overall</td>
<td>5.0 (24.0)</td>
</tr>
<tr>
<td>I / σ</td>
<td>43.38 (4.54)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.6 (97.3)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>6.3 (6.0)</td>
</tr>
</tbody>
</table>

| **Refinement**      |                                                                 |
| Resolution (Å)      | 11.69 - 2.20 (2.25 - 2.20)                                       |
| Reflections         | 3899                                                             |
| R$_{work}$ / R$_{free}$ (%) | 21.6 / 23.1 (29.4 / 19.6)                                       |
| No. of atoms        | 520                                                              |
| Ions                | 2                                                                |
| Water               | 45                                                               |
| Overall B-factor (Å$^2$) | 24.26                                                           |

| **RMS deviations**  |                                                                 |
| Bond-lengths (Å)    | 0.013                                                            |
| Bond-angles (°)     | 0.859                                                            |
| PDB ID              | 3IBK                                                             |

(Values in brackets refer to highest resolution shell, 2.28-2.20 Å)
2.4 RESULTS AND DISCUSSION

2.4.1 CRYSTAL STRUCTURE OF A TELOMERIC RNA G-QUADRUPLEX: OVERALL TOPOLOGY AND GENERAL FEATURES

The structure of an RNA G-quadruplex was determined by X-ray crystallographic methods, revealing a bimolecular parallel-stranded G-quadruplex formed from the human telomeric repeat sequence r(UBrAGGGUUAGGGU) (figure 2.3). The two strands of this bimolecular structure are identical in sequence and largely identical in structure, with only slight differences in loop geometry. The parallel orientation of the strands is maintained by the presence of two propeller-type loops, each composed of the tri-nucleotide sequence, UUA. The core of the G-quadruplex is defined by three G-quartets, which are stabilized by two centrally coordinated potassium ions. The biological unit is described by the asymmetric unit (i.e. one bimolecular quadruplex) however, pseudo-tertiary structure can be seen in the crystal lattice, as two quadruplexes associate via Watson-crick/Hoogsteen basepairing at the 5’ termini. This results in a highly repetitive-stacked structure involving three layers of G-quartets plus two layers of UAUA tetrads plus three further G-quartet layers (figure 2.4).

The final RNA structure was refined to 2.2 Å resolution, producing electron density maps of generally excellent quality, allowing the majority of the RNA to be accurately modelled (figure 2.5). Importantly, the hydroxyl groups of the ribose rings can be accurately placed into electron density (ensuring the ribose puckering is accurate), as well as hydration structure and coordinated cations. The only significant region of the RNA which cannot be placed into electron density is the UUA loop of strand B (described and discussed further below).
Figure 2.3 Overall folding and topology of the telomeric RNA G-quadruplex crystal structure.
The sequence r(U_{Br}AGGGUUAGGGU) folds as a parallel stranded bimolecular G-quadruplex, with
propeller-type loops and two centrally coordinated potassium ions. Top: side-view of the G-quadruplex
(one asymmetric unit), showing the stacking arrangement of the G-quartets (green), with potassium ions
(purple spheres) coordinate bonded between G-quartet layers. Bottom: view onto the 5' surface,
showing the central pore/channel created by the G-quartets. Inset: simple schematic representation of
the structure. The box corresponds to three stacked G-quartet layers. The cartoon representations have
been coloured according to the Nucleic Acid Database colour scheme, which has been used throughout
this thesis: adenine = red; guanine = green; uracil = cyan. The RNA backbone is shown as a grey ribbon.
Unless stated otherwise, atomic 'stick' representations are coloured as follows: carbon = green, oxygen
= red, phosphorous = orange, nitrogen = blue, potassium = purple.
a) G-quadruplex stacking

Asymmetric unit 1

G-quartets

UAUA tetrads

Asymmetric unit 2

b) UAUA tetrad

c) G-quartet

Figure 2.4 Two asymmetric units associating through UAUA tetrads
a) Two quadruplexes associate in the crystal lattice at the 5'-5' interface, joined by two layers of UAUA tetrads. This results in a highly stacked structure involving eight quartets (or tetrads). b) UAUA hydrogen bonding. c) G-quartet hydrogen bonding (all G-quartet hydrogen bonds are between 2.7 and 2.9 Å).

Figure 2.5 Electron density for the RNA G-quadruplex.
2F_o-F_c map contoured at 1.0 σ level. Electron density for the RNA is generally of excellent quality, with the only region of poor density being the UUA loop of strand B. The majority of this loop has been assigned low-to-zero occupancy, and has been assigned geometry based on the strand A UUA loop.


2.4.2 PROPELLER-LOOP STRUCTURE

Tri-nucleotide propeller loops – which link the G₃ tracts – have shown an unexpected propensity in telomeric DNA G-quadruplex structures to rearrange significantly (compared to ‘native’ arrangements) to accommodate ligand binding\textsuperscript{112,114}. For this reason, it is of use to study and understand the factors which are involved in DNA (and RNA) propeller loop structure. The RNA quadruplex described in this work involves two propeller-type loops, however, only one of these loops is accurately modelled into electron density. Analysis of the crystal packing interactions suggests that differential lattice contacts may be responsible for differences in loop arrangement (figure 2.6).

The propeller loop of strand A is stabilized by the π-π stacking interaction of uracil-7 with a symmetry-related 3’ G-quartet (figure 2.6), the remaining nucleotides of this loop then form a stable (based on the high quality of electron density, figure 2.5) stacked arrangement, involving the intercalation of adenine 8 between uracil 6 and uracil 7 (figure 2.3) – an arrangement which is remarkably similar to the propeller loops observed in several human telomeric DNA G-quadruplex structures (PDB IDs, 1KF1, 1K8P and 3CE5). As well as the stabilizing interaction of the strand A propeller loop with a symmetry-related G-quartet, this loop is also stabilized by intramolecular hydrogen bonding interactions and water bridging contacts (figure 2.7). In contrast to the propeller loop of strand A, the strand B loop is presented with a different packing environment in the crystal lattice (figure 2.6) and is not stabilized by any stacking interactions. It is highly likely that this lack of crystal packing interaction is the reason that the loop of strand B is not well ordered and cannot be placed accurately into electron density. This is slightly surprising, as it could be expected that the additional hydrogen bonding afforded by the C2’-OH groups (compared to the equivalent DNA structure) may stabilize the loop structure of strand B to allow it to be resolved in electron density maps. The fact that this is not the case suggests that the typical ‘UAU’/’TAT’ stacked arrangement seen in human telomeric G-quadruplex structures is actually dependent upon crystal packing stabilization, and that these loop regions are likely to be as flexible as implied in the available NMR structures\textsuperscript{39,73,74,77,79}. Justification for this statement comes from the observation that to date, stacked tri-nucleotide loops have only ever been resolved in electron density when at least one of
the residues is stabilized by crystal packing \( \pi \)-stacking interactions – disordered trinucleotide loops of human telomeric G-quadruplex structures can be seen in the human bimolecular structure 1K8P and the human intramolecular quadruplex-complex 3SC8 – in both cases, the disordered loops lack stabilising crystal packing contacts. It should be noted, however, that crystal packing contacts are not the only factor which can influence loop geometry, and if a ligand is included in the system – and is bound on or near any loop residues – then this can significantly affect the loop structure, typically stabilizing the loop in a conformation more favourable for ligand binding\(^{112,114}\).

![Strand A UUA loop stabilized by \( \pi \)-\( \pi \) stacking interactions.](image)

![Disordered strand B UUA loop](image)

*Figure 2.6 Crystal packing interactions: implications for loop structure.*

Three asymmetric units are shown here, revealing the crystal packing contacts available to the two loops of the focal quadruplex (asymmetric unit 2, red). The UUA loop of strand A is stabilised by \( \pi \)-\( \pi \) stacking interactions with the 3' G-quartet of a symmetry related molecule (yellow). In contrast, the UUA loop of strand B does not have such stacking contacts available to it (blue molecule) and hence is disordered in the crystal lattice, as a result of either dynamic or static disorder.
2.4.3 OVERALL COMPARISON TO EQUIVALENT NMR AND DNA STRUCTURES

Before this RNA structure is described further, the equivalent DNA crystal structure and RNA NMR model will be considered, as these impact upon the significance of many of the features of this RNA quadruplex. The RNA quadruplex crystal structure described here is highly similar to that of the equivalent DNA quadruplex 1K8P (formed from the sequence d[UBrAGGGUBrTAGGGT]) and is in agreement with the NMR structure 2KBP (sequence r[UAGGGUUAGGGU]), both of which also fold as parallel-stranded bimolecular G-quadruplexes.

Structural alignment of the RNA and DNA quadruplex crystal structures reveals a high level of similarity between these two molecules (1K8P and 3IBK), which includes correlations in loop structure, G-quartet arrangement and the orientation of the terminal residues (U1 and A2 especially) (figure 2.8a). The similarity of these two structures is reflected in the low root mean square (RMS) deviation value generated from the structural alignment (RMS deviation, 0.99 Å [470 atoms]). Despite sharing general structural features, there are yet subtle differences between the DNA and RNA crystal structures, particularly with regard to sugar puckering, backbone torsion angles and hydrogen bonding contacts, the significance of which will be considered below.

Figure 2.7 Hydrogen bonding within the loop region.
The C2'-OH groups stabilize the UUA loop region of strand A through direct (i.e. intramolecular) contacts with the RNA, as well as through water-bridging contacts. However, these hydrogen bonding interactions do not appear to stabilize the loop of strand B sufficiently enough for it to be resolved in electron density. Note, for clarity not all hydrogen bond lengths are shown. All omitted bond lengths are below 3.0 Å. Hydrogen bonds are shown as black dashes.
Structural alignment of the RNA crystal structure reported here with the equivalent NMR model (2KBP), reveals differences not observed when comparing the RNA crystal structure with DNA crystal structure. The three G-quartets show a reasonable level of similarity (G-quartet RMS deviation, 1.011 Å [246 atoms]), but there are pronounced differences in loop structure and terminal residue orientation when comparing these structures (figure 2.8b). Similarity in G-quartet arrangement is to be expected, as the G-quartet is a particularly stable motif, with eight hydrogen bonds holding four bases together. The differing loop structure is likely to be a consequence of the different techniques used to solve these quadruplexes. In solution, the propeller loops appear to be extremely flexible, with multiple orientations and no definitive single energy minima conformation, whereas crystal packing contacts stabilise the UUA loop (of strand A) in a single ‘UAU’ stacked arrangement. Despite these differences in loop geometry, the NMR structure (2KBP) provides an excellent confirmation of the validity of the crystal structure reported here.

**Figure 2.8** Structural alignments of telomeric G-quadruplex structures.

a) Structural alignment of the telomeric RNA crystal structure described in this work (3IBK – blue) and the equivalent DNA structure (1K8P - red). These quadruplexes show a significant degree of similarity, with overall RMSDs of 0.99 Å and 0.539 Å for the whole model and quartets alone respectively. b) Structural alignment of the telomeric RNA crystal structure described in this work (blue) and the equivalent NMR model (2KBP - yellow). The core regions of these quadruplexes overlay well (i.e. the G-quartets), with deviations mainly seen in the loop and terminal regions. RMSDs of this overlay are 2.971 Å and 1.011 Å for the whole structure and G-quartets alone respectively. RMSD analyses were performed using PyMol.
2.4.4 SUGAR PUCKERING AND BACKBONE TORSION ANGLES

The five membered sugar rings of nucleotides are not planar, and adopt predictable well-studied 'buckled' (or 'puckered') conformations, depending upon sequence identity, base identity and ribose substituents. With a hydroxyl group present at the C2' position of the ribose sugar, RNA ribose rings generally adopt a C3'-endo sugar pucker conformation, with DNA deoxyribose rings generally adopting C2'-endo pucker conformations. These preferences for pucker conformation are not constraints however, exemplified by the RNA structure described here, which displays a mix of C2'-endo and C3'-endo sugar pucker conformations (table 2.2). C3'-endo sugar puckers are seen for the terminal Ugr1 and A2 residues, as well as the guanines of the central G-quartet. The pattern of C3'-endo and C2'-endo puckering is generally consistent with the NMR structure reported for this sequence – both solution and crystal structures have uniform C3'-endo puckers for the guanine residues of the central quartet (G4, G10, G16 and G22), and a mix of C2'-endo and C3'-endo puckers for the guanines of the outer G-quartets (highlight light grey, table 2.2). Sugar pucker consistency between the RNA and DNA crystal structures is also observed to some degree, as the strand A propeller-loops of both quadruplexes adopt the same pattern of sugar puckering (highlighted dark grey, table 2.2).

There are, however, some significant differences between the RNA and DNA crystal structures, primarily with respect to the sugar pucker conformations of non-loop residues (and intramolecular interactions – discussed below). Aside from sharing sugar pucker distribution for the strand A UUA/TTA loop residues, the majority of the sugar puckers for the RNA G-quadruplex differ to those of the DNA. Comparison of the backbone torsion angles of the RNA and DNA quadruplexes shows that, although the backbone torsion angles of these two structures correlate well with each other, there are some significant deviations, which may explain how the two structures are able to adopt the same overall topology, despite differences in sugar pucker preference (and available intramolecular interactions) (figure 2.9). An example is residue G10, which is situated within the plane of the central quartet in both DNA and RNA structures, with a C2'-endo sugar pucker for the DNA and a C3'-endo pucker for the RNA quadruplex. A detailed comparison of the backbone and glycosidic torsion angles of G10 reveals
correlated differences in the backbone dihedral angles $\alpha$, $\gamma$, $\delta$ and $\chi$, suggesting that this difference in pucker is compensated for by changes in the torsion angles to allow the overall orientation of this residue to be conserved in the two structures (figure 2.10).

Table 2.2 Distribution of sugar puckering for three related G-quadruplex structures.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Strand A</th>
<th>Strand B</th>
<th>Strand A</th>
<th>Strand B</th>
<th>Strand A</th>
<th>Strand B</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA (3IBK)</td>
<td>DNA (1K8P)</td>
<td>NMR RNA (2KBP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 U 13</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C2'-exo</td>
<td>C2'-exo</td>
<td>C3'-exo</td>
<td>C3'-exo</td>
</tr>
<tr>
<td>2 A 14</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C2'-endo</td>
<td>C3'-exo</td>
<td>C2'-endo</td>
<td>C3'-exo</td>
</tr>
<tr>
<td>3 G 15</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
<td>C1'-exo</td>
<td>C1'-exo</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>4 G 16</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C4'-exo</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>5 G 17</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
<td>C3'-exo</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>6 U 18</td>
<td>C1'-exo</td>
<td>-</td>
<td>C1'-exo</td>
<td>-</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>7 U 19</td>
<td>C3'-endo</td>
<td>-</td>
<td>C3'-endo</td>
<td>-</td>
<td>C3'-exo</td>
<td>C3'-exo</td>
</tr>
<tr>
<td>8 A 20</td>
<td>C2'-endo</td>
<td>-</td>
<td>C2'-endo</td>
<td>-</td>
<td>C3'-exo</td>
<td>C3'-exo</td>
</tr>
<tr>
<td>9 G 21</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
<td>C2'-endo</td>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>10 G 22</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C2'-endo</td>
<td>C3'-exo</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>11 G 23</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
<td>C2'-endo</td>
<td>C3'-exo</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>12 U 24</td>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
<td>C2'-endo</td>
<td>C3'-exo</td>
<td>C2'-endo</td>
</tr>
</tbody>
</table>

Correlated puckering between the DNA and RNA crystal structures is highlighted dark grey, correlated puckering between the two RNA structures is highlighted light grey. Sugar pucker analysis was performed using 3DNA202.
Figure 2.9 Comparison of backbone torsion angles of RNA and DNA G-quadruplex crystal structures. Backbone and glycosidic torsion angle comparison of RNA (3IBK) (solid black trace with squares) and DNA (1K8P) (grey dashes with diamonds) crystal structures. Dihedral angles are plotted separately; in order from (a) to (g)—alpha, beta, gamma, delta, epsilon, zeta, chi. Significant deviations between the two structures for residue G10 are circled (black). In all the graphs, the x-axis shows residue and the y-axis shows torsion angle. Note, negative torsion angles are converted to positive values to allow meaningful comparison.
The deoxyribose sugar of G10 of the DNA (a) adopts a C2'-endo conformation. A C2'-endo conformation for the ribose of the equivalent residue in the RNA structure (b) causes the C2'-OH group to sterically clash with the O3' oxygen. The G10 ribose of the RNA therefore adopts a C3'-endo conformation, which is accompanied by alterations in backbone dihedral angles which preserves the positioning of the guanine base in within the G-quartet. Dihedrals angles are labelled in (b).

2.4.5 C2' HYDROXYL GROUP INTERACTIONS AND WATER STRUCTURE

The RNA ribose sugars, with hydroxyl groups present at the C2' position, have the potential to make many more inter- and intra-molecular interactions than their DNA counterparts. The RNA structure described here shows the hydroxyl groups to be preferentially forming intramolecular interactions with the quadruplex itself, rather than intermolecular contacts with solvent molecules (figure 2.11). As a result, the RNA G-quadruplex has fewer observed structured water molecules than the equivalent DNA G-quadruplex, and is also accompanied by an increase in intramolecular contacts.

The C2'-OH groups are seen to form hydrogen bonds with a variety of acceptors within the RNA molecule, and interestingly, the type of intramolecular interaction appears to be dependent on the pucker conformation (figure 2.12). For example, within the quadruplex grooves, the hydroxyl groups of the sugars with a C3'-endo pucker are positioned such that they can form a hydrogen bond with the O4' oxygen atom of the neighbouring residue. Alternatively, the hydroxyl groups of the sugars with a C2'-endo pucker preferentially hydrogen bond to the N2 amine group of the guanine base. For example, the C2'-OH group of residue G3 (which has a C2'-endo pucker) forms a hydrogen bond with the N2 amine group of G22 (circled black in figure 2.12). This interaction is also seen between the C2'-OH group of G9 and the N2 amine group.
of G4 (figure 2.13a). Additionally, the C2'-OH groups interact with phosphate oxygens and O5' oxygens of the backbone (figure 2.13b) – these contacts would be expected to further contribute to the overall stability of the RNA quadruplex.

The increased number of intramolecular contacts available to the RNA G-quadruplex (compared to the equivalent DNA) would be expected to result in a more rigid (or stable) structure. The ability of the backbone C2'-OH groups to hydrogen bond directly to guanine bases of the G-quartets may be particularly significant, as this could arguably result in further stabilisation of the G-quartet core.

Figure 2.11 Hydrogen bonding and water structure within DNA and RNA G-quadruplex groove regions. 

a) Hydrogen-bonding interactions within the DNA quadruplex groove region are limited to intermolecular interactions with water molecules. b) Hydrogen bonding within the equivalent region of the RNA quadruplex includes both intramolecular and intermolecular interactions. With many of the hydrogen bonding groups of the RNA quadruplex involved in intramolecular contacts, there is an associated reduction in ordered water molecules within this structure, when compared to the equivalent DNA structure. Water molecules are shown as red spheres and hydrogen bonds are shown as black dashes.
**Figure 2.12** Correlation between sugar pucker and C2' hydroxyl group interaction.
The hydroxyl groups of residues with C3'-endo puckered sugars interact with neighbouring O4' atoms, whereas the hydroxyl group of residue G3, which has a C2'-endo puckered sugar, is seen to interact with the N2 atom of a guanine base (circled black).

**Figure 2.13.** C2'-OH group interactions: G-quartet and backbone stabilisation.
a) Additional G-quartet hydrogen bonds afforded by the C2'-OH groups of residues G3 and G9 (circled black). Increased G-quartet hydrogen bonding could be expected to increase the structural stability of the G-quartet core (relative to the equivalent DNA structure). b) Examples of C2'-OH hydrogen bonding with atoms of the RNA backbone. Hydrogen bonds are shown as black dashes, with bond distances indicated (all G-quartet hydrogen bonds are between 2.7 and 2.9 Å).
NMR and crystallographic studies of human telomeric DNA G-quadruplexes have revealed a vast and unexpected structural diversity, dependent upon cation identity, molecular crowding conditions and sequence context. In contrast, all studies of human telomeric RNA G-quadruplex structures have revealed a universal parallel fold, regardless of the environmental conditions. The structure described here is the first crystal structure of a telomeric RNA G-quadruplex, revealing a parallel-stranded bimolecular quadruplex, which is in agreement with the available NMR structure. The work here builds on this NMR structure by providing details of individual C2'-OH group interactions and water structure, which appear to be important factors in understanding the overall stability and behaviour of TERRA G-quadruplexes.

2.5.1 COMPARISON OF RNA AND DNA G-QUADRUPLEX STRUCTURES

The crystal structure described here provides a unique opportunity to directly compare equivalent DNA and RNA structures at an atomic-scale level of detail, aiding in the further understanding of telomeric G-quadruplexes by providing structural rationalisation of biophysical data (e.g. circular dichroism and UV-melting). Consideration of the available biophysical data for telomeric RNA and DNA G-quadruplexes reveals two interesting differences between DNA and RNA quadruplexes:

1) Telomeric RNA quadruplexes (and RNA quadruplexes in general) adopt a universal parallel fold, and;

2) Telomeric RNA G-quadruplex structures are more thermodynamically stable than equivalent DNA structures.

The first difference here can be explained easily, and is a result of the steric clashing of the C2'-OH group with neighbouring base and backbone atoms, which restricts RNA nucleotides to the anti glycosidic conformation. Anti-parallel quadruplex structures
require a proportion of the nucleotides to adopt a syn conformation (to allow appropriate Hoogsteen hydrogen bonds to be formed with the neighbouring anti-parallel strand). Syn glycosidic angles are not available to RNA, thus, RNA quadruplex structures are sterically restricted to adopt a parallel topology.

The difference in stability between RNA and DNA quadruplexes is perhaps not as easily understood, but can be explained partly based on the information provided by the crystal structure reported here. The C2'-OH groups of the RNA sugar rings are seen in the crystal structure to preferentially interact with the RNA itself, rather than with solvent molecules. The C2'-OH groups interact with a variety of hydrogen bond acceptors, including phosphate and backbone oxygen atoms (O3', O4' and O5') and polar groups attached to the bases (such as the N2 amine group of the guanine base) (see figures 2.11-2.13). These contacts would be expected to result in an overall increase in stability compared to equivalent DNA quadruplex structures, with a resulting loss in ordered water molecules. These expectations are consistent with current biophysical data, for example, Arora and co-workers showed a 24-mer telomeric RNA G-quadruplex (of the sequence r[UUAGGG]_4) to have a ΔT_m value 12.6°C higher than the equivalent DNA quadruplex (in K')\(^{161}\). In the same work, osmotic stress analysis showed RNA G-quadruplexes to have fewer associated water molecules when compared to equivalent DNA structures. Additionally, recent CD analysis and melting studies of various potential G-quadruplex-forming RNA sequences have shown a marked preference for these sequences to form parallel topologies, with higher melting temperatures in potassium when compared to their DNA equivalents\(^{162}\).

2.5.2 IMPLICATIONS FOR DRUG DESIGN

Telomeric DNA G-quadruplex structures have been studied extensively as anti-cancer drug targets, with high-resolution structural techniques (i.e. NMR and X-ray crystallography) playing a major role in this field. The considerable similarity between the RNA structure described here (3IBK) and the equivalent DNA quadruplex structure (1K8P) suggests that designing a drug which can selectively bind DNA G-quadruplex structures over equivalent RNA structures could be rather challenging. Such selectivity may be desirable as it was recently shown that TERRA molecules can bind and inhibit
telomerase activity – the proposed mechanism involves TERRA hybridizing to the intrinsic telomeric-repeat template present in the RNA subunit of telomerase (hTR), thereby inhibiting telomere elongation\textsuperscript{144,148}. It would be expected that TERRA would need to be unfolded to achieve this, consequently, stabilisation of TERRA G-quadruplex structures could conceivably increase telomerase activity (by reducing TERRA-hTR hybridisation). This scenario is yet to be proved experimentally, but it is perhaps worth considering that any ligand designed to bind and stabilize telomeric G-quadruplex DNA would likely bind and stabilize telomeric G-quadruplex RNA in a similar fashion, potentially having the opposite of the intended effect. This is of course speculative, and may be overshadowed by the surprising behaviour of human telomeric RNA quadruplexes when complexed to small molecule compounds, as will be described in the following chapter.
CHAPTER 3

X-RAY CRYSTALLOGRAPHIC STUDIES OF A TELOMERIC RNA G-QUADRUPLEX-ACRIDINE COMPLEX
3.1 BACKGROUND

3.1.1 RNA G-QUADRUPLEXES AS DRUG TARGETS

Until relatively recently, G-quadruplex-targeted drug design has been directed primarily towards G-quadruplexes formed from DNA – the two main areas of DNA G-quadruplex drug-design research being gene promoter region G-quadruplexes and telomeric G-quadruplexes. Of the many regions of the genome in which RNA G-quadruplex structures have been identified (see section 2.1.1), as yet, just one area of these regions has been actively targeted by drug discovery programs – this is the area of 5' untranslated region (UTR) mRNA G-quadruplexes. The targeting of 5' UTR mRNA G-quadruplex structures is similar in principle to DNA promoter region G-quadruplex targeting, and involves designing drugs to selectively bind and stabilize G-quadruplex motifs located upstream of a target gene translation start site, with the aim of disrupting the ribosome complex and thereby inhibiting gene expression. 5' UTR G-quadruplexes (as drug targets) have a considerable potential advantage over DNA promoter-region G-quadruplexes, in that mRNA molecules are exported to the cytoplasm to function (i.e. for translation of protein products), thereby overriding the obstacle of DNA G-quadruplex-directed drugs having to pass the nuclear envelope. Additionally, mRNA G-quadruplex formation does not involve competition with a complementary strand, contrasting with G-quadruplex forming sequences in promoter regions which must compete with duplex formation with the complementary C-rich strand. These two features, which increase the accessibility of RNA G-quadruplexes to small molecule compounds, are applicable to many RNA G-quadruplex structures, making them, in principal at least, more ‘targetable’ than equivalent DNA G-quadruplexes.

G-quadruplex structures located within 5' UTR mRNA sequences represent by far the most studied RNA G-quadruplex structure as a drug target, however, G-quadruplexes formed from human telomeric RNA (TERRA) have recently received attention as drug targets. As mentioned in the concluding comments of the previous chapter, the main reason TERRA G-quadruplexes have been studied as drug targets is in order to investigate how current telomeric DNA G-quadruplex binding compounds...
interact with these recently discovered TERRA molecules. A secondary reason is that it is quite conceivable that TERRA G-quadruplex structures may represent viable drug targets in the future, considering the stability and propensity of TERRA molecules to fold into G-quadruplex structures and the emerging important role TERRA plays in telomere regulation\textsuperscript{144,148}. Following the determination of a native TERRA G-quadruplex (chapter 2), crystallographic studies into TERRA G-quadruplex-drug complexes were undertaken, with a view to providing structural details of RNA-ligand binding. A family of acridine-based compounds were studied – the acridine scaffold representing a well-studied building block in telomeric DNA G-quadruplex drug discovery programs.

3.1.2 TELOMERIC DNA G-QUADRUPLEXES AS DRUG TARGETS

As has been mentioned and discussed in section 1.3.4.4, G-quadruplex structures formed from telomeric G-rich DNA have been studied extensively as drug-targets for many years now. The link between telomeric DNA G-quadruplexes and drug targeting involves the enzyme telomerase, an enzyme which plays a critical role in chromosome maintenance and which is over-expressed in around 85\% of human cancers\textsuperscript{56,57}. It was proposed – and subsequently shown experimentally – that the stabilisation of a telomeric G-quadruplex with the use of a small-molecule compound could result in a reduction in telomerase activity\textsuperscript{92}. It was proposed that this small molecule compound was achieving its effect by sequestering the template of telomerase (i.e. single-stranded telomeric DNA) thereby inhibiting telomerase hybridisation. Following on from this work, a large number of G-quadruplex binding small molecule compounds have been produced, all with the intended use of being effective telomerase inhibitors through telomeric G-quadruplex stabilisation\textsuperscript{68,90,91}. From the outset of telomeric G-quadruplex-directed drug design, structural data has been utilized to inform and guide drug design efforts. While NMR has been critical in providing native G-quadruplex structural information, crystallography has arguably been the workhorse in providing structural data concerning G-quadruplex-drug interactions (although several useful NMR structures of G-quadruplex-drug complexes are available\textsuperscript{104-106}). G-quadruplex-binding small-molecule compounds have been designed and synthesized based on a
wide variety of scaffolds and chemistries, however, this thesis will focus on compounds for which high-resolution structural data exists, and this chapter specifically will focus on the acridine family of compounds.

### 3.1.3 ACRIDINES AS TELOMERIC G-QUADRUPLEX BINDING COMPOUNDS

The acridine core has been exploited extensively as a G-quadruplex binding motif, and a large library of acridine-modified compounds has been assembled over the past 15 years. The acridine core consists of three fused six-membered rings, with a nitrogen substituent at the 10 position of the central ring (figure 3.1).

![Chemical structure of the acridine core.](image)

**Figure 3.1** Chemical structure of the acridine core. Structure and numbering of the acridine core. The 3, 6 and 9 positions have been extensively substituted with carefully designed sidechains.

The acridine core provides an excellent aromatic moiety which is intended to π-stack over an external G-quartet, whilst allowing functionality to be added to the molecule in the form of modified sidechains, which are typically added at the 3, 6, and 9 positions of the acridine core. Sidechains are designed in order to maximize specific contacts between the DNA and the acridine-drug molecule, as well as to improve pharmacological properties, such as solubility and uptake. Another key feature of the acridine core is the presence of a positive charge at the 10 (nitrogen) position at physiological pH. A positively charged substituent has been incorporated into the vast majority of G-quadruplex targeted drugs in order to encourage electrostatic association between the drug molecule and the negative phosphates of the nucleic acid backbone. Although acridine-based molecules contain a positive charge in the acridine core itself, many G-quadruplex binding drugs (including acridine-based compounds) contain positive charges within sidechains. Acridine-based G-quadruplex
binding molecules therefore contain the three main features exhibited by the vast majority of effective G-quadruplex binding drugs:

1) A large aromatic moiety intended to maximize π-stacking contacts with G-quartets;

2) Positive charge(s) to maximize electrostatic contacts with the nucleic acid phosphate groups;

3) Diversely functionalized sidechains to enhance binding strength and selectivity.

Concerning performance and activity, generally, acridine-based compounds are able to selectively bind and stabilize G-quadruplex structures over duplex DNA (as assessed by FRET and UV-melting assays), and many have performed well in anti-cancer in cellulo cytotoxicity assays. Few have progressed to in vivo trials however, with (arguably) the most successful acridine-based compound, BRACO-19, failing due to toxicity. BRACO-19, a tri-substituted pyrrolopropionamido-functionalised acridine, displayed excellent anti-tumour activity in mouse tumour xenograft studies, however, proved to be unacceptably toxic when tested in higher organisms, indicating a problem with off-target effects. Despite this promising acridine-based compound failing so close to beginning clinical trials, the acridine core remains a promising motif for G-quadruplex drug design.

3.1.4 HIGH-RESOLUTION G-QUADRUPLEX-ACRIDINE STRUCTURES

From the outset of telomeric G-quadruplex directed drug design, high-resolution structural data has been used to inform and direct drug development and (ideally) drug improvement. One of the first crystal structures involving a telomeric G-quadruplex-ligand complex was reported in 2003, and involved a 3,6-disubstituted acridine (BSU6039) complexed to an anti-parallel bimolecular quadruplex formed from the *O. nova* telomeric repeat sequence d(GGGGTTTTGGGG) (figure 3.2a). This crystal
structure showed the acridine effectively stacked over an external G-quartet, with the polar groups of the molecule making extensive hydrogen bonding contacts with the DNA. Since this structure, an impressive number of *O. nova* G-quadruplex-acridine complexes have been determined, all displaying an isomorphous mode of ligand binding – i.e. stacking of the acridine molecule onto an external G-quartet along with extensive acridine-DNA polar contacts\textsuperscript{115,116}. In addition to the available *O. nova-*acridine complexes, the structure of the tri-substituted acridine BRACO-19 in complex with a G-quadruplex formed from the human repeat sequence d(TAGGGTTAGGGT) was solved in 2008 (figure 3.2b)\textsuperscript{113}. This was the first crystal structure of a human telomeric G-quadruplex-ligand complex showing an obviously relevant mode of ligand-quadruplex interaction (a previous crystal structure of a complex between a similar human quadruplex and the porphyrin, TmPyP4, revealed a crystal-packing-dependent mode of quadruplex-ligand binding\textsuperscript{112}). BRACO-19 was seen effectively stacked onto the 3’ G-quartet of the human G-quadruplex, with stabilizing interactions provided by a second G-quadruplex in the crystal lattice. As well as these π-stacking interactions, BRACO-19 is held tightly in place by a number of strong direct hydrogen bonds with the DNA, as well as via several water bridges. These DNA-acridine structures allowed the mode of acridine-quadruplex binding to be analysed at an atomic-scale level of detail, and provided information for the further development of these compounds.
3.1.5 DI-SUBSTITUTED TRIAZOLE ACRIDINES: BIOPHYSICS AND BIOLOGY

The work described in this chapter is focused on a member of a recently developed series of di-substituted acridine compounds, which were designed to bind with high affinity (and ideally high selectivity) to human telomeric G-quadruplex DNA\textsuperscript{93}. The molecules of this series include two substitutions of the acridine core at the 3 and 6 positions, with the sidechains sharing common features of triazole and phenyl rings, with differing end-group substituents. A total of 22 compounds in the series were tested in biophysical assays (including FRET and CD) with the best candidates being
tested in cell based assays (figure 3.3). Two members of the series - molecules 9a and 9b - performed particularly well, showing significant and selective stabilisation of human telomeric G-quadruplex DNA over G-quadruplexes formed from sequences found in the ckit proto-oncogene promoter region (at 1 μM concentration, compounds 9a and 9b stabilised human telomeric G-quadruplex DNA by 15.8 °C and 15.5 °C respectively [ΔT_m values], compared to <1 °C [ΔT_m] stabilisation of the ckit1 quadruplex). Compounds 9a and 9b also performed well in cell based growth inhibition assays, showing selective growth inhibition of ovarian (9a and 9b) and pancreatic cancer cells (9b only) over wild-type fibroblast cells. As this series of compounds differs only in sidechain length and end group type, it was possible to draw tentative conclusions regarding the structure-activity relationship for this series. The two sidechains of 9a terminate in pyrrolidine rings, with the sidechains of 9b terminating in diethyl amine groups, suggesting the charge and size of these end groups plays an important role in G-quadruplex binding behaviour.

**Figure 3.3** Triazole-acridine series of G-quadruplex binding compounds.
Structure of the di-substituted triazole-acridine series synthesised and characterised by Sparapani and co-workers. Two members of the series which performed optimally are indicated (yellow box). Compound 9b was used in the crystallographic work described in this chapter, and is subsequently referred to as ‘FD-12T’. Figure modified from Sparapani et al., 2010.⁹³
3.2 AIMS

Based on the extensive and fruitful use of X-ray crystallography in telomeric DNA G-quadruplex drug design, a crystallographic approach was adopted with the aim of providing structural information for TERRA G-quadruplex-ligand binding (specifically, acridine-binding). The series of di-substituted triazole acridines described above was used as the main library of compounds for TERRA G-quadruplex crystallisation screening (figure 3.3). Although this series of compounds was not initially designed to interact specifically with TERRA G-quadruplexes, they were used for crystallisation trials as this series allowed a systematic screening of a relatively diverse set of sidechain end-groups. As with early crystallographic (and NMR) studies into telomeric DNA G-quadruplex-ligand complexes, the main questions set out to be answered were:

1) What are the structural mechanisms by which small-molecule compounds recognise and interact with TERRA G-quadruplex structures?
2) What structural rearrangements of the RNA G-quadruplex occur (if any)?
3) What are the differences between DNA and RNA G-quadruplex-drug binding?
4) What are the implications (if any) for telomeric G-quadruplex-directed drug-design?
3.3 METHODS AND MATERIALS

3.3.1 SAMPLE PREPARATION

The RNA sequence used in these crystallisation experiments, r(UAGGGUUAGGGU), was purchased from Eurogentec (Belgium) (HPLC purified) and used without further purification. The RNA was annealed at 1.8 mM single-strand concentration, in a buffer containing 50 mM potassium chloride and 20 mM potassium cacodylate (pH 6.5), following the annealing protocol described in section 2.3.1. Synthesis of the disubstituted triazole acridine used in these studies — FD-121 — was performed by Sparapani and co-workers and has been described previously93. (N.B. FD-121 corresponds to compound 9b from figure 3.3.) The stock solution of FD-121 used in these experiments consisted of 10 mM acridine dissolved in 100% dimethyl sulfoxide (DMSO), which was stored either at -20°C (long term) or +4°C (when in use). The ligand, generally, had very poor solubility, and it was often necessary to sonicate the stock solution prior to use in order to encourage the compound to dissolve. The RNA-acridine complex was formed by adding FD-121 to the pre-annealed RNA sample to final concentrations of 1.5 mM single-stranded RNA (ssRNA) and 0.5 mM FD-121. The complex was incubated at room temperature for 1 hour prior to crystallisation trials.

3.3.2 CRYSTALLISATION

Crystals of the RNA G-quadruplex-acridine complex were grown in standard-hanging drops consisting of 0.5 μL of the RNA-acridine complex plus 0.5 μL of the crystallisation solution (composed of 20% MPD, 150 mM sodium chloride, 50 mM sodium cacodylate (pH 6.5) and 4 mM spermine). The drop was equilibrated against a well solution composed of 20% MPD. The drop was incubated at 12 °C and crystals appeared after 6 weeks, with final dimensions of approximately 50 μm x 40 μm x 40 μm (figure 3.4).
3.3.3 DATA COLLECTION AND PROCESSING

A single crystal of the RNA-acridine complex was frozen in liquid nitrogen and taken to the Diamond Light Source synchrotron (Oxford, UK) for data collection. Two datasets were collected on a single crystal on beamline I04: a low resolution data set was collected first in order to ensure completeness of data with minimal/no overloaded intensities, followed by a high resolution dataset to in order to obtain the highest quality data possible (the resolution limit for the high resolution dataset was approximately 2.6 Å, see figure 3.5). For both datasets, images were collected with an exposure time of one second per 0.5° phi oscillation, using 0.9763 Å wavelength X-ray radiation. The images were processed and scaled using d*trek in the CrystalClear software package (Rigaku Inc). The crystals belong to the cubic space group P23, with unit cell dimensions of $a = b = c = 56.61$ Å, $\alpha = \beta = \gamma = 90^\circ$. Data collection statistics are shown in table 3.1.
3.3.4 STRUCTURE DETERMINATION AND REFINEMENT

The RNA G-quadruplex acridine complex was solved by molecular replacement, using Phaser (from CCP4) using a single strand of the bimolecular RNA G-quadruplex structure (PDB Id: 3IBK) as a search model. The search model was first reduced to contain guanine residues only, reducing the possibility of introducing phase bias into the initial (and subsequent) electron density maps. The initial 2Fo-Fc maps showed very clear density for the G-quartets and potassium ions, as well as residual electron density for the loop region, particularly residue adenine 8. Importantly, residual electron density was seen above the 5' G-quartet in both 2Fo-Fc and Fo-Fc maps, into which the ligand, FD-121, was eventually modelled. The initial atomic coordinates and geometric restraints for FD-121 were generated using the PRODRG server. Model building and restrained refinement were performed using Coot and Refmac5, following the procedure outlined in section 2.3.4. Individual isotropic B-factors were refined throughout the process, with TLS parameters included towards the end of the
refinement\textsuperscript{199,200}. The majority of the RNA plus the ligand molecule could be confidently and unambiguously fitted into electron density. The final model was refined using data between 16.34 Å and 2.60 Å, with final $R$ and $R_{\text{free}}$ values of 23.6% and 24.7% respectively. The final overall (isotropic) B-factor for the model (including solvent molecules) is 26.46 Å\textsuperscript{2}. Additional refinement statistics are shown in table 3.1.

3.3.5 VALIDATION AND DEPOSITION IN THE PROTEIN DATABANK

The final atomic model was validated using the RCSB web server (www.deposit.rcsb.org) and has no bond lengths or bond angles greater than 6 times the root mean square deviation (RMSD) from standard dictionary values. The RMSDs of bond lengths and bond angles is 0.010 Å and 1.018° respectively.

The atomic coordinates and structure factors have been deposited in the Protein Databank with accession code 3MIJ (deposition date: 10/04/10). This structure was first reported in the Journal of the American Chemical Society in 2011\textsuperscript{207} (see Appendix A).

<table>
<thead>
<tr>
<th>Table 3.1 Data collection and refinement statistics for the RNA G-quadruplex-acridine complex.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
</tr>
<tr>
<td>Sequence</td>
</tr>
<tr>
<td>Ligand</td>
</tr>
<tr>
<td>Space group</td>
</tr>
<tr>
<td>Unit cell dimensions:</td>
</tr>
<tr>
<td>$a$, $b$, $c$ (Å)</td>
</tr>
<tr>
<td>$\alpha$, $\beta$, $\gamma$ (°)</td>
</tr>
<tr>
<td>Resolution (Å)</td>
</tr>
<tr>
<td>$R_{\text{work}}$ (%) overall</td>
</tr>
<tr>
<td>$I / \sigma$</td>
</tr>
<tr>
<td>Completeness (%)</td>
</tr>
<tr>
<td>Redundancy</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
</tr>
<tr>
<td>Resolution (Å)</td>
</tr>
<tr>
<td>Reflections</td>
</tr>
<tr>
<td>$R_{\text{work}} / R_{\text{free}}$ (%)</td>
</tr>
<tr>
<td>No. of atoms</td>
</tr>
<tr>
<td>Ions</td>
</tr>
<tr>
<td>Water</td>
</tr>
<tr>
<td>Overall B-factor (Å\textsuperscript{2})</td>
</tr>
<tr>
<td><strong>RMS deviations</strong></td>
</tr>
<tr>
<td>Bond-lengths (Å)</td>
</tr>
<tr>
<td>Bond-angles (°)</td>
</tr>
<tr>
<td>PDB ID</td>
</tr>
</tbody>
</table>

(Values in brackets refer to the highest resolution shell, 2.49-2.40 Å)
3.4 RESULTS AND DISCUSSION

3.4.1 CRYSTAL STRUCTURE OF A TELOMERIC RNA G-QUADRUPLEX-ACRIDINE COMPLEX: OVERALL TOPOLOGY AND GENERAL FEATURES

The structure of an RNA quadruplex-acridine complex was successfully determined, revealing a parallel stranded bimolecular quadruplex with two acridine molecules bound to the 5' G-quartet surface (figure 3.6a-b). The asymmetric unit contains one RNA strand of the sequence, r(UAGGGUUAGGGU), and one acridine molecule – the complete quadruplex structure is generated by a two-fold crystallographic rotation axis through the central channel of the structure. Consequently, each of the two strands of this quadruplex are identical in both sequence and structure. As with the native telomeric RNA structure described previously (chapter 2), the quadruplex-complex described here is stabilized by two centrally coordinated potassium ions. The parallel strand association is again maintained by two tri-nucleotide propeller-type loops, although, as will be discussed below, these UUA loops have a quite different arrangement to the native RNA crystal structure. The complexed acridine molecules have two identical sidechains attached at the 3- and 6- positions of the acridine core, with each sidechain consisting of a triazole-ring followed by a phenyl-ring and terminating in a diethyl amine functional group, linked to the phenyl ring by an amide group (figure 3.6c).

The structure was refined to 2.6 Å resolution, with the resulting electron density maps being of high quality. All residues of the RNA are accurately modelled into electron density, including those of the UUA loops, which are well ordered in the crystal lattice (figure 3.7). The ligand molecule, potassium ions and ordered water molecules could all be confidently and accurately fitted into appropriate electron density.
Figure 3.6 Overall topology and structure of the RNA quadruplex acridine complex.
a and b) The telomeric RNA sequence r(UAGGGUUAGGGU) folds as a parallel stranded bimolecular G-quadruplex, with propeller-type UUA loops, analogous to the equivalent native structure (3IBK, chapter 2). Two acridine molecules (FD-121 — purple ball-and-sticks) are bound to the 5' G-quartet surface. c) Chemical structure of the acridine (FD-121) used in this study. d) Schematic representation of the G-quadruplex (RNA only), in the same orientation as shown in (a). The cartoon representations have been coloured according to the Nucleic Acid Database colour scheme: adenine = red; guanine = green; uracil = cyan. The RNA backbone is shown as a grey ribbon. Both RNA strands of the G-quadruplex are identical in structure, being related by a crystallographic symmetry axis (indicated).
Figure 3.7 Electron density for the RNA-acridine complex. 
a) 2Fo-Fc map at 0.9 o level. All regions of the RNA are modelled accurately into electron density. b) 2Fo-
Fc map at 1.0 o level. The majority of the acridine ligand is modelled into high-quality electron density. 
The terminal regions of both sidechains are slightly less-well resolved, probably a result of regional flexibility due to few stabilising contacts.

3.4.2 RNA G-QUADRUPLEX-ACRIDINE BINDING: OVERALL MODE OF INTERACTION

The most striking aspect of this complex structure is the presence of two acridine molecules bound to the same G-quartet face (figure 3.6b). This was an unexpected finding, as these G-quadruplex binding molecules (as with almost all types of such molecule) were designed to interact with a G-quartet in a 1:1 stoichiometry. In this structure, two acridine molecules are bound to the same (5') G-quartet, positioned in a laterally displaced back-to-back arrangement, with the nitrogens of the acridine core
orientated away from the central channel. This 'back-to-back' arrangement is almost certainly due to like-charge-repulsion of the acridine nitrogens, which should be protonated in these slightly acidic conditions. Despite a large degree of potential intrinsic flexibility, the ligand is generally planar, with only a slight deviation in planarity occurring at the triazole-phenyl link.

3.4.2.1 RNA G-QUADRUPLEX-ACRIDINE BINDING: π-π STACKING

Interestingly, there are no direct contacts between the two acridine molecules in this complex structure, and very few direct RNA-acridine hydrogen bonds – the primary mode of RNA-ligand interaction is π-π stacking. These π-π stacking interactions originate from two main regions: 1) the ligands themselves (from symmetry-related molecules) and 2) the RNA bases. Two complete RNA-quadruplexes associate in the crystal lattice stacked 5'-5', separated by a total of four acridine molecules (figure 3.8). This unusual crystal packing arrangement serves to secure the acridine molecules in place over the 5' face of the RNA by providing extensive π-π stacking contacts.

Despite the significant surface area provided by the arrangement of these four acridine molecules in the crystal lattice, it appears that an even greater contribution to ligand binding is provided by the RNA itself. In addition to the surface provided by the 5' G-quartet for π-π stacking, the adenine residues of the propeller-loops and terminal regions are flipped up towards the ligands, orientating themselves within the plane of the 5' quartet, thereby providing a platform for ligand binding (figure 3.9). The adenines thus create a novel G₄A₄ octet arrangement, consisting of four guanines from the 5' G-quartet together with the four loop and terminal adenines, which significantly increases the surface area available to the ligand for π-stacking interactions. The majority of the ligand aromatic groups are positioned directly over the purines, with only one phenyl group not being positioned over the purine-octet binding platform (figure 3.9).
**Figure 3.8 Crystal packing interactions.**
a) Two G-quadruplexes associate in the crystal lattice, stacked at the 5'-5' interface, with four acridine molecules mediating the interaction. Four asymmetric units are shown here, each coloured separately.
b) Schematic of the ligand-mediated G-quadruplex packing.

**Figure 3.9 Purine octet binding surface of the RNA G-quadruplex.**
A novel purine octet is present in this RNA G-quadruplex structure, composed of four guanines of the 5' G-quartet (G3 + G9 from strands 1 and 2), plus four adenines from the terminal and loop regions (A2 and A8 respectively, from strands 1 and 2). This octet is stabilised significantly by C2'-OH group interactions - both direct RNA-RNA contacts and water-bridged contacts. Hydrogen bonds for one strand only are shown (black dashes), as these distances are replicated exactly for the second strand. All intra-G-quartet hydrogen bonds are between 2.7 and 2.9 Å. The ligand molecules are shown as transparent spheres. Water molecules are shown as small red spheres.
3.4.2.2 RNA QUADRUPLEX-ACRIDINE BINDING: DIRECT HYDROGEN BONDS

As mentioned above, there are very few direct hydrogen bonding contacts between the acridine molecule and the RNA, which is surprising considering the large number of polar groups present within FD-121 – the triazole group is particularly polar. The only direct RNA-acridine hydrogen bond seen in this crystal structure is between the N3 atom of a terminal uracil residue and the carbonyl group of one of the ligand's sidechains (sidechain 1) (figure 3.10a-b). This hydrogen bond serves to stabilize the acridine in this region, which is reflected in the B-factor distribution, which shows the atomic mobility (within the crystal lattice) of the hydrogen-bonded sidechain of the ligand to be lower than the non-hydrogen bonded sidechain (sidechain 2) (figure 3.10a and c). It is interesting to note that although the hydrogen bonded sidechain of the acridine is more stable than sidechain 2, this region (sidechain 1) is still far less stable (as indicated by the B-factor) than the core region of the acridine, indicating that the acridine core is positioned securely over the G₄A₄ octet. The flexibility of the extreme ends of the ligand is also indicated by the quality of electron density for these regions, which is less defined for the terminal diethyl amine groups compared to the aromatic groups (figure 3.7b, above).

3.4.3 PROPELLER-LOOP REARRANGEMENT

As stated, the primary mode of ligand-RNA binding is π-stacking, in which the G₄A₄ purine octet plays a major role. It should be noted that, although two of the octet-adenines are propeller-loop residues, the remaining two adenines originate from the 5′-UA terminal residues, which can themselves be considered as incomplete loops (see figure 3.11 and legend therein for further explanation). The rearrangement of the loop residues relative to the native RNA structure to create the octet binding platform involves direct hydrogen bonds as well as water bridging contacts. Importantly, several of the loop-stabilising hydrogen bonds appear to be RNA-specific, that is, these interactions involve C2'-OH groups, suggesting this large binding platform may be specific to RNA G-quadruplex structures. Structural alignment and backbone torsion angle comparison of the RNA-acridine complex with the native RNA structure (3IBK)
a) RNA-acridine complex coloured by B-factor. Low B-factor values – such as those of the G-quartets and the RNA in general – indicate low atomic mobility of these regions in the crystal lattice. Regarding the acridine, it is clear that the acridine core region has lower B-factors (and therefore lower mobility) than the sidechains. Sidechain 1, which hydrogen bonds to the RNA (b), has noticeably lower B-factors than sidechain 2, which is not stabilised by any significant hydrogen bonds or interactions with the RNA. c) Plot of the atomic B-factors for the indicated regions of the ligand molecule, highlighting the order of mobility in the crystal lattice, from low to high, being: acridine core < sidechain 1 < sidechain 2.

Figure 3.10 Acridine-RNA binding.
reveals the extent to which the UUA loop is remodelled in the complex structure to accommodate ligand binding (figure 3.12). Adenine 8 rearranges from the intercalated ‘UAU’ arrangement of the native structure to sit within the same plane as the 5’ G-quartet. Uracil 7 is repositioned to maximize π-stacking contacts with adenine 8 – U7 actually appears to be in a very similar position relative to A8 in both structures (i.e. 3IBK and 3MIJ). The largest structural rearrangement of the loop region is seen for U6, which, in the complex structure, is orientated away from the quadruplex, and is in fact Hoogsteen hydrogen bonding to a symmetry related adenine residue. Comparison of the backbone torsion angles of the native and complex loop residues shows a surprising level of similarity, with significant deviations only seen for the γ angle of U6, the α and γ angles of U7, and the α and ζ angles of A8 (figure 3.12c). Surprisingly, the glycosidic torsion angles of the loop residues are highly conserved between structures, suggesting a limited conformational adaptability around the glycosidic bond. The loops of the native and complex structures also display a significant level of conservation in sugar puckering, again surprising, considering the apparent gross conformational differences between these two structures (table 3.2).

**Figure 3.11** From intermolecular quadruplex to intramolecular quadruplex. Bimolecular G-quadruplex structures are primarily used as models for unimolecular (i.e. intramolecular) quadruplexes, as the 5’-UA residues can be (theoretically) connected with the 3’ uracil (U12) to form a complete tri-nucleotide ‘loop’. The 5’ adenines (A2 residues) in the crystal structure described here are positioned in a manner very similar to the ‘true’ loop adenines, and it is an intuitive step to link the 3’ uracil (U12) to the 5’ uracil (U1) to connect the backbone and create a ‘loop’. For the sake of clarity, the A2 residues will be referred to as terminal adenines, but it should be noted that the behaviour and positions of these residues is a model for the equivalent intramolecular structure. Both inter- and intramolecular G-quadruplexes have the potential to form from TERRA sequences in cells, however, experimental evidence suggests intramolecular structures are likely to be the main structural species in a cellular setting.
Figure 3.12 Loop remodelling of acridine complex compared to native RNA.
a) Rearrangement of the UUA loop seen in the acridine-complex structure (3MJU) compared to the
native RNA G-quadruplex (3IBK). b) Structural overlay of the UUA loop regions of the ligand-bound
(blue) and native (red) RNA G-quadruplex structures (non-loop residues are coloured grey). c) backbone
and glycosidic torsion angle comparison of the loop regions of the ligand-complex (black triangles) and
native RNA quadruplex structures (red diamonds). The glycosidic angles show a remarkable level of
conservation between the two structures (indicated by blue dashes). Torsion angle analysis was
performed using 3DNA.²⁰²
Table 3.2. Comparison of ribose puckering of native and acridine-bound RNA G-quadruplexes.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Native (3IBK)</th>
<th>Ligand-complex (3MII)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>C3'-endo</td>
<td>C2'-exo</td>
</tr>
<tr>
<td>A2</td>
<td>C3'-endo</td>
<td>C1'-exo</td>
</tr>
<tr>
<td>G3</td>
<td>C2'-endo</td>
<td>C2'-exo</td>
</tr>
<tr>
<td>G4</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>G5</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>U6</td>
<td>C1'-exo</td>
<td>C3'-exo</td>
</tr>
<tr>
<td>U7</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>A8</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>G9</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>G10</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>G11</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>U12</td>
<td>C1'-endo</td>
<td>C2'-endo</td>
</tr>
</tbody>
</table>

Correlated puckering within the loop residues is highlighted green. G-quartet residue puckers are highlighted grey.

3.4.4 PROPELLER LOOP STABILISATION: HYDROGEN BONDING AND WATER CONTACTS

Both the loop and terminal adenines (A8 and A2 respectively) are stabilized by intramolecular hydrogen bonds, in addition to water-bridged contacts. Significantly, the C2'-OH groups of the ribose rings play a major role in loop-stabilisation, suggesting the purine octet binding platform seen in this complex structure could represent an RNA-quadruplex specific mode of ligand binding. A2 and A8 are both involved in highly similar intramolecular hydrogen bonds involving C2'-OH groups: one hydrogen bond is formed between the N1 of the adenine to the C2'-OH group of a guanine from the 5' G-quartet (A8 N1:G3 C2'-OH and A2 N1:G9 C2'-OH) with another strong hydrogen bond seen between the C2'-OH groups of the adenines and the C2'-OH group of a guanine from the central G-quartet (A8 C2'-OH:G4 C2'-OH and A2 C2'-OH:G10 C2'-OH) (figure 3.13a-b). The loop and terminal adenines are also stabilized within the purine octet by base-water-base bridging contacts involving the N3 nitrogens of the adenines and the N2 nitrogens of guanines from the 5' G-quartet (A8 N3:G3 N2 and A2 N3:G9 N2) (see figure 3.9, above).

These intramolecular and water contacts serve to stabilize the purine octet binding platform, however, one of the main stabilizing features of this octet is quite likely the ligands themselves. In other words, these large and highly aromatic acridine molecules are almost certainly sequestering and stabilizing a transiently formed octet surface, rather than binding to a static 'pre-formed' octet surface. Justification for this
Statement is based on the observation that the purine octet is not observed within the native structure (3IBK), nor is such a motif seen in any of the ten models reported for the equivalent NMR structure (2KBP), suggesting that additional factors (such as a ligand) are required for the formation of this octet. Current DNA-ligand complex crystal data supports this theory, by highlighting a strong propensity for small molecule ligands to alter 'native' quadruplex structures (particularly loop residues) in unexpected ways, and trap these structures in non-native conformations. In crystal structures therefore, a ligand can 'select' a specific quadruplex conformation by binding to a transiently formed surface or motif, which inevitably becomes the main structural arrangement seen in the final atomic model (as crystals are almost always formed of structurally identical repeated molecules).

**Figure 3.13 Stabilisation of the octet adenine residues.**
a) Adenine 8 of the loop is held in place within the octet by two key C2'-OH mediated hydrogen bonds: one between the N1 of A8 and the C2'-OH of G3, and a second between the C2'-OHs of A8 and G4. This pattern of hydrogen bonding is repeated almost exactly for the equivalent region of the terminal region adenine 2 residue (b). Hydrogen bonds are shown as black dashes and bond distances are indicated (in Å). RNA shown as green sticks, with FD-121 shown as purple ball-and-sticks.
3.5 GENERAL DISCUSSION AND CONCLUSIONS

3.5.1 UNPRECEDENTED RNA BEHAVIOUR

The structure described above is the first crystal structure of an RNA G-quadruplex-ligand complex, and has revealed some unexpected features of RNA-ligand binding and recognition. The key features of this complex which were not predicted include:

1) the presence of two acridine molecules bound to the same G-quartet surface;

2) the formation of a G₂A₄ purine octet resulting from significant loop rearrangements (forming a large ligand-binding platform), and;

3) a striking lack of direct hydrogen bonding contacts between the acridine ligand and the RNA.

1) The ability of two acridine molecules to bind to the same G-quartet surface was unexpected, however, a comparable mode of G-quadruplex ligand binding has been observed in a daunomycin-tetraplex structure determined in 2003¹¹⁰. The daunomycin-complex differs significantly from the acridine complex described here in that the daunomycin-bound G-quadruplex is tetramolecular (formed from the sequence, d[TGGGGT]) and therefore lacks the propeller loops of the G-quadruplexes formed from telomeric repeat sequences. No current crystal- or NMR structures of human telomeric G-quadruplex-ligand complexes have displayed multiple ligand molecules bound to a single G-quartet surface. This may be a result of the physical size of the compounds tested, which generally have large cores with three or more sidechains (making these compounds larger than the di-substituted acridine studied here) but it is likely to be due in large part to the presence of the purine octet observed in the RNA-acridine structure described. This purine octet – essentially twice the surface area of a G-quartet alone – provides a ligand-binding surface area large enough for two-acridine cores.
2) The rearrangement of the terminal and UUA propeller-loop adenine residues to form an octet with the 5' G-quartet was an unexpected feature of the structure and has not been observed in any other G-quadruplex crystal- or NMR structure. The pivotal role of the C2'-OH groups in octet stabilisation suggests this may be a motif specific to RNA G-quadruplexes, and therefore makes these motifs of potential use in RNA G-quadruplex-directed drug design. It should be noted that a G4A2 hexad has been observed previously in a DNA G-quadruplex crystal structure, however this extended-G-quartet motif (albeit smaller in size than the octet described above) was not involved in ligand binding, but rather assisted in intermolecular quadruplex stacking interactions.\textsuperscript{114}

3) As stated above, the lack of direct RNA-ligand hydrogen bonding contacts was surprising, considering the multitude of polar groups within the acridine ligand, as well as present within the RNA itself. Additionally, all current DNA G-quadruplex-acridine crystal structures reveal extensive direct hydrogen bonding contacts between the drug molecule and DNA.\textsuperscript{111,113,115,116} A possible reason for the paucity of intermolecular hydrogen bonds between the RNA and this di-substituted acridine may be that this ligand – which was not designed to selectively bind to quadruplex RNA – is simply not the most suitable molecule for binding to RNA G-quadruplex structures through effective hydrogen bonding contacts. A second reason for the limited number of RNA-acridine hydrogen bonds may be the sidechain terminal diethyl-amine groups, which are not suitable for hydrogen bonding. However, the diethyl-amine groups of the sidechain termini – being positively charged – are suitable (and intentionally designed) to interact electrostatically with the negatively charged phosphates (particularly of the quadruplex groove regions). The absence of such electrostatic/dipole-dipole contacts in this structure may be due to the adenines of the G4A4 octet, which, as well as providing a large surface area for ligand binding (thereby partially negating the need for direct hydrogen-bonds and/or electrostatic contacts), are in fact blocking access to the quadruplex groove regions from the 5' direction (figure 3.14). G-quadruplex ligands (such as the di-substituted acridine studied here) are often designed with relatively long and flexible sidechains, which are intended to interact specifically with phosphates and polar groups of the quadruplex grooves. With these electrostatic/hydrogen bonding-rich regions blocked by the orientation of the
loop and terminal adenines, the ligand is not able to make a significant number of hydrogen bonding or electrostatic contacts with the RNA.

**Figure 3.14** G-quadruplex groove inaccessibility.
The positioning of the four adenines from the loop and terminal regions within the 5' purine octet (coloured green) very effectively blocks access of the acridine sidechains to the four groove regions of the G-quadruplex.

### 3.5.2 IMPLICATIONS FOR DRUG DESIGN

Prior to the determination of this structure, it was predicted that a suitable ligand binding to an RNA G-quadruplex would take advantage of the ribose C2'-OH groups for hydrogen bonding interactions\(^{201}\). This is not the case in the structure described here, however, the C2'-OH groups play a crucial role in stabilizing the octet binding platform, and thereby play an indirect but pivotal role in ligand binding. The lack of direct RNA-ligand contacts (hydrogen bonding or electrostatic) may of course be specific to this particular system (i.e. drug type and RNA sequence), however, it is reasonable to assume that the octet binding platform seen in this structure could be a general feature of telomeric RNA G-quadruplexes, or perhaps even RNA G-quadruplexes in general (such as mRNA G-quadruplexes). This motif, therefore, may prove to be an RNA-specific motif which could potentially be exploited in G-quadruplex drug design – either intentionally avoided (if targeting telomeric DNA G-quadruplexes) or actively
targeted (if designing RNA G-quadruplex specific compounds). The observation of two acridine molecules bound to the octet (seen in the structure here) suggests that the use of much larger aromatic cores in drug design may favour RNA G-quadruplex binding and allow such a motif to be exploited. It is, however, difficult to envisage a pharmaceutically acceptable drug molecule with a significantly larger core than the acridine compound used here, which, with a molecular weight of 750 Da, is already approaching the higher end of therapeutically acceptably molecular weights. Ultimately therefore, the exploitation of C2'-OH groups through direct drug contacts may be the most viable avenue for generating RNA vs DNA G-quadruplex drug-selectivity. If this is to be done ‘rationally’, further structural data will be needed, particularly in the form of an RNA G-quadruplex complex involving a ligand able to interact with the RNA through polar contacts. In addition to this, it would be useful to understand how the acridine used in these studies (FD-121) interacts with the equivalent DNA G-quadruplex. Crystallographic investigations into such a complex were indeed pursued, with some success, which will be described and discussed in the following chapter.
CHAPTER 4

X-RAY CRYSTALLOGRAPHY OF A PARTIAL TELOMERIC DNA G-QUADRUPLEX-ACRIDINE COMPLEX
4.1 BACKGROUND

Following the successful crystal structure determination of a complex formed between a telomeric RNA G-quadruplex and a di-substituted acridine molecule (chapter 3), structural investigations into the binding of this acridine (FD-121) to the equivalent DNA G-quadruplex were pursued. The main purpose of this was to identify and structurally characterise the differences between RNA and DNA G-quadruplex ligand recognition, which, as mentioned, would be expected to aid future drug development of both DNA and RNA G-quadruplex-directed small molecule compounds. A key aspect of these studies was to determine whether the unusual mode of ligand binding seen in the RNA-acridine complex was a consequence of the RNA or a characteristic of the ligand itself.

In addition to its application to telomeric RNA G-quadruplex drug binding, crystallographic studies of FD-121 bound to human telomeric G-quadruplex DNA has standalone worth to current telomeric DNA G-quadruplex drug discovery programs. This is because, although eleven high resolution structures of acridine-DNA G-quadruplex structures are available, only one such structure involves human telomeric DNA (this is the BRACO-19 complex). It should not be underestimated that, while structural data for non-human G-quadruplex-drug complexes is undeniably of use for drug development, the binding surfaces presented by (for example) the O. nova G-quadruplex are considerably different to those presented by G-quadruplexes formed from human repeat sequences. For this reason, the generation of further human telomeric G-quadruplex-acridine complexes is likely to be of considerable use for future telomere-directed drug design.
4.2 AIMS

The work described in this chapter can be considered the DNA equivalent of the RNA-complex work described previously (chapter 3). The aim of this study was to generate crystallographic structural data for telomeric DNA G-quadruplex-FD-121 binding. Such data was intended to inform on both telomeric RNA and telomeric DNA G-quadruplex drug design. The key questions set out to be answered were:

1) Does the binding of FD-121 to human telomeric G-quadruplex DNA result in structural rearrangements of the G-quadruplex?

2) Are there any differences between FD-121-RNA and FD-121-DNA G-quadruplex interaction?

3) Can these differences be exploited in drug-design?

4) How does the binding of this acridine compare to previous DNA-acridine complexes?

5) How might such acridine compounds be rationally modified in the future?
4.3 METHODS AND MATERIALS

4.3.1 SAMPLE PREPARATION

The DNA sequence used in these crystallisation experiments, d(TAGGGTTAGGGT), was purchased from MWG (Germany) (HPLC purified) and used without further purification. This sequence is the DNA equivalent of the RNA sequence studied in chapters 2 and 3 (N.B. the RNA used in chapter 2 contained a bromo-uracil at the 5' terminus). The DNA was annealed at 2 mM single-strand concentration in a buffer containing 50 mM potassium chloride and 20 mM potassium cacodylate (pH 6.5), following the annealing protocol as described in section 2.3.1. The di-substituted triazole acridine used in these studies – FD-121 – is the same compound described in chapter 3, and was prepared and handled as described in section 3.3.1. The DNA-acridine complex was formed by adding FD-121 to the pre-annealed DNA to final concentrations of 1.5 mM single-stranded DNA (ssDNA) and 0.75 mM FD-121. The complex was incubated at room temperature for 1 hour prior to crystallisation trials.

4.3.2 CRYSTALLISATION

Crystals of the DNA G-quadruplex-acridine complex were grown in standard-hanging drops consisting of 0.5 μL of the DNA-acridine complex plus 0.5 μL of the crystallisation solution (composed of 20% MPD, 150 mM sodium chloride, 50 mM sodium cacodylate [pH 6.5] and 4 mM spermine). The drop was equilibrated against a well solution composed of 50% MPD. The drop was incubated at 12 °C and crystals appeared after 1 day. The crystals appeared readily in a range of precipitant concentrations but generally had fairly poor microscopic morphology, being irregular in shape with an appearance of possible (non-merohedral) twinning (figure 4.1). When rotated under a double light polarizer, there was a gradual rather than a sharp colour change at the polarization edge, suggesting poor internal ordering.
4.3.3 DATA COLLECTION AND PROCESSING

The apparent poor visual quality of these DNA crystals was reflected in the quality of diffraction data collected. Crystals were frozen directly in liquid nitrogen and screened at the Diamond Light Source synchrotron (Oxford, UK), on beamline I03. The individual quality of the crystals was highly variable, with approximately one crystal out of every twenty producing reasonable diffraction. Figure 4.1a shows the crystal which was used to solve the structure, with a corresponding representative diffraction image. The spots appear to be single, although some spots are smeared, suggesting a high degree of crystal mosaicity (i.e. poor internal ordering of the crystal). Figure 4.1b shows a clearer example of these crystals – the edges are poorly defined and there appears to be several lattices inter-grown (this is non-merohedral or epitaxial twinning). The diffraction for the crystal shown in figure 4.1b is poor quality, with a high degree of crystal mosaicity and evidence of non-merohedral twinning (some spots appear to be split). This crystal was not used for data collection. A dataset was collected on the crystal shown in figure 4.1a, with a collection strategy of one second exposure time per 1° phi oscillation, using X-ray radiation of wavelength 0.9763 Å. Data were processed and scaled using XDS\textsuperscript{208} and Scala\textsuperscript{193}. Merohedral twinning (i.e. when two distinct lattices overlap in three-dimensions) was checked using the SFcheck program in CCP4, with the score of 2.32 (for $<l^2>/<l>^2$) indicating a low likelihood of crystal twinning\textsuperscript{194}. Initially, these crystals could not be unambiguously assigned to a spacegroup, but could be confidently assigned to the Laue group 6/mmm. This dataset was therefore processed in the spacegroup of this Laue group with the lowest internal symmetry, $P6_{2}22$. The final spacegroup was determined to be $P6_{2}22$, with unit cell dimensions of $a = b = 71.67$ Å, $c = 29.37$ Å, $\alpha = \beta = 90^{\circ}$, $\gamma = 120^{\circ}$. Space and Laue group checking and assignment were performed using Pointless\textsuperscript{193}. Additional data collection statistics are shown in table 4.1.
4.3.4 STRUCTURE DETERMINATION AND REFINEMENT

As with the other crystal structures described previously in this thesis, the structure of this DNA-acridine complex was solved by molecular replacement using Phaser\textsuperscript{\textregistered}, using a single strand of the DNA G-quadruplex structure, 1K8P\textsuperscript{\textregistered}, as a search model (the model consisted of the G-quartet guanines and potassium ions only). The Laue group 6/mmm encompasses six space groups, all of which were tested during the molecular replacement structure solution process. Only one space group, P6\textsubscript{2}2\textsubscript{2}, gave a successful molecular replacement solution, with translation and rotation Z scores of 11.3 and 3.8 respectively and a log likelihood gain of 139. Initial electron density maps showed reasonable density for the guanines (which were part of the search model) as well as residual density at the 5’ and 3’ regions of the G-quadruplex. The 5’ terminal adenine and thymine residues and one loop residue were modelled, however, the initial electron density maps could not be greatly improved and much of the DNA

Figure 4.1 X-ray diffraction of DNA-acridine complex crystals.

a) Top – crystal grown from DNA of the sequence, d(TAGGGTTAGGGT), in complex with the acridine, FD-121. Bottom – corresponding X-ray diffraction pattern. This crystal was used for structure determination and refinement. b) Crystal (top) and diffraction pattern (bottom) of a typical crystal grown in these conditions. The crystal appears to be non-merohedrally twinned with poor morphology, resulting in diffuse, weak diffraction. The red box surrounding each crystal represents the diameter of the X-ray beam (100 µm x 80 µm).
remains unmodelled or poorly modelled. Due to the resolution and quality of the data, water molecules could not be placed into electron density. Additionally, the resolution did not allow accurate refinement of the individual (residual) isotropic B-factors, and so overall anisotropic B-factors were refined (including TLS parameters). Although the electron density maps were not of the highest quality, a large, discrete region of electron density could be seen approximately 3.3 Å beneath the 3' G-quartet, in both 2Fo-Fc and Fo-Fc maps, at the start and end of the refinement process. This density could not be accounted for by any components of the DNA and so the ligand FD-121 was modelled into this region. The core of FD-121 (acridine moiety plus triazole groups) fitted into this residual electron density comfortably, and reduced the R and R_free factors by around 4-5% (R and R_free values for ligand omit map calculations are 37.5% and 44.7% respectively). Model building and refinement were performed using Coot and Refmac5. The final structure was refined using data between 9.96 Å and 3.2 Å, with final R and R_free factors of 34.6 % and 37.8 % respectively. These R factors are relatively high and reflect the poor quality of the data. The final atomic model should be viewed and analysed with caution. Refinement statistics are shown in table 4.1.

4.3.5 VALIDATION AND DEPOSITION IN THE PROTEIN DATABANK

The final atomic model was validated using the RCSB web server (www.deposit.rcsb.org) and has no bond lengths greater than 6 times the root mean square deviation (RMSD) from standard dictionary values. One bond angle (thymine 6: O4'-C1'-N1) deviates 2.5° from the standard dictionary value. Residue 12 and the base atoms of residue 7 are not included in the deposited model (discussed in section 4.4.2). The RMSDs of bond lengths and bond angles is 0.006 Å and 0.990° respectively.

The atomic coordinates and structure factors have been deposited in the Protein Databank with accession code 3QCR (deposition date: 17/01/11). This structure was first reported in The Journal of the American Chemical Society in 2011 (see Appendix A).
Table 4.1 Data collection and refinement statistics for the DNA G-quadruplex-acridine complex.

<table>
<thead>
<tr>
<th>Data collection</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>d(TAGGGTTAGGGT)</td>
</tr>
<tr>
<td>Ligand</td>
<td>di-substituted acridine (FD-121)</td>
</tr>
<tr>
<td>Space group</td>
<td>P622 (P6_22)</td>
</tr>
<tr>
<td>Unit cell dimensions:</td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>71.67, 71.67, 29.37</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.00, 90.00, 120.00</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>29.37 – 3.16</td>
</tr>
<tr>
<td>Rint (%) overall</td>
<td>10.2 (11.5)</td>
</tr>
<tr>
<td>I / σ</td>
<td>19.9 (2.8)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>97.5 (100.0)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>16.9 (19.1)</td>
</tr>
<tr>
<td>Refinement</td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>9.96 – 3.20</td>
</tr>
<tr>
<td>Reflections</td>
<td>783</td>
</tr>
<tr>
<td>Rwork / Rfree (%)</td>
<td>34.6 / 37.8</td>
</tr>
<tr>
<td>No. of atoms</td>
<td>278</td>
</tr>
<tr>
<td>Ions</td>
<td>2</td>
</tr>
<tr>
<td>Water</td>
<td>0</td>
</tr>
<tr>
<td>Overall B-factor (Å²)</td>
<td>39.30</td>
</tr>
<tr>
<td>RMS deviations</td>
<td></td>
</tr>
<tr>
<td>Bond-lengths (Å)</td>
<td>0.006</td>
</tr>
<tr>
<td>Bond-angles (°)</td>
<td>0.990</td>
</tr>
<tr>
<td>PDB ID</td>
<td>3QCR</td>
</tr>
</tbody>
</table>

(Values in brackets refer to the highest resolution shell, 3.26 – 3.16 Å)
4.4 RESULTS AND DISCUSSION

4.4.1 PARTIAL CRYSTAL STRUCTURE OF A TELOMERIC DNA QUADRUPLEX-ACRIDINE COMPLEX: OVERALL TOPOLOGY AND GENERAL FEATURES

Despite issues with crystal quality and refinement, it was possible to determine a low-resolution, partial structure of a bimolecular DNA G-quadruplex in complex with the disubstituted acridine, FD-121 (studied in complex with an RNA G-quadruplex in chapter 3). The G-quadruplex structure here is the DNA equivalent of the RNA structure described in chapter 3, and involves a parallel-stranded G-quadruplex formed from two strands of the human telomeric repeat sequence, d(TAGGGTTAGGGT) (figure 4.2). The asymmetric unit contains one DNA strand, with the quadruplex being completed by a two-fold crystallographic rotation axis through the central core of the quadruplex (indicated in figure 4.2a). This makes each strand of the G-quadruplex identical in both sequence and three-dimensional structure, analogous to the RNA-acridine structure described in chapter 3. Although no water molecules were observed (due to the resolution and quality of the data), two potassium ions could be positively identified, centrally coordinated to the O6 atoms of the G-quartets (in accord with all crystallographically determined human telomeric G-quadruplex structures, including those described in chapters 2 and 3). The acridine compound, FD-121 (figure 4.2c), is bound to the 3’ G-quartet surface of the DNA G-quadruplex in a 1:1 stoichiometry (figure 4.2b). However, the acridine is poorly ordered in its G-quartet binding site, allowing two conformations of the ligand in the crystal lattice. The possible causes and implications of this ligand-disorder are discussed below.
Figure 4.2 Overall folding and features of the DNA G-quadruplex-acridine complex.

a) The DNA sequence, d(TAGGGTTAGGGT), folds as a parallel-stranded bimolecular G-quadruplex. A single disordered acridine ligand molecule is bound to the 3' G-quartet surface, occupying two separate positions. b) View onto the 3' G-quartet, highlighting the rotational disorder of the complexed acridine. FD-121 sits over a 2-fold rotation axis, hence two positions for the molecule are seen in the electron density, at relative positions of 0° and 180°. c) Chemical structure of FD-121. d) Schematic representation of the DNA G-quadruplex, in the same orientation as shown in (a). DNA coloured according to NDB scheme: thymine = blue, adenine = red, guanine = green.
4.4.2 QUALITY OF THE ELECTRON DENSITY AND RESULTING DNA-COMPLEX MODEL

As has been mentioned above, although a reliable solution was obtained for this structure, several aspects of the electron density maps (and therefore resulting DNA model) showed areas of concern, and as a result, this structure does not provide a highly detailed view of DNA-acridine interaction. The structure was refined to 3.2 Å, and while electron density for the guanines and backbone is generally of good quality (figure 4.3), several of the DNA bases are poorly modelled, or not modelled at all. The propeller-loop in particular was not stabilized sufficiently by crystal packing interactions and consequently, only adenine 8 of the TTA loop is modelled into reasonable electron density. The 3' terminal thymine (T12) could not be fitted into electron density, and the 5' terminal thymine and adenine (T1 and A2) residues are highly mobile in the crystal lattice (as determined by the quality of the maps). These terminal residues appear to be stabilized by base-stacking forces only. No ordered water molecules could be seen in electron density maps and hence the final PDB model does not contain waters.

The primary reason this complex was studied was to compare the binding mode of the FD-121 acridine when complexed to DNA and RNA. Fortunately, therefore, reasonable electron density can be seen for the acridine molecule. The drug, however, sits on a two-fold rotation axis but lacks rotational symmetry, and therefore disrupts the symmetry of this region. The ligand has been assigned two conformations in the final model (differing by a rotation of 180°), each contributing 50% to the overall electron scattering in this region. The acridine-core, having superficial two-fold rotational symmetry, fits well into electron density, however, the sidechains are poorly ordered, with only one sidechain (up to the phenyl group) of each ligand conformation sitting comfortably in electron density (figure 4.3c).
4.4.3 QUALITY OF THE ELECTRON DENSITY: CONSIDERATIONS

Clearly, the data used to solve and refine this DNA-acridine structure is not ideal. As is often the case in crystal structure determination, the crystals themselves are likely to be one of the main factors affecting (diffraction) data quality. As mentioned above (section 4.3.2), these crystals had poor microscopic morphology, lacking the sharp edges and birefringence of classically ‘good quality’ crystals. This could be due to the
rate of crystal growth – which was relatively high (for a ‘quadruplex-ligand’ system at least), and may not have allowed enough time for stable crystal packing contacts to be formed. These crystals, however, could not be grown at reduced concentration gradients (e.g. closer to 50%), and indeed could not be reproduced easily. It is possible that ligand decomposition could have played a role in crystal quality, as the triazole ring is typically not stable over long periods of time. It is possible, therefore, that a significant percentage of quadruplexes in the crystal lattice were occupied with degraded-ligand molecules – this would be expected to result in a disordered ligand-binding region (i.e. the 3’ G-quartet). It is possible, however, that crystal quality is not the main factor in the quality of the data for this structure, and that the disorder and poor quality of the electron density for the ligand binding site is simply due to a lack of specific, strong contacts formed between the acridine and DNA. If this is the case, then this lack of direct acridine-DNA contacts could have implications for this series of modified-acridine compounds (discussed below).

Despite suboptimal diffraction data quality, this structure can still yield reliable information concerning acridine-DNA interactions. From the electron density, features such as DNA topology, ligand-binding site, gross features of ligand binding mode, loop arrangements and crystal packing interactions can be confidently studied. This structure however does not show accurate hydrogen bonding interactions (as individual atomic positions are not accurately known), B-factor distribution or water structure.

4.4.4 DNA G-QUADRUPLEX-ACRIDINE BINDING

In this crystal structure, a single acridine molecule is bound centrally to the external surface of the 3’ G-quartet. As mentioned above, this single ligand molecule is disordered, and has been assigned two distinct conformations in the model (at relative positions of 0° and 180°). The acridine interacts with the DNA mainly though π-stacking interactions, and there are no visible direct hydrogen bonds or electrostatic interactions between the DNA and the ligand. An absence of direct stabilising contacts between the DNA and ligand could account for the apparent rotational flexibility of the
drug in this binding site. In addition to π-stacking interactions with the G-quartet, the acridine compound is stabilised by symmetry related π-stacking interactions arising from the stacking of two-quadruplex subunits in the crystal lattice, interacting at the 3'-3' interfaces (figure 4.4). This arrangement is quite strikingly similar to the crystal-packing-stabilised mode of ligand binding observed in the FD-121-RNA quadruplex crystal structure (described in chapter 3). However, there are several key differences between these two complexes:

1) The TTA propeller loops of the DNA structure are not remodelled (relative to the native structure) to create an extended ligand binding platform (such as a hexad or octet).

2) Quadruplex-quadruplex association in the crystal lattice involves 3'-3' stacking for the DNA complex, in contrast to the 5'-5' stacking observed for the RNA complex.

3) The acridine, FD-121, interacts with the DNA quadruplex in a 1:1 stoichiometry compared to a 2:1 stoichiometry for the RNA quadruplex complex.

Considering the ability of FD-121 to exploit and stabilise a purine octet in the RNA-acridine complex structure (chapter 3), it was surprising not to see a similar quadruplex-acridine interaction with the equivalent DNA structure. Although the G₄A₄ octet of the RNA quadruplex is dependent upon C2'-OH group interactions (which are obviously not available within the DNA structure), DNA G-quadruplexes are capable of forming similar, albeit smaller, G-quartet-modified platforms. For example, an AGGGGA hexad formed from sheared A-G mismatch basepairing was observed in a DNA-quadruplex complex, involving a bimolecular human telomeric G-quadruplex bound by a tetra-substituted naphthalene diimide compound (3CCO\textsuperscript{114}). The adenine residues of this hexad originate from the TTA propeller loops, and so the naphthalene diimide complex structure displays a mode of extended-G-quartet formation comparable to the RNA structure described previously (chapter 3), (although it should
be noted that this hexad is not involved in ligand binding, but rather mediates 5'-5' quadruplex stacking). The absence of an AGGGGA hexad ligand binding platform in the FD-121-DNA complex described here suggests that such a motif is not suitable for acridine binding – conceivably providing a surface area which is too small to allow two FD-121 molecules to bind and too large for the effective binding of a single FD-121 molecule.

Considering the fact that the 5' G-quartet in this crystal packing environment remains accessible for ligand binding – and yet is not exploited – it appears that the 3' G-quartet represents the optimal site for FD-121-DNA G-quadruplex binding. It is not clear from the crystal structure why this is – although the 3' and 5' G-quartets are not equivalent (e.g. in terms of hydrogen bonding groups and flanking residues), the quality of the data here does not allow the mechanism by which FD-121 recognises such differences to be determined. It is possible that the increased accessibility of the groove regions from the 3' G-quartet direction may be partially responsible, as there is evidence that one sidechain of FD-121 is positioned, to some degree, within in a groove region (figure 4.2a). Higher-resolution crystallographic data would certainly inform on the situation, as would data for an alternative crystal packing environment (i.e. different space group). It is interesting to note that BRACO-19 was also shown to interact with the 3' G-quartet of a bimolecular human telomeric G-quadruplex, whilst packed into a very different lattice environment, suggesting that the 3' G-quartet could be a favourable region for acridine-DNA G-quadruplex binding.

Concerning stoichiometry, the 1:1 FD-121:quartet stoichiometry observed for this structure is probably a consequence of the physical size of the 3' G-quartet binding site, which is approximately half the size of the 5' G₄A₄ octet of the equivalent RNA-acridine complex, for which a 2:1 FD-121:quartet stoichiometry was observed.
Figure 4.4 Crystal packing and ligand binding.

a) Crystal packing interactions of the DNA-acridine complex (3QCR). Four DNA strands plus two ligand molecules associate in the crystal lattice to form a complex of two bimolecular G-quadruplexes stacked at the 3'-3' interface, with two stacked acridine molecules sandwiched between the G-quadruplexes. b) the crystal packing arrangement for the equivalent RNA-complex (3MIJ) shares features with the DNA complex, in that two G-quadruplexes are stacked in the crystal lattice, with ligand molecules stacked between the G-quadruplexes, however, the RNA is arranged in the opposite orientation to the DNA (5'-5' for the RNA) and has twice as many acridine molecules sandwiched between the G-quadruplex subunits (four acridine molecules for the RNA). Schematic representations are shown beneath both structures in (a) and (b). The nucleic acid bases and ligands of each asymmetric unit are coloured separately for the cartoon representations in (a) and (b).
4.5 GENERAL DISCUSSION AND CONCLUSIONS

It is interesting to note that, although there are significant differences in the manner by which the acridine FD-121 interacts with G-quadruplex DNA and RNA, these two crystal structures share the common feature of acridine-mediated G-quadruplex stacking in the crystal lattice. However, the apparent propensity of this acridine to mediate quadruplex-quadruplex associations is perhaps a curiosity of these crystal structures – the key finding of this DNA complex was the contrasting mode of FD-121-quadruplex binding compared to the equivalent RNA crystal structure. The absence of an extended G-quartet binding motif in the DNA-FD-121 complex provides evidence in support of the proposition that the G₄A₄ octet observed in the RNA-acridine complex (chapter 3) is an RNA G-quadruplex specific motif, which may be exploitable in quadruplex-directed drug design.

4.5.1 LIGAND BINDING MODE: COMPARISON TO DNA-ACRIDINE COMPLEXES

As well as differing from the equivalent RNA-acridine complex, the DNA-complex described in this work also differs significantly to the other DNA G-quadruplex acridine complexes which have been previously determined. A significant number of crystal structures are available involving the bimolecular G-quadruplex formed from the O. nova telomeric repeat sequence, d(GGGGTTTTGGGG), in complex with various disubstituted acridine compounds. The structures all share a common binding mode involving an acridine bound between a G-quartet and an external TTTT loop. The acridines are typically stabilised by extensive hydrogen bonds and water bridges with the DNA. These O. nova complexes are perhaps not the most relevant structures for comparison to the DNA-acridine complex described here, they do, however, emphasise the significant number of direct (and water mediated) contacts formed between acridine drugs and these DNA structures, a feature which is clearly not observed in either of the FD-121-quadruplex complexes described here (3MIJ and 3QCR).
An important DNA G-quadruplex-acridine complex of relevance to the structure described here is of that of the human telomeric repeat G-quadruplex formed from the sequence d(TAGGGTTAGGGT) complexed to BRACO-19 (3CE5). As with FD-121, BRACO-19 binds to the 3' G-quartet of the quadruplex, however, this is the extent of the similarity between these two structures, as BRACO-19 is well-ordered and stabilised by significant contacts (direct and water mediated) with the DNA, contrasting significantly to the mode of FD-121-quadruplex binding (figure 4.5). So why is FD-121 disordered when bound to G-quadruplex DNA? It is possible that substitution of the 3- and 6- amide groups of BRACO-19 for the triazole rings of FD-121 is responsible for the lack of direct contacts between FD-121 and the DNA (figure 4.5b-c). The amide groups of BRACO-19 play a major role in forming strong contacts with the DNA (based on the crystallographic data), and it is possible that these groups, in this specific position relative to the acridine core, are more effective at binding to human telomeric G-quadruplex DNA than triazole groups. It is perhaps interesting to note that, although FD-121 was shown to significantly stabilise human telomeric G-quadruplex DNA in a FRET-melting assay (with a ΔT_m of 15.5 °C), BRACO-19 stabilises the same G-quadruplex by a further 10 °C (ΔT_m = 25.9 °C) (table 4.2)\(^9\). Although FD-121 exhibits a considerably higher level of selectivity for human telomeric G-quadruplex DNA than BRACO-19, these data suggest there is a critical structural factor in BRACO-19 which leads to a high binding affinity for human telomeric G-quadruplex DNA. This structural factor could potentially be the location of the amide functionalities, although it should be noted that BRACO-19 and FD-121 differ in many other respects, hence any comparison of the binding of these two molecules to telomeric DNA (i.e. 3CE5 and 3QCR) should be done so with caution.

| Table 4.2 FRET-melting data for acridine-based G-quadruplex binding compounds tested at 1 μM (°C)* |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Ligand          | Human telomeric G-quadruplex | cklt1 promoter G-quadruplex | cklt2 promoter G-quadruplex | duplex DNA |
| BRACO-19        | 25.9             | 20.1             | 25.3             | 11.2           |
| FD-121          | 15.8             | 0.8              | 1.4              | 0.4            |
| FD-121-analogue*| 15.5             | 0.7              | 2.5              | 0.6            |

* Data taken from Sparapani et al., 2010\(^9\). # Diethylamine end groups replaced with pyrrolidine ring.
Figure 4.5 Comparison of acridine-DNA quadruplex binding modes.

a) Structural overlay of the bimolecular G-quadruplex-BRACO-19 complex (3CES: DNA = blue ribbon, BRACO-19 = yellow stick) and the equivalent FD-121 complex (3QCR: DNA = green ribbon, FD-121 = green stick). Both structures involve acridine compounds bound to the external 3' G-quartet surface. BRACO-19 however, is held far more rigidly in place, primarily through the formation of multiple hydrogen bonding contacts.

b) Hydrogen bonds observed in the crystal structure of FD-121 complexed to the RNA sequence r(UAGGGUAGGGU) (3MIJ) (shown as arrows).

c) Hydrogen bonds observed in the crystal structure between BRACO-19 and the DNA (arrows). Figure (c) was taken from Campbell et al., 2008.
4.5.2 IMPLICATIONS FOR DRUG DESIGN

Based on the DNA-acridine complex described in this work, several potentially important factors concerning both DNA and RNA G-quadruplex drug design are highlighted:

1) As the extended G₄A₄-octet binding platform of the RNA-acridine complex described previously (chapter 3) is not seen in the equivalent DNA-acridine complex, is it likely that this octet binding platform is indeed an RNA G-quadruplex-specific motif, and may represent an exploitable feature in quadruplex-drug design.

2) The lack of direct contacts between the acridine compound, FD-121, and either of the quadruplexes with which it was cocrystallised suggests that this is a characteristic specific to this acridine (compared to, for example, BRACO-19). Therefore, the rational targeting of the C²'-OH groups of RNA G-quadruplexes for drug binding interactions remains a viable avenue for deriving RNA vs DNA quadruplex selectivity.

3) It is possible that the lack of direct FD-121-nucleic acid contacts is a result of the absence of amide groups at the 3- and 6- positions of the acridine core. This highlights the potential importance of the amide group positions within BRACO-19 with regard to direct DNA binding.

4.5.2.1 IMPLICATIONS FOR TELOMERIC DNA G-QUADRUPLEX TARGETING

The acridine complex structure reported here, although of suboptimal quality, has provided some useful information concerning the use of acridine based compounds for the targeting of human telomeric DNA G-quadruplexes. It may be worth noting that, of the small number of G-quadruplex-interacting compounds which have performed well in vivo (i.e. in tumour xenograft models) one was acridine based – BRACO-19. The structure reported here – although initially intended simply as a validation of the RNA-
specific binding platform described previously — now represents one of only two crystal structures involving acridine-molecules bound to G-quadruplex structures formed from human telomeric repeat DNA. This crystal structure, therefore, provides data which may be of use for future telomeric G-quadruplex-directed drug design, for example, it is apparent from the structure that the position of amide groups relative to the acridine core could be an important factor in acridine-(human)-quadruplex drug design. Additional structural work involving human telomeric G-quadruplex-acridine complexes would certainly aid rational drug design, particularly regarding functional group positioning (e.g. amide group positioning).

4.5.2.2 IMPLICATIONS FOR TELOMERIC RNA G-QUADRUPLEX TARGETING

As stated above, the DNA-acridine structure reported here was initially intended to inform about RNA-acridine binding, and it has indeed been useful in confirming (or at least in supporting) the RNA-specific nature of the G₄A₄ binding platform observed in the RNA-FD-121 complex. This is fortunate, as this large octet binding platform provides a structural difference between telomeric DNA and RNA G-quadruplex structures which may yet be utilised in drug design. As stated in section 3.5.2 however, the production of a therapeutic compound able to do this is difficult to envisage, but it is certainly within reason. The DNA-acridine complex has informed further on the RNA-FD-121 complex, by exhibiting a similar lack of direct FD-121-quadruplex interactions. This suggests that the paucity of direct FD-121-RNA contacts seen in the previous structure was in fact a characteristic specific to this acridine molecule, rather than a result of the RNA itself. This offers the possibility of exploiting direct C₂'-OH group interactions in RNA G-quadruplex drug-targeting — a goal which appears more achievable than designing a ligand to interact specifically with an extended G₄A₄ octet. Clearly, further structural data would be extremely useful in designing RNA-quadruplex specific ligands, and with the availability now of several high resolution TERRA G-quadruplex structures¹⁵⁹,²⁰¹,²⁰⁷, it might be hoped that some rationally designed RNA G-quadruplex-targeted drugs may be designed and validated. This would overcome a significant obstacle to obtaining a high-resolution structure of an RNA G-quadruplex-drug complex — namely, the absence of ligands which bind TERRA (or 5' UTR mRNA) G-
quadruplex structures with high affinity. Understandably, such compounds do not exist in large numbers as yet, as RNA G-quadruplexes in general are still be investigated and validated as drug targets. It is hoped, therefore, that the structures and findings reported here may be of some use in preliminary (or subsequent) RNA G-quadruplex drug design.
CHAPTER 5

X-RAY CRYSTALLOGRAPHIC STUDIES
OF INTRAMOLECULAR TELOMERIC
DNA G-QUADRUPLEX-NAPHTHALENE
DIIMIDE COMPLEXES
5.1 BACKGROUND

5.1.1 MODIFIED NAPHTHALENE DIIMIDES AS POTENT TELOMERIC DNA G-QUADRUPLEX BINDING COMPOUNDS

The preceding chapters have been concerned primarily with the folding and drug-targeting of G-quadruplexes formed from human telomeric RNA. The focus will now be shifted slightly to concentrate on the interactions of drug molecules with telomeric DNA G-quadruplexes only. As discussed previously (section 1.3.4.5), there is significant therapeutic potential in producing effective telomeric DNA G-quadruplex binding compounds as a route to disrupt telomere elongation and selectively inhibit cancer cell growth. Many ligand-cores have been studied as potential building blocks for effective libraries of modified and diversified G-quadruplex binding compounds, such as the acridine core, which was discussed extensively in previous chapters. The focus of this chapter will be the naphthalene diimide core – a building block which has shown significant potential as an effective G-quadruplex binding motif (figure 5.1).

Naphthalene diimides (and imides) were first investigated as duplex-DNA binding therapeutic agents, with several naphthalene diimide (ND) based ligands reaching anticancer clinical trials. The ND core has progressed from use as a duplex-DNA binding motif to use as a G-quadruplex binding motif. This is because the ND core satisfies, very effectively, the three main requirements of a G-quadruplex-binding ligand, namely:

1) The presence of a large aromatic surface for G-quartet π-π stacking.

2) The potential to allow a high degree of functionality/diversification to be introduced via sidechain addition.

3) The ability to incorporate positive charge(s) (within sidechains).
Previously, an extensive series of di-, tri- and tetra-substituted naphthalene diimides were prepared and assessed for G-quadruplex binding affinity and anti-cancer cellular toxicity \((in\ cellulo)\), with many members of the series performing exceptionally well \(^{211}\). The series involved side chain substitutions of the ND core at the R1 and R2 positions (equivalent and non-equivalent side chains), with sidechains composed of 2-, 3- or 4 carbons terminating in specific functional groups, including dimethylamine, diethylamine, pyrrolidine, morpholine, piperidine and hydroxyl groups. Many of these compounds showed excellent selective telomeric G-quadruplex stabilisation (as measured by the FRET melting assay) as well as effective telomerase inhibitory activity (as measure by the telomerase-repeat amplification protocol \([T R A P]\) assay), often performing better than reference compounds, BRACO-19 and TmPyP4. The structures of the top ranking compounds are shown in figure 5.2, along with FRET and \(in\ cellulo\) data. As well as biophysical performance, several members of this ND series performed well in cell-based anti-cancer assays, showing nanomolar IC\(_{50}\) (inhibitory concentration to reduce growth by 50%) values for a range of cancer cell-lines (including MCF7 and A549). Importantly, toxicity to non-cancerous cells (WI38 human fibroblast cell line) was typically acceptably low. As well as telomerase/telomere disruption, a member of this ND series with pyrrolidine end groups (figure 5.2, ligand 13) was shown to significantly reduce \(c k i t\) expression in gastrointestinal stromal tumour cells\(^{119}\), \(c k i t\) being an important overexpressed oncogene in this cell line/cancer type. The anti-\(c k i t\) effects of this compound were proposed to be a result of the stabilisation of a G-quadruplex motif located within the promoter region of the \(c k i t\) gene, which would be
expected to disrupt ckit expression (this hypothesis was supported with additional biophysical data).^119^

**Ligand 24**
FRET $\Delta T_m$: 23.7 °C (3.5 °C)
IC$_{50}$ MCF7: 287.7 nM (10 μM)

**Ligand 3**
FRET $\Delta T_m$: 33.2 °C (4.5 °C)
IC$_{50}$ A549: 28.7 nM (292.3 nM)

**Ligand 13**
FRET $\Delta T_m$: 29.7 °C (3.5 °C)
IC$_{50}$ MCF7: 10.2 nM (63.4 nM)

Figure 5.2 Representative members from a series of tetra-substituted naphthalene diimide compounds. Example data are given for the ability of the compounds shown to stabilise human telomeric G-quadruplex DNA (of the sequence d[GGGG(TTAGGG)]$_3$) in a FRET melting assay (0.5 μM ligand concentration). The values in brackets refers to the $\Delta T_m$ of the same compound for a control duplex DNA sequence. IC$_{50}$ (inhibitory concentration for 50% cell kill) values are given for specified cancer cell lines, along with the equivalent IC$_{50}$ concentration for a non-cancerous control cell line (WI38 – human fibroblast cells). Data are taken from Cuenca et al., 2008.^211^

5.1.2 CRYSTAL STRUCTURES OF NAPHTHALENE DIIMIDE G-QUADRUPLEX-COMPLEXES

The biophysical and biological data for the series of ND compounds described above quite convincingly highlights the significant potential of ND compounds in G-
quadruplex based drug design. As is often the case with drug development programmes, structural studies can often inform about drug-design and mode of action, and such studies were indeed performed on several members of this ND series. Four high-resolution crystal structures have been solved involving ND compounds bound to human telomeric DNA G-quadruplex structures. These crystal structures include complexes formed between ligand 24 (figure 5.2) bound to a bimolecular and an intramolecular quadruplex (figure 5.3a and 5.3b, respectively), a complex between compound 3 and an intramolecular quadruplex and a complex between an analogue of compound 3 (with diethyl terminal substitutions) and a bimolecular quadruplex. Despite the apparent variability in quadruplex type and ND functionality, all four complexes display a surprisingly similar mode of ligand binding. The key findings from the complexes were:

1) The ND ligands associate with both the 3' and 5' G-quartet surfaces.

2) The ND ligands also interact with binding sites formed exclusively from TTA loop residues.

3) The ND sidechains form very few contacts with the quadruplex groove regions.

The binding of these compounds to both 5' and 3' G-quartets was expected, considering the large aromatic surface area provided by the ND core, however, it was not expected that these compounds would bind to loop-generated platforms (i.e. binding sites composed of thymine and adenine residues only). A second unexpected finding was the lack of interaction between the ND sidechains and the quadruplex groove regions, as these compounds were developed specifically to utilize sidechain-groove hydrogen bonding/electrostatic contacts, to enhance G-quadruplex binding affinity and selectivity. Although these crystal structures did not show the expected mode of ND-quadruplex interaction, the structural information provided fuelled further hypotheses and investigations into the improvement of ND compounds as effective telomeric G-quadruplex binding agents.
5.1.3 A RECENT SUBSET OF MODIFIED NAPHTHALENE DIIMIDE COMPOUNDS

Based on the crystallographic data described above, it was hypothesized – and shown through molecular modelling studies – that several modifications of the previous ND sidechains could yield more effective compounds\(^\text{94}\). The suggested changes included:

1) The increase of sidechain length to include 4 and 5 carbons linking the ND core and terminal functional group of the ligand;

2) The use of the bulky, positively charged piperazine end-groups.

Subsequently, a short series of three tetra-substituted (N)methyl-piperazine ND compounds were synthesized and assessed in FRET and cancer-cell based assays – performing well in both areas (figure 5.4 for chemical structures and data). All three
compounds, which differ only in sidechain length, significantly stabilised human telomeric G-quadruplex DNA, with negligible stabilisation of a test duplex structure. These compounds also showed nanomolar toxicity towards a panel of cancer cell-lines (MCF7, A549, MIA-Pa-Ca-2, PANC-1, HPAC, BxPc-3), with acceptably low toxicity to non-cancer cells (WI38) (figure 5.4b). Investigations into the mode of action of these compounds identified POT1 and topoisomerase IIIα displacement as key cellular effects, suggesting telomeric G-quadruplex stabilisation to be a likely mode of action.

Based on the activity of this subset of ND compounds – and considering the utility of structural data in quadruplex drug-design – crystallographic investigations into the binding of these compounds to human telomeric DNA G-quadruplexes were undertaken, the results of which are described below.

![Chemical structure](image)

<table>
<thead>
<tr>
<th></th>
<th>n = 3</th>
<th>n = 4</th>
<th>n = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human tel</td>
<td>28.3</td>
<td>24.7</td>
<td>23.8</td>
</tr>
<tr>
<td>ckit1</td>
<td>1.8</td>
<td>4.9</td>
<td>1.5</td>
</tr>
<tr>
<td>ckit2</td>
<td>15.2</td>
<td>16.7</td>
<td>1.5</td>
</tr>
<tr>
<td>T-loop</td>
<td>1.3</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>IC50 (nM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCF7</td>
<td>167</td>
<td>100</td>
<td>196</td>
</tr>
<tr>
<td>A549</td>
<td>108</td>
<td>69</td>
<td>258</td>
</tr>
<tr>
<td>WI38</td>
<td>9042</td>
<td>5500</td>
<td>8430</td>
</tr>
</tbody>
</table>

**Figure 5.4** Methyl-piperazine modified ND compounds.  
(a) Chemical structure of the ND-modified compounds synthesised by Hampel and co-workers and corresponding FRET and *in cellulo* data (b). FRET stabilisation is shown for the human telomeric G-quadruplex ("Human tel"), G-quadruplexes formed from *ckit* promoter region sequences (*ckit1* and *ckit2*) and a control duplex sequence (T-loop). All three ND analogues significantly stabilise human telomeric G-quadruplex DNA at 0.5 μM concentration. Anti-cancer cell based assay data is also shown for all three compounds (IC50), highlighting the potency and selectivity of these compounds in killing cancer cells. The data in (b) were taken from Hampel et al., 2010.
5.2 AIMS

The work described in this chapter was aimed at providing high-resolution crystallographic data for (ideally) all three members of the subset of (N)methyl-piperazine tetra-substituted naphthalene diimide compounds (shown in figure 5.4). As the compounds of this short series differ only in sidechain length, it was hoped they could be studied in equivalent systems (i.e. bound to the same G-quadruplex) to allow reliable conclusions to be drawn regarding optimal ND-modification. The key questions sought to be answered were:

1) **What is the mode by which these piperazine-modified ND compounds recognise and bind human telomeric G-quadruplex DNA?**

2) **Are there large differences regarding quadruplex binding site location within these series?**

3) **What effect does sidechain length have on specific quadruplex-ligand contacts?**

4) **What effect does sidechain length have on core localization?**

5) **How does the binding of these ND compounds compare to previous ND-G-quadruplex interactions?**

6) **Can we use this information to improve these compounds as selective telomeric G-quadruplex-binding agents?**
5.3 METHODS AND MATERIALS

5.3.1 SAMPLE PREPARATION

The DNA sequence used in these crystallisation experiments, \text{d[AGGG(TTAGGG)_3]}, was purchased from MWG (Germany) (HPLC purified) and used without further purification (referred to as 'DNA 22-mer'). The DNA was annealed at a single-strand concentration of 2 mM, following the ionic conditions and annealing protocol described in section 4.3.1. Synthesis of the tetra-substituted naphthalene diimide compounds used in these studies (BMSG-SH3, BMSG-SH4 and BMSG-SH5) was performed by Hampel and co-workers and has been described previously\(^9\). Stock solutions of the ligands used in this work consisted of 10 mM of the compound dissolved in water. Ligand-complex solutions used for crystallization experiments consisted of 1.5 mM DNA plus 1.5 mM ligand (for the BMSG-SH3 complex) and 1 mM DNA plus 1 mM ligand (for the BMSG-SH4 and BMSG-SH5 complexes). The complexes were incubated at room temperature for 1 hour prior to crystallisation trials.

5.3.2 CRYSTALLISATION

Crystallisation of the DNA 22-mer-ligand complexes described in this chapter was performed with the assistance of Rossella Promontorio, a visiting Erasmus student from the University of Cagliari, Italy.

Crystals of the DNA 22-mer plus BMSG-SH3 were grown at 12 °C in hanging drops consisting of 0.5 uL of the complex solution plus 0.5 uL of the crystallization reagent (composed of 20 % PEG400, 100 mM Li\textsubscript{2}SO\textsubscript{4} and 50 mM sodium cacodylate, buffered at pH 6.5). Square crystals of the complex appeared after 1 week and were left to grow for 1 month, with final approximate dimensions of 0.1mm x 0.1mm x 0.05mm (figure 5.5a).

Crystals of the DNA 22-mer BMSG-SH4 complex were grown at 10 °C in hanging drops consisting of 0.7 uL of the complex solution plus 0.7 uL of the crystallization reagent (composed of 15 % PEG400, 300 mM KBr, and 50 mM sodium cacodylate...
buffered at pH 6.5). Hexagonal crystals of the complex appeared after 10-14 days and were left to grow for 2 months (figure 5.5b).

Crystals of the DNA 22-mer BMSG-SH5 complex were grown at 10 °C in hanging drops consisting of 0.7 uL of the complex solution plus 0.7 uL of the crystallization reagent (composed of 20 % PEG400, 300 mM KBr, and 50 mM sodium cacodylate buffered at pH 6.5). Small, poorly ordered crystals appeared after three weeks and were left to grow for 2 months before data collection (figure 5.5c).

![DNA 22-mer + BMSG-SH3](image1)

![DNA 22-mer + BMSG-SH4](image2)

![DNA 22-mer + BMSG-SH5](image3)

**Figure 5.5 DNA G-quadruplex-naphthalene diimide complex crystals.** Examples of the DNA-naphthalene complex crystals used for data collection and structure determination. The deep-blue colour of the crystals arises from the ligand molecule. Crystals of the DNA-BMSG-SH3 and DNA-BMSG-SH4 complexes were particularly well formed (a and b, respectively). The hexagonal shape of the DNA-BMSG-SH4 complex crystal reflects the space group symmetry (P6/hexagonal). Crystals of the DNA 22-mer + BMSG-SH5 complex were of much poorer quality in comparison, being much smaller with less well-defined edges (however, these crystals were surprisingly birefringent).

### 5.3.3 DATA COLLECTION AND PROCESSING

Crystals of the DNA 22-mer-BMSG-SH3 complex were frozen in liquid nitrogen and taken to the Diamond Light Source synchrotron (Oxford, UK) for data collection. Ten crystals were screened on beamline I04-1, with all crystals producing good quality diffraction data. A dataset was collected on the crystal shown in figure 5.5a, with a collection strategy of one second exposure time per 1° phi oscillation, using X-ray radiation of wavelength 0.9173 Å. This crystal diffracted to around 2.2 Å resolution (figure 5.6a). The images were processed using XDS\textsuperscript{208} and scaled in Scala\textsuperscript{193}. These crystals belong to space group $P3_121$, with unit cell dimensions $a = b = 50.87$, $c = 52.46$, $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$. 

---

155
A single crystal of the DNA 22-mer-BMSG-SH4 complex was soaked in a cryoprotection solution for 60 seconds before flash freezing in liquid nitrogen. The cryoprotection solution consisted of 15% glycerol, 7.5% PEG400, 150 mM potassium bromide and 25 mM sodium cacodylate (pH 6.5). A dataset was collected at the Diamond Light Source synchrotron on beamline I03, following a similar collection strategy as described above. Data were processed and scaled using Xia2\textsuperscript{12} and Scala\textsuperscript{193}. The structure belongs to the space group \textit{P6} with unit cell dimensions: 
\[ a=b=63.03\AA, c=42.39\AA, \alpha=\beta=90^\circ, \gamma=120^\circ. \]

Diffraction data were collected on a single crystal of the DNA 22-mer-BMSG-SH5 complex on a microfocus synchrotron beamline (I24, Diamond Light Source) by Dr. Ambrose Cole (Birkbeck, University of London, UK). Unfortunately, the data were of poor quality and the structure could not be solved, however the space group and unit cell constants could be confidently determined – the crystals belong to space group \textit{P6} with unit cell dimensions of 
\[ a=b=64.25\AA, c=43.1\AA, \alpha=\beta=90^\circ, \gamma=120^\circ. \]

5.3.4 STRUCTURE DETERMINATION AND REFINEMENT

The DNA-G-quadruplex-BMSG-SH3 complex was solved by molecular replacement using Phaser\textsuperscript{195}, using the native human telomeric DNA G-quadruplex structure (1KF1) as a start model\textsuperscript{40}. The start model was truncated to include the guanines and the two internally coordinated potassium ions only. Initial electron density maps were of good quality, allowing the majority of the DNA residues to be accurately fitted into electron density after several rounds of model building and restrained refinement. As residues were modelled, a region of discrete residual electron density above the 3' G-quartet became visible in both 2Fo-Fc and Fo-Fc maps. The naphthalene diimide core of the ligand could be easily fitted into this residual density, with electron density for the sidechains resolving towards the end of the refinement process. Model building and refinement were performed using Coot\textsuperscript{196} and Refmac5\textsuperscript{197}, following the procedure described in section 2.3.4. Individual isotropic B-factors were refined throughout the process, with TLS parameters included towards the end of the refinement\textsuperscript{199,200}. A starting atomic model of BMSG-SH3 plus initial geometry values were generated using
the PRODRG server\textsuperscript{206}. The final model was refined using data between 10.30 Å and 2.3 Å, with final R and R\textsubscript{free} values of 25.8% and 28.8% respectively. The final overall isotropic B-factor of the model (including solvent molecules) is 45.26 Å\textsuperscript{2}.

The DNA 22-mer-BMSG-SH4 structure was solved following a similar procedure as described for the BMSG-SH3 complex. All 22 residues of the DNA could be easily fitted into electron density, however, electron density for the ligand (above the 3' G-quartet) was more diffuse than for the DNA 22-mer-BMSG-SH3 complex, suggesting BMSG-SH4 to be less rigidly bound to the DNA. The electron density for the ligand core was very strong however, allowing the naphthalene-diimide region of BMSG-SH4 to be easily identified and modelled. The quality of the electron density for the DNA and solvent components is generally excellent. The final model was refined using data between 10.20 Å and 2.1 Å, with final R and R\textsubscript{free} values of 24.5% and 28.2% respectively. The final overall isotropic B-factor of the model (including solvent molecules) is 40.69 Å\textsuperscript{2}. Additional refinement statistics for both complexes are shown in table 5.1.

\textbf{5.3.5 VALIDATION AND DEPOSITION IN THE PROTEIN DATABANK}

The final atomic models for the DNA 22-mer-BMSG-SH3 and DNA 22-mer-BMSG-SH4 complexes were validated using the RCSB web server (www.deposit.rcsb.org). The respective RMSDs of bond lengths and bond angles is 0.007 Å and 1.048° for complex 1 and 0.008 Å and 1.623° for complex 2.

The atomic coordinates and structure factors have been deposited in the Protein Databank with accession codes \textbf{35C8} (complex 1) and \textbf{3T5E} (complex 2) (deposition dates: 07/06/2011 and 27/07/2011, respectively). This work has been submitted for publication in the Journal of the American Chemical Society (in peer review).
Figure 5.6 X-ray diffraction patterns of the DNA-quadruplex-naphthalene complex crystals. Diffraction images produced from exposure of the crystals shown in figure 5.5 to high-intensity synchrotron radiation. The DNA 22-mer-BMSG-SH3 and 22-mer-BMSG-SH4 crystals diffracted well in all phi orientations (a and b, respectively), with a maximum diffraction limit of around 2.2 Å and 2.0 Å, respectively. In contrast, the DNA 22-mer-BMSG-SH5 crystals were highly anisotropic and diffracted poorly (c), with a maximum resolution of around 5.0 Å.
Table 5.1 *Data collection and refinement statistics for three telomeric DNA-ligand complexes.*

<table>
<thead>
<tr>
<th>Data collection</th>
<th>Complex 1</th>
<th>Complex 2</th>
<th>Complex 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence</strong></td>
<td>d(AGGG[TTAGGG]₃)</td>
<td>d(AGGG[TTAGGG]₄)</td>
<td>d(AGGG[TTAGGG]₅)</td>
</tr>
<tr>
<td><strong>Ligand</strong></td>
<td>BMSG-SH3 (n =3)</td>
<td>BMSG-SH4 (n = 4)</td>
<td>BMSG-SH5 (n = 5)</td>
</tr>
<tr>
<td><strong>Space group</strong></td>
<td>P 3₁2₁</td>
<td>P 6</td>
<td>P 6</td>
</tr>
<tr>
<td><strong>Unit cell dimensions:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>a, b, c (Å)</em></td>
<td>50.87, 50.87, 52.46</td>
<td>63.03, 63.03, 42.39</td>
<td>64.25, 64.25, 43.1</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.00, 90.00, 120.00</td>
<td>90.00, 90.00, 120.00</td>
<td>90.00, 90.00, 120.00</td>
</tr>
<tr>
<td><strong>Resolution (Å)</strong></td>
<td>26.23 – 2.17 (2.23 – 2.17)</td>
<td>54.59 – 1.94 (1.99 – 1.94)</td>
<td>Data not integrated</td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>10.30 – 2.30</td>
<td>10.20 – 2.10</td>
<td></td>
</tr>
<tr>
<td><strong>Reflections</strong></td>
<td>3467</td>
<td>5372</td>
<td></td>
</tr>
<tr>
<td>R&lt;sub&gt;work&lt;/sub&gt; / R&lt;sub&gt;free&lt;/sub&gt; (%)</td>
<td>25.8 / 28.8</td>
<td>24.5 / 28.2</td>
<td></td>
</tr>
<tr>
<td><strong>No. of atoms</strong></td>
<td>541</td>
<td>572</td>
<td></td>
</tr>
<tr>
<td><strong>Ions</strong></td>
<td>2.5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td><strong>Water</strong></td>
<td>29</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td><strong>Overall B-factor (Å&lt;sup&gt;2&lt;/sup&gt;)</strong></td>
<td>45.26</td>
<td>40.69</td>
<td></td>
</tr>
<tr>
<td><strong>RMS deviations</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond-lengths (Å)</td>
<td>0.007</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Bond-angles (°)</td>
<td>1.048</td>
<td>1.623</td>
<td></td>
</tr>
<tr>
<td><strong>PDB ID</strong></td>
<td>3SC8</td>
<td>3T5E</td>
<td></td>
</tr>
</tbody>
</table>
5.4 RESULTS AND DISCUSSION

5.4.1 OVERVIEW OF RESULTS

Crystallographic studies into the binding of three ND compounds (shown in figure 5.4 and table 5.2) to telomeric DNA G-quadruplex structures resulted in the successful structure determinations (to high resolution) of two separate complexes: one formed between the ND compound BMSG-SH3 and the intramolecular G-quadruplex formed from the human telomeric repeat sequence d(AGGG[TTAGGG]_3) and a second structure formed between the same G-quadruplex and the ND compound BMSG-SH4 (table 5.2). Additionally, limited crystallographic data was obtained for a complex formed between this same G-quadruplex and the ND compound BMSG-SH5. This data, however, allowed only for the assignment of unit cell constants. The main work described in this chapter will be based around the two high-resolution crystal structures formed from human G-quadruplex DNA and the ND compounds BMSG-SH3 and BMSG-SH4, with a focus on the utility and potential importance of these findings to telomere-directed drug design.

Table 5.2 Summary of crystallographic results for the three ND compounds used in these studies

<table>
<thead>
<tr>
<th>Naphthalene diimide compound used in these studies</th>
<th>Sidechain length: n (carbons) =</th>
<th>Compound name</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image.png" alt="Diagram" /></td>
<td>3</td>
<td>BMSG-SH3</td>
<td>High-resolution crystal structure (2.3 Å)</td>
</tr>
<tr>
<td><img src="image.png" alt="Diagram" /></td>
<td>4</td>
<td>BMSG-SH4</td>
<td>High-resolution crystal structure (2.1 Å)</td>
</tr>
<tr>
<td><img src="image.png" alt="Diagram" /></td>
<td>5</td>
<td>BMSG-SH5</td>
<td>Unit cell and space group assignment only</td>
</tr>
</tbody>
</table>
5.4.2 CRYSTAL STRUCTURES OF TWO TELOMERIC DNA G-QUADRUPLE LIGAND COMPLEXES: OVERALL TOPOLOGY AND GENERAL FEATURES

The two crystals structures solved for the ND compounds, BMSG-SH3 and BMSG-SH4, bound to the much-studied human telomeric repeat G-quadruplex d(AGGG[TTAGGG]_3), display a highly similar overall arrangement, in terms of quadruplex association and ligand binding. Both structures involve parallel stranded G-quadruplexes, in accord with all current crystallographically determined telomeric G-quadruplex structures. Both structures show two quadruplexes associating in the crystal lattice via 5'-5' G-quartet stacking, an interaction which is mediated and stabilised by a coordinate-bonded potassium ion (figure 5.7). This arrangement is almost identical to that of the equivalent native G-quadruplex structure (IKF1). It is perhaps interesting to note that these two structures do not share the same spacegroup, suggesting that the potassium-mediated 5'-5' stacking association seen in these structures (as well as the native structure, IKF1) could be considered as a model for higher-order G-quadruplex structure at the telomere.

As well as sharing topology and general crystal-packing characteristics, both structures described here contain a single ND compound bound to the 3' G-quartet surface, resulting in a 1:1 quadruplex:ligand stoichiometry. This is in contrast to the previously reported complexes involving ND compounds bound to telomeric G-quadruplexes, all of which display multiple ND-binding sites, including the unexpected presence of binding sites created exclusively from TTA-loop residues. The singular ligand binding sites in the methyl-piperazine-ND structures reported here therefore suggests a certain degree of specificity in the mechanism of quadruplex recognition by these compounds. Significantly, many of the methyl-piperazine-ND sidechains in both structures can been seen penetrating the quadruplex grooves and interacting strongly with components of these regions, including strong charge-charge interactions. This is the first structural data showing clear interactions of drug molecules with human telomeric G-quadruplex groove regions.
Figure 5.7 Overall folding and features of complexes formed between a human telomeric G-quadruplex and the ND compounds BMSG-SH3 and BMSG-SH4.
(a) Complex formed between the ND compound, BMSG-SH3 (purple sticks), and a G-quadruplex formed from the human repeat sequence, \(\text{d(AGGG[TTAGGG])}\). The DNA folds as an intramolecular parallel-stranded G-quadruplex with three propeller-type loops. Two asymmetric units are shown, involving two G-quadruplexes stacked at the 5'-5' interface, with an ND compound bound to each of the external 3' G-quartets. b) Equivalent complex involving the ND compound, BMSG-SH4 (purple sticks), bound to the same G-quadruplex. This complex displays a highly similar mode of DNA folding and ligand binding to (a). Each structure contains a 'channel' of five potassium ions (purple spheres), which is indicated by the dashed line. c) Schematic of a single asymmetric unit, showing the DNA only. Applicable to either complex structure. Colour key (used throughout thesis): adenine = red, guanine = green, thymine = blue. The DNA backbone is shown as a grey ribbon.
5.4.3 QUALITY OF ELECTRON DENSITY AND RESULTING MODELS

The quality of the diffraction data and electron density for both structures is generally of excellent quality – almost all regions of the DNA and ligand molecules of both structures are modelled accurately into high quality electron density (figures 5.8 and 5.9). The main regions of DNA not modelled accurately (or at all) relate to the BMSG-SH3 structure, wherein the 5’ terminal adenine and one of the TTA propeller loops (T17, T18, A19) are not modelled into high-quality electron density (figure 5.8a). The poor ordering of this TTA loop relates to a previous point (chapter 2.4.2) regarding propeller loop geometry and stability, suggesting again that the typical ‘TAT’ stacked geometry of the TTA propeller loops relies upon crystal packing contacts for some degree of stabilisation.

Importantly, the electron density for the bound ND compounds is of good quality in both structures, allowing the ND core and sidechains to be confidently modelled into density. The electron density for BMSG-SH3 is particularly well defined, removing any concern regarding the rotational orientation of the compound (figure 5.8b). This concern arises from the high degree of pseudo-rotational symmetry of these tetra-substituted diimide compounds, which can result in the ligand adopting multiple conformations with the crystal lattice\(^97,114\). This is indeed the case for the BMSG-SH4 complex, for which electron density within the ND core regions allows more than one rotational conformation of the ligand to be permitted (figure 5.9b). This will be discussed in detail below, however it should be noted that although BMSG-SH4 molecule has been assigned a single-conformation in the final PDB model (which is the highest-occupancy conformation), there remains a high possibility of this compound occupying several different rotational conformations in crystal.
Figure 5.8 Electron density for the BMSG-SH3-G-quadruplex complex.

a) $2F_o-F_c$ map (1.0 $\sigma$ level) for a single asymmetric unit (ASU) of the DNA 22-mer BMSG-SH3 complex (a single asymmetric unit represents one complete intramolecular G-quadruplex plus one ligand molecule). The electron density for the DNA and ligand is generally of excellent quality, with the only area of poor quality density being the T17, T18, A19 propeller loop region (indicated). Additionally, the 5’ terminal adenine (A1) cannot be placed in electron density and hence has not been included in the final model. b) View onto the 3’ G-quartet showing a $2F_o-F_c$ map (0.9 $\sigma$ level) for the ND compound only. The electron density for BMSG-SH3 is of high quality. All carbons are coloured green in (a). Carbons in (b) are coloured yellow (DNA only) or green (ligand only). All figures: nitrogen = blue, oxygen = red, phosphorous = orange.
**Figure 5.9** Electron density for the BMSG-SH4-G-quadruplex complex.

a) $2F_o - F_c$ map (1.0 σ level) for a single asymmetric unit of the DNA 22-mer BMSG-SH4 complex (ASU = one intramolecular G-quadruplex plus one ligand molecule). All regions of the DNA are modelled into high-quality electron density. All 22 residues of the DNA sequence are present in the final model. b) View onto the 3' G-quartet showing a $2F_o - F_c$ map (0.9 σ level) for the ND compound only. While the electron density for BMSG-SH4 is of reasonable quality, there is 'excess' density in the ND core region (indicated), suggesting this compound to occupy several conformations in the crystal lattice. Additionally, the quality of electron density for the sidechains, while reasonable, is of slightly poorer quality than those of the BMSG-SH3 ligand (figure 5.8).
5.4.4 NAPHTHALENE DIIMIDE-G-QUADRUPLEX BINDING MODE – GENERAL

The ND-quadruplex complex structures determined in this work share several features of ligand binding mode, including:

1) the binding of a single ligand molecule to the external 3' G-quartet, resulting in a 1:1 quadruplex:ligand stoichiometry (figure 5.10);
2) extensive interactions between the methyl-piperazine sidechains and the quadruplex groove regions;
3) ligand-stabilisation from symmetry-related TTA loop residue stacking interactions.

Despite the similarity in overall ND-quadruplex binding between these two structures, there are in fact several differences in the binding of BMSG-SH3 and BMSG-SH4 to the human telomeric DNA G-quadruplex. Each ND-complex will be considered in turn therefore, followed by a discussion concerning the implications and consequences of the differential binding of these two closely related ND analogues.

![Figure 5.10 General mode of DNA 22-mer – naphthalene diimide interaction.](image)
Cartoon (a) and equivalent schematic (b) representations showing the mode of intramolecular G-quadruplex-ND binding seen in both complex crystal structures reported here – two intramolecular G-quadruplexes associate in the crystal lattice stacked at the 5'-5' interface, with a naphthalene diimide ligand bound to the two exposed 3' G-quartet surfaces.
5.4.5 TELOMERIC DNA G-QUADRUPLEX-BMSG-SH3 COMPLEX

5.4.5.1 G-QUARTET STACKING AND SYMMETRY STABILISATION OF BMSG-SH3

The ND compound BMSG-SH3 interacts with the telomeric DNA G-quadruplex through significant π-π stacking interactions – both with the 3’ G-quartet (as mentioned) as well as with symmetry related base-stacking interactions. These symmetry-related stacking interactions arise mainly from TTA propeller loop residues – particularly thymine-6 (figure 5.11). The π-stacking surfaces generated by the 3’ G-quartet and symmetry-related loop residues serve to effectively ‘sandwich’ the ND compound securely over the G-quartet.

![Figure 5.11 Symmetry-related base stacking interactions: BMSG-SH3.](image)

BMSG-SH3, however, is not positioned centrally over the G-quartet as one might expect (considering the equal length of the four sidechains), but is instead positioned off-centre, in the direction of the T5, T6, A7 loop (figure 5.12a). This appears to be a consequence of a combination of direct symmetry-hydrogen bonding, and strong electrostatic interactions between the sidechains of BMSG-SH3 and the quadruplex phosphates (described in the following section) (figure 5.12). The ND core and sidechains of BMSG-SH3 are stabilized significantly by three hydrogen bonds with two symmetry related loop residues – thymine-12 and adenine-13 (figure 5.12b and c). The
N3 and O4 atoms of the thymine-12 base hydrogen bond with an imide oxygen of the ND core and a nitrogen atom of the piperazine moiety, respectively. The piperazine group of the neighbouring sidechain is stabilized by an electrostatic interaction between its positively charged nitrogen and the phosphate group of adenine-13. In this way therefore, two of the sidechains (as well as the core) are stabilized by symmetry-related contacts.

a) **View onto 3’ G-quartet: BMSG-SH3 complex**

![Sidechain 1](image1)

![Sidechain 2](image2)

![Sidechain 3](image3)

![Sidechain 4](image4)

**Figure 5.12** Hydrogen bonding contacts between BMSG-SH3 and the DNA G-quadruplex.  
a) View onto the 3’ G-quartet of the BMSG-SH3-DNA complex, showing the localisation of the ligand sidechains to the quadruplex groove regions. Hydrogen bonds formed between the ligand (purple carbons) and the DNA (green carbons) are shown as black dashes, with bond distances labelled. The distances in red for sidechain 4 represent electrostatic contacts (discussed below). ND sidechains have been given arbitrary numbers for ease of reference and description. b and c) Detailed views of sidechains 1 and 2 respectively. Green carbons = DNA, purple carbons = BMSG-SH3, yellow carbons = symmetry-related atoms/residues, red spheres = water molecules.
Although sidechains 1 and 2 are interacting strongly with symmetry related components, (raising the issue of whether these interactions are likely to occur in a biologically relevant scenario), sidechains 3 and 4 of BMSG-SH3 are positioned deep within the quadruplex grooves, and interact strongly with a number of quadruplex substituents (5.12). This is in full agreement with the predicted (and hoped) mode of quadruplex interaction for this ND compound – i.e. to stack on to a G-quartet surface, with the sidechains interacting with the quadruplex groove regions. In this way therefore, this structure represents the first example of a rationally designed G-quadruplex binding compound which significantly exploits quadruplex-groove regions in its mode of binding. The possible reasons for such ligand-groove binding will be considered below, but it is likely to be a result of ligand sidechain length and (perhaps more importantly) the size and charge of the sidechain end groups (i.e. the methylpiperazine moieties).

Of the four sidechains, sidechain 3 appears to have the strongest interaction with the DNA (based on distance), as the methyl-piperazine end-group is involved in a strong electrostatic interaction with the phosphate group of thymine-11 (figure 5.12a and 5.13a). Sidechain 4 binds to the DNA in a slightly different manner, as the methyl-piperazine end group here is situated equidistant between the phosphates of thymine-5 and guanine-9 (figure 5.13b). This groove region is highly electronegative, suggesting sidechain 4 to be held in this position by electrostatic contacts, despite the relatively large distances between the charged groups. Sidechain 4 is additionally bonded to the DNA through an intricate hydrogen bonded water network (figure 5.13b).
Figure 5.13 BMSG-SH3-phosphate electrostatic interactions.
(a) A close contact can be seen between sidechain 3 (purple carbons) and the phosphate group of T11. Based on the bond distance and opposite charges of the bonding substituents, this is likely to be a strong electrostatic contact. Panel (b) shows the positioning of sidechain 4 (purple carbons) within the quadruplex groove region, sitting equidistant between two phosphates (of T5 and G9). While these distances are too large to be hydrogen bonds, based on the electronegativity of this region, it is likely that the positively charged N-methyl-piperazine is interacting electrostatically with the DNA.

5.4.6 TELOMERIC DNA G-QUADRUPLEX-BMSG-SH4 COMPLEX

As stated above, although the two ND compounds studied in this work have the same overall mode of quadruplex-binding, there are subtle (as well as more pronounced) differences between BMSG-SH3 and BMSG-SH4 quadruplex binding. These two compounds differ only in the length of sidechains, having three and four carbons linking the methyl-piperazine end groups to the ND core, for BMSG-SH3 and BMSG-SH4 respectively. The length of sidechain would be expected to influence the type and number of hydrogen bonding and electrostatic contacts available to the ligand — which is indeed the case — however, the most significant difference between these to ND compounds is the apparent rotational flexibility of the BMSG-SH4 binding site/binding mode. BMSG-SH4 is bound to the 3' G-quartet of the G-quadruplex, however, based on the electron density for this region, it appears more than one conformation of the compound can be accommodated at this binding site (figures 5.9b and 5.14). Rotation of BMSG-SH4 by 90° accounts for the excess electron density which can be seen around the core region. R and R_free calculations based on a model involving two ligand conformations (at 90° relative to each other) at half occupancy, reveals little change in the overall agreement between the calculated and observed structure factors. From
the maps however, it is clear that there is one ligand conformation with a higher occupancy (in the crystal lattice) – this conformation has therefore been assigned to the ligand in the final model (figure 5.14a).

5.4.6.1 G-QUARTET STACKING AND SYMMETRY STABILISATION OF BMSG-SH4 BINDING

Although the rotational orientation of BMSG-SH4 cannot be unambiguously determined, it is clear that the ligand is still positioned off-centre relative to the G-quartet, albeit to a lesser degree than BMSG-SH3 (figure 5.14a). As with the BMSG-SH3-quadruplex complex, BMSG-SH4 interacts with the G-quadruplex through extensive π-stacking contacts – both with the 3′ G-quartet, as well as with symmetry related TTA loop residues (figure 5.15). However, there are no direct hydrogen bonding contacts between symmetry related residues and BMSG-SH4, which is likely to be a significant contributing factor to the rotational flexibility of this ND within its binding site.
a) View onto 3' G-quartet: BMSG-SH4 complex

![Diagram showing BMSG-SH4 binding to DNA G-quadruplex](image)

**Figure 5.14** Hydrogen bonding contacts between BMSG-SH4 and the DNA G-quadruplex.

a) View onto the 3' G-quartet, showing BMSG-SH4-DNA hydrogen bonding contacts (black dashes). (Bond distances are given in figure 5.16 below.) All four of BMSG-SH4's sidechains are localised to the quadruplex groove regions. (DNA = green carbons, BMSG-SH4 = purple carbons.) b) There appears to be a level of rotational conformational flexibility of BMSG-SH4 in its binding site, with electron density (figure 5.9b) providing evidence of two conformations existing in the crystal lattice. These conformations are shown in (b), with two ligand molecules (yellow and purple carbons) superimposed, separated by a 90° rotation. (DNA: grey carbons.)
Figure 5.15 Symmetry-related base stacking interactions: BMSG-SH4.
In a similar (but not identical) manner to BMSH-SH3, BMSG-SH4 (green carbons) is sandwiched between π-π stacking surfaces generated from the 3' G-quartet plus loop residues from symmetry related molecules (T12 and A1) (DNA carbons coloured yellow). Electron density from a 2Fo-Fc map at 0.9 σ level is shown.

5.4.6.2 BMSG-SH4-GROOVE INTERACTIONS WITHIN THE ASYMMETRIC UNIT

In contrast to BMSG-SH3, all four of the sidechains of BMSG-SH4 are positioned within the quadruplex groove regions and interact directly with the DNA within the same asymmetric unit. Based on the quality of the electron density however, the sidechain contacts appear to be less secure or rigid than those formed between the sidechains of BMSG-SH3 and the DNA. Sidechains 2 and 4 both interact with guanine base nitrogens of residues from the 3' G-quartet binding surface, utilizing the charged nitrogens of their methyl-piperazine end groups (figure 5.16). Sidechain 1 binds to the quadruplex in a similar manner (i.e. by forming a hydrogen bond between its piperazine nitrogen and a guanine amine nitrogen), however, this interaction is mediated by a water bridge. Sidechain 3 is the only sidechain which appears to form a strong electrostatic contact with the quadruplex – a close contact is seen between the charged nitrogen of sidechain 3’s methyl-piperazine group and the phosphate group of thymine-11 (figure 5.16). Interestingly, this mirrors the strongest interaction between a sidechain of BMSG-SH3 and the quadruplex (see figure 5.13a, sidechain 3).

Although all four sidechains of BMSG-SH4 are interacting with the quadruplex groove regions, it is clear from the moderate quality of the electron density of these
regions (as well as the B-factor distribution – see below), that the sidechains are not rigidly held in these positions, and that many other sidechain-groove contacts are available. This is not too surprising, considering the high degree of intrinsic flexibility of these sidechains – being composed of linker-lengths of four carbons. This apparent flexibility of the sidechains is likely to be the reason for the rotational flexibility of BMSG-SH4 within its binding site, as these flexible sidechains would be expected to allow ligand-stabilising contacts to be made with the quadruplex in any of the four rotational conformations of the ligand.

![Sidechain Diagrams](image)

**Figure 5.16** Interactions between BMSG-SH4 sidechains and quadruplex groove components. Hydrogen bonds formed between BMSG-SH4’s sidechains (purple carbons) and the DNA groove regions (green carbons) are shown as black dashes. Water bridges are also shown (waters shown as red spheres). The only electrostatic contact formed between BMSG-SH4 and the DNA involves the positively charged nitrogen of sidechain 3 and the phosphate group of thymine-11 (an equivalent interaction is seen within the BMSG-SH3 complex (figure 5.13a)). Surprisingly, all sidechain-DNA interactions involve an exocyclic N2 nitrogen of a guanine base (labelled).
5.4.7 COMPARISON OF BMSG-SH3 AND BMSG-SH4 BINDING MODES

It is quite clear that BMSG-SH3 and BMSG-SH4 interact with this telomeric DNA G-quadruplex in a similar manner, the key similarities being: 1) 3' G-quartet-ligand binding (in a 1:1 stoichiometry); 2) extensive interactions between ligand sidechains and the quadruplex groove regions and 3) significant stabilisation of the ligand from symmetry-related stacking interactions. There are however, some important differences in quadruplex-binding between these two ND compounds, namely:

1) BMSG-SH3 adopts a more constrained position above the 3' G-quartet, whereas BMSG-SH4 appears to possess a degree of rotational (and more general) flexibility.

2) BMSG-SH3 interacts directly (through hydrogen bonding and electrostatic contacts) with symmetry related components, whereas the sidechains of BMSG-SH4 interact exclusively within a single asymmetric unit (i.e. to a single G-quadruplex).

Naturally, any differences observed between the binding mode of these two compounds must be related to the only significant difference between them – i.e. the length of sidechains. As alluded to above, the extended length of the sidechains of BMSG-SH4 confers greater flexibility upon them – this is indeed shown by comparison of the B-factor distribution between the two complexes (figure 5.17), which reveals the B-factors of BMSG-SH4 to be noticeably higher than those of BMSG-SH3, particularly towards the terminal piperazine end-groups. This sidechain flexibility has two interrelated consequences – more contacts can be made between the sidechains and the quadruplex, but fewer specific contacts. This allows the ligand to sit on the 3' G-quartet in a number of orientations (probably two main orientations, out of a possible four in total, using a 90° rotation sampling).

Concerning the issue of differential crystal-packing contacts, this again is likely to be a result of the different sidechain lengths. It is probable that one of the main reasons all four of the sidechains of BMSG-SH4 are able to interact with the grooves
(rather than with symmetry related residues) is that they are simply long enough to do so. The sidechains of BMSG-SH3 are not of sufficient length to allow all four of the quadruplex grooves to be occupied, and hence, several of these sidechains interact with closer symmetry-related atoms.

\[
\begin{align*}
\text{a) B-factors: BMSG-SH3} & \quad \text{b) B-factors: BMSG-SH4} \\
\end{align*}
\]

![Diagram showing B-factors for BMSG-SH3 and BMSG-SH4](image)

**Figure 5.17** Ligand B-factor distribution.

Isotropic B-factor (temperature factor) distribution for BMSG-SH3 (a) and BMSG-SH4 (b). High B-factors indicate high atomic mobility in the crystal lattice. The mobility of the terminal regions of BMSG-SH4 is significantly higher than the equivalent regions of BMSG-SH3.

### 5.4.8 Diffraction Data for the BMSG-SH5-Quadruplex Complex

Although only limited diffraction data was obtainable for crystals formed from the DNA sequence d(AGGG[TTAGGG]₃) in complex with compound BMSG-SH5 (with sidechain lengths of n = 5 carbons), the determination of the space group and unit cell dimensions can yield a certain degree of information. The BMSG-SH5 complex belongs to the same space group (i.e. internal symmetry and packing arrangement) as the BMSG-SH4 complex, and has highly similar unit cell dimensions. This suggests, or at least supports, a mode of ligand binding for BMSG-SH5 as being very similar to BMSG-SH4 — i.e. stacking onto the 3' G-quartet. It is quite likely that the poor quality of diffraction data for the BMSG-SH5 complex is a result of the poor quality of the
crystals, and that optimal conditions for the growth of high-quality crystals have not been found. However, considering the relationship between sidechain length and ND conformational flexibility, it is possible that, with five carbons, the sidechains of BMSG-SH5 are not able to make enough suitable strong contacts with the DNA, thereby reducing crystal packing interactions and crystal quality. This is of course speculative, but if BMSG-SH5 was able to bind to the quadruplex in a tight and rigid manner, then crystal and diffraction quality might be expected to be far better than observed (e.g. comparable to the other ND complexes).
5.5 **GENERAL DISCUSSION AND CONCLUSIONS**

Concerning the development of ND compounds as telomeric G-quadruplex-stabilising agents, the structures reported here highlight two important characteristics of ND-quadruplex binding which have not been seen in previously determined ND-quadruplex crystal structures, these being:

1) The active involvement of the ND sidechains in groove binding, and;

2) A single and therefore potentially ‘specific’ binding site within the quadruplex.

These are two key features of ND-quadruplex binding which, although intended, have not been observed in previous crystal structures\(^{97,114}\), all four of which involve loop-only ND-binding platforms (in addition to ND-G-quartet binding) as well as an absence of sidechain-groove interactions. All ND compounds used in these previous crystallographic studies consisted of tetra-substituted ND cores with 3-carbon sidechain lengths (figures 5.2 and 5.3) – differing from the BMSG-SH3/4 compounds described here only in sidechain terminal functionality. The (N)methyl-piperazine functional groups of the BMSG-SH3 and BMSG-SH4 compounds are therefore likely to be key factors in conferring binding site specificity and affinity. The use of these functional groups in future ND-based quadruplex binding compounds might therefore increase the chances of producing active G-quadruplex binding compounds.

Concerning the future development of this series of methyl-piperazine-substituted NDs specifically, the structures reported here provide some useful information. Clearly, sidechain length is important, as this affects sidechain flexibility, consequently affecting the number and type of interactions available to the terminal functional groups (i.e. the positively charge methyl-piperazine groups). This in turn can affect the rigidity of the ND over the G-quartet binding site, allowing, arguably, multiple conformations of the ND core to be adopted. Sidechain length, at least in this case, seems to have an upper limit regarding optimal length, as the progression from \( n = 3 \) to \( n = 4 \) (i.e. comparing the BMSG-SH3 and BMSG-SH4 structures) leads to conformational flexibility, with a progression from \( n = 4 \) to \( n = 5 \) resulting in poor...
quality diffraction data, which is potentially a result of poor crystal packing arising from unfavourable ligand-quadruplex binding. Linker length, therefore, needs to be considered carefully when designing such G-quadruplex binding compounds, however, it would appear that a linker length of three carbons may be optimal.

G-quadruplex groove regions – being rich in potential electrostatic and hydrogen bonding contacts – have been extensively targeted through the rational modification of ligand sidechains in quadruplex drug-discovery programs. However, based on the crystal structures determined here, it could be suggested that the targeting of G-quadruplex groove regions may not in fact be such an effective method for achieving favourable G-quadruplex binding (and inter-quadruplex selectivity). There are two main considerations in support of this proposition:

1) G-quadruplex grooves do not contain any single functional group which may be rationally selected for in a ligand-interaction. In other words, all quadruplex grooves contain multiple polar groups and multiple phosphate groups – considering the flexibility of the loop regions (which affects groove dimensions) and the flexibility of the ligand sidechains (in this case and many other documented cases), it would be difficult to envisage ligand-induced stabilisation of the quadruplex in these conditions.

2) The strongest sidechain-quadruplex interaction seen in both of these structures (based on distance and quality of the electron density) is formed between the positively charged methyl-piperazine nitrogen of sidechain 3 and the phosphate group of thymine-11. This phosphate is not strictly located within a quadruplex groove region (sitting almost within the same plane as the 3’ G-quartet), and, considering the apparent strength of the interaction, the phosphate groups of the equivalent residues (i.e. thymine-5 and thymine-17) could be excellent candidates for intentionally targeting with positively charged sidechain groups. If this is to be achieved, it may be prudent to reduce flexibility in the ligand region which links the functional end-group (i.e. methyl-piperazine) to the ND core, as this would reduce unspecific contacts and could potentially force strong electrostatic contacts to be made between the ligand end-groups and
these highlighted phosphate groups. The use of non-equivalent sidechain lengths may be the best method to achieve such interactions.

Concerning the significance of these structures more generally to G-quadruplex drug design, it is perhaps worth noting that, of the limited number of high-resolution structures available involving human telomeric G-quadruplex-ligand complexes, only two have been solved involving intramolecular G-quadruplex structures\(^{97,114}\). The significance of studying an intramolecular (i.e. unimolecular) quadruplex over an intermolecular (e.g. bimolecular) quadruplex may seem trivial, however, when designing drugs and assessing drug-binding, it is important to study as biologically-relevant a structure as possible – currently the most biologically relevant telomeric G-quadruplex model is indeed the intramolecular G-quadruplex (such as that studied here), as this arguably best represents the repeat quadruplex motif expected to exist at the telomere. It may also be worth noting that the complex structures described here are in fact the first G-quadruplex-ligand complexes involving the human 22-mer G-quadruplex structure – a sequence which has been studied extensively in biophysical work\(^{215}\) and which was used to solve the first native NMR\(^{39}\) and crystal structures\(^{40}\) of human telomeric DNA G-quadruplex formation. For this reason, the complex structures reported here involve particularly relevant G-quadruplexes, which can be correlated to additional findings reported by others.
CHAPTER 6

BIOPHYSICAL STUDIES OF G-QUADRUPLEXES FORMED FROM HUMAN TELOMERIC RNA AND DNA
6.1 BACKGROUND

While X-ray crystallography excels at providing atomic-scale details of macromolecular folding - including details of intra- and intermolecular bonding, water structure and ligand-binding interactions - the addition of alternative structural data provided by biophysical techniques can often complement crystallographic data, and provide a more complete view of a molecule's folding and behaviour. For example, although possible, it is technically difficult to study thermodynamic stability or kinetics using crystallographic methods. Additionally, due to the nature of macromolecular crystallisation, it is often difficult to study the effects of changes in ionic conditions in a predictable manner, meaning that the influence of cation-type on G-quadruplex folding cannot be easily investigated.

In this chapter, results from biophysical studies on a range of G-quadruplexes formed from human telomeric RNA and DNA will be reported. The primary techniques used in these studies were electrospray ionisation mass spectrometry (ESI-MS), circular dichroism (CD) and polyacrylamide gel electrophoresis (PAGE). The main impetus for these studies was to provide data on two aspects of telomeric G-quadruplexes:

1) RNA and DNA G-quadruplex tertiary structure (and stability), and;

2) RNA and DNA G-quadruplex-ligand interactions.

6.1.1 TERTIARY STRUCTURE OF HUMAN TELOMERIC G-QUADRUPLEXES

The study of telomeric G-quadruplex tertiary structure - or 'higher-order' structure - refers to the mechanism by which multiple G-quadruplex subunits can self-associate through π-π interactions to form assemblies of stacked G-quadruplexes. The association of multiple subunits is considered an important aspect of telomeric G-quadruplex research, as both telomeric RNA and DNA exist as extended single-stranded polymers composed of repeated G-quadruplex forming sequences (i.e \( r[UUAGGG]_x \) and \( d[TTAGGG]_y \), where \( x = 10-1000 \) and \( y = 20-35 \)), meaning that individual telomeric RNA and DNA molecules both have the potential to accommodate
multiple discrete G-quadruplex structures. The manner by which these G-quadruplex units can self-associate (if at all) has implications concerning:

1) **Protein (and chromatin) binding** – a number of different proteins have been identified which bind specifically to telomeric DNA and RNA sequences (e.g. shelterin or TRF2, for telomeric DNA and RNA, respectively\(^{59,147}\)). Although challenging, it is important to understand whether such proteins are binding to unfolded or folded (i.e. G-quadruplex) telomeric nucleic acid, as this could have important implications concerning telomere structure and function. This is also applicable to understanding the interaction between TERRA and chromatin\(^{144}\).

2) **Ligand binding** – the ligand binding site generated from single G-quadruplexes (e.g. exposed external G-quartets) is fundamentally quite different to the binding site generated from the association of multiple G-quadruplexes (e.g. intercalation sites between the G-quartets of a quadruplex-quadruplex stacking interface). The importance of G-quadruplex multimerisation is arguably an underappreciated aspect to telomeric DNA G-quadruplex drug design.

### 6.1.1.1 MODELS OF TELOMERIC DNA G-QUADRUPLEX TERTIARY STRUCTURE

Surprisingly, a fairly limited number of experimental studies have been conducted on human telomeric G-quadruplex forming sequences composed of more than a single G-quadruplex unit\(^{216-218}\) – the typical human telomeric DNA G-quadruplex forming sequence studied is the four-repeat sequence, d(AGGG[TTAGGG])\(_4\), which represents a single complete G-quadruplex structure. This is probably due to the difficulty in obtaining the required length of oligonucleotide samples – for example, in order to study two contiguous G-quadruplexes formed within the same sequence, a 45-mer oligomer is required, d(GGG[TTAGGG])\(_7\)). Despite the lack of experimental data, many theoretical models for higher-order DNA G-quadruplex assembly have been proposed, differing mainly in the topology of the repeated G-quadruplex considered. Figure 6.1,
adapted from Renciuk et al., 2009, very nicely summarizes the three main models for G-quadruplex self-assembly at the telomere:


2) Limited end-to-end stacking of the ‘3+1’ mixed G-quadruplex – inter-G-quadruplex π-π stacking is possible, although the edgewise loops would be expected to interfere with close quadruplex-quadruplex stacking.

3) Complete end-to-end 5′-3′ stacking of the parallel G-quadruplex – the propeller loops, being external to the G-quartet core, allows a compact arrangement of multiple parallel G-quadruplexes. This is arguably the most compact and stable model.

Figure 6.1 Models of telomeric DNA G-quadruplex tertiary structure. Summaries of the main models of telomeric DNA G-quadruplex higher-order (or tertiary) structure (adapted from Renciuk et al., 2009). The models here are based upon the topology of the repeated G-quadruplex monomer: 1) all anti-parallel; 2) mixed; 3) all parallel. Further tertiary arrangements are generated when non-equivalent monomer topologies are considered, e.g. alternating parallel and ‘3+1’ folded G-quadruplexes.
6.1.1.2 MODELS OF TELOMERIC RNA G-QUADRUPLEX TERTIARY STRUCTURE

Considering the relatively recent discovery of TERRA sequences, several studies into longer G-quadruplex forming sequences have already been performed on TERRA transcripts. These have been discussed in chapter 1, section 1.5.7, but briefly, it appears that TERRA G-quadruplexes are certainly amenable to higher-order structure formation\(^{159,163,164}\). An interesting recent study into TERRA G-quadruplex higher-order structure has been performed by Martadinata and co-workers\(^{166}\), who used T1 RNase digestion of long (96-mer) TERRA sequences and modelling studies to propose a model of TERRA G-quadruplex subunit assembly in long transcripts, which involves a ‘beads-on-a-string’ model, with the beads comprising alternate single- and double (stacked) unimolecular G-quadruplexes (figure 6.2). Interestingly, this model allows the interaction of G-quadruplex subunits via quartet-quartet stacking (as in model 3 for the telomeric DNA G-quadruplex assembly, figure 6.1) as well as ‘side-by-side’ associations, via UUA-loop interactions. The figure below, adapted from the Martadinata et al., 2011\(^{166}\), very clearly illustrates their proposed model, and describes a higher order arrangement for TERRA G-quadruplexes quite different from those proposed for telomeric DNA G-quadruplexes.

![Figure 6.2 Proposed model of TERRA G-quadruplex higher-order structure.](image)

Model proposed by Martadinata and co-workers for higher-order/tertiary structure within long RNA sequences. Each G-quadruplex subunit here represents a four-repeat sequence (e.g. r[UUAGGG]\(_4\)). The quadruplex subunits are shown stacked 5′-5′, with additional interactions between the UUA propeller-loop residues. (Figure adapted from Martadinata et al., 2011\(^{166}\).)
6.1.2 HUMAN TELOMERIC G-QUADRUPLEX-LIGAND INTERACTIONS

The need to study the interactions between small molecule ligands and telomeric G-quadruplexes has been emphasised many times in the preceding chapters. While telomeric DNA G-quadruplex targeted ligands have the potential to be effective anticancer targets, the study of ligand interactions with the equivalent RNA G-quadruplexes has utility by providing details off-target drug binding. Additionally, TERRA G-quadruplexes may yet be validated as drug targets, considering the role of TERRA in telomerase inhibition and telomere biology\(^{143,144,148}\). For these reasons, it is of use to compare the binding of specific ligands to both DNA and RNA G-quadruplexes, in order to understand the structural mechanisms and features of telomeric G-quadruplex-drug binding. Ideally, a compound with the ability to selectively bind either telomeric DNA or telomeric RNA G-quadruplexes would be highly useful as a tool for elucidating the role of telomeric RNA G-quadruplexes in a cellular or \textit{in vitro} environment.

6.1.3 PRACTICAL CONSIDERATIONS

The techniques used in this work to study telomeric G-quadruplexes were chosen to provide complementary data to the crystallographic results. As with crystallography, the biophysical techniques used here, while informative, are accompanied with limitations and drawbacks, which will be considered briefly below.

6.1.3.1 ELECTROSPRAY IONISATION MASS SPECTROMETRY – UTILITY AND VALIDATION

A mass spectrometer, fitted with an electrospray source and used in the negative ion mode, allows the exact mass (to \(< 1\) Da) of nucleic acid samples to be determined. Importantly, the relatively ‘soft’ ionisation conditions of the ESI-MS system (compared to, for example, MALDI-ToF) allows non-covalent interactions to be preserved and studied. This method, in nucleic acid and G-quadruplex research, has been used successfully to study (relative) structural stability, kinetics, strand/molecule
stoichiometry and quantitative ligand binding interactions (in the form of dissociation constants)²¹⁹-²²⁷.

Sample analysis in an ESI-mass spectrometer is performed in the gas phase, and as such, it is important to consider whether the gas-phase environment of the ESI-MS system is comparable to the solution-phase behaviour of the target molecules. If the target molecule under analysis is an obligate multimer — such as duplex DNA — then the detection of peaks corresponding to twice the molecular weight of the single strand are assumed to be structured duplex. This assumption is applied to G-quadruplex research also, which often involves analysis of tetramolecular and bimolecular structures. Importantly, studies into the validity of G-quadruplexes analysed by ESI-MS have been performed — for example, using ESI-MS, CD, ion mobility and modelling. Baker and co-workers showed that the structures of several quadruplex forming sequences were retained in the solvent free environment of the ESI-MS system²²⁸. The DNA sequences used for this study were the same or similar to the sequences used in the work reported here.

Concerning unimolecular G-quadruplexes — which cannot be detected in mass spectra based upon stoichiometry — an alternative method is used to detect and follow G-quadruplex formation, based upon the retention of specific coordinate bonded cations. Monovalent cations are known to coordinate bond to the O₆ oxygens of the G-quartets, resulting in stabilisation of the G-quadruplex structure (for example, coordinate bonded potassium cations can be visualised directly in crystal structures, see chapters 2-5). Although potassium ions are not compatible with ESI-MS instruments due to sensitivity, ammonium ions, which have a comparable ionic radius, are perfectly compatible (ion radii: $K^+ = 1.52 \text{ Å}$; $\text{NH}_4^+ = 1.43 \text{ Å}^{16}$). Consequently, in order to confirm the presence of a unimolecular quadruplex structure from its ESI-MS spectra, we would expect to see a peak corresponding to the mass of the quadruplex plus n-1 $\text{NH}_4^+$ adducts (where n = the number of G-quartets) (see figure 6.3).
Figure 6.3 Identification of G-quadruplex formation in mass spectra: example. Spectra of the G-quadruplex forming sequence r(AGGG[UUAGGG]₃) at the -5 charge state. At low voltage (150 V, bottom spectra) the main species has mass of 1452.69, corresponding to the mass of the monomer (7234.1 Da) plus two ammonium cations (+2 x 17 Da), divided by the charge (-5): (7234.1 + 34) / -5 = 1453.62 Da. As the voltage is increased to 180 V (top spectra), the structure is destabilised, resulting in a peak corresponding to the mass of the RNA molecule alone (1445.87), corresponding to un-coordinated (and therefore unstructured) RNA species.

6.1.3.2 CIRCULAR DICHROISM – UTILITY AND LIMITATIONS

Circular dichroism (CD) is a spectrophotometric technique involving the measurement of absorption of left and right circularly polarised light when directed towards a sample solution. This technique is used routinely to detect protein secondary structure, and has been used to an equal degree in G-quadruplex research. CD can be used to detect the presence of G-quadruplex structure, as well as to study the effect of ions, the effect of temperature and to study G-quadruplex-ligand interactions. Parallel and anti-parallel G-quadruplex topologies each generate distinctive CD spectra, with parallel topologies producing a peak at 260 nm and trough at 240 nm, and anti-parallel topologies producing a peak at 290 nm. In this way, G-quadruplex formation can be detected in CD spectra, along with limited information concerning topology, which is determined empirically based on the spectra of known and validated G-quadruplexes. The differential spectra produced from parallel and anti-parallel topologies is in fact generated from the differential π-π overlaps created by anti-anti
or syn-anti guanine base stacking interactions – the topology is then inferred based on the understanding that a G-quadruplex with all guanine glycosidic angles in the anti range must result in a parallel topology, with any guanines with syn glycosidic angles resulting in an anti-parallel strand orientation. This raises one of the main caveats of G-quadruplex CD analysis, in that detection of an ‘anti-parallel’ spectra does not provide information concerning loop arrangements, and, as pointed out by Webba da Silva, a unimolecular G-quadruplex forming sequence with four G-tracts and three linking loops can theoretically adopt 26 different topologies, one of which is parallel, with the remaining 25 arrangements being anti-parallel\textsuperscript{36,37}. Clearly, therefore, care must taken when interpreting CD spectra of G-quadruplex samples.

6.1.3.3 POLYACRYLAMIDE GEL ELECTROPHORESIS – UTILITY AND LIMITATIONS

Polyacrylamide gel electrophoresis (PAGE) allows the separation of biomolecules based on size alone (denaturing PAGE) or based on size and three-dimensional shape (native PAGE). Native PAGE experiments allow the rapid identification of secondary or tertiary structure in a sample solution, using relatively small amounts of material (between 50 ng and 1 μg). The addition of salts to native PAGE gels and run buffers allows the effects of ionic conditions to be studied fairly easily. The main limitations of native PAGE experiments are: 1) the extremely low resolution of structural data obtained and 2) as the precise molecular weight of each species cannot be unambiguously determined, care must be taken when correlating PAGE bands with structural species. Native PAGE allows G-quadruplex formation within given sequences to be rapidly identified, as folded unimolecular G-quadruplexes have higher mobility than unfolded sequences of the same molecular weight, due to the compact, globular shape of folded G-quadruplex structures\textsuperscript{164}.
6.2 AIMS

The aim of the work reported here was to generate additional structural data for human telomeric G-quadruplexes – including ligand-binding data – using a range of biophysical techniques, including ESI-MS, circular dichroism and gel electrophoresis. The main questions set out to be answered by ESI-MS were:

1) Can ESI-MS be used to detect differences in stability between telomeric RNA and DNA G-quadruplexes?
2) Can higher-order assemblies (i.e. tertiary structure) of telomeric RNA and DNA G-quadruplexes be detected using ESI-MS?
3) What are the differences (if any) between telomeric RNA and DNA tertiary structure?
4) Can current telomeric DNA G-quadruplex binding compounds bind to the equivalent RNA structures?
5) What implications might differential ligand binding have for drug design?

Additionally, CD and PAGE experiments were performed with the aim of validating ESI-MS data. The main questions set out to be answered by CD and PAGE were:

1) Does the behaviour of telomeric RNA and DNA G-quadruplexes in the ESI-MS system match the behaviour of these molecules in physiologically relevant salt conditions?
2) What is the topology of these structures?
3) Can more extensive higher order assemblies be detected using alternative methods?
6.3 METHODS AND MATERIALS

6.3.1 ELECTROSPRAY IONISATION MASS SPECTROMETRY

All mass spectrometry and circular dichroism experiments were conducted at the Mass Spectrometry Laboratory of the University of Liege in Belgium, under the expert guidance of Drs. Valerie Gabelica and Frederic Rosu. I am sincerely grateful for their support and guidance.

6.3.1.1 SAMPLE PREPARATION

All oligonucleotide sequences were purchased from Eurogentec (Belgium) and used without further purification (see table 6.1). The annealing protocol used here for G-quadruplex formation differed slightly from crystallographic sample annealing: the oligonucleotides were heated for 5 minutes at 85 °C, followed by cooling at a rate of 1 °Cmin⁻¹ to 75 °C, followed by cooling at a rate of 0.2 °Cmin⁻¹ to 20°C. All oligonucleotide sequences were annealed in 100 mM ammonium acetate to a final single-strand concentration of 1 mM, unless stated otherwise.

6.3.1.2 MASS SPECTROMETER PARAMETERS AND DATA COLLECTION

All ESI-MS experiments were carried out on a Q-TOF Ultima Global mass spectrometer, fitted with an electrospray source. The instrument was used in the negative ion mode, with the capillary voltage set to 2.2 kV, the cone voltage set to 100 V and the collision voltage set to 10 V. The RF lens voltage was varied between 60 V and 210 V. The source temperature was set to 40 °C and the desolvation temperature was set to 60 °C. G-quadruplex samples were injected into the mass spectrometer at final concentrations of 20 μM (for the 12-mers), 10 μM (for the 22-mers), and 5 μM (for the 45-mers), at a rate of 4.0 μLmin⁻¹. In order to improve the electrospray signal, methanol was added to all samples prior to injection, to a final concentration of 20%.

For the ligand binding studies, solutions of the nucleic acid-ligand complexes were prepared at ligand:quadruplex ratios of 2.5:5, 5:5 and 10:5 (μM) (including 20%
methanol). Equilibrium dissociation constants \( (K_d) \) were determined from the integrated peak areas following the procedure outlined by Rosu et al., 2008\(^{226}\). Data were processed using the MassLynx 4.0 and Sigma-Plot software packages. Details of the ligands studied are shown in Table 6.2.

**Table 6.1** Oligonucleotide sequences used for ESI-MS, CD and PAGE analysis.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Molecular mass / Da</th>
<th>Annealing [ ] / mM</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>d(TTAGGTTAGGG)</td>
<td>3756.5</td>
<td>1.0</td>
<td>DNA12</td>
</tr>
<tr>
<td>d(AGGG[TTAGGG]₃)</td>
<td>6966.6</td>
<td>1.0</td>
<td>DNA22</td>
</tr>
<tr>
<td>d(GGG[TTAGGG]₇)</td>
<td>14290.3</td>
<td>0.3</td>
<td>DNA45</td>
</tr>
<tr>
<td>r(UUAGGGUUAGGG)</td>
<td>3892.4</td>
<td>1.0</td>
<td>RNA12</td>
</tr>
<tr>
<td>r(AGGG[UUAGGG]₃)</td>
<td>7234.4</td>
<td>1.0</td>
<td>RNA22</td>
</tr>
<tr>
<td>r(GGG[UUAGGG]₇)</td>
<td>14813.9</td>
<td>0.08</td>
<td>RNA45</td>
</tr>
</tbody>
</table>

**Table 6.2** Details of small-molecule ligands used for ESI-MS analysis.

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Molecular mass / Da</th>
<th>Solvent</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRACO-19</td>
<td>594</td>
<td>DMSO</td>
</tr>
<tr>
<td>FC4ND-01</td>
<td>638.66</td>
<td>H₂O (distilled)</td>
</tr>
<tr>
<td>FC4ND-06</td>
<td>684.89</td>
<td>H₂O (distilled)</td>
</tr>
<tr>
<td>FC4ND-10</td>
<td>581.4</td>
<td>DMSO/30 mM HCl</td>
</tr>
</tbody>
</table>

**6.3.2 CIRCULAR DICHROISM**

CD experiments were performed on a Jasco J-810 spectropolarimeter. CD spectra of annealed nucleic acid samples were recorded at room temperature in 0.4-mm path quartz cuvettes. Annealed samples were prepared as described above (for the ESI-MS experiments). Spectra of the annealed sequences were recorded over a range of 200 to 400 nm, using a scan rate of 100 nm/min, with an average of 10 scans recorded per spectra.

**6.3.3 POLYACRYLAMIDE GEL ELECTROPHORESIS**

Native polyacrylamide gel shift studies were performed on the human telomeric RNA sequence, r[AGGG(UUAGGG)₃] (RNA22) and the equivalent DNA sequence, d[AGGG(TTAGGG)₃] (DNA22). Home-made 20% polyacrylamide gels were run at room temperature in standard tris-borate-EDTA run buffer, which was supplemented with either 50 mM NaCl, 50 mM KCl or 50 mM NH₄OAc. Samples were studied both with
and without annealing steps. Additionally, some samples were pre-heated prior to loading (to remove pre-formed secondary structure and investigate G-quadruplex stability) – this is indicated in the text and figure legends where appropriate. Samples were annealed as described above, at either low (50 μM) or high (1 mM) strand concentration, with 100 mM monovalent salt (either NaCl, KCl or NH₄OAc). Individual sample preparation conditions are indicated in figure legends. Typically, 1 μg of the quadruplex sample was loaded per lane. Note: ethidium bromide appears to have a much weaker interaction with folded monomer quadruplexes compared to stacked quadruplexes (dimer, trimer, tetramer), hence monomer quadruplex bands appear much weaker on acrylamide gels than the stacked structures.

6.3.4 ADDITIONAL

This work was first reported in Chemical Communications in 2009 and in the Journal of the American Chemical Society in 2010 (see Appendix A).
6.4 RESULTS AND DISCUSSION

6.4.1 ELECTROSPRAY IONISATION MASS SPECTROMETRIC STUDIES OF TELOMERIC RNA AND DNA SEQUENCES

ESI-MS experiments were performed on telomeric RNA and DNA sequences composed of two-, four- and eight- G-tracts (corresponding to the sequences DNA12/RNA12, DNA22/RNA22 and DNA45/RNA45, respectively [see table 6.1]). The expected G-quadruplex arrangements for such sequences involve: bimolecular G-quadruplex formation for the 12-mers (i.e. formed of two oligonucleotide strands); unimolecular (i.e. intramolecular) G-quadruplex formation for the 22-mers and the formation of two stacked G-quadruplexes linked through UUA/TTA linkers for the 45mers. All G-quadruplex forming sequences analysed by ESI-MS were annealed in ammonium acetate (NH₄OAc) solution. A key principle of ESI-MS analysis of G-quadruplex forming sequences is that, due to the ability of NH₄⁺ ions to coordinate to the O6 atoms of the central core of G-quadruplexes, the presence of G-quadruplex structures within sprayed samples can be detected and followed based on specifically coordinated NH₄⁺ ions. Specifically coordinated NH₄⁺ ions can be selected for by increasing the collision energies used in the mass spectrometer, which typically results in the loss of the more weakly-bound non-specific NH₄⁺ adducts (e.g. those which may be bound to the phosphates of the backbone). As well as identification of G-quadruplex structures, this method also allows comparative G-quadruplex stabilities to be studied, as G-quadruplexes which retain specific (i.e. O6 coordinated) NH₄⁺ ions at higher internal voltages than a comparator sequence can be considered to be more stable (albeit in the gas phase).

6.4.1.1 ESI-MS ANALYSIS OF TELOMERIC DNA – STRAND STOICHIOMETRY AND G-QUADRUPLEX DETECTION

ESI-MS spectra recorded for the G-quadruplex forming sequence, DNA12, showed the main species to be monomeric (i.e. single-stranded, unfolded DNA), with a limited amount of dimer species visible (figure 6.4a). This represents a low level of bimolecular
G-quadruplex formation, indicating such structures to be unstable in these conditions. The DNA22 sequence was detected mainly as a monomer, with a very low level of dimer species detectable when the spectra is magnified (figure 6.4b). Dimer formation for a unimolecular G-quadruplex forming sequence would be expected to consist of two-stacked G-quadruplexes. The DNA45 sequence was detected exclusively as a monomer (figure 6.4c).

6.4.1.2 ESI-MS ANALYSIS OF TELOMERIC RNA – STRAND STOICHIOMETRY AND G-QUADRUPLEX DETECTION

It was apparent from the first spectra obtained for a telomeric RNA G-quadruplex forming sequence that these molecules behaved very differently to the equivalent DNA sequences. The RNA12 sequence forms a dimer (i.e. a bimolecular G-quadruplex) far more readily than the DNA12 sequence, however, the most striking difference between these two sequences is the presence of tetrameric species in the RNA12 spectra (figure 6.4d). Indeed, the tetramer is the main species detected for the RNA12 sequence, and is expected to correspond to the stacking of two discrete G-quadruplex units, with each G-quadruplex unit being composed of two RNA12 strands (i.e. each G-quadruplex unit is a bimolecular G-quadruplex).

The RNA22 sequence was detected as both a monomer (i.e. intramolecular G-quadruplex) and a dimer, the predominant species being the dimer (figure 6.4e). As with for the DNA22 dimers, dimers of the RNA22 sequence are proposed to correspond to two complete intramolecular G-quadruplexes, interacting through π-π stacking interactions. The RNA22 dimers were detected even at low annealing strand concentrations (10 μM), indicating a concentration-independent effect. Similar to the DNA45 sequence, the RNA45 was also detected as a monomer in the spectra, however the ammonium ion distribution showed important differences (figure 6.4f).
Figure 6.4 ESI-mass spectra: Strand stoichiometry and G-quadruplex detection for telomeric RNA and DNA sequences.
Spectra of six different telomeric RNA and DNA sequences annealed in 100 mM ammonium acetate. Final (injected) strand concentrations are: DNA12/RNA12 = 20 pM; DNA22/RNA22 = 10 pM; DNA45/RNA45 = 5 pM. RF lens voltage was set to 100 V (a-c), 120 V (d-e) and 140 V (f). The peak annotations, \([n]^z\), indicate the number of strands (n) and the total charge (z). G-quadruplex dimer species are indicated by the green dashed box in (d) and (e). Note the 20-fold magnification of the DNA12 dimer in (a) and the 10-fold magnification of the DNA22 dimer in (b).

6.4.1.3 ANALYSIS OF RELATIVE STABILITIES OF RNA AND DNA G-QUADRUPLEXES USING ESI-MS: MONOMER G-QUADRUPLEX SPECIES

As mentioned above, by adjusting the internal energies of the mass spectrometer, G-quadruplex stabilities can be studied based on the presence and loss of specific (i.e. O6 coordinated) ammonium ions.

Peaks corresponding to the DNA12 dimer plus two coordinated ammonium ions (at the -5 charge state) were detected at 100 V, which were then lost as the internal collision energy was increased to 140 V (figure 6.5a). The bimolecular G-quadruplex formed from the RNA12 sequence was seen to be far more stable, with two ammoniums preserved at 140 V (figure 6.5b). The difference in stability was even more striking when comparing the DNA22 and RNA22 sequences, for which it was impossible to preserve two ammoniums even at 60 V for the DNA22 sequence (figure 6.5c), whereas the RNA22 retained two ammonium ions up to 140 V (figure 6.5d). These results strongly suggest that RNA G-quadruplex monomers in the gas are considerably more stable than the equivalent DNA structures.
G-quadruplex monomers

- **a)** [DNA12]_2
- **b)** [RNA12]_2
- **c)** [DNA22]_1
- **d)** [RNA22]_1

Figure 6.5 Stability of monomeric G-quadruplex species.

ESI-mass spectra of monomeric telomeric G-quadruplex species showing ammonium ion distribution as a function of RF lens voltage (voltages indicated). Peak annotation numbers refer to the number of coordinated ammonium ions. Cartoons representing the stability of the structures, as measured by the influence of voltage changes on ammonium coordination, are shown below the relevant spectra. Each parallelogram represents a G-quartet (blue for DNA, red for RNA). Note – for clarity, loop sequences have been omitted from the cartoon structures. Checked spheres = labile/unstable ammonium ions; solid black spheres represent stable ammonium ions. The RNA22 sequence is exceptionally stable.

6.4.1.4 ANALYSIS OF RELATIVE STABILITIES OF RNA AND DNA G-QUADRUPLEXES USING ESI-MS: HIGHER-ORDER G-QUADRUPLEX ASSEMBLIES (TERTIARY STRUCTURE)

Analysis of the G-quadruplex dimer peaks at 100 V (at charge state -8) shows a broader distribution of ammonium adducts compared to the single G-quadruplex species, which is indicative of non-specific cation association (e.g. cations binding to the nucleic acid backbone phosphates). In order to study the stability of the higher-order assemblies, the collision voltage was increased to 140 V. The main peaks for the tetramer formed from the RNA12 sequence at 140 V correspond to species comprising four and five ammonium ion adducts (figure 6.6b). This is proposed to be indicative of two stacked G-quadruplexes (each with two coordinated ammonium ions), with a fifth
more labile ammonium mediating the quadruplex-quadruplex stacking interaction (figure 6.6b). Dimers formed from the RNA22 sequence were seen to be most stable with four coordinated ammoniums, with a more labile fifth ammonium ion species also observed, suggesting (as for the RNA12 ammonium ion distribution) two ammonium ions per G-quadruplex subunit, with a highly labile ammonium ion mediating the interaction between two stacked G-quadruplexes (figure 6.6c).

Ammonium ion adducts for the DNA45 sequence were all fairly labile, with all of the 3-5 ammoniums seen at 100 V being lost as the voltage was increased to 140 V (figure 6.6a). In contrast, the RNA45 sequence revealed the species with five ammonium ions to be the most stable at 140 V (figure 6.6d). This is proposed to be evidence of the cation-mediated stacking of two intramolecular G-quadruplexes, as five highly-stable ammoniums suggests two ions per G-quadruplex, with a fifth ammonium ion coordinating between the G-quartets of two separate G-quadruplex units. This is some of the strongest data reported so far concerning higher-order three dimensional arrangements of telomeric (RNA) G-quadruplexes.
**G-quadruplex dimers**

![G-quadruplex dimers](image)

Figure 6.6 Stability of dimeric G-quadruplex species.

ESI-mass spectra of dimeric telomeric G-quadruplex species showing ammonium ion distribution as a function of RF lens voltage (voltages indicated). Peak annotation numbers refer to the number of coordinated ammonium ions. Cartoons representing the stability of the structures, as measured by the influence of voltage changes on ammonium coordination, are shown below the relevant spectra. Each parallelogram represents a G-quartet (blue for DNA, red for RNA). Note – for clarity, loop sequences have been omitted from the cartoon structures. Checked spheres = labile/unstable ammonium ions; solid black spheres represent stable, rigid ammonium ions. The RNA45 sequence is exceptionally stable.

### 6.4.2 ANALYSIS OF TELOMERIC RNA AND DNA G-QUADRUPLEXES USING CIRCULAR DICHROISM

All six telomeric samples shown in table 6.1 were analysed by circular dichroism in order to confirm that G-quadruplex structures were indeed forming in the ionic and annealing conditions used in the ESI-MS experiments. CD experiments were also performed in order to assess the overall topology of these sequences in ammonium acetate.
CD spectra for the DNA12 indicated a largely unstructured species, supporting the ESI-MS data by showing this sequence to form a G-quadruplex of very low stability. With a maximum peak at 295 nm, the CD spectra for the DNA22 and DNA45 sequences were indicative of predominantly anti-parallel structures (although, as mentioned, an 'anti-parallel' assignment from CD spectra permits a large number of different anti-parallel loop arrangements).

CD spectra for all RNA sequences studied indicated a predominantly parallel fold (with maximum peaks at 260 nm and minimum troughs at 240 nm). However, there appeared to be significant peaks (or 'shoulders') at the 300 nm range for the RNA22 and RNA45 sequences, indicative of anti-parallel strand arrangements. This contrasts to all previously reported telomeric RNA G-quadruplex CD spectra, all of which have shown a clear all-parallel distribution\textsuperscript{153,154,150–162}. Previously, we proposed two possible reasons for the peak seen in the 300 nm region of the CD spectra:

1) the bases of telomeric RNA G-quadruplexes are in fact able to adopt \textit{syn} glycosidic angles, allowing anti-parallel topologies for such structures and meaning telomeric RNA G-quadruplexes are more structurally polymorphic than previously considered, or;

2) the 300 nm signal is generated from base-base interactions of the loops.

While loop-loop interactions of telomeric G-quadruplexes have recently been proposed to be important in telomeric RNA tertiary structure\textsuperscript{166}, the first explanation given now appears insufficient, as it is difficult to conceive how the required \textit{syn} glycosidic torsion angles could be sterically permitted within an RNA G-quadruplex. It is more likely that the 'anti-parallel' signal (i.e. at 300 nm) is being generated from the association of two G-quadruplexes via quartet-quartet stacking interactions. As the anti-parallel CD signal arises simply from the stacking of \textit{syn} bases on \textit{anti} bases (the anti-parallel arrangement is then inferred), it is likely that this signal could be replicated (or mimicked) by the base-base associations of, for example, two RNA G-quadruplexes stacked at the 5'-5' interface. This hypothesis needs experimental confirmation.
Figure 6.7 Circular dichroism spectra for annealed telomeric RNA and DNA sequences. Spectra for telomeric DNA (solid lines) and RNA (dashed lines) sequences annealed in 100 mM ammonium acetate. Concentrations are the same as those of initial ESI-MS experiments: 12-mers = 20 μM; 22-mers = 10 μM; 45-mers = 5 μM. DNA12 is predominantly unfolded, DNA22 and DNA45 are assigned as mainly anti-parallel G-quadruplexes and all RNA sequences are assigned as predominantly parallel G-quadruplexes. (It should be noted that these experiments were repeated several times to confirm this finding.)

6.4.3 NATIVE POLYACRYLAMIDE GEL ELECTROPHORESIS STUDIES

6.4.3.1 NATIVE POLYACRYLAMIDE GEL ELECTROPHORETIC ANALYSIS OF TELOMERIC RNA AND DNA G-QUADRUPLEX TERTIARY STRUCTURE

Native polyacrylamide gel electrophoresis (PAGE) studies were performed on the RNA22 and DNA22 sequences with the main aim of validating the G-quadruplex multimerisation observed for the RNA22 sequence in ESI-MS experiments under more physiologically relevant salt conditions — namely, potassium and sodium. Although native PAGE provides extremely low resolution structural data, it allows rapid and clear identification of multimer formation for a given sequence or molecule.

To investigate the formation of stacked quadruplex structures, oligonucleotides were prepared in water (with no ions) and loaded onto polyacrylamide gels with no pre-annealing step. For these experiments, monovalent cations for quadruplex formation were sourced from the gel and run buffer only.
In KCl conditions, the RNA22 sequence was seen to form stable dimer, trimer and tetramer structures, at both low and high strand concentrations (figure 6.8a, lanes 5 and 6). These higher order structures were removed upon heating (figure 6.8a, lane 7) (a heating step was included in order to assess relative G-quadruplex stability). In KCl conditions, very faint bands could be seen for the DNA22 sequence at the ‘tetramer’ mass region, indicating weak multimer formation at low and high strand concentration (figure 6.8a, lanes 2 and 3). The DNA22 sequence predominantly formed folded monomer quadruplex under KCl conditions. In contrast, little or no folded monomer quadruplex can be seen for the RNA22 sequence.

In NaCl conditions, the RNA22 behaves in a similar manner to KCl conditions, with stable dimer, trimer and tetramer structures seen at both low and high strand concentrations (figure 6.8b, lanes 5 and 6), which are removed following heating of the sample (figure 6.8b, lane 7). Very little folded quadruplex monomer is seen for the RNA22 sequence in NaCl conditions. In NaCl conditions, the DNA22 sequence behaved in a manner similar to KCl conditions, with weak trimer bands seen at low and high strand concentrations (figure 6.8b, lanes 2 and 3). Unique to NaCl conditions, the DNA22 sequence was seen to form a strong and distinct band at the ‘dimer’ mass region (figure 6.8b, lane 3), for high strand concentration only, indicating intramolecular telomeric DNA quadruplexes may dimerise in a concentration dependent manner. RNA dimer and multimer formation appears to be strand concentration independent.

Samples were also analysed on NH₄OAc native gels (figure 6.8c), showing higher order structure formation for the RNA22 sequence (figure 6.8c, lanes 5 and 6) (similar to NaCl and KCl conditions), and very little higher order structures for the DNA22 (which predominantly shows monomer folded quadruplex structures – figure 6.8c, lanes 2-4).

These PAGE experiments validate the higher-order assemblies detected by ESI-mass spectrometry and show that the RNA22 sequence forms even larger multimer assemblies, composed of up to four associated G-quadruplexes (i.e. tetramers), with possibly even larger assemblies present also (figure 6.8a-c, lanes 5 and 6).
Figure 6.8 Native polyacrylamide gels of telomeric RNA and DNA in various monovalent cations. DNA22 and RNA22 sequences were analysed on native PAGE gels supplemented with either 50 mM KCl (a), 50 mM NaCl (b) or 50 mM ammonium acetate (c). Gels were run at room temperature for 120 minutes at 100 V. Lane 1 in all gels corresponds to 22-nt and 43-nt molecular weight markers. Lanes 2-4 = DNA22 samples, lanes 5-7 = RNA22 samples. Details of individual sample conditions are provided in separate panels. *'Pre-heated' samples were incubated at 90 °C for 3 mins prior to loading on the gel.
Considering the high stability noted for telomeric RNA G-quadruplexes structures (and higher-order assemblies thereof), the RNA22 sequence was analysed on denaturing and partially denaturing polyacrylamide gels (under various conditions) in order to further test the relative stability of these structures. The 'denaturing' components of the denaturing gels involves the inclusion of urea (a chaotropic molecule) at high concentration (7 M) in the gel and running buffer, as well as the use of formamide in the sample loading buffer. Additionally, some samples loaded onto denaturing gels were heated briefly to 90 °C prior to loading, in order to disrupt preformed structures.

Higher-order G-quadruplex structures formed from the RNA22 sequence showed quite remarkable stability when analysed under denaturing conditions on polyacrylamide gels. All RNA22 samples shown in figure 6.9 were annealed (in various salt types, as well as without salt) as described above. Surprisingly, the trimers/tetramers formed from this RNA sequence are maintained under the denaturing conditions of the urea-gel, even when a stabilising cation is omitted from the annealing procedure (figure 6.9, lane 6). Analysis of samples which were annealed in either KCl, NaCl, ammonium acetate or water, and subsequently heated prior to loading (3 minutes at 90 °C), reveals KCl to be the optimal cation for stabilisation of higher-order telomeric G-quadruplex structures (figure 6.9, lane 2). It was quite surprising to see the retention of higher-order structure under such harsh denaturing conditions – i.e. high temperatures and the use of chaotropic agents – highlighting the structural stability of these (presumably) stacked G-quadruplex structures.
Figure 6.9 Partially denaturing PAGE of RNA22 sequence.
Denaturing and partially denaturing conditions were used to test the stability of multimers formed from the RNA22 sequence. The gel and run buffer contain 7 M urea. The gel was run at room temperature at 180 V for 120 mins. Samples annealed in KCI, NaCl, ammonium acetate or water showed multimers formed in potassium conditions to be significantly stable, surviving even when heated to 90 °C prior to loading on the gel. Individual sample conditions are shown in the table. *'Pre-heated' samples were incubated at 90 °C for 3 mins prior to loading on the gel. "Refers to the salt type used for the annealing procedure (annealing salts were used at 100 mM concentration).

6.4.4 TELOMERIC G-QUADRUPLEX LIGAND BINDING STUDIES USING ESI-MS

In addition to the study of telomeric G-quadruplex stability and higher-order assemblies, ESI-MS was also used to investigate the interactions of four different ligands with the RNA22 and DNA22 sequences. The aim of this work was to obtain (albeit limited) structural data concerning telomeric RNA G-quadruplex-ligand interactions, as there is currently very little data available concerning telomeric RNA G-quadruplex (or mRNA G-quadruplex) ligand binding. The ligands used in this study include three different four-substituted naphthalene diimides (NDs) (figure 6.10) and BRACO-19.

Figure 6.10 Details of naphthalene diimide compounds used in this study.
Naphthalene diimide core (left) with ligand names and sidechain details shown in the table on the right.
**6.4.4.1 LIGAND-QUADRUPLEX DISSOCIATION CONSTANTS**

Dissociation constants ($K_d$) were determined for the compounds BRACO-19, FC4ND-01, FC4ND-06 and FC4ND-10 in complex with G-quadruplexes formed from the DNA22 and RNA22 sequences (figure 6.11). $K_d$ values obtained for the BRACO-19-quadruplex complexes show BRACO-19 to bind to the DNA 22-mer with higher affinity (two-fold) than the RNA 22-mer. This finding is in accord with FRET data and modelling predictions for the binding of BRACO-19 to telomeric RNA and DNA G-quadruplex structures, which predicted that the C2'-OH groups of the RNA would restrict the access of the BRACO-19's sidechains to the groove regions (as seen in the DNA crystal structure).

All three ND ligands bind to the DNA22 G-quadruplex with reasonably strong affinity, whereas the RNA22 G-quadruplex appears only to form a strong complex with the FC4ND-10 ligand (magenta arrow, figure 6.11). These ND ligands were designed as DNA G-quadruplex specific compounds (see chapter 5), which is illustrated by the $K_d$ values seen here. The ability of only one of these compounds (FC4ND-10) to form a strong complex with an RNA quadruplex highlights a structural feature in the ligand which may be important for quadruplex-targeted ligand design. FC4ND-10 and FC4ND-01 share a core domain and $R_1$ side chains (dimethylamine with a three carbon linker), differing only in the terminal functional groups of the $R_2$ substitutions, which involve dimethylamine functional groups for FC4ND-01 and hydroxyl groups for FC4ND-10. It appears, therefore, that the $R_2$ hydroxyl groups of FC4ND-10 are responsible for the relatively strong complex formed between this compound and the RNA22 quadruplex. In this case, the substitution of a dimethylamine group for a hydroxyl group increases RNA quadruplex-ligand binding 15-fold, but increases DNA quadruplex-ligand binding just three-fold. The significant effect these hydroxyl groups have on RNA quadruplex-ligand binding highlights a structural feature which may be worth considering when designing RNA (and DNA) G-quadruplex interacting compounds.
Figure 6.11 ESI-MS analysis of telomeric RNA and DNA ligand interactions.

a-d) Spectra for complexes formed between compounds FC4ND-01, FC4ND-06, FC4ND-10 and BRACO-19 and the RNA22 and DNA22 G-quadruplex forming sequences. The only strong ligand-RNA complex detected is indicated by the magenta arrow. Dissociation constants determined from the integrated peak areas are shown in the table below the spectra (e).

<table>
<thead>
<tr>
<th></th>
<th>RNA22 r[AGGG(UUAGGG)₃]</th>
<th>DNA22 d[AGGG(TTAGGG)₃]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_{d1}$</td>
<td>$K_{d2}$</td>
</tr>
<tr>
<td>FC4ND-01</td>
<td>63.1 (± 8.5)</td>
<td>-</td>
</tr>
<tr>
<td>FC4ND-06</td>
<td>25.1 (± 2.0)</td>
<td>12.6 (± 2.9)</td>
</tr>
<tr>
<td>FC4ND-10</td>
<td>4.0 (± 0.5)</td>
<td>5.0 (± 1.4)</td>
</tr>
<tr>
<td>BRACO-19</td>
<td>15.8 (± 1.0)</td>
<td>4.0 (± 0.5)</td>
</tr>
<tr>
<td></td>
<td>12.6 (± 3.0)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>5.0 (± 1.0)</td>
<td>12.6 (± 2.0)</td>
</tr>
<tr>
<td></td>
<td>4.0 (± 1.0)</td>
<td>12.0 (± 2.2)</td>
</tr>
<tr>
<td></td>
<td>7.9 (± 1.4)</td>
<td>5.3 (± 1.0)</td>
</tr>
</tbody>
</table>
6.5 GENERAL DISCUSSION AND CONCLUSIONS

6.5.1 KEY FINDINGS FROM BIOPHYSICAL STUDIES

The biophysical work involving human telomeric G-quadruplexes described above was aimed at providing an additional structural perspective to the crystallographic data reported in previous chapters. The two main aspects of telomeric G-quadruplexes studied here include investigation of tertiary G-quadruplex structure and analysis of quadruplex-ligand binding. The key findings include:

1) Detection of significant multimerisation of telomeric RNA G-quadruplexes by ESI-MS. RNA multimers formed far more readily and were considerably more stable than equivalent DNA structures.
2) Strong evidence was provided by ESI-MS for the role of (ammonium) cations in RNA G-quadruplex multimerisation.
3) PAGE analysis confirmed the presence of telomeric RNA G-quadruplex higher-order assemblies in physiologically appropriate ionic conditions, i.e. Na⁺ and K⁺ conditions.
4) Additionally, PAGE analysis showed the presence of further multimerisation beyond G-quadruplex dimers, which were not detected in ESI-MS experiments.
5) CD analysis of the samples studied by ESI-MS confirmed the formation of G-quadruplex structures and provided additional topological data.
6) Additionally, CD spectra for the RNA revealed the presence of peaks at 300 nm, corresponding either to anti-parallel RNA G-quadruplex structures or perhaps evidence of 5'-S' (or 3'-S') G-quadruplex stacking.
7) Ligand-quadruplex studies of the 22-mer DNA and RNA sequences revealed a compound (FC4ND-10) with the ability to bind tightly to telomeric G-quadruplex RNA.
6.5.2.1 TERTIARY STRUCTURE OF HUMAN TELOMERIC RNA AND DNA

One of the most surprising findings of this work was the discovery that human telomeric RNA sequences very readily form stable higher-order/tertiary assemblies of (presumably) stacked G-quadruplex subunits. In contrast, the equivalent DNA sequences studied showed a reluctance to multimerise, or indeed to even form stable G-quadruplex structures (e.g. very little bimolecular quadruplex formation was observed in the spectra for the DNA12 sequence, see figure 6.4a). There are two main points to consider concerning such differences:

1) The stability of high-order G-quadruplex assemblies will depend to a significant degree upon the stability of the individual monomers (i.e. single G-quadruplex structures). Based on previous biophysical studies performed by others, it is clear that telomeric RNA G-quadruplex monomers are indeed more thermodynamically stable than equivalent DNA structures\(^{161,162}\). This work has been rationalised by crystallographic work reported in this thesis on a native bimolecular telomeric RNA G-quadruplex (Chapter 2), which highlighted a key role of the RNA's C2' OH groups in forming (potentially) stabilising intramolecular hydrogen bonds with a number of components of the RNA (e.g. exocyclic N2 nitrogens of guanines of the G-quartets and O4' oxygens and backbone phosphate groups). It is clear, therefore, that telomeric RNA G-quadruplex monomers are more stable than equivalent DNA structures, which would be expected to afford a certain degree of stability to any G-quadruplex multimers formed from such monomers.

2) The apparent structural monomorphism of telomeric RNA G-quadruplexes may be an important factor in higher-order structure formation and stability. The stacking interactions of G-quadruplex monomers is expected to be more stable for parallel-stranded topologies\(^{216}\), which allows effective inter-G-quadruplex \(\pi\)-\(\pi\) stacking interactions (via the G-quartets). Telomeric RNA G-quadruplexes are expected to be exclusively parallel stranded, permitting the formation of stable higher-order assemblies. In contrast, equivalent DNA sequences have been
shown to be far more polymorphic, with DNA22 sequence studied here shown by CD to form predominantly anti-parallel G-quadruplex structures (figure 6.7). Based on modelling predictions, it is clear to see that anti-parallel G-quadruplex structures are not able to stack via quartet-quartet surfaces as effectively as parallel-stranded G-quadruplexes, because the lateral and diagonal TTA loops sterically block G-quartet surfaces (see figure 6.1).

Understanding the mechanism by which telomeric RNA and DNA G-quadruplexes can self-associate to form tertiary complexes is important for basic research as well as drug discovery, as the surfaces presented by such structures to either proteins or therapeutic compounds would be expected to differ significantly for different tertiary arrangements (e.g. stacked anti-parallel G-quadruplexes vs stacked parallel G-quadruplexes). The biophysical data reported here provides important insight into the propensity and degree of tertiary structure formation for telomeric RNA sequences, but is limited in the resolution of data provided. For example, based on this data, we cannot tell the direction of quadruplex-quadruplex orientation in the higher-order assemblies detected (by ESI-MS and PAGE) – the possible orientations being: 5'-5', 3'-3', 3'-5' and 5'-3' – all of which would be expected to generate different surfaces and binding sites (for proteins or small-molecule ligands). In order to more fully understand such tertiary arrangements, data from high-resolution techniques would certainly be highly informative.

6.5.2.2 TERTIARY STRUCTURE OF HUMAN TELOMERIC DNA AND RNA – CORRELATION BETWEEN BIOPHYSICAL RESULTS AND CRYSTALLOGRAPHIC DATA

Although crystallographic (or NMR) data is not available for long TERRA sequence (i.e. sequences capable of forming two-contiguous G-quadruplexes, e.g. r[UUAGGG]₉), the crystal structures reported in chapters 2 and 3 provide a certain degree of high-resolution insight into TERRA G-quadruplex tertiary structure. The crystal structure of a bimolecular G-quadruplex formed form the TERRA sequence, r(U₈AGGGUUAGGGU), described in chapter 2, provides some details of telomeric RNA G-quadruplex higher-order structure, as two quadruplexes are seen in the crystal lattice, stacked 5'-5'. This
association, however, is based upon two layers of stacked UAUA tetrads, which in fact arise from the Watson-Crick/Hoogsteen base-pairing interactions made available through the terminal residues. In other words, the RNA quadruplex dimers seen in this crystal structure are quite likely artefacts of the technique and sequence used. A more relevant correlation between the biophysical data reported here and previous crystallographic data can be seen in the telomeric RNA-acridine complex reported in chapter 3. This structure shows the ligand-mediated stacking of two G-quadruplexes at the 5'-5' interface. The arrangement seen here could easily be applicable to two-stacked unimolecular RNA G-quadruplexes formed within the same molecule, and provides a potentially valuable insight into the interaction of drug molecules with tertiary assemblies of G-quadruplexes within long TERRA sequences (e.g. r[UUAGGG]_{90-100}). That is, it is highly likely that the interfaces between stacked TERRA G-quadruplexes in higher-order assemblies could be effectively targeted with appropriately designed and highly specialised intercalative agents.

Further crystallographic (or NMR) data on longer (i.e. ≥ r[UUAGGG]_{8}) TERRA sequences would allow a far better understanding of the mechanism by which TERRA G-quadruplexes self-associate. It is unfortunate, therefore, that a high-resolution structure has not been determined for a TERRA sequence larger than 12 nucleotides in length. This is surprising, considering the vast number of 22-24-mer length telomeric DNA structures which have been solved. Based on the biophysical findings reported here, it seems likely that such structure determinations (i.e. of four-repeat TERRA sequences, such as r[AGGG[UUAGGG]_{3}]) are in fact being hindered by the propensity of unimolecular TERRA G-quadruplexes to multimerise in solution and the subsequent stability of such assemblies. It is far easier to crystallise (or assign peaks in NMR spectra) a single, well-folded, structurally stable monomer than it is to perform the same task for a significantly larger, repetitive polymeric structure. With this in mind, in the future, careful experimental setup may permit the successful structure determination of a unimolecular G-quadruplex formed from telomeric RNA.
6.5.3 HUMAN TELOMERIC RNA AND DNA G-QUADRUPLEX-LIGAND BINDING

Very few studies have been reported so far concerning TERRA G-quadruplex-ligand binding. This is surprising, as, considering the importance of telomeric DNA G-quadruplexes as anti-cancer drug targets, it is important to understand how current (and future) telomeric DNA G-quadruplex binding compounds can interact with the equivalent RNA structures (which may have very different biological roles). ESI-MS derived dissociation constants for four DNA G-quadruplex binding compounds in complex with 22-mer DNA and RNA G-quadruplexes were reported here. The results showed all four compounds to bind to the DNA G-quadruplex with reasonably high-affinity (as expected), and importantly, showed a single member of the ND family to bind to the RNA G-quadruplex with relatively strong affinity. This compound, FC4ND-10, differing from an analogue which exhibited poor RNA G-quadruplex binding by the substitution of two dimethylamine groups for hydroxyl groups, highlights a possible structural feature of the ligand which may prove useful in the design of strong telomeric RNA (as well as non-TERRA) G-quadruplex binding compounds. Although not displaying RNA vs DNA ‘selectivity’, the ability of FC4ND-10 to bind telomeric RNA G-quadruplex suggests designing a small molecule with true RNA vs DNA selectivity is certainly possible. Such compounds would be valuable tools for elucidating the role of TERRA G-quadruplex structures in cells, as well as being of use therapeutically, as TERRA sequences are certainly linked to a range of key roles at the telomere, including the ability to inhibit telomerase activity.

From the ESI-MS data determined here, the mechanism by which FC4ND-10 binds telomeric RNA G-quadruplexes cannot be determined. Considering the effect the inclusion of two polar hydroxyl groups have on RNA binding, it is reasonable to propose that such functional groups are exploiting hydrogen bonding groups within the RNA, which may include the C2'-OH groups (such as those located in the quadruplex groove regions). As mentioned previously, although the only high-resolution structural data available for telomeric RNA G-quadruplex-ligand binding showed an indirect role for the C2'-OH groups in drug-binding (chapter 3), these groups may yet represent the most viable structural feature of RNA G-quadruplexes for designing RNA vs DNA selectivity in G-quadruplex binding compounds. As always,
high-resolution structural data regarding FC4ND-10-RNA binding would be highly useful in determining the role (if any) of the C2'-OH groups of the RNA in direct ligand binding. Indeed, following these ESI-MS results, crystallographic studies into this complex were pursued – and complex crystals obtained. Disappointingly however, problems with the quality of crystals and diffraction data meant that the structure could not be solved, leaving the role of the RNA’s C2'-OH groups in ligand recognition as mere speculation (figure 6.12).

**Figure 6.12 Crystallographic data for a telomeric RNA G-quadruplex-FC4ND-10 complex.**

a) Crystals were obtained for the telomeric sequence r(UAGGGUAGGGU) in complex with the naphthalene diimide compound, FC4ND-10. The blue colour originates from the ligand, implying FC4ND-10 is almost certainly bound to the RNA. These crystals were grown in standard hanging drops, using MPD as a precipitant and spermine as an additive, with an ssRNA concentration of 1.5 mM and a ligand concentration of 1 mM. b) Data collection statistics obtained from X-ray diffraction experiments for the complex crystals shown in (a). Although the statistics appear to be of reasonable quality, it is clear from the diffraction images themselves (c-d) that the crystals are highly anisotropic (i.e. lack order along one axis). Typically, a large degree of systematically absent data causes significant problems for molecular replacement methods. Because of this, among other reasons, this structure could not be solved. Data were collected at the Diamond Light Source on beamline I03.
CHAPTER 7

SUMMARY, CONCLUSIONS AND FUTURE PROSPECTS
7.1 SUMMARY OF FINDINGS

The telomeric ends of mammalian chromosomes serve to protect the coding regions of the genome, and play a pivotal role in regulating cellular senescence and apoptosis. Due to the involvement of telomeres in such fundamental cellular process, dysregulation of these regions is a key factor in cancer progression, and consequently, telomeres have been the subject of intense research as anti-cancer drug targets. While telomeres are composed of a complex array of both proteins and nucleic acids, it is the ability of the nucleic acid components (both DNA and RNA) to form specialised G-quadruplex structures which has been the focus of the work presented here. Such telomeric G-quadruplex structures have been studied extensively as anti-cancer drug targets, however, high-quality structural data is currently lacking. The primary goal of the work undertaken here has been to address this issue, by generating high-quality X-ray crystallographic and biophysical data for a range of human telomeric G-quadruplexes, with an emphasis on the importance and potential application of such structures to anti-cancer drug design. The research and findings can be divided into four main areas:

1) human telomeric RNA G-quadruplex folding;
2) human telomeric RNA G-quadruplex-ligand recognition and binding;
3) human telomeric DNA G-quadruplex-ligand recognition and binding, and;
4) higher order/tertiary structure of human telomeric RNA and DNA G-quadruplexes.

The first crystal structure of a human telomeric RNA G-quadruplex, presented in chapter 2, provides atomic-scale details of telomeric RNA folding, and permits a rare opportunity to directly compare equivalent RNA and DNA structures. Such RNA vs DNA comparison has provided structural rationalisation of current biophysical findings reported by others concerning telomeric G-quadruplexes, which have shown telomeric RNA G-quadruplexes to be significantly more thermodynamically stable and monomorphous than the equivalent DNA structures\textsuperscript{160-162}. Analysis of the crystal
Structure reported here suggests the C2'-OH groups of the RNA to be largely responsible for such differences, by:

1) restricting the RNA bases (through steric hindrance) to the anti glycosidic conformation, thereby restricting RNA G-quadruplexes to parallel topologies, and;

2) by preferentially interacting intramolecularly with the RNA (rather than with solvent molecules), resulting in a significant increase in stabilising intramolecular contacts within the RNA (compared to the equivalent DNA structure).

While this native RNA structure provides understanding within the domain of basic research, data concerning telomeric RNA G-quadruplex-drug binding was a key goal of the work reported here – this was provided by the successful crystal structure determination of a complex formed between a human telomeric RNA G-quadruplex and a di-substituted acridine compound (chapter 3). This crystallographic data highlights the difficulty in accurately predicting RNA G-quadruplex-drug binding mechanisms, as the crystal structure reveals the formation of a novel purine-octet ligand-binding platform, formed from the 5’ G-quartet and adenine residues of the propeller-loop regions. In addition to the surprise of discovering the octet arrangement, this structure involves further unexpected features of two ligand molecules bound to a single G-quartet surface, as well as a striking absence of direct RNA-drug interactions. Following from this, a crystal structure was determined of a complex formed between the same acridine drug and the corresponding DNA G-quadruplex, again, permitting an excellent opportunity to directly compare equivalent DNA and RNA structures (chapter 4). Importantly, this DNA-acridine complex reveals a significantly different mode of drug-binding compared to the equivalent RNA-acridine structure, including:

1) an absence of an extended G₄A₄ octet ligand-binding platform (as seen in the RNA structure);
2) DNA-acridine binding occurs at the 3' G-quartet surface (in contrast to the 5' G-quartet-acridine binding in the RNA structure);

3) the acridine binds the DNA in a 1:1 stoichiometry (contrasting to the 2:1 stoichiometry seen within the RNA structure).

Crystallographic data was also provided concerning intramolecular G-quadruplex-ligand binding, in the form of two high-quality crystal structures of complexes involving human telomeric DNA G-quadruplexes bound by potent methyl-piperazine modified naphthalene diimide compounds. These structures add to the relatively limited number of intramolecular G-quadruplex drug complexes available, and reveal an unprecedented involvement of the ligand's sidechains in quadruplex-groove binding – a characteristic in a telomeric G-quadruplex binding ligand which has been long sought after yet not previously visualised.

In addition to crystallographic data, biophysical studies were performed in order to provide a more complete structural picture of telomeric G-quadruplex behaviour. A key finding of this work was the unexpected propensity of telomeric RNA G-quadruplexes to multimerise and form higher-order assemblies of (arguably) stacked G-quadruplex subunits. Additionally, ligand-binding studies were performed, revealing a single member of a small series of compounds with the ability to bind telomeric RNA with reasonable affinity.
7.2 IMPLICATIONS OF FINDINGS AND CONCLUSIONS

7.2.1 TELOMERIC RNA G-QUADRUPLEX DRUG DESIGN

RNA G-quadruplexes have considerable potential as drug-targets, as such structures have been implicated in a significant number of diseases, for example, in the regulatory regions of a number of oncogenic mRNA transcripts. Telomeric RNA G-quadruplexes are also of relevance to drug design, due to the role of unfolded TERRA transcripts in telomere biology, as well as the need to understand the manner by which current DNA G-quadruplex directed compounds interact with the equivalent RNA structures.

The first crystal structure of an RNA G-quadruplex-ligand complex, reported here, reveals the presence of a novel purine octet-motif, which forms an RNA-specific binding platform for the acridine ligand studied. This unexpected finding has two important implications concerning RNA G-quadruplex drug design:

1) The significant and unpredictable rearrangement of native G-quadruplexes to accommodate ligand binding suggests that native G-quadruplex structures may not (at present) be such reliable tools for designing drug molecules.

2) The presence of an RNA-specific extended octet offers the opportunity to exploit such a motif for the design of RNA vs DNA G-quadruplex selective ligands (either as therapeutic drugs or as molecular probes). While designing a drug to selectively bind to such a large motif may be technically challenging, it remains feasible. Additionally, the presence of such a motif is not restricted to telomeric RNA G-quadruplexes, and could reasonably be applicable to G-quadruplexes formed within mRNA untranslated regions.

It should be noted that the utility of the octet motif in deriving RNA vs DNA selectivity does not preclude the direct exploitation of C2'-OH groups for the same such purpose. The C2' hydroxyl groups of the RNA represent the primary difference between RNA
and DNA, and could quite viably be exploited for direct ligand-RNA hydrogen bonding interactions. Additional RNA G-quadruplex-ligand complex structures would certainly inform on this matter. The determination of such structures would be considerably aided by the generation of strong telomeric RNA G-quadruplex binding compounds, which are currently lacking. NMR-derived RNA G-quadruplex ligand complex structures would be particularly informative, as this would (hopefully!) validate the crystallographic work, as well as providing models which take into account the apparent local flexibility of such structures (e.g. within loop regions). Based on ESI-MS derived binding data reported here – showing the ND compound, FC4ND-10, to bind an intramolecular telomeric RNA G-quadruplex with reasonable affinity – the design of an RNA G-quadruplex selective compound certainly appears achievable.

7.2.2 TELOMERIC DNA G-QUADRUPLEX DRUG DESIGN

Hundreds of telomeric G-quadruplex interacting small molecule compounds have been reported in the literature, primarily aimed at selectively inhibiting cancer cell growth through the disruption of telomerase activity. Rational design of such compounds has relied on a relatively limited number of crystallographic and NMR structures – a number which has now been increased (albeit slightly!) with the addition of the three DNA G-quadruplex-ligand crystal structures reported in this work.

The crystal structure of a DNA-acridine complex – although of suboptimal quality – reveals two structural features of potential importance for the future design of acridines as telomere-directed anti-cancer compounds:

1) The absence of direct DNA-acridine contacts (in contrast to the interaction of BRACO-19 with the same G-quadruplex) suggests the relative positions of the amide functionality may be a critical factor in strong acridine-DNA binding. Careful consideration of the placement of such hydrogen bonding groups in relation to the acridine core would be recommended in future drug design.
2) The binding of FD-121 to the 3' G-quartet mimics the binding mode (in this respect) of BRACO-19 – the 3' G-quartet, therefore, may be the preferred binding site for acridine-DNA G-quadruplex recognition, of potential significance to future acridine drug development.

The successful determination of two high-quality crystal structures involving intramolecular human telomeric DNA G-quadruplexes and two N-methyl-piperazine modified naphthalene diimide compounds reveals three important considerations regarding drug-binding:

1) Positively charged methyl-piperazine end-groups are able to make effective electrostatic contacts with DNA phosphate groups and other components of the quadruplex groove regions. These end groups would be highly recommended for future telomeric G-quadruplex binding drugs.

2) Ligand sidechain length has a significant influence on the overall mode of ligand binding, with longer sidechains resulting in a reduction in specific contacts with the DNA, along with an increase in binding site mobility.

3) Telomeric G-quadruplex grooves are mobile regions – the rational targeting of phosphate groups which lie adjacent to an external G-quartet (i.e. T5, T11 and T17) may be a more effective method for deriving strong G-quadruplex binding affinity, as these phosphates are geometrically restrained by their proximity to the G-quartet core. The use of sidechains with non-equivalent linker lengths (e.g. \( R_1 = 2 \) carbons, \( R_2 = 3 \) carbons) may be a viable route to achieving this.

The acridine and naphthalene diimide families of G-quadruplex binding compounds have performed exceptionally well in anti-cancer \textit{in cellulo} assays, and there is substantial evidence suggesting such molecules to be acting through a telomeric G-quadruplex stabilisation mechanism. It is hoped that the DNA complexes...
reported here will be of use for the further development of these promising anti-cancer agents.

7.2.3 TELOMERIC G-QUADRUPLEX HIGHER-ORDER/TERTIARY STRUCTURE

The manner by which multiple contiguous telomeric G-quadruplexes can self-associate is an aspect of G-quadruplex research which has received minimal attention, despite being of considerable relevance to both basic research and drug discovery interests. The significance of such research arises from the intrinsic polymorphism of G-quadruplex tertiary structure, as the surfaces presented by different higher-order arrangements could vary significantly, having implications concerning both protein and drug molecule recognition.

Biophysical studies into telomeric G-quadruplex higher-order structure, reported here, revealed telomeric RNA sequences to multimerise far more readily than equivalent DNA sequences, with the resulting RNA assemblies being far more stable than the corresponding DNA assemblies. The resolution of the biophysical data, however, does not allow a detailed view of quadruplex-quadruplex association, thus, it is not known whether telomeric RNA (or DNA) G-quadruplexes associate in a 3'-5', 5'-5', 3'-3' or 5'-3' manner. Based on the RNA-acridine complex described in chapter 3, it could be predicted that the formation of an extend aromatic surface area at the 5' G-quartet (in the form of the purine octet) would make 5'-5' association highly favourable. Further work into this issue could certainly clarify the matter – fluorescence-resonance energy transfer experiments involving careful fluorophore modifications of telomeric RNA sequences could be a viable method for generating useful data.
7.3 FUTURE PROSPECTS

As a structural biologist I am perhaps prone to over-using the phrase, “further structural data would inform on the issue”, however, with respect to the role of structural biology in the future development of telomeric G-quadruplex directed drugs, I feel such a statement cannot be avoided. This is because there remains a considerable lack of high-quality structural data available involving human telomeric G-quadruplex-drug complexes — the two intramolecular-ND complexes described above bring the total number of intramolecular telomeric G-quadruplex-ligand complex structures to four. Additionally, these ND-complexes represent the only ligand-complex structures determined thus far involving the much-studied human 22-mer sequence. The RNA G-quadruplex-acridine complex presented here represents the sole RNA-quadruplex-ligand complex determined to date. Clearly, this is not enough data to allow meaningful trends or structure-activity relationships to be inferred. Of course, it is much easier to say, “further structural data would inform on the issue”, than it is to perform the task practically, perhaps mainly due to, (among a number of reasons), the difficulty in generating high-quality crystals of such complexes.

With regards to future telomeric DNA G-quadruplex structural work, the determination of complex structures (X-ray or NMR) involving compounds which have performed well in vivo — such as telomestatin, RHPS4 and quarfloxin — would be particularly useful. The field of telomeric RNA G-quadruplex structure determination is, understandably, far behind in comparison, although it is surprising that a native intramolecular G-quadruplex structure formed from telomeric RNA has not been solved, considering the significant number of such structures which have been solved for equivalent DNA sequences. The propensity of TERRA G-quadruplexes to multimerise (as reported here) may be responsible for this observation, as the formation of large multimeric assemblies could feasibly hinder both NMR and crystallographic structure determination methods. Careful experimental design may therefore permit the successful solution of an intramolecular RNA G-quadruplex structure — for example, such a strategy could involve chemical modification of the 5’ or 3’ terminal residues to block the external G-quartet surfaces, and thereby prevent π-π mediated G-quadruplex stacking.
A future goal combining the two fields of DNA and RNA G-quadruplex research might indeed involve a combination of these two molecules, as the determination of a telomeric DNA-RNA hybrid G-quadruplex could be of considerable importance to understanding the role of TERRA and G-quadruplexes at the telomere. Telomeric DNA-RNA G-quadruplex hybrids have been detected in vitro, and have been proposed to be a possible mechanism of telomere length regulation in vivo. However, working with such a hybrid entity is technically challenging, as it is not a trivial matter to prevent RNA-RNA and DNA-DNA hybridisation in a test sample, whilst preserving the DNA-RNA species. The formation of a single DNA-RNA chimeric molecule may be the only viable route to achieving such a goal.

Although DNA G-quadruplexes have generally received considerably more research attention than RNA G-quadruplexes, considering the increased number of cellular process and diseases in which RNA G-quadruplex structures have been identified, it is possible that research into RNA G-quadruplex structures will overtake DNA G-quadruplex research in the not-too-distant future, particularly with regard to drug design. It is hoped that the crystallographic and biophysical findings reported in this thesis will be of value to such future research.
REFERENCE LIST


(74) Dai, J.; Carver, M.; Punchihewa, C.; Jones, R. A.; Yang, D. Structure of the Hybrid-2 type intramolecular human telomeric G-quadruplex in K+ solution:


(201) Collie, G. W.; Haider, S. M.; Neidle, S.; Parkinson, G. N. A crystallographic and
modelling study of a human telomeric RNA (TERRA) quadruplex. Nucleic Acids 
Res. 2010, 38, 5569-5580.


Selectivity in small molecule binding to human telomeric RNA and DNA quadruplexes†

Gavin Collie,a Anthony P. Reszka,a Shozeb M. Haider,a Valérie Gabelica,b Gary N. Parkinsona and Stephen Neidlea,a

Received (in Cambridge, UK) 29th January 2009, Accepted 24th October 2009
First published as an Advance Article on the web 13th November 2009
DOI: 10.1039/b901889a

Quadruplex RNAs are less well understood than their DNA counterparts, yet of potentially high biological relevance. The interactions of several quadruplex-binding ligands with telomeric RNA quadruplexes are reported and compared with their binding to the analogous DNA quadruplexes.

DNA quadruplexes (G4s), both within gene promoter sequences and at telomeres, have been extensively studied as potential small-molecule therapeutic targets.1 A variety of topologies have been found for both inter- and intramolecular DNA G4s, and detailed structural studies by X-ray crystallography, molecular simulation and NMR methods have highlighted the role played by the loops in small-molecule recognition.2 By contrast, there have been few studies until recently on RNA quadruplexes. Putative RNA G4 sequences have been found in 5'-untranslated sequences of a number of genes,3 and may be plausible therapeutic targets for appropriate small molecules. Most recently surprising observations have been made that telomeric DNA can be transcribed into telomeric RNA sequences,4 which appear to play an important role in regulating telomerase function. It is plausible that in cellular conditions they can spontaneously fold into G4 arrangements.5 A binuclear G4 formed from two separate strands, each having two telomeric RNA repeats, has a parallel topology in both Na+ and K+ solutions.6 An NMR analysis7 has shown that the structure is analogous to the parallel tetramolecular DNA G4 crystal structures.4

Both 23- and 24-mer four-repeat telomeric RNA sequences also form a stable parallel topology in either Na+ or K+ solution.8,9

We report here on biophysical and molecular modeling studies of small-molecule interactions with the four-repeat intramolecular human telomeric RNA sequence d(TAGGGTUAGGG). We have used the experimental telomeric G4 drug BRACO-19 (Fig. 1) and three tetrasubstituted naphthalene diimides10 (compounds 1–3). In each case the side-chains comprise two or more CH3 groups terminating in an amine group. The binding of BRACO-19 to G4 DNAs has been previously studied in vitro and in vivo by a variety of biophysical and biological methods.11 A crystal structure12 of a complex between a bimolecular human telomeric G4 of sequence d(TAGGGTAGGGT) and BRACO-19 has shown that this DNA G4 has a parallel topology in the crystalline state, in common with the native 12-mer and 22-mer telomeric DNA sequences,8 and that the substituents interact extensively with the TTA loop regions in the complex, promoting their conformational remodelling. The naphthalene diimide ligands have high affinity for telomeric DNA G4 as well as some promoter G4s.13 If telomeric quadruplexes are to be useful as druggable targets then binding to (as well as selectivity between) G4 DNAs and RNAs is required. Telomeric RNAs inhibit telomerase activity, possibly by directly binding to the telomerase RNA template, so their stabilization as quadruplex-ligand complexes would enhance this inhibition and thus would be a viable therapeutic strategy.

Circular dichroism (CD) spectra show that the 22-mer RNA telomeric G4 has a characteristic parallel topology15 in K+ solution (Fig. 2). The spectrum of the RNA 22-mer shows a positive band at 265 nm and a minimum at 245 nm, in accord with that reported for the bimolecular telomeric G4 RNA 12-mer8 and the four-repeat 23- and 24-mer's,9 telomeric G4 RNAs. This conservation of topology is in striking contrast with the highly polymorphic behaviour of four-repeat telomeric DNA G4s, which depends on factors such as the precise nature of the flanking sequences, molecular crowding and concentration.14 We conclude that the parallel topology of RNA G4s, now observed in 22-, 23- and 24-mers, is an intrinsic feature that does not depend on flanking sequence.
The melting behaviour of the 22-mer sequence was examined by UV and CD methods, in K⁺ solution, and shows a hypochromic shift characteristic of quadruplex formation and a Tm of 74.7 °C. The Tm for the DNA 22-mer is 66.5 °C in identical conditions (50 mM K⁺ buffer). Interactions of the 22-mer RNA quadruplex with BRACO-19 were examined using CD, and no significant spectral changes were observed (Fig. 2). The melting behaviour of the 22-mer in 50 mM K⁺ solution with BRACO-19 at various molar ratios was monitored by a FRET method, and compared with the behaviour under identical conditions with the 22-mer DNA quadruplex. The ΔTm values at a 1 μM ligand concentration for the DNA and RNA G4s are 27.8° and 11.3°, respectively; this preferential stabilization of the DNA G4 is maintained over a range of ligand concentrations.

We have used electrospray mass spectrometry methods to quantitate the interactions of all four ligands with DNA and RNA G4s. Spectra corresponding to stable G4 monomers with two NH₃⁺ ions were observed with both intramolecular G4s (Fig. 3). Ligands were added to G4 solutions prior to injection and incubated at room temperature for 1 min at ligand:G4 ratios of 2.5:5, 5:5 and 10:5 (μM). Equilibrium dissociation constants (Kd) were determined from the peak areas (Table 1).

The ranking order for the Kd values obtained by ESI-MS for the two BRACO-19 G4s is in qualitative accord with the FRET data (and with the modeling studies—see below) as BRACO-19 was found to bind to the telomeric DNA G4 with ca. 2-fold higher affinity than to the RNA one. All three naphthalene diamide ligands bind to the DNA G4 with high affinity, whereas the RNA G4 forms a strong complex only with ligand 3.

A molecular dynamics approach has been used to simulate the structures of all eight DNA and RNA intramolecular G4 ligand complexes. The crystal structure of the native 22-mer DNA G4 (PDB id 1KF1) and the bimolecular G4 complex with BRACO-19 (PDB id 3CE5) were used as starting-points for the study. The RNA 22-mer G4 was generated from the DNA structure by direct addition of 2'-OH groups. Simulations for the native DNA and RNA quadruplexes and the eight ligand complexes all employed identical protocols.

The slightly lower RMSD value (1.90 Å) of the native 22-mer RNA G4 following its simulation suggest that it is more stable than that of the DNA G4 ((RMSD) of 2.01 Å) in accord with the Tm data presented here, and with thermodynamic studies on DNA and RNA tetramolecular G4s. In addition to the eight H-bonds that form the co-planar array in each G-quartet, the presence of the 2'-OH groups in the G4 RNA was shown by the MD simulations to impart rigidity to the backbone by invoking interactions with the O5' groups and water molecules in the grooves. The overall structures of the 22-mer DNA G4 and RNA are closely similar, in accord with the recent NMR analysis and more detailed X-ray studies of a bimolecular RNA G4 (Collie et al., to be published).

On the other hand, the significantly lower RMSD values of the 22-mer DNA G4 complexes with ligands 1, 2 and BRACO-19 (1.77 Å vs. 2.18 Å for BRACO-19) suggest that they are more stable than their RNA G4 complexes, and are in qualitative accord with the binding data presented here. The primary interactions of all the ligands with both G4s are π-π stacking interactions of the aromatic chromophore with a terminal G-quartet and electrostatic interactions between the positively-charged ligand side-chain and the negatively-charged G4 backbone in the loops. These strong electrostatic interactions are responsible for stabilizing the otherwise highly mobile loops. The lower affinity shown by ligands 1, 2 and BRACO-19 for RNA G4 is explained by the presence of the additional 2'-OH groups which constrict the space available to the ligand side-chains to interact with the loops, by reducing the depth and the width of the UUA loops.

The structural stability of the BRACO-19-DNA G4 complex is enhanced by the interactions of the terminal 3' thymine O4 atom with a side-chain amide group in BRACO-19, as observed in the co-crystal structure. This maintains BRACO-19 asymmetrically on one half of a G-quartet face. Analogous interactions have been observed in antiparallel crystal structures of an Oxytricha nova telomeric DNA G4 in complexes with disubstituted acridine ligands. In the RNA G4-BRACO-19 simulated complex, the side-chain amide group is sandwiched between the terminal uracil O4 atom and the 2'-OH group of a guanine (G17) on which the drug is stacked, so that an amide nitrogen atom of BRACO-19 interacts with the carbonyl oxygen atom (O4) from U24. However the amide carbonyl oxygen atom is also directly opposite 2'-OH from G17. The close proximity of the two oxygen atoms causes unfavourable electrostatic repulsion. This results in the side chain of the drug oscillating during the simulation between the amide nitrogen–O4 and 2'-OH stable interactions, and repulsion from oxygen atoms at the opposite end. This behaviour further destabilises

Table 1 Equilibrium dissociation constants (in μM), with esds calculated from repeat experiments

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Kd,1 (μM)</th>
<th>Kd,2 (μM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>d[AGGG(UUAGGG)]</td>
<td>63.1 ± 8.5</td>
<td>3.6 ± 0.5</td>
</tr>
<tr>
<td>2</td>
<td>25.1 ± 2.0</td>
<td>12.6 ± 2.9</td>
</tr>
<tr>
<td>3</td>
<td>4.0 ± 0.5</td>
<td>5.0 ± 1.4</td>
</tr>
<tr>
<td>BRACO-19</td>
<td>15.8 ± 1.0</td>
<td>4.0 ± 0.5</td>
</tr>
<tr>
<td>d[AGGG(TTAGGG)]</td>
<td>12.6 ± 3.0</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>5.0 ± 1.0</td>
<td>12.6 ± 2.0</td>
</tr>
<tr>
<td>3</td>
<td>4.0 ± 1.0</td>
<td>12.0 ± 2.2</td>
</tr>
<tr>
<td>BRACO-19</td>
<td>7.9 ± 1.4</td>
<td>5.3 ± 1.0</td>
</tr>
</tbody>
</table>
The naphthalene diimide ligands were originally designed as high-affinity G4 DNA binding molecules. A surprising observation is that only one of them (3) forms a strong complex with a tetrameric RNA G4 monomer (Table 1). The substitution of -NMe2 for an -OH group increases RNA G4-ligand binding 15-fold, but gives only a 3-fold increase in DNA G4-ligand binding. The effect of -OH groups on RNA vs. DNA G4-ligand interactions is an important structural feature to consider when designing selective G4 interacting compounds. This could be exploited, if targeting RNA G4s (e.g. the 5' UTR of oncogene promoters) or avoided, if targeting DNA G4s (e.g. telomeric DNA).

In summary, this study has shown that small-molecule ligands can discriminate between DNA and RNA G4s. It also confirms that four-repeat 22-mer telomeric RNA sequences fold into a parallel G4 in solution that is more stable than their DNA counterparts, and that flanking sequence changes do not affect their ability to form parallel G4s. Ligand binding selectivity is a consequence of the 2'-OH groups in the RNA and their effects on groove and loop widths so that ligand side-chains interact less effectively than with the DNA G4. One can anticipate that enhanced selectivity between DNA and RNA G4s may be achieved by appropriate ligands; their detailed design will require experimental structural data on RNA G4 complexes.

This work was supported by Cancer Research UK (Programme Grant No. C129/A449), by European COST (Programme Grant No. G129/A4489), by National Cancer Institute (Grant No. CA 105669), and by the European COST (Programme Grant No. G129/A4489), by National Cancer Institute (Grant No. CA 105669).
Electrospray Mass Spectrometry of Telomeric RNA (TERRA) Reveals the Formation of Stable Multimeric G-Quadruplex Structures

Gavin W. Collie, Gary N. Parkinson, Stephen Neidle, Frédéric Rosu, Edwin De Pauw, and Valérie Gabelica

CRUK Biomolecular Structure Group, The School of Pharmacy, University of London, WC1N 1AX London, United Kingdom, and Physical Chemistry and Mass Spectrometry Laboratory, Department of Chemistry, University of Liège, Belgium

Received January 14, 2010; E-mail: v.gabelica@ulg.ac.be

Abstract: We report on the self-assembled structures formed by 12-mer, 22-mer, and 45-mer telomeric RNA (telRNA/TERRA) sequences compared to their DNA analogues, as studied by electrospray mass spectrometry, circular dichroism, and thermal denaturation. The major difference between telomeric RNA and DNA sequences is the ability of telomeric RNA to form higher-order dimeric assemblies, initiated by cation-mediated stacking of two parallel G-quadruplex subunits. The 5'-5' stacking had been observed recently by NMR for the r(GGGUUAGGGU) 10-mer (Martadinata, H.; Phan, A. T. J. Am. Chem. Soc. 2009, 131, 2570); the present work shows that stacking also occurs for the 22-mer containing four G-tracts and for the 45-mer containing eight G-tracts, suggesting a general structural feature of telomeric RNA. The importance of kinetic effects in multimer formation, unfolding, and structural rearrangements is also highlighted.

Introduction

The telomeric ends of mammalian chromosomes comprise tandem repeats of the sequence d(TTAGGG), together with associated proteins. Mammalian telomeric DNA is typically 5–8 kb in length, with a 3' single-stranded overhang of 100–200 bases. These G-rich sequences are able to form four-stranded G-quadruplex structures, which have as their core the stacking of at least two guanine quartets, a planar arrangement of four guanine bases held together by eight Hoogsteen hydrogen bonds. Mammalian telomeres are systematically shortened by at least two guanine quartets, a planar arrangement of four guanine bases held together by eight Hoogsteen hydrogen bonds. The telomeric ends of mammalian chromosomes comprise tandem repeats of the sequence d(TTAGGG), together with associated proteins. Mammalian telomeric DNA is typically 5–8 kb in length, with a 3' single-stranded overhang of 100–200 bases. These G-rich sequences are able to form four-stranded G-quadruplex structures, which have as their core the stacking of at least two guanine quartets, a planar arrangement of four guanine bases held together by eight Hoogsteen hydrogen bonds. Mammalian telomeres are systematically shortened by at least two guanine quartets, a planar arrangement of four guanine bases held together by eight Hoogsteen hydrogen bonds. This is counteracted in most cancer cell types by the enzyme telomerase, a heterodimeric reverse transcriptase, which catalyzes the synthesis of telomeric DNA repeats. One approach to inhibit telomerase activity involves stabilizing telomeric DNA as G-quadruplex structures by the use of small-molecule G-quadruplex-binding ligands.

The assumption that the telomere was transcriptionally silent has been overturned by recent findings that telomeric DNA can be transcribed, the resulting RNA transcripts ranging in size from ∼100 nucleotides up to at least 9 kb, composed of tandem repeats of the sequence r(UUAGGG). It has been shown by circular dichroism (CD) and X-ray crystallography that these G-rich telomeric RNA sequences are able to form parallel G-quadruplex structures in sodium as well as in potassium ionic conditions, and are significantly more stable than their DNA counterpart in the presence of potassium, but not sodium ions.

The role that telomeric RNA (and by implication, RNA quadruplexes) plays in telomere biology is yet to be fully understood.
(26) Beck, J.; Coll grave, M. L.; Ralph, S. F.; Shell, M. M. R. D. J. room temperature (RT). The starting time of G-quadruplex forma-
min to disrupt any preformed structures, and then equilibrated at
otherwise, the single-strand concentrations upon annealing were 1
I\text{TAGGG)}\text{r}, M W  =  14290.3 \text{ Da}). For the kinetics experiment
were purchased from Eurogentec (Belgium) and used without
Materials and Methods
have therefore studied telomeric RNA sequences with two to
tries. Moreover, because G-quadruplexes accommodate
cations between consecutive G-quartets, counting the number
unambiguous determination of the number of strands in non-
coherent assemblies, and can resolve mixtures of stoichiome-
TERRA Reveals Formation of Stable G-Quadruplex Structures
The assembled structure is a tetramer in ammonium acetate. This quadruplex much more readily than DNA 12, but the most G-tracts with their DNA counterparts. All samples were annealed in ammonium acetate. The smallest G-quadruplex that can be formed with the 12-mers is dimeric, since each strand contains two G-tracks and four G-tracks are needed to form a tetrameric G-quadruplex. The 22-mers, which contain four G-tracks, are expected to form intramolecular (monomeric) G-quadruplexes, and the 45-mers, which contain eight G-tracks, are expected to form two intramolecular G-quadruplex subunits connected by a (TAJ) linker.

The ESI-MS spectra recorded for all six sequences are shown in Figure 1. The two shorter DNA sequences used for the study were either identical (DNA12) or very similar (DNA22) to the sequences studied previously by ESI-MS and ion mobility spectrometry. Optical DNA12 (Figure 1a) was found to form a dimer, but the dimer is much less abundant than the monomer. DNA22 (Figure 1b) was detected mainly as a monomer, but broad peaks corresponding to a dimer can also be seen when zooming on the spectra. The presence of dimer of DNA22 was confirmed by experiments on an electrospray Synapt HDMS instrument, where ion mobility separation allows better differentiation of monomer and dimer signals (data not shown). The DNA45 sequence was also detected exclusively as a monomer (Figure 1c).

The RNA telomeric sequences behave very differently from their DNA counterpart. RNA12 (Figure 1d) forms a dimeric quadruplex much more readily than DNA12, but the most striking difference between them is that the major RNA12 self-assembled structure is a tetramer in ammonium acetate. This sequence has been previously studied by Xu and coauthors using MALDI-TOF mass spectrometry in sodium chloride. The spectra showed mainly monomer and some dimer. Two reasons can explain this difference: either the tetramer can form in ammonium but not in sodium solution, or the tetramer can form in sodium as well, but does not survive the MALDI ionization conditions (50% acetonitrile in the sample preparation, and higher internal energies in MALDI than in electrospray).

In the case of RNA22 (Figure 1e), we found both monomeric (intramolecular) and dimeric G-quadruplex structures, the predominant species being the dimer. A trimer was also detected. The trimer became undetectable when the annealing was performed at 300 μM strand concentration or lower, but the dimer remained the major species even if the annealing was performed at 10 μM strand concentration. The dimer formation by the same RNA22 sequence in KCl has previously been reported by Qi and Shafer, from gel electrophoresis and equilibrium ultracentrifugation experiments. We also performed native gel electrophoresis experiments that confirmed the formation of dimers by RNA22 in sodium, potassium, and ammonium ionic conditions (Supporting Information Figure S1).

The DNA45 (Figure 1c) and RNA45 (Figure 1f) sequences were both detected as monomers. These monomers can contain two G-quadruplex units; the formation of these G-quadruplex subunits can be determined by counting the number of ammonium ions selectively trapped in the structure. Native gel electrophoresis experiments were performed on samples left over from the ESI-MS analysis of Figure 1. The gels (Supporting Information Figure S2) reveal the presence of even higher-order assemblies of RNA than could be detected by ESI-MS (either using the Q-TOF Ultima mass spectrometer, or using a Waters Synapt HDMS equipped with an 8 kDa quadrupole, data not shown). Note that the sensitivity of ethidium bromide staining in gel electrophoresis is proportional to the number of bases in the assembly, while the sensitivity in ESI-MS is proportional to the molar concentration of each assembly. Moreover, the electrospray response of noncovalent assemblies usually decreases when the size increases. Another explanation could be

![Figure 1](image-url)
that higher-order assemblies survive in the gel, but fragment into dimeric assemblies even in the soft ESI-MS conditions used here.

The Number of Trapped Ammonium Ions Suggests the Cation-Mediated Interaction of Two G-Quadruplex Subunits. Figure 2 shows the distribution of the number of ammonium ions preserved inside each G-quadruplex form, at selected charge states, as a function of the source RF lens 1 voltage. Low voltage (60 or 100 V) corresponds to soft conditions where inner ammonium ions are more likely to be retained in the interior of the G-quartet core. Increasing the voltage induces more energetic collisions in the source, which cause the loss of these inner ammonium ions. The exact mechanism of ammonium ion loss in the gas phase is still unknown, but it involves proton transfer from the ammonium ions to the oligonucleotide that results in the loss of neutral ammonia (there is no change in total charge of the oligonucleotide ion).

The first row (Figure 2a—d) shows the strand stoichiometries corresponding to a single quadruplex unit (dimeric for the 12-mers, monomeric for the 22-mers), at charge state 5−. Each quadruplex subunit has three contiguous G-quartets, and is expected to preserve two ammonium ions. For the DNA 12 dimeric G-quadruplex (Figure 2a), two ammonium ions are preserved at 100 V, and lost at 140 V, while for the RNA counterpart (Figure 2b), the two ammonium ions are much more stable. The difference is even more striking when comparing DNA22 (Figure 2c), for which it is impossible to preserve the two ammonium ions even at 60 V, and RNA22 for which two ammonium ions are still preserved up to 140 V (Figure 2d) and even 180 V (not shown). A recent comparison of RNA and DNA quadruplexes, using crystallographic and modeling methods, provides a possible explanation for the significant differences in stability of the RNA and DNA quadruplexes seen in the ESI-MS spectra.18 The study identified the C2' hydroxyl groups of the RNA to play a key role in RNA quadruplex stability, which were seen to preferentially form intramolecular hydrogen bonds with the RNA itself (as opposed to intermolecular interactions with water molecules) thereby stabilizing the quadruplex through an extended intramolecular hydrogen bonding network.

The second row (Figure 2e—h) shows the strand stoichiometries corresponding to two quadruplex units, at charge state 8−. For these larger species, there is a broad distribution of ammonium ion adducts at low voltage, suggesting either the presence of alternative folds with additional specific coordination sites, or nonspecific cation adduction which is typical of ESI-MS. The number of ammonium ion adducts and how they evolve in response to increased voltages does not give information on the ammonium ion binding sites, but it allows the number of weakly bound and tightly bound ammonium ions to be determined. At 140 V, only the most tightly bound ammonium ions can be preserved. For the DNA 45-mer (Figure 2e), all ammonium ions are quite labile. The preferred stoichiometry for the ammonium ions in the native state lies between three and five, but cannot be precisely determined. Ammonium ion stability is much higher in the RNA counterparts containing two G-quadruplex subunits. For the unexpected higher-order RNA assemblies formed by RNA12 and RNA22, the most stable adduct peak (i.e., the one which is preserved at high voltage) corresponds to four ammonium adducts. At charge state 8−, the adduct with five ammonium ions seems to be significantly

Figure 2. Distribution of the number of ammonium ions preserved in the structure of (a) dimeric DNA12, (b) dimeric RNA12, (c) intramolecular DNA22, (d) intramolecular RNA22, (e) intramolecular DNA45, (f) tetrameric RNA12, (g) dimeric RNA22, and (h) intramolecular RNA45. The distribution is shown at two RF lens voltages for each species, and the number of ammonium ions are indicated on each peak.

J. AM. CHEM. SOC. • VOL. 132, NO. 27, 2010 9331
stable (Figure 2g at 100 V), but increasing the collision energy in the source leads to one ammonia loss, leaving the dimer with four ammonium ions (Figure 2g at 140 V). Further voltage increase leads to the complete loss of the ammonium ions, two at a time. The four-ammonium ion stoichiometry suggests two ammonium ions per subunit. To check this structural interpretation, product ion spectra of the RNA22 dimer were recorded for the more highly charged dimer \([M+2\text{NH}_4]^+\). The separation into the monomers is observed instead of ammonia loss owing to greater Coulombic repulsion (Supporting Information Figure S3). Remarkably, a significant fraction of the monomers keep two ammonium ions. These observations are consistent with RNA22 comprising two rather distinct quadruplex subunits, each containing two ammonium ions, and hence presumably three G-quartets.

Finally, for the RNA45 sequence (Figure 2h), the adduct with five ammonium ions has particularly high stability, even at 140 V and higher. Importantly, this particular stoichiometry was observed for all charge states, up to 10− (Supporting Information Figure S4). As for the RNA22 dimer, the ammonium ion loss pattern for RNA45 is 5 → 4 → 2 → 0, consistent with two G-quadruplex subunits, each having two ammonium ions, plus one ammonium ion tightly stacked between the two subunits. The results therefore suggest a structural model for the higher-order assemblies involving the cation-mediated stacking of G-quadruplex subunits (Figure 3).

Higher-Order RNA Assemblies Formed by Annealing in Ammonium May Not Be Strictly Parallel G-Quadruplexes. The remarkable preservation of the ammonium ions at high voltage in each RNA G-quadruplex subunit, and not in each DNA G-quadruplex subunit, is one further example of what appears to be a general trend in the ESI-MS of G-quadruplexes. When sprayed from ammonium acetate solutions, ammonium ions inside mostly parallel G-quadruplexes (e.g., \([TGGGGT]_n\)28,32 Pu22myc and Pu27myc intramolecular quadruplexes33) survive in the gas phase even at relatively harsh conditions, whereas ammonium ions inside antiparallel G-quadruplexes (e.g., \([G_3\cdots T_2G_1\cdots T_2]_n\)28,29 telomeric DNA34,35) are less stable and more easily lost upon gentle collisional activation.

This can be explained as follows. In solution, one factor influencing the parallel/antiparallel topology of a telomeric quadruplex structure is the nature of the cation.38 Potassium ions may favor parallel structures because they remain between the G-quartets, whereas sodium ions favor antiparallel structures because they can more readily move through the G-quartet stem. Ammonium ions show intermediate behavior.39 In ESI-MS, the probability of ammonia loss in the gas phase is dependent on the probability that a proton-accepting group from the oligonucleotide is located at the exit of the G-quadruplex stem. This can explain why ammonia loss is more likely in antiparallel structures, because the loops, which are stacked onto 3' or 5' ends of the quadruplex stem, can serve as proton acceptors.

Because the inner ammonium ions are exceptionally stable in all detected RNA G-quadruplexes, and because the structure of telomeric RNA has been reported to be parallel in both sodium and potassium solutions,12,13,18 we anticipated a CD signature of purely parallel-strand orientation for the RNA sequences with ammonium ions. The CD spectra of the annealed structures are shown in Figure 4, and the thermal difference spectra of the same samples are shown in Supporting Information Figure S5. The CD confirms that DNA12 is predominantly in an unstructured form, whereas DNA22 and DNA45 are forming mixed or hybrid, predominantly antiparallel G-quadruplexes (highest maximum at 295 nm).

All RNA sequences adopted predominantly parallel structures (highest maximum at 260 nm, minimum at 240 nm), but surprisingly, a significant maximum at 300 nm was also observed for RNA22 and RNA45, suggestive of a non-negligible fraction of base steps in alternating syn—anti conformations. This could be due to a partially antiparallel arrangement of the guanine steps, or to particular arrangements and/or stability of the UUA loops. The relative intensity of the CD band at 295 nm is higher than in all previously published CD spectra of annealed RNA22, or of closely related 21-nt or 23-nt sequences.
on the self-assembly kinetics of the tetramolecular G-quadruplex zation. To obtain further insight into the mechanism of dimer formation by RNA22, we investigated the kinetics of dimer formation by annealing. This shows that, contrary to the common belief that they form exclusively parallel structures, telomeric RNA sequences are also significantly polymorphic under these conditions. This also concurs with the NMR study of a 23-nt telomeric RNA in potassium, which showed broad and unresolved spectra in the imino proton region.16

The substantial effect of the kinetics of multimer formation is also important for studying the thermal stability of telomeric RNA G-quadruplexes by thermal denaturation experiments.12,13,16,21 Previous studies have shown that telomeric RNA G-quadruplexes are significantly more stable than their DNA counterpart in potassium,13,16,21 but not in sodium solution.16,21 Substantial hysteresis in the denaturation curves of RNA22 in potassium has previously been noted,21 but the origin of this hysteresis was not elucidated. Here we performed thermal denaturation experiments on the structures annealed in ammonium acetate solution, and we can use ESI-MS to interpret the results in terms of multimer formation.

Hysteresis was observed in ammonium solution for all three telomeric RNA sequences (Figure 6). No hysteresis is observed for their DNA counterparts (Supporting Information Figure S6). For RNA22, in the light of the kinetic study at room temperature, the hysteresis can be attributed to slow dimer formation. The bimodal cooling curve can be interpreted by assuming a fast monomer refolding, followed by slower dimerization and rearrangement. The fact that the heating curves of the first cycle (curve 1) and the second cycle (curve 3) are not superimposed indicates that the structure formed by 0.2 °C/min cooling of the dilute solution (here 10 μM final concentration) is different from that formed by annealing at 300 μM strand concentration, but it can also be due to partial strand degradation.

The hysteresis is even more pronounced for RNA12. The starting structure is tetrameric (Figure 1d), and curve 1 leading to the unfolding into single strands appears to be multimodal, indicating a complex dissociation mechanism. There is very large hysteresis between curves 1 and 2, and refolding only occurs at low temperature. Surprisingly, curve 3 superimposes

\[ \text{(dTG}_{3}T)_{n}. \]

The ammonium ion count in the dimer (Figure 5b) reveals another difference with the annealed structure. In the dimer formed by incubation at room temperature, the adducts with five ammonium ions and more are the only detectable species, while in the annealed dimer (Figure 5c), the ammonium adduct distribution shows also a four-ammonium ion adduct. These results suggest that the dimer that is initially formed at room temperature consists of a cation-mediated stacking of two monomers, and that the dimer formed by annealing has undergone a more complex rearrangement of the strands.

The ESI-MS results show that, by varying the incubation time, it is possible to study the structure of the monomer (<10 min), or the stacked dimer (longer times at room temperature), and compare them to the annealed structure. Figure 5d shows the evolution of the CD spectra as a function of the incubation time at room temperature. The monomer which is folded after 3 min adopts an almost purely parallel structure. As the incubation time increases, the band indicative of parallel G-quadruplex arrangements (260 nm) further increases, and the smaller band indicative of antiparallel arrangement (300 nm) also grows at the same rate. However, in all cases, the monomers and dimers formed at room temperature are more parallel than those formed by annealing. This shows that, contrary to the common belief that they form exclusively parallel structures, telomeric RNA sequences are also significantly polymorphic under these conditions. This also concurs with the NMR study of a 23-nt telomeric RNA in potassium, which showed broad and unresolved spectra in the imino proton region.16

Figure 5. Kinetics of RNA22 folding and dimerization. (a) Time evolution of the dimer concentration followed by ESI-MS (140 V on the RF lens 1, total strand concentration = 10 μM, no methanol). Time t = 0 corresponds to ammonium acetate addition to the RNA solution. The red line is a fit with first-order kinetics (dimer formation rate of 0.017 min⁻¹). (b) Ammonium ion distribution on the M²⁺ ion of the dimer, obtained after 50 min incubation at room temperature. (c) Ammonium ion distribution on the M²⁺ ion of the annealed dimer, with the identical source conditions as in panel b. (d) Time evolution of the CD spectrum of RNA22 after cation addition to 10 μM unfolded monomer.
with curve 2, indicating that the structure formed by the cooling cycle is very different from the starting structure. We used ESI-MS to follow the G-quadruplex formation from RNA12 when incubated in ammonium acetate at room temperature (Supporting Information Figure S7). At 5 µM strand concentration, the dimer is formed in 20 min, but formation of the tetramer has not yet been initiated. We can therefore conclude that curve 1 corresponds to a multistep dissociation of the tetramer, and curves 2 and 3 correspond to the folding/unfolding of the dimeric G-quadruplex.

Interestingly, in the case of RNA45, although the sequence only folds intramolecularly, hysteresis is also observed. Notably, the denaturation is bimodal, while the renaturation is monomodal. Time-resolved ESI-MS analysis of the cation incorporation in RNA45 revealed that all five ammonium ions are incorporated in less than 1 min (Supporting Information Figure S8). It can, therefore, be concluded that the refolding and cation-mediated stacking is fast, and therefore that the hysteresis arises from denaturation which is slower than the temperature gradient step. This emphasizes the extra stability provided by the cation-mediated stacking of the neighboring G-quadruplex subunits.

Conclusions

Telomeric RNA sequences form dimers of G-quadruplex subunits more readily than the equivalent DNA sequences under the conditions of the experiments reported here. A 5'-5' stacking of two G-quadruplex subunits was previously reported for the telomeric RNA 10-mer GGUAGGAGGU,44 and our observations suggest that the formation of higher-order assemblies by telomeric RNA is a more general phenomenon, occurring also in longer sequences. ESI-MS also shows that the dimers in the longer sequences are stabilized by cation-mediated stacking of two G-quadruplex subunits. We note that 5'-5' quadruplex–quadruplex stacking has been observed in the crystal structures of both an intramolecular parallel and a dimeric parallel human telomeric DNA quadruplex,40 as well as in the crystal structure of a parallel dimer of the human telomeric RNA quadruplex sequence 5'GAGGGUGUAGGU.41 Suggesting that such stacking is also a consequence of the crowded conditions in the crystalline state, and possibly also in a cellular environment.41,42 However, here stacking was found in dilute aqueous solutions, in the absence of crowding agents or of alcohols.

RNA G-quadruplexes are usually considered to form only parallel G-quadruplexes with all-propeller loops, because the C3'-endo pucker for the ribose sugars constrains the guanosine nucleotides to be in the anti-conformation. Our ESI-MS and CD studies of annealed and RT incubated structures of RNA22 show that the structures adopted are more polymorphic than initially assumed, multimer formation being one of the features of polymorphism. This explains why telomeric RNA analogues with four G-quartets give unresolved NMR spectra. We also suggest that, because of the formation of stable intramolecular dimers, telomeric RNA sequences containing at least eight guanine repeats could be better models for analyzing the structure of long telomeric RNA. The 45-mer should be the shortest sequence to be considered for mimicking the structural arrangement of long telomeric RNA. Future work on longer sequences is needed to determine how these 45-mer subunits interact with each other.

Our results also highlight the influence of the sample preparation procedure on the structures obtained. We therefore suggest that future reports on RNA G-quadruplexes should include detailed information on the concentration–temperature-time profiles during sample preparation. The influence on dimerization of the bases which are flanking the G-quadruplex subunits also merits further investigation.

Acknowledgment. This work was supported by the European COST action MP0802, by Cancer Research UK (Programme Grant C129/A10902 to S.N), by the School of Pharmacy (research studentship to G.W.C.), by the Fonds de la Recherche Scientifique-FNRS (CC 1.5.286.09. F to V.G., research associate position to V.G., postdoctoral fellowship to F.R.), and by the University of Liège (starting grant FRSD-08/10 to V.G.).

Supporting Information Available: Native polyacrylamide gel electrophoresis; CID experiments; UV spectrophotometry; thermal denaturation of annealed DNA sequences; kinetics data for RNA12 and RNA45. This material is available free of charge via the Internet at http://pubs.acs.org.

JA100345Z

A crystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex

Gavin W. Collie, Shozeb M. Haider, Stephen Neidle and Gary N. Parkinson*

Cancer Research UK Biomolecular Structure Group, The School of Pharmacy, University of London, 29-39 Brunswick Square, London WC1N 1AX, UK


Received February 24, 2010; Revised March 26, 2010; Accepted March 29, 2010

ABSTRACT

DNA telomeric repeats in mammalian cells are transcribed to guanine-rich RNA sequences, which adopt parallel-stranded G-quadruplexes with a propeller-like fold. The successful crystallization and structure analysis of a bimolecular human telomeric RNA G-quadruplex, folded into the same crystalline environment as an equivalent DNA oligonucleotide sequence, is reported here. The structural basis of the increased stability of RNA telomeric quadruplexes over DNA ones and their preference for parallel topologies is described here. Our findings suggest that the 2'-OH hydroxyl groups in the RNA quadruplex play a significant role in redefining hydration structure in the grooves and the hydrogen bonding networks. The preference for specific nucleotides to populate the C3'-endo sugar pucker domain is accommodated by alterations in the phosphate backbone, which leads to greater stability through enhanced hydrogen bonding networks. Molecular dynamics simulations on the DNA and RNA quadruplexes are consistent with these findings. The computations, based on the native crystal structure, provide an explanation for RNA G-quadruplex ligand binding selectivity for a group of naphthalene diimide ligands as compared to the DNA G-quadruplex.

INTRODUCTION

G-quadruplexes are non-canonical nucleic-acid structures with unusually high stability. This stability is derived in part from the stacking together of G-quartets, which are planar arrangements of four guanines held together by eight hydrogen bonds. These G-quartets stack through π-π interactions to form stable quadruplex motifs (1,2). The recently identified non-coding telomeric RNA's are composed of extended tandem r(UUAGGG) repeats, transcribed from telomere DNA sequences located at the terminal ends of chromosomes (3-7). The G-rich telomeric repeat-containing RNA sequences (TERRA/telRNA) have been shown to form stable parallel-stranded G-quadruplex structures in solution (8-10), analogous to their single-stranded DNA counterparts. TERRA molecules have regulatory roles in telomere maintenance and other regulatory functions in both yeast and mammalian cells (3-7). They were recently shown to directly associate with two core proteins of the Shelterin complex—telomere repeat factors 1 and 2 (TRF1 and TRF2) (11). The Shelterin complex is an important multimeric complex involved in telomere maintenance, and is located proximal to the 3'-end of the chromosome (12). Composed of several proteins, it forms a tight complex with double-stranded G-rich telomeric DNA, and is directly involved in recruitment of the enzyme telomerase to single-stranded 3'-ends (12). Additionally, G-rich TERRA sequences have the potential to directly interfere with telomere extension and maintenance: by hybridizing to the complimentary single-stranded C-rich telomeric DNA. This is transiently exposed during the replication process, thereby interfering with the replication machinery; by hybridizing to the C-rich template region of the RNA subunit of telomerase (hTR), it can directly disrupt telomerase function.

The current interest in G-rich telomeric DNA folded into quadruplex structures stems from its attractiveness as an anti-cancer target, linked to the ribonucleoprotein telomerase, an enzyme that maintains chromosomal integrity and is up-regulated in >85% of various human cancers types (13). Telomerase catalytic function requires hybridization between its RNA template sequence and the single-stranded 3'-end of telomeric DNA; this association can be impeded by the stabilization of this DNA into a higher order-structure. The hybridization equilibrium can be shifted by addition of small-molecule ligands acting as stabilizers of these higher-order structures. By inhibiting substrate binding, telomerase is down-regulated, so interfering with overall telomere maintenance. Selective molecules that bind and topologically trap G-quadruplex...
structures can prevent other important telomere regulatory proteins such as POT1 from binding single-stranded DNA, resulting in telomere attrition and eventually senescence or cell death (14–16). Critically, the extreme ends of mammalian chromosomes terminate with a single-stranded 3' overhang of ~120–150 nt (17). With no complimentary C-rich strand, this region is free to form G-quadruplex structures and influence overall genomic stability, cellular division and cellular replicative lifespan (18). Thus, the overhang has been extensively studied as a therapeutic target. TERRA molecules have the same sequence as the repeats of single-stranded telomeric DNA, and are similarly free to self-associate and form higher order G-quartet based quadruplex motifs. An understanding of ligand selectivity between RNA and DNA quadruplexes, and thus of RNA quadruplex architecture, will be important for the future design of selective telomere targeting agents.

We report here the first crystal structure of a quadruplex formed from human telomeric RNA (TERRA). This provides fine detail of RNA quadruplex folding, such as water structure, groove widths and specific hydroxyl group interactions, as well as being a basis for the design of selective RNA quadruplex-targeted ligands. The chosen sequence contains two G4 runs, linked by a UUA sequence and flanked on the 5’- and 3’-ends by uracil and adenosine bases. Previous crystallographic analysis of the DNA sequence d(UgrAGGGUAgrTAGGG T) revealed a folded, parallel-stranded quadruplex topology, with the two strands associating to form a bimolecular quadruplex (19). We hypothesized that an analogous RNA sequence r(UgrAGGGUAgrGGU) should also crystallize; this was synthesized and indeed has crystallized. This RNA structure displays the same overall topology and structural arrangement as its DNA analogue, with the expected C3'-endo sugar puckers for many of the ribose sugars, albeit with a modified hydration structure. The influence of the hydroxyl groups on the ribose sugars in this context will be described in detail, particularly in the light of recent modelling and mass spectrometry studies which have identified several quadruplex-binding ligands with RNA versus DNA selectivity (20).

MATERIALS AND METHODS

Crystallization and data collection

The oligonucleotide sequence, r(UgrAGGGUAgrGGU), was purchased from Eurogentec (Belgium) and used without further purification. Mass spectrometry showed it to comprise a single species. A stock solution was prepared by dissolving the RNA in RNase-free water to a final single-strand concentration of 3 mM. For crystallization trials, the RNA solution was further diluted to 1.5 mM (ssRNA) and annealed in 30 mM KCl and 30 mM potassium cacodylate (pH 6.5) by heating to 90°C for 5 min, followed by cooling to room temperature overnight. RNA quadruplex crystals were grown by the hanging drop vapour diffusion method. One microlitre of the annealed RNA quadruplex was mixed with 1 µL reagent solution, composed of 15% MPD, 150 mM NaCl, 50 mM sodium cacodylate (pH 6.5) and 5 mM spermine. The hanging drop was equilibrated against a well containing 15% MPD. Crystals appeared overnight and were left to grow for 1 week at 12°C. A dataset was collected at 105 K on a single flash-frozen crystal, on an Oxford Diffraction Xcalibur NovaT X-ray diffractometer. The data were processed and scaled using CrystalPro (Oxford Diffraction) and Scala (from the CCP4 suite) (21). The crystals were assigned to space group P3121, with unit cell dimensions a = b = 57.58 Å, c = 38.38 Å, α = β = 90°, γ = 120°.

Structure determination and refinement

The structure was solved by molecular replacement using the Phaser program (22) (CCP4), using the bimolecular DNA quadruplex structure, PDB 1K8P, as a model, and refined using Refmac5 (23). Data collection and refinement statistics are shown in Table 1. The RNA model was generated by modification of the DNA structure (addition of hydroxyl groups at the C2' position and deletion of methyl groups from the C5 position of thymine). The three guanine quartets, bromine atoms and K+ ions could be clearly seen in the initial σA-weighted 2Fo-Fc electron density maps, while the initial Fo-Fc difference electron density maps showed residual density for the bromine atoms and the propeller loop of strand A. The RNA model was initially built into maps calculated from low-resolution data, and extended and refined into maps calculated with progressively higher-resolution data, through iterative cycles of manual adjustment with maximum-likelihood restrained refinement, using Coot (24) and Refmac5 (23). Electron density for the hydroxyl groups could also be clearly seen in the initial σA-weighted 2Fo-Fc electron density maps, allowing these

<table>
<thead>
<tr>
<th>Data collection</th>
<th>Sequence</th>
<th>r(UgrAGGGUAgrGGU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cell dimensions: a, b, c (Å)</td>
<td>57.58, 57.58, 38.38</td>
<td></td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.00, 90.00, 120.00</td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>49.87–2.20</td>
<td></td>
</tr>
<tr>
<td>Rw (%) overall</td>
<td>5.0 (24.0)</td>
<td></td>
</tr>
<tr>
<td>Values in brackets refer to the highest resolution shell, 2.28–2.20 Å.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
features to be accurately placed at an early stage of the solution process, and ensuring the sugar puckers and backbone could be modelled accurately. A translation/rotation/screw (TLS) motion determination (25) approach was used in the final stages of refinement in an attempt to fit the propeller loop of strand B (residues U18, U19, A20) into density. The RNA model was divided into four groups (two per strand) for the TLS refinement. Although the TLS refinement improved the accuracy of the model, the majority of the UUA loop of strand B could not be fully fitted into density. The occupancy of this loop has therefore been set to 0% and has been assigned geometry based on the A-strand linking loop (U6, U7, A8). The phosphate group (atoms P, OP1 and OP2) for residue U18 could be fitted into density and has been given full occupancy in the model. The final model (including solvent molecules) was refined using data between 11.69 and 2.20 Å, with final R and R_free values of 21.6 and 23.1% respectively.

Modelling

In order to explore the effects of dynamics on structural stability and inter- and intra- molecular interactions, multiple multiananosecond molecular dynamics (MD) simulations were performed on the crystal structure of this 12-mer RNA quadruplex, sequence U8AGGGUUA GGGU. The X-ray structure has a vertical column of consecutive K+ ions within the central core, sandwiched equidistant between G-quartets, which was retained for the simulations. Explicit solvent (TIP3P; 20829 waters) and counter ions (20 K+ ions) were added such that the total charge was zero. Energy minimization (steepest descent algorithm) was run for 5000 steps, followed by a 500-ps equilibration phase during which the quadruplex was harmonically restrained with a force constant of 1000 KJ mol⁻¹ nm⁻² at 300 K, which was gradually relaxed until no restraints were applied. During the entire process, the ions and the solvent were allowed to equilibrate. The final production run was carried out without any restraint on the system using the GROMACS program, version 4.0.2 (26) (http://www.gromacs.org) with the AMBER ff99 and parmBsc0 forcefield (27) that has been ported into the GROMACS MD suite (28). Additional details of modelling methods can be found in the Supplementary Data.

The specificity of ligand interactions with the RNA quadruplex was examined by means of three additional MD simulations with bound ligands (Supplementary Figure S1). Molecular models of the napththalene diimide series of quadruplex-stabilizing ligands (29) were constructed; minimized and partial charges calculated semi-empirically using the MOPAC program (30) as implemented in the Insight II suite software (http://www.accelrys.com). Prior to docking, terminal uracil and adenine residues were removed in order to fully expose the flat planar quartet to the ligand. Several uracil and adenine terminal 3' and 5' residues are involved in crystal packing interactions but are not integral to the stabilization of this quadruplex arrangement. Of the two uracils at the 3-ends, one was not observed while the second is observed stacking through crystal packing with the loop of a symmetry related quadruplex. The ligand structures were minimized and docked on the 3' surface of the quadruplex structure using the AFFINITY docking program (http://www.accelrys.com) employing the grid docking method available with AFFINITY. This approach has been previously validated in our studies on the rational design of quadruplex ligands (31,32). The G-quartets were frozen during the entire docking procedures since they have been shown to be exceptionally stable and structurally rigid, as confirmed by several experimental and simulation studies (19,33-42). The automated docking process identifies and ranks positions based on interaction energies within the binding site. The final conformation of the complexes was then subjected to a further 500 steps of unrestrained molecular mechanics minimization (steepest descent algorithm). The force-field parameters for ligands were generated using the AMBER force field f99 (43) and subsequently ported to GROMACS format using amb2gmx perlscript (44). The three production runs were each performed for 20 ns.

A MD simulation of the equivalent 12-mer DNA quadruplex (PDB 1K8P; resolution 2.40 Å) was performed for 18 ns in order to highlight the differences between RNA and DNA quadruplexes. The simulation protocols were consistent for all systems. Trajectories were analysed using GROMACS and locally written scripts and the data visualized by means of the VMD program (45) with graphs plotted using the Xmgrace program (http://plasma-gate.weizmann.ac.il/Grace/).

RESULTS

Topological conservation between RNA and DNA G-quadruplexes

In the crystal packing arrangement for the P321 space group, the asymmetric unit contains two RNA strands (A and B) of sequence r(U8AGGGUUA GGGU). Each RNA strand contains two G-rich repeats, folding to form a parallel four-stranded bimolecular G-quadruplex. The intervening UUA nucleotides link the phosphate backbone through an external propeller-type arrangement, keeping the phosphate backbone strand associations parallel. The structure contains three stacked G-quartets, with two clearly identified K+ ions in the centre of the channel, positioned between the three G-quartets (Figure la and b) and coordinated in a bipyramidal antiprismatic manner to the O6 atoms of each guanine. The overall quadruplex fold seen here is identical to that of the equivalent DNA quadruplex (19) (1K8P) and is in agreement with the NMR structure reported for an unmodified RNA sequence (10). A comparison of the root mean square (rms) deviations between the X-ray derived RNA quadruplex and the NMR derived structure is shown in Table 2. There is very little difference between the central G-quartets of the three structures (Figure 1c and d), however there are significant deviations between the structures when comparing the coordinates for the overall structures. These differences are primarily related to the diverse set of conformations adopted by the
Figure 1. Cartoon representation of a bimolecular quadruplex formed from human telomeric RNA contained within one asymmetric unit. The UUA loop for strand B is disordered, with the connecting backbone atoms modelled on the weakly observed residual electron density. The colouring scheme is based on nucleotide type, with guanines green, modified and non-modified uracils blue and adenines red. (a) Highlights the stacking of the UUA loops with the central uracil perpendicular to the stacked guanines, (b) shows the stacking of the uracils and adenines on to the G-quartets. (c) Overlay of the bimolecular RNA quadruplex (PDB 3IBK) (blue) with the equivalent bimolecular DNA quadruplex (PDB 1K8P) (red), (d) overlay of the bimolecular RNA quadruplex (PDB 3IBK) (blue) with the NMR model of the same sequence (PDB 2KBP, model I) (yellow). The molecular graphics program PyMOL was used for structural alignments, calculation of rms deviations and visualization (http://www.pymol.org).

Table 2. Comparison of the rms deviation between the two RNA models [2KBP determined by NMR (model 1) and 3IBK determined by X-ray diffraction], and between the two crystal structures ([1K8P-DNA quadruplex and 3IBK-RNA quadruplex)

<table>
<thead>
<tr>
<th></th>
<th>Whole model</th>
<th>Quarts only</th>
</tr>
</thead>
<tbody>
<tr>
<td>3IBK v 1K8P (DNA)</td>
<td>0.990 (470 atoms)</td>
<td>0.539 (232 atoms)</td>
</tr>
<tr>
<td>3IBK v 2KBP (NMR)</td>
<td>2.971 (473 atoms)</td>
<td>1.011 (246 atoms)</td>
</tr>
</tbody>
</table>

The rms deviations are generated from structural alignments, which can be seen in Figure 1c and d.

UUA loops of strand A and B for the NMR structure as reported in the ten deposited NMR-derived coordinate sets (10).

Crystal packing interactions

The two 5'nts U6 and A2 associate together in the crystal lattice through a 2-fold crystallographic symmetry element forming two stacked U6/AU6/A* tetraads, to generate a 5' to 5' stacked arrangement of two quadruplexes. The bromine atom attached at position C5 on the uracil base is a required modification, enhancing crystal diffraction quality by holding the dimer interface tightly together. The two UUA loops have different packing interactions, where the trinucleotide linking loop of strand A (U6, U7 and A8) is well ordered, folding as a stacked UUA loop, similar to strand A in the bimolecular DNA quadruplex (1K8P), whereas the UUA loop from strand B is not sufficiently ordered to model the position of the three nucleotides (Figure 1a and b). This overall lack of ordering in the B strand loop is not the result of an unfavourable crystal packing arrangement as space is available for folding into a conformation seen in the strand A loop.

Sugar puckering variation

As expected for a RNA with 2'-OH groups, the majority of the RNA quadruplex ribose sugar puckers can be classified as being in the C3'-endo conformation, including the 5'nts U6,1 and A2, and the central guanines of the G-quartets (Table 3). Uniform C3'-endo puckering is not observed throughout the structure however, as the ribose sugars of the loop residues and the guanines within the terminal G-quartets adopt a variety of pucker conformations. This mix of C3'-endo and C2'-endo puckering is consistent with the NMR model and with the DNA crystal structure (2KBP and 1K8P respectively), and was also observed in the r(UGGGGU) intermolecular
Table 3. Sugar pucker conformations for the 12-mer DNA crystal structure (1K8P), the 12-mer RNA crystal structure (31BK) and the 12-mer NMR structure (2KBP, model 1)

<table>
<thead>
<tr>
<th>Residue</th>
<th>Crystal structures</th>
<th>NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RNA (31BK)</td>
<td>DNA (1K8P)</td>
</tr>
<tr>
<td></td>
<td>Strand A</td>
<td>Strand B</td>
</tr>
<tr>
<td>1 U 13</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>2 A 14</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>3 G 15</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>4 G 16</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>5 G 17</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>6 U 18</td>
<td>C4'-exo</td>
<td>C4'-exo</td>
</tr>
<tr>
<td>7 U 19</td>
<td>C4'-endo</td>
<td>C4'-endo</td>
</tr>
<tr>
<td>8 A 20</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>9 G 21</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>10 G 22</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>11 G 23</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>12 U 24</td>
<td>C1'-endo</td>
<td>C3'-endo</td>
</tr>
</tbody>
</table>

Linking U6 U7 and A8 residues for the B strands have been excluded for the crystal structures. Sugar puckers consistent between 31BK and 2KBP are shaded light grey. Sugar puckers consistent between 31BK and 1K8P are shaded dark grey. The sugar puckers are defined by the combined pseudorotation phase angle.

Hydroxyl group interactions

The RNA ribose sugars, with their 2'-OH groups, have the potential to make many more inter- and intra-molecular interactions than their DNA counterparts. In this quadruplex arrangement, the hydroxyl groups are seen to be preferentially forming intramolecular interactions with the quadruplex itself, rather than intermolecular contacts with solvent molecules. The resulting RNA quadruplex has fewer observed structured water molecules than the equivalently packed DNA quadruplex, and is also accompanied by an increase in intramolecular contacts. The 2'-OH hydroxyl groups are seen to form hydrogen bonds with a variety of acceptors within the RNA molecule, and interestingly, the type of intramolecular interaction appears to be dependent on the pucker conformation (Figure 4a). For example, within the quadruplex grooves, the hydroxyl groups of those sugars with a C3'-endo pucker are positioned such that they can form a hydrogen bond with the O4' oxygen atom of the neighbouring residue. Alternatively, the hydroxyl groups of the sugars with a C2'-endo pucker preferentially hydrogen bond to the N2 amine group of the guanine base. For example, the hydroxyl group of residue G3, with a C2'-endo pucker, forms a hydrogen bond with the N2 amine group of G22 (circled black in Figure 4a). This interaction is also seen between the hydroxyl group of G9 and the N2 amine group of G4.

MD simulations: stability and ligand binding

In order to assess the comparative conformational stability of the telomeric RNA quadruplex, a 20ns MD simulation was performed on both the RNA 12-mer and the DNA 12-mer quadruplex structures (PDB IDs 31BK and 1K8P respectively). The root mean square deviations were monitored during the course of the simulations, as a measure of structural stability. The results of these simulations indicate that the 12-mer RNA quadruplex...
Figure 2. Comparison of torsion angles in bimolecular quadruplex crystal structures. RNA (PDB 3IBK), solid black trace with squares; DNA (1K8P), grey dashed trace with diamonds. Dihedral angles are plotted separately; in order from (a) to (g) alpha, beta, gamma, delta, epsilon, zeta, chi. Significant deviations between the two structures are circled (black). In all the graphs, x-axis shows residue and the y-axis shows torsion angle.

Figure 3. Comparison of sugar puckering and backbone dihedral angles of residue G10 between DNA (a) and RNA (b) quadruplexes. The deoxyribose sugar of the DNA quadruplex adopts a C2'-endo conformation. A C2'-endo conformation for the RNA at this position would cause the C2' hydroxyl group to sterically clash with the O3' oxygen atom. The sugar of the G10 residue within the RNA structure therefore adopts a C3'-endo conformation (b), accompanied by alterations in backbone dihedral angles which preserves the positioning of the guanine base within the quartet.
structure is more stable than the 12-mer DNA analogue, with rms deviations of 1.90 Å and 2.62 Å for the RNA and DNA respectively (Supplementary Figure S1). It is notable that the propeller topology of the UAU loop in the RNA quadruplex was maintained throughout the simulation, which was in contrast to the DNA simulation, during which the loops un-stack and open up. The C2' hydroxyl groups can be seen during the course of the RNA MD simulation to interact with O3' atoms (within the same residue), O4' and O5' atoms (of the adjacent residue in the backbone) and with N2 amine groups of the guanine bases (Supplementary Figure S2). In accordance with the crystal structure, the MD simulations show that the 2'-OH-N2 interactions are dependent upon the sugar adopting a C2'-endo pucker, and also show that this interaction is lost as the sugar conformation changes (Figure 5 and Supplementary Figure S3). The MD simulations of the RNA G-quadruplex further highlight the ability of the 2'-OH groups to make multiple interactions within the loop during the 20 ns of the simulation. These interactions impart rigidity to the overall sugar-phosphate backbone, and also explain the reduced flexibility observed within the UUA loops.

A recent study using electrospray-ionization mass spectrometry identified several small-molecule compounds with the ability to select between telomeric DNA and RNA quadruplexes (20). In order to rationalize these experimental findings, we have performed a series of MD simulations on the bimolecular RNA quadruplex structure (31B2) bound with members of a group of three naphthalene diimide ligands. Details of the ligands used for the studies are shown in Supplementary Figure S4. MD simulations identified a ranking order for RNA quadruplex-ligand binding strength as follows: Ligand 3 > Ligand 1 > Ligand 2 that is in accord with the experimental data (20). The MD studies provide insight into the ligand binding mode, and can offer explanations for differences in binding affinity. The three naphthalene diimide compounds used for the present studies share a common core ring system, which is seen to stack effectively over the 3' G-quartet. The 2'-OH groups of the RNA quadruplex make effective interactions with the carbonyl groups of the central chromophore, as well as with nitrogen groups within the side chain of the ligands. The interactions between the chromophore and 2'-OH groups at the apices of the G-quartets effectively locks the ligands in place and maximizes π-π stacking interactions (Figure 6). A total of six hydrogen bonds are formed between the ligand chromophore and 2'-OH groups. As the naphthalene diimide ring system is a common feature
Figure 5. C2'-endo 2'-OH-N2 interactions. (a) A 20-ns snapshot of interactions formed between 2'-OH and N2 atom of guanine residues when C2'-endo sugar puckering is present. (b) 2'-OH-N2 interactions calculated over 20 ns simulations. Over the course of the 20 ns simulations, the 2'-OH groups makes multiple interactions with O4', O3', O5' and phosphate oxygen atoms. These interactions impart rigidity to the sugar phosphate backbone and are likely to contribute to the higher stability of RNA G-quadruplexes over DNA ones.

Figure 6. RNA quadruplex-ligand interactions studied by MD. Interactions between the side chain of ligands (a) Ligand 1, (b) Ligand 2 and (c) Ligand 3 and the 12-mer RNA G-quadruplex. The interactions are shown as dashed lines. The ligand (cyan) is stacked onto the 3' G-quartet (carbons shown in green). Comparison of the surface charge distribution between the RNA and DNA bimolecular quadruplexes reveals areas of the RNA structure with increased polarity, which may be of significance for ligand binding (Supplementary Figure S5).

of these ligands, differences in quadruplex-ligand binding affinities are a result of differing side-chain functionalities. These side-chains interact with the quadruplex loops, rather than the G-quartets themselves, providing an explanation for the experimental data, which show that all three ligands do bind to DNA quadruplexes with high affinity, but only one ligand (Ligand 3) binds to the RNA quadruplex with comparable affinity (20). MD
Correlations between sugar puckering and backbone geometries preserve G-quartet stacking

In stacked RNA structures, the ribose sugar puckers generally adopt the C3'-endo conformation, as seen in the present structure, where the majority of the sugars of the terminal residues have puckers in the C3'-endo conformation. However, for the stacked G-quartets, a more mixed puckering arrangement is observed, which is consistent with that observed in an intermolecular parallel stacked quadruplex of sequence r(UGGGGU) (46) albeit without the connecting loops. This mixing of ribose sugar puckers seen for the stacked guanines may reflect a conformational preference that is finely balanced between the requirements for stacking of the G-quartets and the steric hindrance caused by the addition of hydroxyl group at the C2' position on the ribose sugar. This balance of sugar pucker between C3'-endo and C2'-endo may be shifted by the addition of the connecting propeller loops and constraints in geometry in the connecting backbones.

Hydroxyl group interactions are linked to sugar pucker conformation, and appear to affect the hydration structure and overall stability of the RNA quadruplex

The sugar pucker arrangement in our structure for the guanine G3 stack has C2'-endo puckers at the ends and a C3'-endo for the central guanine, an arrangement consistent with the RNA quadruplex (PDB ID 2KBP) in NMR solution, but quite different from the two equivalent DNA quadruplex structures (PDB ID 1K8P and PDB ID 1KF1). The parallel-stranded quadruplex topology with a tightly folded propeller loop may impose additional steric constraints that need to be accommodated, such as inducing a C2'-endo conformation and being stabilized by additional 2'-OH–N2 hydrogen bonding. The modelling studies indicate that the propeller loops in the RNA have lower rms deviations over the time course of the simulation when compared to the modelling carried out on the 12-mer DNA, implying greater rigidity (Supplementary Figure S1). We infer that the additional hydrogen bonding arrangements derived from the 2'-OH groups are important contributors to stabilizing specific structural arrangements in this parallel-stranded topology.

The 2'-OH groups interact with a variety of hydrogen bond acceptors, including phosphate and backbone oxygens (O3', O4' and O5') and polar groups attached to the bases (such as the N2 amine group of the guanine base). With the additional hydroxyl groups, much of the internal hydrogen bonding of the RNA quadruplex is notably different to the equivalent DNA structure. In addition, these multiple intramolecular contacts of the RNA, which are observed both in the crystal structure and in our modelling studies (Supplementary Figure S2), would be expected to result in a reduction in associated water molecules, compared to the DNA (Figure 7). This data suggests to us that the extensive intramolecular contacts made by the C2'-OH groups makes a major contribution to the increased stability of the RNA quadruplexes over equivalent DNA structures. The overall loss of structural waters, the additionally constrained waters through hydrogen bonding and an increase in direct contacts through hydroxyl groups is consistent with recent biophysical studies (49) which showed the 24-mer telomeric RNA quadruplex (r(UUAGGG)) to have a ΔTm value 12.6°C higher than the equivalent DNA quadruplex (in K+).

DISCUSSION

Similarity of topologies

The high quality of the diffraction data in this 2.20 Å resolution structure has enabled a detailed view of the RNA core structure to be obtained, with its stacked G-quartets, along with one of the two UUA loops, hydration structure and detail for all other ordered components in the crystal lattice. In contrast to the diverse folded topologies seen for human telomeric DNA quadruplex sequences, especially with changes in ionic environment, the RNA fold observed in this X-ray structure is consistent with the solution NMR data in K+ buffer (10) and in Na+ buffer, as confirmed by NMR, CD and MALDI-TOF MS (9). The striking similarity between the RNA and DNA structures is most apparent when they are overlaid (Figure 1c). This consistency in topology can be explained by the marked preference in RNA for an anti glycosidic bond angle, a consequence of steric constraints imposed by the C2' hydroxyl groups (47); syn conformations for RNA nucleotides have been observed but only when contained within a very specific structural context, which includes quartet arrangements (48) for example with 8-bromo-guanine.

Correlations between sugar puckering and backbone geometries preserve G-quartet stacking

Simulations of the RNA showed that the 2'-OH groups make multiple interactions within the loop, thereby reducing the depth and width of loop dimensions when compared to the equivalent DNA structure (Figure 4b and c). As a result, the RNA quadruplex is less amenable to bind ligands with side-chains terminating in bulky and/or inflexible functional groups. The lack of flexibility of the pyrrolidine group of Ligand 2, combined with the inaccessibility of the loop, makes this ligand the weakest binder in the series. The substitution of the pyrrolidine group by -NHMe2 (Ligand 1) results in a functional group with greater flexibility, but which is still ineffective in participating in stabilising 2'-OH interactions with the loop. Substitution of -NHMe2 with the -OH group (Ligand 3) increases functional group flexibility and also contributes towards hydrogen bonding interactions with the loop. An MD simulation snapshot of the ligand side-chain-loop interactions shows that a maximum of two hydrogen bonds per loop can be formed between Ligand 2 and the loop (Figure 6b). Additional flexibility of the -NHMe2 group in Ligand 1 increases the ability of this molecule to pick up dynamic interactions during the simulation (Figure 6a). Replacement of the -NHMe2 groups within two side-chains having -OH groups (Ligand 3) increases the hydrogen-bonding ability of this ligand significantly, as the -OH group is seen to be involved in up to three hydrogen bonds (per side chain) with the quadruplex (Figure 6c).

Hydroxyl group interactions are linked to sugar pucker conformation, and appear to affect the hydration structure and overall stability of the RNA quadruplex
same work, osmotic stress analysis showed RNA quadruplexes to have fewer associated water molecules when compared to the DNA ones (49). Additionally, recent CD analysis and melting studies (50) on various putative quadruplex-forming RNA sequences have shown a marked preference for these sequences to form parallel topologies, with higher melting temperatures in potassium when compared to their DNA equivalents. It is notable that the DNA quadruplex crystals diffracted to a significantly lower resolution compared to the RNA ones which also has overall considerably lower individual atomic temperature factors, implying greater ordering in the lattice and a higher rigidity.

Modelling studies combined with structural data identify significant differences between RNA and DNA quadruplexes which may be useful in the design of selective quadruplex interacting ligands

Telomeric DNA quadruplexes have been extensively investigated as a target for anti-cancer therapies by the use of small molecule ligands (16), thus creating a wealth of biophysical and structural data accumulated over the last twelve years. This ligand-quadruplex data base will now have to be expanded to take into account an expanding diversity of targets generated by the discovery of TERRA sequences. The design of ligands that take into account DNA versus RNA quadruplex selectivity will have to be addressed. The studies presented here using differential analysis of dynamic interactions between ligands and RNA G-quadruplexes highlight some of the important structural features that should be considered when designing selective G-quadruplex stabilising ligands. For example, when comparing the naphthalene diimide ligands to DNA versus RNA selectivity, as well as binding affinities, the differences in association can be directly linked to the 2'-OH groups of the RNA. We observed that the 2'-OH groups make extensive interactions, both within the RNA molecule itself, by altering groove distances compared to the equivalent DNA structure, as well as with the ligands themselves, the 2'-OH groups being directed upwards onto the 3'-quartet external surface (Figure 6). This highlights the importance of hydrogen bonding groups within quadruplex-binding small-molecule ligands.

Based on our observations, selectivity between DNA and RNA quadruplexes may be achievable through the appropriate selection of chromophores and side-chains. We therefore propose that, with the recent discovery of TERRA and telomeric RNA quadruplexes, it may now be necessary to screen potential telomeric DNA quadruplex stabilizing ligands in parallel with the equivalent RNA sequences, to assess this specificity. The application of structural data has played a major role in the understanding of interactions between ligands and DNA quadruplexes, particularly with the observations of large conformational changes in loop structures to accommodate diverse families of ligands. For example, the extensive loop modification induced by TmPyP4 binding to a parallel-stranded DNA quadruplex (51) may also be relevant to parallel-stranded RNA quadruplexes and may explain the observed similarity in binding affinity and stoichiometry of TmPyP4 to RNA and DNA human telomeric parallel-stranded quadruplexes (49).

This similarity of loop structure for the connecting UUA loops in RNA and DNA in the crystal structures extends to almost identical backbone dihedral angles (Figure 2) even with the addition of the 2'-OH groups.
The loops are stabilized in this arrangement with the inclusion of a U6(2')-A8(7') hydrogen bond. However, the dimensions of the loops, as defined by width and depth, are significantly reduced with the addition of the 2'-OH groups (Figure 4b and c), which will affect the available space for ligand binding, at least in the absence of disruptions to the loop arrangement.

As the majority of RNA quadruplexes studied to date exhibit a parallel topology (8–10,49,50), small molecule selectivity between RNA quadruplex types [e.g. TERRA quadruplex versus mRNA quadruplex (52,53)] offers a particular challenge in quadruplex targeting. This may be an important consideration, as ligands designed to stabilize mRNA quadruplexes (e.g. for anti-cancer therapies) would need to be carefully designed to not bind to TERRA quadruplexes, as inadvertent stabilization of TERRA quadruplexes could potentially increase telomerase activity (by reducing the amount of hTR-template-RNA which is maintained in duplex form), although this is yet to be demonstrated experimentally. Sequence-dependent connective loop structures, and their accessible conformational diversity, may be the main avenue available for ligands to discriminate between the many similar parallel-stranded RNA structures, and we suggest therefore that this should be the main consideration in the design of inter- and intra-molecular RNA quadruplex selective ligands.

DATA BANK ACCESSION CODES
The atomic coordinates and structure factors have been deposited in the Protein Data Bank (PDB) with ID code 3BK

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
Oxford Diffraction is thanked for the use of their collection facilities.

REFERENCES
Structural Basis of Telomeric RNA Quadruplex—Acridine Ligand Recognition

Gavin W. Collie, Silvia Sparapani, Gary N. Parkinson, and Stephen Neidle*

CRUK Biomolecular Structure Group, The School of Pharmacy, University of London 29-39 Brunswick Square, London WC1N 1AX, United Kingdom

ABSTRACT: Human telomeric DNA is now known to be transcribed into noncoding RNA sequences, termed TERRA. These sequences, which are believed to play roles in the regulation of telomere function, can form higher-order quadruplex structures and may themselves be the target of therapeutic intervention. The crystal structure of a TERRA quadruplex—acridine small-molecule complex at a resolution of 2.60 Å, is reported here and contrasts remarkably with the structure of the analogous DNA quadruplex complex. The bimolecular RNA complex has a parallel-stranded topology with propeller-like UUA loops. These loops are held in particular conformations by multiple hydrogen bonds involving the O2' hydroxyl groups of the ribonucleotide sugars and play an active role in binding the acridine molecules to the RNA quadruplex. By contrast, the analogous DNA quadruplex complex has simpler 1:1 acridine binding, with no loop involvement. There are significant loop conformational changes in the RNA quadruplex compared to the native TERRA quadruplex (Collie, G. W.; Haider, S. M.; Neidle, S.; Parkinson, G. N. Nucleic Acids Res. 2010, 38, 5569—5580), which have implications for the future design of small molecules targeting TERRA quadruplexes, and RNA quadruplexes more generally.

INTRODUCTION

The DNA component of telomeres in eukaryotic organisms, comprising tandem repeats of simple guanine-rich sequences, has long been assumed to be transcriptionally silent. In humans and other mammals this repeat is TTAGGG. The recent findings that the C-rich strand of telomeric DNA is transcribed into lengths of telomeric RNA (termed TERRA) were therefore totally unexpected. TERRA is composed of tandem rUAGGG repeats that vary between 100 and 1000 bp in length and may be involved in a number of key cellular processes, including chromatin regulation and remodelling. It has also been demonstrated that TERRA sequences are potent and direct inhibitors (and thus negative regulators) of telomerase function. Several TERRA-interacting proteins have been identified, notably the telomeric duplex DNA binding proteins TRF1 and TRF2, pointing to a significant role of TERRA in more general chromosome biology. The POT1 protein contains an OB fold, which binds to the single-stranded overhang repeats at the terminus of telomeric DNA and is unable to bind to TERRA. The telomere repeat sequences show a high propensity to form stable compact structures in vitro, analogous to the behavior of the equivalent G-rich telomeric DNA sequences and consistent with the formation of G-quartets as the packing motif. Solution studies on short (comprising two and four telomeric repeats) TERRA sequences have shown that these readily form highly stable G-quadruplex structures in both sodium and potassium ion containing solution. NMR analysis of the two-repeat 12-mer sequence rUAGGGUUAAGGGU has shown that this forms a highly stable parallel-stranded intermolecular bimolecular quadruplex, with the topology and propeller loop arrangement that has been previously observed for telomeric DNA quadruplexes in the crystalline state. A subsequent crystal structure of the same sequence has found the same topology and has shown that the presence of the O2' hydroxyl groups, together with the preference of the ribosugars for a C3'-endo pucker, plays a significant role in stabilizing the parallel arrangement. This consistency between crystal and solution topologies is in striking contrast to the situation with telomeric DNA quadruplexes, where there is marked topological variability depending on the environmental conditions.

The stabilization of telomeric DNA quadruplex structures with small-molecule ligands can result in the inhibition of the activity of the telomere-maintenance enzyme telomerase, whose expression is up-regulated in the majority of human cancer cells types. A large number of such ligands have been characterized, and several have been used to demonstrate that telomerase inhibition in cancer cells can result in selective inhibition of cell growth, and indeed in antitumor activity in vivo. The structures of several quadruplex DNA complexes with acridine-based ligands have been determined by X-ray crystallography.
crystallography and NMR methods, and these have facilitated rationalizations of the biophysical and biological data for these compounds.

These and other telomerase and telomere-targeting ligands have been designed and developed to interact exclusively with telomeric DNA quadruplex structures. The emerging pivotal roles of TERRA in telomere biology suggest that ligands with high affinity and selectivity to TERRA G-quadruplexes may be of therapeutic use. Biophysical data on a number of ligands has shown that selective binding between RNA and DNA telomeric quadruplexes is achievable.

Potential RNA G-quadruplex-forming sequences may be encoded elsewhere in genomes other than at telomeres. They have been identified in the 5' untranslated regions of a number of eukaryotic and prokaryotic genes, pointing to a regulatory role of such folded RNA structures in translational inhibition of gene expression. Translational inhibition may be enhanced by binding a small molecule to a 5'-UTR quadruplex in order to enhance its stability, as has been demonstrated for the human NRAS gene. Quadruplex-binding ligands can also inhibit dicer quadruplex RNA processes. There is as yet very little structural information on these RNA quadruplexes, but circular dichroism data suggests that a parallel fold is common to many of these structures, together with increased stability for RNA quadruplexes over the equivalent DNA structures.

We present here the crystal structure of a TERRA RNA quadruplex complexed with an acridine ligand extended in length by triazole-based click chemistry. This ligand is symmetrically substituted at the 3- and 6-positions with triazole-phenyl-diamine side chains (Figure 1a) and was originally designed to bind optimally to telomeric DNA quadruplexes. It shows selectivity over several other DNA quadruplexes, notably those found in the promoter sequence of the c-kit gene. Crystal structures of several ligand–DNA quadruplex complexes, together with theoretical calculations, have shown that the ligand planar surfaces are stabilized by interaction with a terminal G-quartet surface and that the loops can play a secondary role by interacting with substituents. In general loop conformations are highly conserved in native telomeric DNA quadruplexes, whereas ligands can trap particular conformations. The RNA quadruplex complex crystal structure presented here reveals a pattern of active loop involvement in ligand recognition that involves several major RNA-dependent changes in loop conformation, which have relevance to the future design of ligands.
targeting various categories of defined DNA and RNA quadruplexes.

■ RESULTS

Overall Arrangement of the Ligand Complex. The crystallographic asymmetric unit contains one human telomeric RNA strand of the sequence r(UAGGGUAGGGU) and one triazole-acridine ligand molecule. The biological unit is formed by a crystallographic twofold rotation axis and comprises a bimolecular G-quadruplex formed from two strands of the RNA sequence, complexed to two molecules of the ligand (Figure 1a). The quadruplex is parallel-stranded with linking UUA propeller-like chain-reversal loops. The two triazole-acridine ligand molecules are stacked side-by-side onto a single terminal S' G-quartet face (Figure 1b). Two complete intermolecular quadruplexes are arranged in the crystal structure in a S' to S' manner, with a sandwich of two layers of ligand molecules between them (Figure 1, parts b and c). There are very few direct RNA—ligand hydrogen-bond contacts, the primary mode of interaction being \( \pi-\pi \) stacking. The central nitrogen atoms of the acridine rings are orientated away from the center of the quadruplex, such that the two ligand molecules comprising the dimer on each quartet face are arranged in a laterally displaced back-to-back manner (Figure 1b) across the crystallographic twofold axis and around the central potassium ion channel. There are no direct contacts between the two coplanar ligand molecules.

In addition to the nonpolar surface provided by the S' G-quartet for \( \pi \)-stacking, the adenine residues of the propeller-like loops are flipped up toward the ligands, orientating themselves within the plane of the S' quartet and providing a platform for the triazole rings of the ligand (Figure 2a). The adenines are thus participants in a novel pseudo-four-fold all-purine G4A4 octet arrangement comprising four guanines from the G-quartet together with the four loop adenines (Figure 2d), which significantly increases the surface area available to the ligand for \( \pi \)-stacking interactions. The majority of the ligand aromatic groups are positioned directly over the purines, with only one phenyl group not being positioned over the purine-octet binding platform (Figure 1b). Both of the highly polar triazole rings are
Figure 3. (a) $2F_{o} - F_{c}$ electron density map of a loop, contoured at 1.0 $\sigma$, showing the stacking between U7 and A8 together with the G4...A8 O2'...O2' and A8...G4 N1...O2' hydrogen bond. (b) $2F_{o} - F_{c}$ electron density map of the internal loop, contoured at 1.0 $\sigma$, showing the G4...A8 O2'...O2' hydrogen bond. (c and d) Detailed views of the hydrogen bonding in the internal and terminal loops, respectively.

stacked over purines; one end of the ligand is held in a pseudointercalated manner (Figure 2a), stacked effectively between one adenine and a terminal uracil base. The overlap between triazoles and adenine/uracil is minimal at the other end of the ligand, but here instead there are several hydrogen-bond contacts. The involvement of the loop adenine bases in the G4A4 octet is a consequence of the hydrogen bonding in the loops (described in detail below).

The ligand itself is involved in few hydrogen-bonded contacts with the RNA. There is a weak hydrogen bond (3.3 Å) between the amide carbonyl group at the minimal stacked end of the ligand side chain and the N3 nitrogen of the 5' terminal uracil residue (Figure 2b). A water molecule mediates a pair of strong hydrogen bonds between the amide nitrogen atom and the O4' ribose sugar ring atom of an adjacent guanosine (Figure 2c). The arrangement in this region is tightly ordered since there is also a hydrogen bond between the O2' of this guanosine and the 5' adenine (confirmed by an omit map, shown in Figure 1d); this base also hydrogen bonds to the neighboring terminal uracil. The ligand, despite having a potentially high degree of intrinsic conformational flexibility, adopts a predominantly planar conformation, with only a slight deviation from planarity occurring at the triazole-phenyl link, where the ligand side chains curve up toward the 5' face of a symmetry-related biological unit. This planar arrangement is due in large part to the high degree of overlap between the aromatic groups of the ligand and the purine octet.

The Purine Octets Involve Water Molecule Bridges. The core of the G4A4 octet is the 5' G-quartet. The four adenines surrounding it and in the same plane (Figure 2d) originate from the UUA loop (loop 1) and the 5' terminal UA sequence, which may be considered to be an incomplete second trinucleotide loop (loop 2). The loop 1 adenine is strongly hydrogen-bonded via base atom N1 to the 2' hydroxyl group of the adjacent guanosine, and in addition there is a water-mediated pair of hydrogen bonds linking the adenine N3 atom and the N2 of this guanine. This pattern is repeated across the other side of the G-quartet as a consequence of the crystallographic twofold symmetry. The two 5' adenines form the final components of the octet. These are involved in similar interactions to the first adenine pair, notably a strong hydrogen bond to a guanosine 2' hydroxyl group and a hydrogen bond to a water molecule that similarly sits between the adenine and the adjacent guanine. The four adenine bases are slightly tilted out of the mean octet plane, by $-6^\circ$ to $8^\circ$.

Loop Conformations. The UUA loop and the incomplete 5'-UA loop are both held in particular orientations by a combination of base...base stacking and intraloop hydrogen bonding, some of which is RNA-specific. The UUA loop (Figure 3, parts a and b)
Figure 4. (a) Schematic showing the native and ligand-bound loops in cartoon form, highlighting the conformational changes between them. The arrows indicate those bases that have altered glycosidic angles. (b) Superposition of the native (red) and ligand-bound (blue) loops. (c) Plot of the loop backbone angles, showing the native loop values in red and the ligand-bound ones in black.

has the 5' U6 swung out from the loop, and the adjacent bases, U7 and A8, are approximately parallel yet poorly stacked on each other at a separation of >3.6 Å. There are several cross-loop hydrogen bonds that serve to stabilize the loop conformation (Figure 3c). The A8 2' hydroxyl group is hydrogen-bonded with a phosphate oxygen atom from the adjacent U. It also hydrogen bonds, more weakly, with the 2' hydroxyl group from residue G4. This hydroxyl group is also the terminus of a two-water molecule bridge to the N3 atom of G4 and N3 of A8. The incomplete 5'-UA loop has three short cross-loop hydrogen bonds involving the two 2' hydroxyl groups, from A2 and G10, and between O2' of G9 and N1 base atom of A2 (Figure 3d). A water molecule is bridging, using two hydrogen bonds, between the N2 of G10 and a G4 phosphate oxygen atom.

Loop Conformational Reorganization. Comparison of the conformational angles for the loop residues with those of the native RNA quadruplex highlights the significant reorganization in loop structure that has taken place on ligand binding. The orientations of all three bases in the native loop, U6, U7, and A8, which are almost perpendicular to the G-quartet planes in the native structure, are changed so that U7 and A8 become aligned with the G-quartets (Figure 4, parts a and b). In particular A8 becomes part of the extended G4A4 octet so that it can stack with part of the ligand. Examination of the individual backbone torsion angles (Figure 4c) shows that the changes needed to produce the ligand binding site involving the A8 platform are not uniform along the UUA loop sequence. The angle \( \gamma \) in U6 increases by ca. 140°, whereas few other angles in this nucleotide are significantly altered. Angle \( \alpha \) increases in U7 by ca. 90°, and angles \( \beta \) and \( \gamma \) increase by lesser amounts, up to 60°. There are two significant changes in A8 in the ligand-complexed structure, decreases in \( \alpha \) and \( \zeta \) of ca. 75°. It is notable that the glycosidic angles have closely similar values in the native and ligand complexes, suggesting that this feature of RNA quadruplexes remains constant, along with sugar pucker.

## DISCUSSION

The topology of telomeric RNA quadruplexes is rather rigid and appears to remain in the all-parallel fold under a variety of environmental conditions, in striking contrast to the polymorphism observed for telomeric DNA quadruplexes. This has been ascribed to the structure-stabilizing role of the 2' hydroxyl substituents in RNA quadruplexes. It is thus at first sight paradoxical that these groups also play a key role in changing the UUA loop conformation in the present ligand–quadruplex complex. However, it is also apparent from this structure that the loop structural reorganization is driven by the requirement to maximize interactions with the ligand, especially with its planar component. Hence, the G4A4 purine octet motif is stabilized both by stacking interactions with the planar ligand substituents and by a series of pivotal hydrogen bonds with 2' hydroxyl groups (Figure 2, parts c and d). Such an arrangement is not obviously available to a DNA quadruplex, although a A2G4 hexad has previously been observed in a DNA quadruplex–ligand complex. So although there are no direct interactions with the ligand from the 2' hydroxyl groups in the present structure, RNA structural selectivity has been achieved in a more indirect manner.

The observation here of induced ligand modification of loop conformation reinforces and extends to RNA quadruplexes the
earlier conclusions from an analysis of their observed conformations that rational design and optimization of such compounds cannot be based on structural knowledge of native quadruplexes alone, even though in all the crystal structures reported to date the loops retain their overall propeller topology. The 12-mer DNA quadruplex—BRACO-19 ligand complex shows loop modifications consequent to ligand binding that optimize non-bonded interactions with the substituent side chains of this particular ligand. The 12-mer and 23-mer DNA telomeric quadruplexes with a bound napththalene diamide ligand and the 12-mer with bound porphyrin ligand TMPyP4 also show loop conformations differing from those in the native crystal structures. The loop conformations in most of these DNA quadruplex complexes are stabilized by stacking interactions with ligand molecules on the exterior of the quadruplexes, separate from the G-quartet interactions, and are thus not involved in the recognition of tightly bound ligand molecules. In no instance is a DNA loop stabilized by internal hydrogen bonding, a key difference between these and the RNA loops in the present structure. This has been confirmed by a medium-resolution (3.20 Å) crystal structure of the analogous DNA bimolecular quadruplex co-crystallized with the same acridine ligand. The overall structure of this quadruplex complex is closely similar to that observed in previous crystal structures, with a single (disordered) ligand molecule bound on a G-quartet face (Figure S) and the complex having 1:1 stoichiometry, in contrast to the 2:1 observed in the present RNA complex. The acridine core in the DNA complex is positioned over the central ion channel of the quadruplex, as seen in previous acridine quadruplex structures, contrasting with the offset positioning in the RNA complex. There are no loop—acridine contacts in the DNA complex and no G4A4 octet. We conclude that these differences between the DNA and RNA complexes are due to the presence of the Z′ hydroxyl groups in the RNA and their active involvement in UUA loop conformation remodeling and ligand recognition.

The recent demonstration of the presence of discrete TERRA RNA quadruplexes in cells, together with the finding that it is a natural inhibitor of the telomerase enzyme complex, suggests that selective and direct targeting of TERRA by small molecules may help to clarify the role of TERRA molecules, especially in protecting telomeres. Targeting of the Z′ single-stranded telomeric DNA overhang by sequestering it into quadruplex DNA is an effective way to inhibit telomerase processivity, with the potential for therapeutic use in human cancers. It may then be important to ensure that ligands are able to discriminate between DNA and RNA telomeric quadruplexes, on the basis of their structural features. The distinctive features in the present structure suggest that such discrimination can be achievable. Other types of RNA quadruplexes, especially those derived from untranslated genomic regions, may be amenable targets for artificial regulation of translation under the influence of an RNA quadruplex stabilizing small-molecule ligand. The present structure also suggests that such ligands may similarly affect the conformations of the non-telomeric loops in these quadruplexes.

 METHODS

Crystallization and Data Collection. The RNA sequence used for crystallization trials, r(UAGGGUUAGGGU), was purchased from Eurogentec (HPLC purified). Synthesis of the triazole-acridine ligand used in these studies has been described previously. Crystals of the complex between the RNA and the triazole-acridine were grown in standard hanging drops. The RNA was annealed at a 1.8 mM single-stranded concentration by heating to 90 °C for 5 min, followed by slow cooling to room temperature overnight in an annealing buffer of 50 mM KCl and 20 mM potassium cacodylate (pH 6.5). RNA and ligand were mixed to final concentrations of 0.75 mM (quadreples) and 0.5 mM ligand, respectively, and incubated at room temperature for 1 h before mixing with crystallization solution (20% MPD, 150 mM NaCl, 50 mM sodium cacodylate (pH 6.5) and 4 mM spermine). The RNA—ligand complex plus reagent was equilibrated against a well containing 20% MPD at 12 °C, and crystals appeared after 6 weeks. A suitable crystal was screened and characterized on an in-house Oxford Diffraction X-ray machine, and a final data set collected at the U.K. Diamond Light Source synchrotron (beamline I.04). Crystals were found to be in space group P2₁ with cell dimensions and other crystallographic data given in Table 1.

Data Processing, Structure Solution, and Structure Refinement. Images were processed and scaled using the d*trek program (Rigaku Inc.). The structure was solved by molecular replacement using the Phaser program from the CCP4 software package. The native bimolecular RNA quadruplex structure (PDB id 3B8K) was used as a search model, with loop and terminal residues removed. Electron density for the three G-quartets and K⁺ ions could be clearly seen in the initial 2Fo−Fc maps, as well as electron density for the UUA loop and residual density above the 5′ G-quartet into which the ligand core could be accurately placed. The K⁺ ions were located midway between adjacent G-quartet planes, in a pyramidal prismatic coordination to O6 atoms of the guanine bases. Model building and restrained structure refinement (plus TLS parameters) were performed using Coot and Refmac5. As the phases were improved, the UUA loop could be accurately fitted into electron density, as well as the majority of the ligand (an initial PDB file plus geometric parameters for the ligand was generated using the PRODRG server: http://davcpl.bioch.dundee.ac.uk/prodrg/). Electron density observed beneath the 3′ G-quartet allowed the terminal uracil residues to be placed with reasonable accuracy. Further crystallographic details are given in Table 1.

Figure 5. Cartoon representation of the crystal structure of the DNA quadruplex—acridine 1:1 complex. The bases in the loops are shown as solid polygons (red for adenine, blue for thymine), and the backbone is shown in ribbon form. A single acridine molecule is shown (as a yellow stick representation) bound on the G-quartet surface as two disordered half molecules. The sequence crystallized is d(TAGGGTTAGGGT), and crystals of the complex are in the space group P6₃22 with cell dimensions a = b = 71.67, c = 29.37 Å. Diffraction data were collected on the U.K. Diamond Light Source, and the structure has been refined to an R of 34.6% for 783 reflections to 3.20 Å resolution. The structure has been deposited (PDB id 3QCR) and will be discussed elsewhere in more detail.

dx.doi.org/10.1021/ja109767y}
Journal o f th e A m e ric an C hem ical S ociety

Ta ble 1. C ry s ta llo g ra p h ic D ata fo r the C o m p le x "
sequence

r(U A G G G U U A G G G U )

space group

m
U nit Cell D im ensions

a, b, c (A )

56.61, 56.61, 56.61

resolution (A )

1 7 .9 0 -2 .4 0
5.5 (36.8)

(% ) overall

Uo

16.1 (4.3)

completeness {‘X

100.0 ( 100.0)

redundancy

6.6 ( 6.8)
Refinem ent

resolution (A )

1 6 .3 4 -2 .6 0

reflections

1901

Rw.,k/R„cc(%)

2 3.6 /2 4 .7

no. o f atoms

333
2

water

16

overall B factor (A ")

26.46
rm s D eviations

bond lengths (A )

0.010

bond angles (deg)

1.018

PD R ID

3MIJ

“ Values in parentheses refer to the highest resolution shell, 2.49—2.40 A.

■ AUTHOR INFORMATION
Corresponding Author
S tephen.neidle^pharm acy.ac.uk

■ ACKNOW LEDGM ENT
W e are grateful to C R U K (Program m e G rant N o. C l 2 9 /
A 4489), the E U (FP7 G rant fo r M olecular Cancer M edicine),
and the School o f Pharmacy fo r support. S.S. is a M aplethorpe
Fellow o f the U niversity o f London. T h is w o rk was carried out
w ith the support o f the D iam o nd Lig h t Source.
■ REFERENCES
[Online early access]. Doi: 10.1101/cshperspect.a003558. Published
Online: July 21, 2010. http://cshperspectives.cshlp.org/content/early/
2010/07/19/cshperspect.a003558.1ong.
(2) Azzalin, C. M.; Reichenbach, P.; Khoriauli, L ; Giulotto, E.;
(5) Caslini, C.; Connelly, J. A ; Serna, A.; Broccoli, D.; Hess, J. L.


130, 11179-11184.
(15) Parkinson, G. N.; Lee, M. P. H.; Neidle, S. Nature 2002, 417,
876-880.
Acids Res. 2010, 38, 5569-5580.
(17) Sun, D.; Thompson, B.; Gathers, B. E.; Salazar, M.; Kerwin,
1997, 40, 2113-2116.
(18) De Cian, A ; Lacroix, L.; Douarre, C.; Temime-Smaali, N.;
(19) Burger, A. M.; Dai, P.; Schultes, C. M.; Reszka, A P ; Moore,
M. J.; Double,;. A ; Neidle, S. Cancer Res. 2005, 65, 1489-1496.
(20) Salvati, E.; Leonetti, C.; Rizzo, A ; Scarsella, M.; Mottolese, M.;
Galati, R.; Sperduti, 1.; Stevens, M. P. G.; D ’lncalci, M.; Blasco, M.;
Chiorino, G.; Bauwens, S.; Horard, B.; Gilson, E.; Stoppacciaro, A ; Zupi,
(21) Campbell, N. H.; Parkinson, G. N.; Reszka, A. P.; Neidle, S.
(22) Campbell, N. H.; Patel, M .; Tofa, A. B.; Ghosh, R.; Parkinson,
(23) Gavathiotis, E.; Heald, R. A ; Stevens, M. P. G.; Searle, M. S.
(24) Collie, G.; Reszka, A. P.; Haider, S. M .; Gabelica, V.; Parkinson,
Guittat, L.; Riou,J.-L.; Trentesaux, C.; Sacca, B.; Lacroix, L ; Alberti, P.;
(26) Rzuczek, S. G.; Pilch, D. S.; Liu, A ; Liu, L ; LaVoie, E.J.; Rice,
(27) Kumari, S.; Bugaut, A ; Huppert,J. L.; Balasubramanian, S. Nat.
(28) Huppert, J. L ; Bugaut, A ; Kumari, S.; Balasubramanian, S.
(30) Haider, K.; Wieland, M.; Hartig,J. S. Nucleic Acids Res. 2009,37,
6811-6817.

(31) A ora, A ; Dutkiewicz, M.; Scaria, V.; Hariharan, M.; Maiti, S.;
(33) Balkwill, G. D.; Derecka, K.; Garner, T. P.; Hodgman, C.; Flint,
A. P. P.; Searle, M. S. Biochemistry 2009, 48, 11487-11495.
Teulade-Pichou, M.; Calsou, P. Nucleic Acids Res. 2010, 38,7187-7198.
(36) Bugaut, A ; Rodriguez, R ; Kumari, S.; Hsu, S. D.; Balasubra­
(37) Henn, A ; Joachimi, A ; Gonçalves, D. P. N.; Monchaud, D.;
9, 2722-2729.
(39) Joachimi, A ; Benz, A ; Hartig,J. S. Bioorg. Med. Chem. 2009,17,

(6) Redon, S.; Reichenbach, P.; Lingner, J. Nucleic Acids Res. 2010,
38, 5797-5806.

6811-6815.

(7) Deng, Z.; Norseen, J.; Wiedmer, A.; Riethman, H.; Lieberman,

Soc. 2010, 132, 7231-7233.

(8) Deng, Z.; Campbell, A. E.; Lieberman, P. M. Cell Cycle 2010, 9,
69-74.
(9) Feuerhahn, S.; Iglesias, N.; Panza, A ; Porro, A ; Lingner, J. FEBS
Lett. 2010, 584, 3812-3818.
U.SA. 2010, 107,651-656.

(41) Sparapani, S.; Haider, S. M.; Doria, P.; Gunaratnam, M.; Neidle,
(42) Parkinson, G. N.; Cuenca, P.; Neidle, S. }. Mol. Biol. 2008, 381,
1145-1156.
(44) Parkinson, G. N.; Ghosh, R ; Neidle, S. Biochemistry 2007, 46,
2390-2397.
dx .d o i.o rg /1 0 .1 0 2 1 /ja l0 9 7 6 7 y \J. Am. Chem. Soc. XXXX, XXX, 0 0 0 - 0 0 0

