Molecular modelling and biological implications of non-canonical structures in RNA viruses and long non-coding RNAs

Simon Chapman BSc, MSc, MSc

A thesis submission to the Open University for the degree of Doctor of Philosophy

Faculty of Science, Technology, Engineering and Mathematics

School of Life, Health and Chemical Sciences

05 01 2023
Declaration

Declaration The material in this thesis is the result of research undertaken in the School of Life Sciences at the Open University between October 2019 and October 2022. It is the original work of the author except where acknowledged by reference and has not been submitted for any other degree.

Statement of Copyright The copyright of this thesis lies with the author. No quotations from it should be published without the author’s prior written consent, and information derived from it should be acknowledged.
Abstract

Guanine quadruplexes folding in single-strand DNA have been extensively studied for some time. But far less attention has been paid to systems arising in duplex DNA or RNA. Furthermore, the literature shows an apparent absence of quadruplex systems generated exclusively by other nucleic acid bases. With guanine quadruplexes (G4s) as a benchmark, a comparison is made here with equivalent complexes folding in DNA and RNA when derived from cytosine, thymine and adenine. Molecular dynamics simulations determined cytosine and thymine models in DNA as relatively fragile systems, but all non-guanine RNA models were found to be stable into biologically relevant times. Uniquely biplanar or triplanar adenine quadruplexes (A4s) have ostensibly not been described. Adenine models constructed for this work were resolved in silico with stabilities comparable to known guanine equivalents. These complexes achieved further significance with the advent of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).

RNA viruses are characterised by a poly-adenylated structure capping the genomic terminus. This poly(A) tail is crucial to a cascade of viral replicative activity occurring both extra- and intra-cellular during infection. As a route to proposing potential chemotherapy, this study suggests simple biplanar A4s may fold in this poly-adenylated domain. Notably, mRNA configured as a biplanar A4, shows less dynamic activity than DNA equivalents. This observation may be especially relevant in a physiological context. In contrast to well-characterised guanine quadruplexes, co-ordination with cations appears not to impact on stability. These conclusions may apply to SARS-CoV-2, its variants and other pathogenic RNA viruses.

The thesis also infers models of potential interactions between small nucleolar RNAs and long non-coding RNAs may have biological relevance. The function of some RNAs are not currently well understood particularly the two species investigated here. Non-canonical base-pairing in some instances suggests a molecular mechanism for dysfunction in the development of the embryonic pre-frontal cortex.
Acknowledgements

I am most grateful for the continued support and constructive advice from my Lead Supervisor Dr. Yao Xu. His expert guidance in the formative stage of my proposal, and indeed for the entirety of the work, were invaluable. I gained much useful insight into the chemistry of nucleic acids from him and have emerged as a slightly wiser individual predicated on Yao’s worldview. Moreover, my self-confidence and communication skills have benefited greatly from Yao’s salient commentary. His positivity helped develop resilience in the face of practical difficulties and heartened me when staring at regular impasses in the research. I was most fortunate to gain his mentorship.

As well as important practical commentary on organic chemistry throughout my research, Professor Peter Taylor’s acuity furnished me with an improved discernment of relevance. At our very first meeting, his perspicacity highlighted the resources needed for my research. His careful re-renderings of my initial diagrams of chemical reactions are invaluable. Furthermore, where I struggled with the arcane maths of thermodynamics, Peter’s corrections were Damascene. Finding a route through PhD research is a complex and often baffling experience. Peter’s affable and courteous approach is a panacea for student bewilderment.

Dr. Clare Sansom’s observations regarding computational chemistry are especially appreciated. Clare’s involvement as my Tutor-Supervisor for MSc. Medicinal Chemistry led directly to this current work. Towards the end of those studies, Clare advocated my introduction to Yao. Throughout this thesis, her gentle reprimands for my occasional schoolboy howlers and her enduring encouragement, have maintained my academic buoyancy. Furthermore, Clare’s astonishing proof-reading capability would warrant the nom de plume ‘Hawkeye’.

Peaks and troughs characterise a thesis. The elation of my first research paper being published more than compensated for the months of near-despondency. Dr. Nick Power’s continued advice in its construction and identification of relevant papers continually fortified my incentive. Nick’s acumen pinpoints ambiguity of interpretation: I am much obliged to him for assimilating some of this essential skill. Additionally, his accomplished networking found Dr. Pierre Ghesquiere. Pierre reinforced aspects of my understanding in the methodology of molecular dynamics. Moreover, Nick devoted considerable effort in acquiring the Linux server requested for the work. Everyone should have a Nick Power in their corner.
I am very happy to acknowledge my Supervisors’ exceptional contributions individually, but their collective sagacity is compelling and their good-humour carried me along in the slipstream.

Exterior to my direct supervision but significant nevertheless, Dr. Elaine Moore’s detailed advice and suggestions regarding Molecular Dynamics software, was greatly appreciated. Moreover, Elaine very kindly arranged for the AMBER16 licence fee to be waived. Yao’s pre-module help in expediting contact with Elaine was timely. Also, as an acolyte of Dr. Christine Gardner, I am very pleased to signpost my continued contact with her. Chris was the first tutor to help me navigate the previously uncharted world of molecular biology.

For technical support, I am obliged to several contributors outside of the Open University. Dr. Sandro Bussi of Trieste University clarified the complexities of eRMSD, Dr. Matt Hackbart of Washington University provided extensive details of viral infectivity only marginally addressed in his paper. Professor Kate Pollard at the University of California kindly responded to my inquiries regarding species-specific mutations in non-coding RNA. Michael Bareford at Edinburgh University was long-suffering in his help with ARCHER2 operation after progressing a licence waiver for access. My friends Peter Ford and Martin Gautier frequently talked me through the vicissitudes of Linux severs upon my early encounters with Unix systems.

All graciously maintained contact and advice despite my naïve pestering.

These paragraphs are less about priorities of acknowledgment than a sincere token of overall endorsement. Tertiary education relies on generous non-academic support. Family and friends need patience, understanding and trust when ‘volunteered’ to share the PhD experience. Foremost is my wonderful Diane. Her fortitude is the cornerstone for all the work that followed. Her advocacy for higher learning was shared by my Mother. After gaining my BSc., she declared I should ‘go as far as possible with this’. That her comment is now posthumous doesn’t detract at all from its sentiment.

To understand Nature, all we need is curiosity. During this research, mine has been satisfied and provoked in equal measure.
Conference attendances and publications


Nucleic Acid Forum: poster accepted July 2022 (Supplementary Material)

Professor Tina Duren, Cardiff University “Molecular simulation of adsorption in metal-organic frameworks”. November 2019

Student conference Presentation November 2022 (annotated slides in Supplementary Material)

Published papers archived at oro.open.ac.uk: https://oro.open.ac.uk/

Journal of Computational Chemistry and Biology, September 2021, DOI: 10.1142/S273741652150040X “Adenine quadruplexes show surprising stability: potential implications for SARS-CoV-2”

Epigenomics 13(7), October 2021. DOI:10.2217/epi-2021-0069 “HAR1: an insight into IncRNA genetic evolution”

Journal of Computational Chemistry and Biology, December 2022, DOI: 10.1142/S2737416523500400X “SARS-CoV-2 main protease inhibitors: Structure-based enhancement to anti-viral pre-clinical GC376 encourages further development.

Sparrow: Online popular science digest: 30 07 2020
https://www.sparrow.science/covid-19-faq-can-we-design-drugs-that-slow-down-viral-replication-in-our-own-cells/

Supplementary Material archived at Open Research Data Online:
https://doi.org/10.21954/ou.rd21817890v1

Table of contents
General Introduction........................................................................................................p1

Chapter 1 Methods and metrics .............................................................p9
  1.1 Molecular modelling .........................................................................................p9
  1.2 Molecular Dynamics............................................................................................p9
    1.2.1 introduction
    1.2.2 preparation of models in NAMD and Chimera
    1.2.3 Periodic Boundary Conditions
    1.2.4 MD production
    1.3.1 Metrics for molecular dynamics analysis
    1.3.2. Root Mean Square Deviation
    1.3.3 ellipsoidal Root Mean Square Deviation
    1.3.4 Radius of gyration
    1.3.5 Frequency of hydrogen-bond formation
    1.3.6. Torsional analysis

Chapter 2 Guanine quadruplexes ..........................................................p26
  2.1.1 Introduction
    2.1.2 Guanine quadruplexes in single-stranded telomeric DNA
    2.1.3. Pathogenicity and intervention
    2.1.4 Molecular dynamics
  2.2. Results..............................................................................................................p29
    2.2.1 Results overview: 1KF1 non-modified
    2.2.2 Results overview: 1KF1 O6-methylation
    2.3.1 Results overview: 2KF8_non-modified
    2.3.2 Results overview: 2KF8 O6 methylation
2.3.3 Results overview: DNA single strand unfolded 22mer

2.3.4 Results overview: DNA single strand unfolded 22mer

2.4 Theoretical melting temperature of double strand DNA 22-mer........p45

2.5 Potential impact of thionation and methylation at guanine O6.......... p45

2.5.1 Introduction

2.5.2 Results

2.5.3 Discussion

2.5.4 Enzymatic repair mechanism for O6 methylation

2.5.4.1 introduction

2.5.4.2 Methods

2.5.4.3 Results

2.5.4.4 Discussion

2.6 Occupancy of waters in guanine quadruplexes 1KF1 and 2KF8...........p55

2.6.1 Introduction

2.6.2 Methods

2.6.3 Results

2.7 Ribose conformations and potential compromise of function.............p59

2.7.1 Introduction

2.7.2 Methods

2.7.3 Results

2.7.4 O6-methylation of biplanar guanine and N6-methylation of adenine models

2.7.5 Discussion of C2’-endo and C3’-endo configurations

2.8 Biplanar, triplanar DNA quadruplexes co-ordinated by K+, Na+ or Li+....p67

2.8.1 Methods

2.8.2 Results
2.8.2.1 models derived from 1KF1
2.8.2.2 models derived from 2KF8

2.9 Executive summary of Chapter 2.................................................................p69
2.10 General discussion of guanine quadruplexes.............................................p69
2.11 Future work guanine quadruplexes.............................................................p72

Chapter 3. Cytosine- and thymine-based quadruplexes ......................p73

3.1 Introduction: Cytosine

3.2 Cytosine dimers.........................................................................................p74

3.3 Cytosine quartets.......................................................................................p76
  3.3.1 Cytosine quartets co-ordinated with Ba^{2+}
  3.3.2 Cytosine quartets co-ordinated with Pb^{2+}, Tl^{+}, Rb^{+}, Hg^{2+}
  3.3.3 Cytosine quartets co-ordinated with K^{+}, Na^{+}, Li^{+}
  3.3.4 Guanine-cytosine and guanine-adenine hybrid quadruplexes

3.4 Cytosine quadruplexes...............................................................................p79

3.5 Methods........................................................................................................p79

3.6 Results............................................................................................................p80
  3.6.1 Results overview: Triplanar cytosine quadruplex (CDN12) K^{+} co-ordination
  3.6.2 Results overview: Triplanar cytosine quadruplex (CDN12) Na^{+} co-ordination
  3.6.3 Results overview: Triplanar cytosine quadruplex (CDN12) Li^{+} co-ordination
  3.7.1 Results overview: Biplanar cytosine quadruplex (CDN8) No ion
  3.7.2 Results overview: Biplanar cytosine quadruplex (CDN8) K^{+} co-ordination
  3.7.3 Results overview: Biplanar cytosine quadruplex (CDN8) Na^{+} co-ordination
  3.7.4 Results overview: Biplanar cytosine quadruplex (CDN8) Li^{+} co-ordination
3.8 Occupancy of waters in cytosine quadruplexes CDN12 and CDN8....p84

3.8.1 Introduction and methodology

3.8.2 Results (S68 – S74)

3.9 Ribose conformations of biplanar and triplanar cytosine models....p86

3.9.1 Introduction and methodology

3.9.2 Results (S75 – S80)

3.10 Discussion of cytosine quadruplexes....................................................p89

3.11 Thymine quadruplexes: Introduction.................................................p91

3.11.1 Methods..........................................................................................p92

3.11.2 Results............................................................................................p92

3.11.2.1 Results overview: Triplanar thymine quadruplex (TDN12) K⁺ co-ordination

3.11.3.1 Results overview: Triplanar thymine quadruplex (TDN12) Na⁺ co-ordination

3.11.4.1 Results overview: Triplanar thymine quadruplex (TDN12) Li⁺ co-ordination

3.11.5.1 Results overview: Biplanar thymine quadruplex (TDN8) No ion

3.11.6.1 Results overview: Biplanar thymine quadruplex (TDN8) K⁺ co-ordination

3.11.7.1 Results overview: Biplanar thymine quadruplex (TDN8) Na⁺ co-ordination

3.11.8.1 Results overview: Biplanar thymine quadruplex (TDN8) Li⁺ co-ordination

3.12 Occupancy of waters in thymine quadruplexes TDN12 and TDN8......p96

3.12.1 Introduction and methodology

3.12.2 Results (S115 – S121)

3.13 Ribose conformations of biplanar and triplanar thymine models....p98

3.13.1 Introduction and methodology

3.13.2 Results (S122 - S127)
4.9.6 Results overview: RMSD, eRMSD, Radius of Gyration and energy histograms for ADN12-N1 in all co-ordinating ion conditions

4.10 Results overview: Biplanar N6-methylated adenine quadruplexes ARN8-N1, ARN8-N3 and ARN8-N7

4.11 Results overview: Biplanar adenine quadruplex ARN8-N1 no ionic co-ordination

4.12 Results overview: Biplanar adenine quadruplexes ARN8-N1 cationic co-ordination

4.13 Results overview: Biplanar adenine quadruplexes ARN8-N1 Ca$^{2+}$ and Mg$^{2+}$ co-ordination

4.14 Results overview: Biplanar adenine quadruplexes ARN8-N1 NH$_4^+$ and H$_3$O$^+$ co-ordination

4.15 Potential of Mean Force, Torsional analysis..................................................p141

4.15.1 Introduction

4.15.2 Methods

4.15.3 Results

4.15.3.1 PMF results summary

4.15.3.2 results: concatenated data and per base analysis

4.16 Discussion of PMF.................................................................p148

4.17 Occupancy of waters in adenine quadruplexes ADN12-N1 and ARN8-N1.............................................................p150

4.17.1 Introduction and methodology

4.17.2 Results

4.18 ribose and deoxyribose conformations of biplanar and triplanar adenine systems..........................................................p151

4.18.1 Introduction and methodology

4.18.2 results

4.18.3 N6-methylation of ADN12-N1 and ARN8-N1
4.19 Executive summary of Chapter 4.........................................................p157
4.20 Discussion of adenine quadruplexes.................................................p159
4.21 Future work..........................................................................................p162

Chapter 5.  Purine quadruplexes in RNA viruses ......................p165

5.1 Introduction

5.2 Genomic 3’ polyadenylation.................................................................p166

5.3 Function of the viral poly(A) tail.........................................................p167
  5.3.1 Introduction
  5.3.2 Viral responses to host cell destruction of the poly(A) tail
  5.3.3 Implications of viral poly(A) tail proliferation

5.4 Adenine N6-methylation in the polyadenylated viral genome.........p168

5.5 RNA virus infection cascade...............................................................p172
  5.5.1 Introduction
  5.5.2 polyadenylation in RNA viruses and the host cell response

5.6 Purine quadruplex assembly in RNA viruses...............................p174
  5.6.1 Introduction
  5.6.2 Relevance of biplanar quadruplexes to SARS-CoV-2
    5.6.2.1 Introduction
    5.6.2.2 Materials and Methods
    5.6.2.3 Results

5.7 Guanine and adenine-guanine hybrid quadruplexes in SARS-CoV-2...p180
  5.7.1 Introduction
  5.7.2 Results
  5.7.3 Discussion
Chapter 6. Quadruplexes and compromise of cellular function...p191

6.1 Introduction

6.2 Concept of quadruplexes as molecular obstacles.................................p192

6.2.1 Ribosome pausing

6.2.2 Lagging strand cytosine quadruplexes

6.2.3 conjectural quadruplex-induced dislocation of Shelterin

6.2.4 Discussion of Section 6.2

6.3 Challenges of interpreting purine quadruplex contribution to aetiology.................................................................p196

6.3.1 Introduction

6.3.2 Vaccination

6.3.3 Targeting purine quadruplexes as chemotherapy

6.4 Purine quadruplex ligands........................................................................p200

6.4.1 Telomestatin

6.4.2 Adapted porphyrins

6.4.2.1 Introduction
6.4.2.2 Methods
6.4.2.3 Results
6.4.2.4 Telomestatin, AP2 and AP3 docking with 1KF1
6.4.2.5 Discussion
6.4.2.6 N-meso methylporphyrin
6.4.2.7 Methods
6.4.2.8 Results
6.4.2.9 Discussion
6.5 Conjectural approach to destabilise pathogenic purine quadruplexes .................................................................................................................. p210
   6.5.1 Introduction
   6.5.2 peptide nucleic acids
      6.5.2.1 Introduction
      6.5.2.2 Methods
      6.5.2.3 separating loop nucleotides: type restriction enzymes
6.7. Executive summary of Chapter 6........................................................... p215
6.8. Future perspectives.................................................................................. p216

Chapter 7. Relevance to long non-coding RNA ................................. p218
   7.1 Introduction
   7.2 RNA-RNA interaction: HAR1A,HAR1 and small nucleolar RNAs........ p222
Conclusion .................................................................................................... p255
Thesis retrospective...................................................................................... p257
List of figures

**Figure 1.1a: guanine quadruplex structure.** Loop pattern in image B is the parallel type; DNA allows this type and an anti-parallel variation to fold. RNA guanine quadruplexes fold only with parallel loops.

**Figure 1.1b: biplanar RNA guanine quadruplex (eg 2KF8), in cellulo relevance and location**

**Figure 1.2. Cytosine substituted for all guanine.** Image created in Chimera and the ‘swapna’ instruction in command line used for the substitutions. The command re-writes the PSF and PDB files by default with parameters for the substituents.

**Figures 1.3a – 1.3c: Adenine quartets found experimentally to be stable.** Designed in BioVia Draw and adapted from Scheme 1, van der Wijst (2010).

**Figure 1.3d: biplanar RNA adenine quadruplex tertiary structure.** The interacting quartet planes are depicted in green and clearly show the biplanar structure central the complex. Designed in Discovery Studio Client.

**Figure 1.4: co-ordination of eRMSD base-pairing.** Exemplified here with guanine and cytosine. The metric calculates interplanar distances between C2, C4 and C6 of bases thus providing specificity to an analysis. Values from sugars and the phosphate backbone are ignored. This allows a better insight into the overall activity in a quadruplex when compared alongside RMSD data.

**Figure 1.5: outline of workflow in MD analysis for the investigation.**

**Figure 2.1. (PDB) 1KF1: triplanar guanine quadruplex.** Phosphate backbone, deoxyribose and bases depicted in line format. Co-ordinating cations and guanine-O6 oxygens shown as ball and stick.

**Figure 2.2. Orthodox and non-canonical H-bond acceptors for guanine:** illustrating the difference between Watson-Crick and Hoogsteen binding. The latter interface allows guanines to self-associate into quartets which subsequently build into quadruplexes.

**Figures 2.5a and 2.5b: 1KF1 methylated at guanine O6.** Images designed to emphasise structural symmetry before molecular dynamics simulation and after 500ns. the Cartesian relationship between O6 methyl group and co-ordinating cation is also clarified.

**Figures 2.6a and 2.6b: Distances (left) and angles (right) between K+ cation and guanine O6 for 1KF1.**

**Figures 2.7a, 2.7b, 2.7c and 2.7d: frequency of H-bonds for residues of the central plane in GDN-N1_meth K**
Figures 2.8a, 2.8b, 2.8b and 2.8d: N6 – N1(H1) distances for residues of the central quartet GDN12-N1_meth, K⁺ co-ordination

Figure 2.9: RMSD for all bases in the quartet pair of 2KF8 without cationic co-ordination

Figures 2.10a, 2.10b, 2.10c and 2.10d: frequency of H-bonds for residues of the central plane in GDN8-N1 no cation

Figures 2.11a, 2.11b, 2.11b and 2.11d: N6 – H1(N1) distances for residues of the central quartet GDN8-N1, no ionic co-ordination

Figures 2.12a and 2.12b: 2KF8 methylated at guanine O6. Modification enabled in Discovery Studio. The PSF and PDB files were corrected manually for downstream molecular dynamics simulations.

Figures 2.13a, 2.13b, 2.13c and 2.13d: frequency of H-bonds for residues of the central plane in GDN8-N1 O6-methylation

Figures 2.14a, 2.14b, 2.14b and 2.14d: N6 – N1(H1) distances for residues of the central quartet GDN8-N1, O6-methylated

Figures 2.15a and 2.15b. DNA single strand unfolded 22mer. A model constructed for in silico comparison with an identical sequence folded into a quadruplex. Single strand DNA 22-mer at t = 0ns and single strand DNA 22-mer at t = 500ns

Figure 2.16. RMSD single strand DNA 22mer. The sequence of bases is identical to that of telomeric 1KF1.

Figures 2.17a and 2.17b. RNA single strand unfolded 22mer. A model constructed for in silico comparison with an identical sequence folded into a quadruplex. Single strand RNA 22-mer at t = 0ns and single strand RNA 22-mer at t = 500ns (right)

Figure 2.18. RMSD trajectory for single strand RNA 22mer. The sequence of bases is identical to that of telomeric 1KF1 with uridine substituted for thymine. Chimera used for the substitutions.

Figure 2.19a – 2.19c: modifications to guanine (left). Methylation at the O6 position (centre) and thionation at O6 (right).

Figures 2.20a and 20b: methylation and thionation at guanine O6 of 1KF1. Non-polar interactions in red. K⁺ model shown, a very similar network of bonds was obtained for Na⁺ and Li⁺.

Figures 2.21a – 2.21c: rejection of K⁺ co-ordinating cation from O6- methylated guanine quadruplex 1KF1. The relocation of cations occurs early in a simulation.

Figures 2.22a – 2.22c: interatomic distances between co-ordinating cations and modified guanine O6 in 1KF1. Interatomic distances between cation and modified group for were typically less for Li⁺ than for K⁺ or Na⁺.
Figure 2.23: RMSD for K'-co-ordinated guanine quadruplex with and without single O6-methylation or O6-thionation.

Figure 2.24: single O6-methylation or O6-thionation to base 15 for each quartet of 1KF1. 3D plot created in OriginLab.

Figure 2.25: docking protocols for 1KF1 and O6-methylated 1KF1 ligated with MGMT.

Figure 2.26: H-bond plots for 1KF1 and O6-methylated 1KF1 ligated with MGMT.

Figure 2.27: docking protocols for 2KF8 and O6-methylated 2KF8 ligated with MGMT.

Figure 2.28: H-bond plots for 2KF8 and O6-methylated 2KF8 ligated with MGMT.

Figures 2.29a and 2.29b: residues H-bonding to form quartets in 2KF8 (A) and 1KF1 (B). Predictions for waters within 2 Å of guanine O6 for each residue were plotted.

Figure 2.30: number of waters over time within 2 Å of guanine O6 for the three quartets of 1KF1 and 1KF1 methylated at O6.

Figure 2.31: histogram of figure 2.20 data. Both 1KF1 and 1KF1 methylated at O6 show an optimum of three waters allowed but the maxima differ noticeably.

Figure 2.32: number of waters over time within 2 Å of guanine O6 for biplanar quartets. This model would be expected to have a lower population than the DNA equivalent due to less available interplanar space.

Figure 2.33: histogram of figure 2.22 data. The models indicate an optimum population of two waters and a maximum of five.

Figure 2.34: interatomic distances between co-ordinating K+ cations and single O6-methylation or O6-thionation.

Figure 2.35: C2'-endo and C3'-endo conformations.

Figures 2.36a and 2.36b: trajectories for C1'-C2'-C3'-C4' dihedrals in DNA and RNA guanine models.

Figure 2.37: distribution histograms from data for figures 2.36a and 2.36b data GDN8-N1 (blue) and GDN12-N1 dihedrals (pink).

Figures 2.38a and 2.38b: trajectories for biplanar RNA adenine quadruplexes with and without N6 methylation.

Figure 2.39: distribution histograms from figures 27a and 29b data for ARN4 and ARN4_meth dihedrals.

Figure 3.1: for pH values < 7.0, cytosines can be protonated to associate into a dimer.

Figure 3.2: cytosine – X dimers
Figure 3.3: possible alignment of a cytosine dimer pair

Figure 3.4. Water molecules mediating Ba” co-ordinating a cytosine quartet

Figure 3.5. Potential cytosine quartet formed on graphene

Figure 3.6. Re-aligned paired cytosine dimers to allow H-bond angles <20°.

Figure 3.7: Histogram of interplanar water occupancy in triplanar cytosine quadruplex models co-ordinated by K’, Na’ or Li’.

Figure 3.8: Histogram of interplanar water occupancy in biplanar cytosine quadruplex models with no cationic co-ordination or co-ordinated by K’, Na’ or Li’.

Figures 3.9a, 3.9b and 3.9c: trajectories for C1’-C2’-C3’-C4’ dihedrals in DNA triplanar cytosine models

Figures 3.10a and 3.10b, trajectories for C1’-C2’-C3’-C4’ dihedrals in DNA biplanar cytosine models

Figure 3.11: RMSD trajectory for triplanar thymine quadruplex K’ co-ordination.

Figure 3.12: Histogram of interplanar water occupancy in triplanar thymine quadruplex models co-ordinated by K’, Na’ or Li’.

Figure 3.13: Histogram of interplanar water occupancy in triplanar thymine quadruplex models co-ordinated by K’, Na’ or Li’.


Figures 3.15a and 3.15b: trajectories for C1’-C2’-C3’-C4’ dihedrals in DNA biplanar thymine models, residue 7 no ion and K’ co-ordination

Figure 3.16: distribution histograms from figures 3.15a and 3.15b data for TRN8-N1 no ionic co-ordination and TDN12-N1 K’ co-ordination dihedrals.

Figure 4.1: planar representation of an experimentally validated adenine quartet with H-bonding between N1 and N6 (H1) of an adjacent amino group.

Figure 4.2: planar representation of an experimentally validated adenine quartet with H-bonding between N3 and N6 (H1) of an adjacent amino group.

Figure 4.3: planar representation of an experimentally validated adenine quartet with H-bonding between N7 and N6 (H1) of an adjacent amino group.

Figure 4.4. Interactions between bases and co-ordinating ions for a guanine-adenine hybrid quadruplex GAG12.

Figure 4.5 Tetrahedral co-ordination of Li’ in GAG12

Figures 4.6a and 4.6b: ADN12-N1, ADN12-N3, ADN12-N7 RMSD values and histograms.
Figures 4.7a and 4.7b: ADN12-N1, ADN12-N3, ADN12-N7 eRMSD values and histograms

Figures 4.8a and 4.8b: ADN12-N1, ADN12-N3, ADN12-N7 Radius of gyration values and histograms.

Figures 4.9a and 4.9b: ADN12-N1, ADN12-N3, ADN12-N7 frequency of H-bond formation.

Figure 4.10: Frequency of H-bond formation for base 1 of ADN12-N1 K⁺.

Figures 4.11a and 4.11b: compilation of energies for K⁺ co-ordinated ADN12-N1, ADN12-N3 and ADN12-N7.

Figure 4.12: base-specific population of H-bonds, ADN12-N7 base 3.


Figures 4.15a and 4.15b: Na⁺ co-ordinated RMSD data for ADN12-N1, -N3 and -N7.

Figures 4.16a and 4.16b: Na⁺ co-ordinated Radius of gyration data for ADN12-N1, -N3 and -N7.


Figure 4.18: H-bonds between N6(H1) and N1 of an adjacent base in the central plane of ADN12-N7 Na⁺.

Figures 4.19a and 4.19b: ADN12-N1, ADN12-N3, ADN12-N7 Li⁺ RMSD values and histograms

Figures 4.20a and 4.20b: ADN12-N1, ADN12-N3, ADN12-N7 Li⁺ eRMSD values and histograms.

Figures 4.21a and 4.21b: Li⁺ co-ordinated Radius of gyration data for ADN12-N1, -N3 and -N7.

Figures 4.22a and 4.22b: ADN12-N1 Li⁺, ADN12-N3 Li⁺, ADN12-N7 Li⁺ frequency of hydrogen bond formation.

Figures 4.23a. and 4.23b: compilation of energies for Li⁺ co-ordinated ADN12-N1, ADN12-N3 and ADN12-N7.

Figure 4.24 H-bonds between N6(H1) and N1 of an adjacent base in the central plane of ADN12-N7 Li⁺.

Figures 4.25a and 4.25b: RMSD (left) and eRMSD (right) data for ADN12-N1 co-ordinated by Ca²⁺ or Mg²⁺.

Figures 4.26a and 4.26b: Radius of gyration (left) and energies (right) data for ADN12-N1 co-ordinated by Ca²⁺ or Mg²⁺

Figures 4.27a and 4.27b: RMSD (left) and eRMSD (right) data for ADN12-N1 co-ordinated by H₂O⁺ or NH₄⁺.
Figures 4.28a and 4.28b: Radius of gyration (left) and energies (right) data for ADN12-N1 co-ordinated by $\text{H}_2\text{O}^+$ or $\text{NH}_4^+$.  
Figures 4.29a, 4.29b, 4.29c and 4.29d: RMSD (a), eRMSD (b), Radius of gyration (c) and energies (d) data for all ADN12-N1 models with ionic co-ordination.

Figures 4.30a and 4.30b: RMSD trajectories and histogram compilation of N6-methylated ARN8 models.

Figures 4.31a and 4.31b: eRMSD trajectories and histogram compilation of N6-methylated ARN8 models.

Figures 4.32a and 4.32b: radius of gyration trajectories and histogram compilation of N6-methylated ARN8 models.

Figures 4.33a and 4.33b: trajectories and histogram compilation for energies in N6-methylated ARN8 models.

Figures 4.34a and 4.34b: RMSD trajectories and histogram compilation of ARN8 models with no ionic co-ordination.

Figures 4.35a and 4.35b: eRMSD trajectories and histogram compilation of ARN8 models with no ionic co-ordination.

Figures 4.36a and 4.36b: Radius of gyration trajectories and histogram compilation of ARN8 models with no ionic co-ordination.

Figures 4.37a and 4.37b: trajectories and histogram compilation for energies of ARN8 models with no ionic co-ordination.

Figures 4.38a and 4.38b: RMSD and eRMSD data for ARN8-N1 co-ordinated by $\text{K}^+$, $\text{Na}^+$ or $\text{Li}^+$.  
Figures 4.39a and 4.39b: Radius of gyration (left) and energies (right) data for ARN8-N1 co-ordinated by $\text{K}^+$, $\text{Na}^+$ or $\text{Li}^+$.

Figure 4.40 (and S6a – S6h): Frequency of H-bond formation base 1 ARN8_N1 K+. This figure should be compared with data for the triplanar equivalent (figure 4.5).

Figures 4.41a and 4.41b: RMSD (left) and eRMSD (right) data for ARN8-N1 co-ordinated by $\text{Ca}^{2+}$ or $\text{Mg}^{2+}$.  
Figures 4.42a and 4.42b: Radius of gyration (left) and energies (right) data for ARN8-N1 co-ordinated by $\text{Ca}^{2+}$ or $\text{Mg}^{2+}$.

Figures 4.43a and 4.43b: RMSD and eRMSD data for ARN8-N1 co-ordinated by $\text{H}_2\text{O}^+$ or $\text{NH}_4^+$.  
Figures 4.44a and 4.44b: Radius of gyration (left) and energies (right) data for ARN8-N1 co-ordinated by $\text{H}_2\text{O}^+$ or $\text{NH}_4^+$.  

xxi
Figures 4.45a, 4.45b, 4.45c and 4.45d: RMSD (a), eRMSD (b), Radius of gyration (c) and energies (d) data for all ARN8-N1 models with ionic co-ordination.

Figure 4.46 Glycosidic (χ) and phosphate backbone torsions.

Figure 4.47. Global PMF trajectories for χ values of all ADN12-N1, ADN12-N3 and ADN12-N7 bases for a 500ns simulation.

Figures 4.48a, 4.48b and 4.48c: compilation of per base torsions for ADN12-N1 (a), ADN12-N3 (b), ADN12-N7 (c).

Figure 4.49: The glycosidic angle minimum for the ARN8-N7 system.

Figures 4.50a, 4.50b and 4.50c: compilation of per base torsions for ARN8-N1 (a), ARN8-N3 (b), ARN8-N7 (c).

Figure 4.51: the N9 (blue) environment in adenine quadruplexes.

Figure 4.52: Histogram of interplanar water occupancy in ADN12-N1, ADN12-N3 and ADN12-N7 co-ordinated by K⁺.

Figure 4.53: Histogram of interplanar water occupancy in ARN8-N1, ARN8-N3 and ARN8-N7 without cationic co-ordination.

Figures 4.54a, 4.54b and 4.54c: trajectories for C1'-'C2'-'C3'-'C4' dihedrals in DNA biplanar adenine models in bases 1, 7 and 15.

Figures 4.55a, 4.55b and 4.55c: trajectories for C1'-'C2'-'C3'-'C4' dihedrals in RNA biplanar guanine models for base 1, 7 and 15.

Figure 4.56: interconversion of C2'-'endo to C3'-'endo progresses via the O4' endo conformation.

Figures 4.57a and 4.57b: C1'-'C2'-'C3'-'C4' dihedrals for ADN8 and ADN8-meth.

Figure 4.58: distribution histograms for ARN8-N1 (pink) and ARN8-N1_meth dihedrals (blue).

Figure 4.59: variation of amino angles for N6-methylated adenine.

Figure 4.60: amino-meth angles relative to the heterocyclic plane of N6-methylated adenine.

Figure 5.1: Positive sense single-stranded RNA viruses are capped at the 5’ terminus and further stabilised by the addition of a series of adenines at the 3’ terminus.

Figure 5.2: theoretical renovation of viral poly(A) tail targeted for destruction.

Figure 5.3: the poly(A) tail is crucial to RNA virus replication and suppression of the interferon early-immune response.
Figure 5.4: predicted locations of potential biplanar adenine quadruplex-folding sequences in SARS-CoV-2

Figure 5.5: graphic representation of data in figure 5.4.

Figure 5.6: PA4 locations for SARS-CoV-2 variants MW185823.1 and MW185825.1.

Figure 5.7: identified biplanar RNA G4s (blue) and putative G4s (green) in SARS-CoV-2.

Figure 5.8: predicted locations of potential biplanar guanine quadruplex-folding sequences in the SARS-CoV-2 reference genome.

Figure 5.9: predicted locations for adenine-guanine hybrid quadruplexes show similarities with the SARS-CoV-2 reference genome and variant MW185823.1.

Figure 5.10 Silicon-rhodamine (SiR)

Figure 5.11: Active (A) and inactive (B) forms of SiR-PyPDS

Figure 5.12: C5, small planar molecule

Figure 5.13 1,6 diaminohexane (A) and N-BoC (B).

Figure 5.14: C5 and N-BoC-cadaverine prior to coupling (A) and with an amide link (B)

Figure 5.15 Trifluoroacetic acid (TFA) removes N-BoC to reveal a terminal amino group available for coupling

Figure 5.16: C5-spirofluorene

Figure 5.17: screenshot of C5-spiro-fluorene docking with ARN8-N1.

Figure 6.1: the RNA molecule TERC provides template reconstruction for a telomere reduced by cell division

Figure 6.2: step 1 of shelterin capping end of telomere

Figure 6.3: step 2 of shelterin capping end of telomere

Figure 6.4: telomestatin

Figure 6.5: 1KF1 ligation with telomestatin

Figure 6.6: incremental modifications to porphyrin intended to develop improved ADMET data.

Figure 6.7: Chimera interpretation of AP2 docked to 1KF1 with a VDW surface imposed on the quadruplex.

Figure 6.8: RMSD for AP2 docked with 1KF1 K⁺ co-ordination

Figure 6.9: Radius of gyration data for AP2 docked with 1KF1 K⁺ co-ordination
Figure 6.10: N-meso methylporphyrin (N-MM) and analogues N-MM2 – N-MM6

Figure 6.11: basic unit of peptide nucleic acids.

Figure 6.12: serial modifications to a thymine-PNA complex.

Figure 6.13: proposed mechanism for compound 6 interpolation of a single base exemplified by guanine

Figure 7.1: potential sequences for biplanar purine quadruplex folding in HAR1.

Figure 7.2: HAR1 guanine-adenine biplanar hybrid RMSD.

Figure 7.3: HAR1 guanine-adenine biplanar hybrid eRMSD

Figure 7.4: HAR1 guanine-adenine biplanar hybrid Radius of gyration.

Figure 7.5: HAR1 guanine-adenine biplanar hybrid frequency of hydrogen bond formation.

Figure 7.6: RMSD data for guanine and adenine quadruplexes derived from HAR1 sequence.

Figure 7.7: eRMSD data for guanine and adenine quadruplexes derived from HAR1 sequence.

Figure 7.8: Radius of gyration data for guanine and adenine quadruplexes derived from HAR1 sequence.

Figure 7.9: H-bond frequency of formation for guanine and adenine quadruplexes derived from HAR1 sequence.

Figure 7.10: Interplanar occupancy of waters for G4, A4 and hybrid models derived from HAR1 sequences

Figure 7.11. Secondary structures of HAR1A for Homo and Pan

Figure 7.12: secondary structure of HAR1 (Homo) and (Pan) predicted by base-pairing probabilities

Figure 7.13. 3D structures of HAR1 for Homo (A) and Pan (B).

Figure 7.14: square root transform of p values for query strands comprising 13 – 20 nts

Figure 7.15: predicted number of complements for nts in query strands comprising 12 to 20 nts

Figure 7.16: snora29 and snord35b alignments with HAR1A_Pollard


Figure 7.18: predicted complementarity for HAR1A and snord116-26

Figure 7.19: microdomains in HAR1A_Pollard

Figure 7.20: conserved bases in HAR1, snord115 and 5ht(2C)R

Figure 7.21: Mg²⁺ ligation with phosphate oxygens of 5UNE
Figure 7.22: AutoDock calculation for snord35b ligation with the HAR1 fragment located between helices C/H4 and H3 (5UNE).

Figures 7.23a and 7.23b: Histograms for melting temperatures of potential G4 sequences and non-potential G4 sequences

Figure 7.24. Duplex RNA strands with guanine-uridine pairs

Figure 7.25. Ca$^2+$ binds with 6 hydrogen bonds in Homo duplex RNA G-U pairs

List of Supplementary figures

Available in Supplementary Material

List of tables

Table 1.1: non-guanine quadruplexes. Descriptor suffixes ‘8’ and ‘12’ define the number of bases forming planar quartets within a system.

Table 2.1: docking characteristics for biplanar and triplanar guanine quadruplexes non-methylated and O6-methylated. Molecular Docking Server and AutoDock Vina were used as comparative methodologies. Although the former analytical software relies on some aspects of AutoDock Vina for calculation, it provides an extended interface for visualizing results. AutoDock has no GUI facility.

Table 2.2: fluctuations in deoxyribose configurations for triplanar guanine quadruplex K$^+$ co-ordination

Table 2.3: fluctuations in ribose configurations for RNA biplanar guanine quadruplex.

Table 2.4: fluctuations in deoxyribose conformations for triplanar non-methylated guanine quadruplex and O6-methylated guanine quadruplex.

Table 2.5: fluctuations in ribose configurations for biplanar non-methylated guanine quadruplex and O6-methylated guanine quadruplex. Data were compiled from visual inspection of .pdb structures opened in Chimera after applying Script 3 (1.5.6.2).

Table 3.1: fluctuations in ribose configurations for triplanar cytosine quadruplexes all residues with K$^+$ co-ordination.

Table 3.2: fluctuations in ribose configurations for biplanar cytosine quadruplexes all residues with no co-ordinant or with K$^+$ co-ordination.

Table 3.3: fluctuations in ribose configurations for triplanar thymine quadruplexes all residues with K$^+$ co-ordination.

Table 3.4: fluctuations in ribose configurations for biplanar thymine quadruplexes all residues with no co-ordinant or with K$^+$ co-ordination.
Table 4.1: sequence for triplanar guanine-adenine-guanine hybrid complex.

Table 4.2: melting temperatures, enthalpy and entropy values for triplanar guanine, cytosine, thymine and adenine 22-mers.

Table 4.3: Global PMF minima for all bases in ADN12 systems

Table 4.4: Global PMF minima for all bases in ARN8 systems

Table 4.5: fluctuations in deoxyribose configurations for triplanar non-methylated adenine quadruplex.

Table 4.6: fluctuations in ribose configurations for biplanar non-methylated adenine quadruplex

Table 4.7: fluctuations in deoxyribose configurations for biplanar non-methylated adenine quadruplex and N6-methylated adenine quadruplex.

Table 4.8: fluctuations in ribose configurations for biplanar non-methylated adenine quadruplex and N6-methylated adenine quadruplex.

Table 5.1: N6-methylation sequences for adenine in SARS-CoV-2 mRNA

Table 5.2: locations identified for potential supramolecular complexes and purine or pyrimidine quartets in SARS-CoV-2

Table 5.3: description of potential adenine quadruplex locations

Table 5.4: C5 binding energies and number of carbons in aliphatic linker.

Table 6.1. ADMET data for telomestatin, porphyrin and porphyrin analogues

Table 6.2: binding data predicted by ADT and MDS for telomestatin, AP2 and AP3 docked to 1KF1.

Table 6.3: ADMET data for APT2-APT2.

Table 6.4: ADMET data for N-meso methylporphyrin and analogues

Table 7.1: ribose configurations for G4 A4 and guanine-adenine biplanar hybrid quadruplex.

Table 7.2: $p$ values for nucleotides 13-20 ($n$) in query strand

Table 7.3: Bonferroni-corrected data for snords with $p \leq 2.0 \times 10^{-4}$, $\alpha = 0.005$

Table 7.4: snords with $p \leq 2.0 \times 10^{-4}$, $\alpha = 0.005$ prior to data correction

Table 7.5: Autodock optimum binding affinities for divalent metal cation ligation with SUNE or HAR1.

Table 7.6: Guanine / cytosine ratios and calculated melting temperatures for potential and non-potential G4-folding sequences
List of computer scripts

All scripts are detailed in Supplementary Material under Section 1.6.

**Script 1 (1.6.2.1): annotated command line script for NAMD preparation.** A remark is indicated by #; text drafted in American English. Submitted molecule, referred to as ‘test’, is protonated and a protein structure file written. This model is solvated with a 15Å water box and neutralised with the requisite number of Na⁺ counterions. The solvated, ionised.psf and protonated, ionised.pdb files are submitted for MD calculation.

**Script 2 (1.6.2.2): configuration file.** Input files (ionised.psf and ionised.pdb) are generated in the NAMD preparation file. Parameters, force-field, temperature, pressure, timings and periodic boundary conditions parameters are annotated. Amongst numerous post-simulation files generated, the trajectory file (extension dcd), is required for all subsequent analyses when downloaded in VMD with the ionized protein structure file.

**Script 3 (1.6.3.2): Single frame visualisation.** After ionised.psf and .dcd files are uploaded, the number of frames in the simulation is set as a variable ($nf). Each frame in $nf is set as a second variable ($i) which can be written as a single image visualized in Chimera. If $i is advanced incrementally, each frame of the entire simulation is available for inspection as a single Chimera file. Detailed structural comparisons and monitoring at any point is subsequently possible. The output folder is large.

**Script 4 (1.6.3.3): Radius of Gyration Tcl files.** Scripts 4a and 4b set parameters for the calculation based on centre of mass for the molecule and co-ordinates for each atom in the system. Script 4c calculates the molecular radius of gyration from these parameters. Script 4d initiates the calculation and output is returned as a text file which can be transferred to a suitable plotting programme.

**Script 5 (1.6.3.4): Calculate waters within a specified distance of specified residues.** This is adapted from script 3. The variable $i is set for individual waters in each frame which is advanced incrementally as $nf. The protein structure file only accepts “OH2” to identify waters. The script has a maximum tolerance of 2 Å for all waters relative to residues 1 2 7 8 14 15 19 20 of guanine O6 or adenine N6 in b poultry quadruplexes.

**Script 6 (1.6.3.5): Calculate angles or dihedrals.** A simple command after ionized.psf and dcd files are uploaded for all specified angles or dihedrals in a system. Individual angles or dihedrals can be calculated by including a frame number instead of all frames. Data are returned as text files which can be subsequently opened in suitable plotting software.

**Script 7 (1.6.3.6): run ellipsoidal RMSD.** This script bypasses the extensive option in VMD to a format compliant with Barnaba eRMSD calculation. A log file is created which submits the original test molecule as a reference .pdb file which is then compared with the topology of each frame in the simulation. Output is returned as a text file which can be transferred to a suitable plotting programme.
ARCHER2 access protocol (1.6.4): the script is written for Unix systems and assumes a public passkey and private passkey have been agreed with the ARCHER SAFE account and a MobaXterm interface has been downloaded.

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADT</td>
<td>AutoDock tools</td>
</tr>
<tr>
<td>DSV</td>
<td>Discovery Studio Visualiser</td>
</tr>
<tr>
<td>eRMSD</td>
<td>ellipsoidal RMSD</td>
</tr>
<tr>
<td>MDS</td>
<td>molecular docking server</td>
</tr>
<tr>
<td>NAMD</td>
<td>nanomolecular dynamics</td>
</tr>
<tr>
<td>nsp</td>
<td>non-specific protein</td>
</tr>
<tr>
<td>NPT</td>
<td>constant pressure simulation</td>
</tr>
<tr>
<td>NVT</td>
<td>constant volume simulation</td>
</tr>
<tr>
<td>PAGQ</td>
<td>potential guanine quadruplex-forming sequence</td>
</tr>
<tr>
<td>PBC</td>
<td>periodic boundary conditions</td>
</tr>
<tr>
<td>PBD</td>
<td>protein data bank</td>
</tr>
<tr>
<td>PMF</td>
<td>potential of mean force</td>
</tr>
<tr>
<td>Rg</td>
<td>radius of gyration</td>
</tr>
<tr>
<td>RMSD</td>
<td>root mean square deviation</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>VDW</td>
<td>Van der Waals</td>
</tr>
<tr>
<td>VMD</td>
<td>visual molecular dynamics</td>
</tr>
</tbody>
</table>
Overview

(i) Overview of DNA and RNA quadruplexes
(ii) Thesis overview
(iii) Value of *in silico* inquiry
(iv) Metrics
(v) Purine and pyrimidine quadruplexes
(vi) Relevance to RNA viruses
(vii) Relevance to long non-coding RNA
(viii) Possible pathogenicity proposals
(ix) Future perspectives: quadruplex-targeting ligands
General Introduction

The thesis reports an original in silico investigation into molecular structures folding in DNA and RNA. These molecules do not reflect classical Watson-Crick binding and when derived from guanine, have implications for chemotherapy. Analogues in cytosine, thymine and adenine are modelled in this research and molecular dynamics applied for analysis. Were adenine structures to fold, stability would compare favourably with guanine equivalents. Association with these complexes in mRNA viruses and long non-coding RNA are proposed which further suggest therapeutic options.

(i) Overview of DNA and RNA quadruplexes

Gels comprised exclusively of four guanines self-associating into a planar quartet were definitively identified in the early 1960s (1). Research in vitro later showed quartets could further associate into a supramolecular structure developed from two, three or more quartets stacking vertically (2). The system is stabilised by co-ordination with a metal cation, typically K⁺. Various topologies are known and derive primarily from the length of the primer sequence d(TTAGGG)n (3). Possible in vivo interest regarding these structures, guanine quadruplexes (G4s), has been greatly amplified by macro-scale analysis of vertebrate genomes (4). Numerous sequences capable of folding into G4s were found to be prevalent (5). Subsequently, bacterial and mitochondrial genomes were also found to contain similar sequences (6). Implicitly therefore, in vivo G4s could have a biological role. Research into this aspect has generally focussed on selective chemotherapy contingent on the association of G4s with oncogenes (7). Terminal regions of chromosomes, telomeres, are crucial for cell maintenance of recombination and degradation.

If unregulated, telomeres would undergo adverse reiterative extension. Such a runaway process could be interpreted by cellular oversight mechanisms simply as damaged DNA and flagged accordingly as deleterious. Enzymatic repair in this circumstance would be inappropriate. In effect, G4s as telomeric structures, signal an alternative to repair. Where possible therefore, a d(TTAGGG)n sequence folding into G4s disallows telomerase from inconsonant extension of the chromosome. Although telomeric G4s are closely associated with telomerase inhibition (8), since telomerase affinity is preferential to linear DNA, the enzyme may also have the capacity to dismantle G4s (9).

The original proposal of the d(TTAGGG)n sequence as a sine qua non of G4 folding was displaced as the possibility of non-telomeric G4s also became apparent (10). Some 350 000 genomic sites capable of supporting G4 folding were predicted with numerous other guanine-rich sequences (11). However, understanding of G4 function is confounded by observations suggesting both adverse and beneficial roles for the structures (12). This paradox is exemplified by telomeric extension remaining unchecked and subsequent oncogenesis despite the presence of G4s. The issue is currently unresolved with the implicit positive and negative roles of G4s as persistently controversial (13).
Despite noted adverse aspects, G4s are strongly associated with suppression of telomerase (14). Accordingly, drug design has often focussed on stabilising ligands. Porphyrin ring structures resemble guanine quartets and some analogues are characterised by substantial binding affinity (15). Both characteristics suggest an encouraging anti-cancer propensity. Poor selectivity remains a challenge however, as ligands do not discriminate between double-strand (duplex) DNA and neoplastic telomeres. Specificity is therefore a major concern of drug design.

Whilst G4s folding in DNA has received widespread attention, research surrounding RNA quadruplexes is still inchoate. Both nucleic acids have prime importance in the replication of all living organisms and viruses. These macromolecules are also ultimately responsible for cell maintenance. DNA and RNA have many similarities, but also significant differences: in particular their location and function. Moreover, a significant proportion of RNA transcripts do not code for proteins, implying a majority are involved in other processes (16). RNA strands can vary greatly in length and are predominantly single-strand.

However, apparent similarities between complexes arising in these nucleic acids have to be qualified. Apart from obvious structural differences, the two species show varied biochemical and environmental characteristics. DNA remains in the nucleus and harbours genetic information; cytoplasmic and nuclear RNAs are involved in protein synthesis. Whilst both locate in mitochondria, only DNA is generally bound to histones (17) (18). RNA is found in other organelles such as the endoplasmic reticulum. RNA quadruplexes show less hydration and may be more thermodynamically stable than DNA equivalents (19) (20).

Possible mediation of gene regulation by G4s is implicit in folding sequences located in promoter regions (21) (22). Because oncogenes exemplified by C-myc, Bcl 2 and K-ras are expressed in these domains, better understanding of quadruplex form and function is a continued research aspiration. As noted, telomeric DNA is a substrate for telomerase (23) and runaway activity of the enzyme underpins tumorigenesis (24). It is now known that telomeric RNA is transcribed in similar loci and observed in the cell lines of numerous species (25) (26). Subsequent consideration has been given to a possible role for these molecules in senescence and mutagenesis. RNA quadruplexes have subsequently been characterised comprising a uniquely parallel loop system and the RNA sequence has enzymatic immunity from disintegration (27). Implicit with this finding is that RNA quadruplexes might have chemotherapeutic value if any interaction with telomerase has equivalence with DNA counterparts. This further suggests that not only RNA supramolecular structures per se are worthy of consideration, but also systems derived from other nucleic acid bases. A well-characterised quadruplex (PDB 1KF1) was identified as a prototype for modification and analysis in this study. The complex is a 22mer derived directly from the telomeric sequence noted earlier.
(ii) Thesis overview

This thesis presents *in silico* investigation of non-canonical supramolecular purine or pyrimidine structural reconfigurations in nucleic acids which may be influential in human physiology and viral anatomy.

Potential chemotherapeutic value of RNA quadruplexes mitigates assessing their thermodynamic stability and becomes an important adjunct to the thesis. Research to date appears to essentially explore guanine-based quadruplexes. However, this study develops the concept that other bases may self-associate into similar structures and elicits further possibilities of intervention. Consequently, potential ligands extending quadruplex stability pertinent to telomerase inhibition in human physiology are considered. Moreover, some speculation extends to compromising the replication mechanism of RNA viruses.

The computational chemistry herein endeavoured to address several points: quadruplex integrity following various base-substitutions, influence of cationic co-ordination, different starting geometries and disruptive effects of N6-methylation. Given single-stranded RNA is much more susceptible to chemical stress and variation than DNA complexes, an explanatory mechanism was also sought. Some reflection was given to why the actual process of folding into an RNA quadruplex may stabilise the molecule. Particular attention was paid to the additional 2'-OH moiety of ribose. A comparative study of DNA and RNA quadruplexes assessing any influence of different core bases and various co-ordinating cations is timely. There is increasing interest in RNA complexes thought to fold in non-coding regions of the genome, especially as that folding may be mediated by cationic ligation (28). A collateral interest lies in their potential response to pathogenesis (29) (30).

Molecular dynamics analysis was used to elucidate properties of these structures and gain insight into specific biological contexts. The work considers two main facets of inquiry: the stability of such structures, and their plausible impact on biochemical systems. Prior to the main text, an Introduction outlines the primary rationale of the thesis in eight descriptors. Each delineates a framework for individual Chapters but also as a commonality underpinning all. Chapter 1 describes methodologies and metrics. Chapter 2 contextualises well-characterised guanine-based macromolecules self-associating into unorthodox conformations. Analogues predicated on these structures and not previously proposed were modelled and their physico-chemical characteristics assessed. Chapter 3 covers cytosine- and thymine-based complexes. The majority were found to be relatively transient, but a mechanism is introduced to clarify a paradoxical feature associated with guanine-based systems. Chapter 4 describes adenine-based analogues. Notably, these systems configured in mRNA show less dynamic activity than DNA equivalents and were non-transient into biologically relevant times. This observation may be especially relevant
in a physiological context, particularly relating to specific sequences of RNA virus genomes. Chapter 5 details the application of Molecular Dynamics and interpretation of data. Possible mechanisms compromising cellular functions, discussion of pathogenesis and concepts in rational drug design are proposed in Chapter 6. Implications of some results are considered from the perspectives of both viral replication and also long non-coding RNAs for genetic evolution in Chapter 7. A thesis retrospective is followed by brief concluding remarks.

Various folding patterns are known for both DNA and RNA quadruplexes (31). However, because the quadruplexes functioning as scaffolds for this work are so-called anti-parallel structures (32), this feature was uniquely analysed for the derived models. Moreover, this structure is generally observed in the majority of biplanar and triplanar quadruplexes (3) although some RNA complexes show both parallel and anti-parallel looping (33). DNA and RNA systems can be discriminated by folding directions of loops. Since widespread and comprehensive data for DNA quadruplexes are available, deriving RNA analogues from such models is cautiously feasible. Because uridine substitution for cytosine alters local topology, some care is needed (34) (35). A model with the lowest free energy would be preferable. A facility is available in the modelling software accessed in this study to find the lowest energy configuration.

Results were assessed in the light of specific biological circumstances. Some supramolecular structures in DNA paradoxically appear to have both beneficial and pathogenic attributes; some may be involved in viral RNA activity. This study also describes original work suggesting previously non-characterised adenine quadruplexes may have relevance to RNA viruses and non-coding RNA.

(iii) Value of in silico inquiry

Historically, many areas of scientific inquiry have benefitted from the contributions of mathematical modelling. Its value in molecular biology has a specific and direct correlation with the exponential growth of computing power. Consequently, the paradigms of modelling have been increasingly applied in a molecular biological context to obviate the inherent complexity of cellular systems and interactions (36). Propitiously, in silico investigation with this virtual approach allows rapid data returns. Moreover, enhanced resolution and detailed monitoring are often available with high specificity. As noted previously, mathematical predictions of non-linear biological systems require an appropriate understanding of the assumptions made. This caveat underpins the ergonomic design of direct experimentation.

Hitherto, investigating guanine quadruplexes (G4s) has combined experimental techniques with in silico prediction. Detailed structural insights have been provided from
NMR (37), X-ray crystallography (38) and fluorescent tags for visualisation (39), both *in vitro* and *in vivo*. A well-characterised human telomeric G4 (PDB 1KF1) has provided detailed insights into molecular stability under various conditions of ion concentration, hydration models, chromophore ligation and loop topology (40). Higher order 45-mer and 93-mer G4 models have also been investigated with proportionate increasing stability as calculated by free energy, a finding that resonates with experimental observation (41). Since direct experimentation at a genomic scale can be very limited, identifying sequences *in silico* has complemented biophysical G4 characterisation. The two approaches support and reinforce each other. Previous research has suggested that RNA G4s can display enhanced stability compared to DNA equivalents and may strongly influence gene expression if located at crucial sites (55). Computational methodology and metrics supporting research to date into guanine quadruplexes underpin this thesis. The technique continues to show advances in reliability, speed of analysis and accessibility, prompting improved understanding of key aspects in quadruplex form and function. Targeted chemotherapy benefits from accurate information.

(iv) Metrics

Scripts were drafted for Molecular Dynamics analyses of supramolecular models for simulation times of 500ns. Several Linux servers were accessed for molecular dynamics and molecular work: 256-core allocation at the ARCHER2 National Supercomputer, on-campus PowerEdge T640 96-core, 16-core and 8-core personal Desktops. All servers were supported by Linux OS. Data were collated with a Microsoft OS 8-core personal Desktop. NanoMolecular Dynamics (NAMD) and Visual Molecular Dynamics (VMD) were the primary research tools, supported by Discovery Studio (DSV,) only accessible in Microsoft), and Chimera modelling software. Docking protocols were submitted to AutoDock and AutoDock Vina accessed in both Linux and Microsoft. Metrics used to evaluate trajectories were Root Mean Square Deviation, Radius of gyration, frequency of hydrogen bond formation. For the latter, a donor-acceptor distance was set to the defaults of a maximum 3 Å and an angle cut-off at 10° either way from a notional perpendicular bond. To evaluate the strength of binding between interacting atoms of adjacent bases in participating quartets, frequency of hydrogen bond formation and distance between N6 and H1(N1) were measured.

Potential of Mean Force torsion of glycosidic angles analysis was applied to adenine systems only. Ellipsoidal Root Mean Square Deviation (eRMSD) is more specific to nucleic acid structures, and was evaluated with Anaconda3 and Barnaba software. Plotting programmes were XmGrace, Multiplot, OriginLab and GraphPad Prism. Data returns in this research indicate RNA quadruplexes derived from single-strand helices are not simpler analogues of their double-strand DNA counterparts. For example, some folding in RNA were shown to be more thermodynamically stable than in DNA. Co-ordination of
complexes with various cations was also assessed in Molecular Dynamics simulations. RNA models were methylated at adenine N6 to further assess stability.

Data for eRMSD and Radius of gyration are returned as text files and visualised in GraphPad Prism. Frequency of hydrogen bond formation and RMSD data are formatted by NAMD as trajectories in Multiplot. Transferring these to XmGrace allows further analysis such as histogram formatting. Empirical values for Frequency of hydrogen bond formation and RMSD are not accessible as text files or similar: data returns can only be opened in downstream analytics.

A licence waiver for the National Supercomputer (ARCHER2) was granted for 6 months. Remote access allowed much shorter calculation times for 500ns molecular dynamics simulations and docking protocols for long non-coding RNA interactions with purine quadruplexes.

(v) Purine and pyrimidine quadruplexes

Whilst guanine-based DNA complexes have been very well characterised for some time, limited attention has been paid to RNA structures [42]. Ostensibly, there has been little or no less consideration of cytosine-, thymine-, uridine- and adenine-based quadruplexes for both DNA and RNA. To compare and contrast these variants, a series of quadruplexes uniquely derived from these bases were modelled and analysed for both nucleic acids. This approach has evidently not been previously proposed. Most cytosine and thymine systems were found to be relatively transient. However, a conjectural context was outlined in which transient cytosine quadruplexes might compromise the beneficial aspects of telomeric guanine quadruplexes. Adenine quadruplexes potentially folding in DNA and RNA have implicit stability comparable with guanine equivalents. This result precipitated much further work, particularly with a biological perspective.

(vi) Relevance to RNA viruses

The unanticipated finding that adenine quadruplexes are non-transient in silico was applied to the 3’polyadenylated region capping the genome of SARS-CoV-2. This sequence is associated with viral replication [43]. Given the imminent onset of global infection, thesis work readily focused on this observation. Consideration was given to a provision for differentiating adenine quadruplexes in 3’polyadenylated RNA viruses and those possibly folding in human RNA. Methylation at N6-adenine was identified as a tentative route because this modification stabilises 3’polyadenylation [44]. There appears to be
very little similarity between host and viral mRNA in the binding site for the enzyme prompting polyadenylation. Consequently, there may be an exploitable difference for feasible intervention. Current research appears not to document compromising host immune response to viral infection predicated on 3’polyadenylation and endo-enzymes. Inspection of the genome is also suggesting other potential routes for chemotherapy. Because replication mechanisms remain similar across many species, the data from this study could be applicable to other RNA-viruses such as HIV and Hepatitis C. Species diversity and high mutation rates amongst RNA viruses confound vaccine efficacy: alternative avenues to intervention are implicitly propitious.

(vii) Relevance to long non-coding RNA

Thermodynamic properties of a guanine-adenine hybrid model closely resembled quadruplexes entirely derived from either base. The hybrid structure found unexpected significance as the thesis progressed investigating singular aspects of an exceptional long non-coding RNA strand. A specific sequence was found to be capable of folding into a hybrid quadruplex at a critical location. Furthermore, sequences upstream and downstream of the site were noted as regions in which guanine quadruplexes could fold. Numerous small molecules are now known to bind with complexes, some of which may enhance stability (45). Taken together, these two observations were suggested as reinforcing current contention that the long non-coding RNA strand may accord an evolutionary advantage between Homo sapiens and Pan troglodytes. Both developments would imply enhanced stability. Since the long non-coding RNA strand and small nucleolar RNAs are both nuclear, possible interactions were surveyed. Apparently having no direct precedent, the work found occasional non-canonical base-paring that might underpin ionic ligation moderating RNA function and folding. Only a small fraction of small nucleolar RNAs analysed were predicted to interact with the long non-coding RNA strand. However, several intriguing results were noted including two indicative of cogent hydrogen bonding with 3’polyadenylation.

(viii) Pathogenicity

The extensive interest in DNA guanine quadruplexes has often eclipsed an inherent paradox: whilst single strand telomeric systems suppress an oncogenic enzyme, other genomic models closely correlate with biological dysfunction. Current research finds RNA structures are also in this dichotomous orbit. The challenge for medicinal chemistry is to address stabilising quadruplexes beneficial to human physiology but disrupt complexes favourable to viral anatomy. Selectivity is therefore the sine qua non of rational drug design. Chapter 5 outlines a nascent synopsis which may be relevant to RNA viruses.
Chapter 6 considers differences in the chemistry of DNA and RNA which might partially address the challenge. The Chapter also proposes three speculative scenarios in which quadruplexes could compromise cellular routine. Chemotherapeutic options analysed and discussed are quadruplex-binding ligands and intercalation of synthetic nucleic acids.

(ix) Future perspectives: quadruplex-targeting ligands.

Further work investigating these unorthodox molecules could contribute to understanding of their biological context and relevance in medicinal chemistry. Guanine quadruplexes are known to arise with a maximum abundance during viral replication within the host nucleus (46). Given the proportion of viral to host G4 population may be several orders of magnitude (47), this suggests promising selectivity as only infected cells will have greatly increased numbers of G4s. Furthermore, viral G4s are known to regulate for some viruses during lytic and latent conditions (47). This hints at the intriguing possibility of intervention at all stages of the viral life-cycle and a dividend for rational design of small neoteric ligands targeting supramolecular complexes. For example, carboxy pyrimidostatin has high affinity for guanine quadruplexes; docking protocols in this work established adenine quadruplexes also have this property. Guanine quadruplexes folding in DNA have also been promising targets for anti-cancer intervention (48). But paradoxically, some G4s have recently been linked to neurodegeneration (49) (50). This observation implies current understanding of supramolecular complexes in nucleic acids is far from complete; a partial resolution may lie in ancillary assessment of non-guanine quadruplexes. Direct visualisation of guanine complexes in vivo has recently been achieved (39); a route to accomplishing a similar effect for adenine quadruplexes is outlined in Section 5.8.
Chapter 1   Methods and Metrics

1.1 Molecular Modelling

Molecular models were constructed in Discovery Studio Visualiser (Client) v16.1.0.15350, BioVia Draw 4.1 (BVD) and UCSF Chimera 1.12 (build_41623). Some models were visualised in PyMol 2.2.1. Software for molecular dynamics analyses were UCSF Chimera 1.12 (build_41623) and NanoMolecular Dynamics NAMD_2.13_Linux-x86_64-multicore; both were accessed via UBUNTU command line. Because theoretical constructs can vary in molecular dynamics software, two methodologies were used to compare and contrast data. Initial stages of molecular dynamics preparation in Chimera accessed the Antechamber sub-routine of the AMBER19 toolkit (AmberTools19.tar.bz2) downloaded from https://ambermd.org. Chimera has an integral Graphical User Interface, but visualisation in NAMD relies on an external programme: Visual Molecular Dynamics (VMD 1.9.3) (51). Molecular dynamics preparation in NAMD requires a configuration file and command-line script (Scripts 1 and 2 respectively). Force field CHARMM27 was applied for both NAMD and Chimera as a recent study concludes it is optimal for nucleic acid simulations (52). Conversion of a trajectory file to a Protein Structure File in VMD for some analyses (eg. eRMSD) is limited: the metric supplies an overall macro perspective.

(PDB) 1KF1 acted as a scaffold for triplanar DNA and RNA models (figure 1.1).
Figure 1.1a. Guanine quadruplex (1KF1) located in single-strand telomeric DNA. A: planar guanine tetramer co-ordinated by a monovalent cation, M⁺. B: three such quartets stack vertically linked by thymine-thymine-adenine loops and co-ordinated by at least two K⁺ ions. The ions re-locate from intramolecular to interplanar. C: structure in plan as determined by Chimera software. Cations are attracted to the very electronegative central column of oxygens shown in red. D: structure C rotated through 90° around an arbitrary vertical axis shows how the three planar quartets are stacked within the complex. Interplanar loops are depicted as ribbon in C and D.

Biplanar complexes folding in RNA have some differences and limited commonality with DNA equivalents (figure 1.1b). Initial self-association of four RNA guanines into a planar quartet does not coalesce around a co-ordinating cation and there is currently no confirmed evidence that helicases exist to dismantle a biplanar RNA quadruplex (53). Moreover, the folding mechanism is also uncertain whereas DNA structures rely on cationic attraction. DNA systems are exclusively nuclear, but mRNA complexes can arise both in the nucleus (detailed in Chapter 5) and as nuclear or cytosolic complexes in long non-coding RNAs (IncRNAs, described in Chapter 6).

Figure 1.1b: biplanar RNA guanine quadruplex (eg 2KF8), in cellulo relevance and location. The non Watson-Crick (Hoogsteen) interaction is shown by red dotted lines.

Until recently, RNA guanine quadruplexes were considered globally unfolded in vivo therefore casting doubt on any biological relevance (54). Direct visualisation of DNA equivalents has been achieved (39). Section 5.6 explores this finding in further detail. Implicitly, similar characteristics may apply to RNA models and if so, suggesting these nucleic acid supramolecular species follow an iterative dynamic folding process. Folding may only occur under specific circumstances such as cellular stress or a specific stage in the cell cycle (55). Change in the binary folded / unfolded equilibrium may underpin the shift of a quadruplex from relatively inert to pathogenic (56) (57). This conundrum might be partially resolved if an RNA quadruplex is dismantled by binding a helicase, but methylation at adenine N6 (m6A), compromises helicase activity (58). m6A modification is discussed in detail throughout this thesis but particularly in Chapters 4 and 5.
Substitutions in 1KF1 of guanine for adenine, thymine or cytosine were made for triplanar DNA models. Cytosine substitutions illustrate the archetype in figure 1.2:

![Figure 1.2](image)

**Figure 1.2. Cytosine substituted for all guanine.** The 22-mer complex 1KF1 folds from the sequence AGGG(TTAGGG)$_3$, hence the cytosine structure has ACCC(TTACCC)$_3$. Substitutions for triplanar adenine and thymine quadruplexes follow the same template sequence. Quartets derive from x2-x8-x14-x20, x3-x9-x15-x21 and x4-x10-x16-x22 where x = adenine, cytosine or thymine.

Recent research concludes DNA biplanar guanine quadruplexes are more stable than triplanar analogues and cation co-ordination appears to be absent (33). Accordingly, biplanar quadruplexes were constructed from (PDB) 2KF8 as scaffold. The complex is located in single-strand human telomeric DNA. Due to the coincidental onset of the SARS-CoV-2 pandemic and relevance of 3’polyadenylation to RNA viruses, only adenine-based DNA and RNA structures were modelled. For the latter, cytosines were replaced by uridine.

All models constructed for both DNA and RNA are itemised in Table 1.1 and identified by the terminology used for molecular dynamics simulations and analysis. Models for adenine and cytosine triplanar quadruplexes and adenine biplanar quadruplexes co-ordinated with physiologically relevant cations were also constructed.
Table 1.1: non-guanine quadruplexes in DNA (blue) and RNA (red) constructed for molecular dynamics analysis. Triplanar models derived from (PDB) 1KF1; biplanar models from (PDB) 2KF8. N1, N3 and N7 suffixes are explained in Chapter 4.

<table>
<thead>
<tr>
<th>Triplanar</th>
<th>adenine</th>
<th>cytosine</th>
<th>thymine</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADN12</td>
<td>CDN12</td>
<td>TDN12</td>
<td></td>
</tr>
<tr>
<td>ADN12_Ca²⁺</td>
<td>CDN12_K⁺</td>
<td>TDN12_K⁺</td>
<td></td>
</tr>
<tr>
<td>ADN12_H³O⁺</td>
<td>CDN12_L⁺</td>
<td>TDN12_L⁺</td>
<td></td>
</tr>
<tr>
<td>ADN12_K⁺</td>
<td>CDN12_N⁺</td>
<td>TDN12_N⁺</td>
<td></td>
</tr>
<tr>
<td>ADN12_L⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADN12_Mg²⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADN12_Na⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADN12_NH₄⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biplanar</th>
<th>adenine ADN8</th>
<th>cytosine CDN8</th>
<th>thymine TDN8</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARN8_N1</td>
<td>ARN8_N3</td>
<td>ARN8_N7</td>
<td>CDN8</td>
</tr>
<tr>
<td>ARN8_N1_Ca²⁺</td>
<td>ARN8_N1_meth</td>
<td>ARN8_N3_meth</td>
<td>ARN8_N7_meth</td>
</tr>
<tr>
<td>ARN8_N1_H³O⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARN8_N1_K⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARN8_N1_L⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARN8_N1_Mg²⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARN8_N1_Na⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARN8_N1_NH₄⁺</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adenine quadruplexes were modelled in both DNA and RNA, and derived from three experimentally-validated adenine quartets (59). These are labelled A₄-N1, A₄-N3 and A₄-N7 in Figures 1.3a, 1.3b and 1.3c. The suffix refers to three possible configurations contingent on H-bonding between N6 and N1, N3 or N7 of an adjacent base. Biplanar RNA_N1 complexes were also analysed with N1 methylation for each configuration. Figure 1.3d shows an exemplar RNA model with two A₄-N1 quartets centrally located.

Adenine cytosine or thymine/uridine substitutions were made for guanines in each quartet. Loops in DNA models matched 1KF1 sequences. The sequential pattern for these triplanar quadruplexes is noted in Figure 1.2. Thymine in loops for RNA models was replaced by uridine; similar substitutions also apply to biplanar adenine and cytosine structures.
Figures 1.3a – 1.3c: Adenine quartets found to be experimentally stable. H-bonding occurs between N6 and either N1, N3 or N7 of adjacent bases. Figure 1.3d: bilobar RNA quadruplex of an oligonucleotide containing 22 adenines. Quartets (shown in green) comprise bases A1-A8-A14-A20 and A2-A7-A15-A19. Dotted lines indicate predicted intraplanar hydrogen bonds. Selected bilobar RNA models were methylated at N6.

RNA models were derived from a 22-mer bilobar guanine quadruplex (PDB: 2KF8), as a scaffold. The complex is located in human telomeric DNA. The sequence studied and capable of folding is: GGGATAGGGTAGGGATAGGGT. Algorithms applied by open resource packages such as Quadparser and G4Hunter, search for much longer potential sequences (5) (60). Since models analysed here derive essentially from the telomeric sequence, extended algorithms were not accessed. Furthermore, comparison of results from different algorithms can be challenging to interpret accurately. Substitution of
guanine bases by adenine and uridine for cytosine in the 2KF8 model was achieved with the swapna instruction available in Chimera command line. The modified sequence was: AAAAAUAAAUAATAAAAUU where bases highlighted in the same colour form a quartet.

1.2 Molecular Dynamics

1.2.1 Introduction

Atomic and molecular motion can be studied in a virtual environment with the simulation field of molecular dynamics. Where atoms and molecules have freedom of interaction for a given period, the progress of thermodynamic change can be specified mathematically and geometrically. Atomic and molecular trajectories derived from imposing a force field can be solved numerically as classical Newtonian mechanics. Potential energy and interatomic forces are calculated within the force field parameters. The numerical complexity of molecular systems may vary between hundreds to many thousands of particles. Time scale for a simulation is an iterative process requiring millions of incremental steps. Since each time step is scripted in the very low femtosecond range, the duration of simulation in real time can vary from days to months contingent on computing capacity. The physical dimensions of a molecule are also a major contributive factor. The simulated time however, generally ranges from nanoseconds to milliseconds.

The computational effort is almost exclusively devoted to calculating electrostatic forces. Were these values to be reckoned as separate quantities between pairs of individual particles, the data would scale excessively. The problem of naivety is overcome by algorithmic computation for a field of interactions considered as a periodic system. Particle Mesh Ewald, originally constructed for crystallographic analysis, provides a suitable method for calculating such long-range interactions. The algorithm requires the molecular system to be located in a virtual box of solvent and charge neutrality imposed for precise computation. Water was the solvent used exclusively for all simulations. Where Periodic Boundary Conditions are applied to this infinitely-replicated neutral cube (or truncated octagon) of charged particles, a Particle Mesh Ewald summation presents a mathematical interpretation of the molecular system’s electrostatic energy as a matrix. Application of the field to molecular dynamics is therefore eminently suitable. As noted in Section 1.1, two different methodologies were accessed for the study, Chimera and NAMD.
1.2.2 Preparation of models in NAMD and Chimera

Given its broader and simpler approach Chimera was used for preliminary assessment of potential complexes. More detail and analysis derived from molecular dynamics simulations run in NanoMolecular Dynamics NAMD_2.13_Linux-x86_64-multicore (NAMD). The default CHARMM27 force field applied for both methodologies. A hydration cube with 15 Å clearance was applied, giving a total of 19 374 atoms in DNA systems and 16 282 in RNA models. The Particle Mesh Ewald summation set to a cut-off at 8 Å.

For Chimera analyses, a truncated octagonal box of 17 230 waters with a 15Å clearance was imposed. Models in both methodologies were neutralised with 22 K+ counter-ions. The default neutralisation option for NAMD was applied via VMD Auto-ionize Plug-in v.1.5 balancing the system with Na+. Counter-ions were added during the system preparation stage in Chimera. For both methodologies, NaCl concentration at 140 mM was set to reflect a cellular environment at pH 7.3 (61). Protonation can be achieved in the integrated Chimera Prep Structure option rather than having recourse to the H++ facility required by AMBER.

1.2.3 Periodic Boundary Conditions

Water is significant in MD simulations: molecular biology operates in an aqueous environment. To more accurately reflect in vivo conditions, periodic boundary conditions (PBCs) were applied. PBCs regard a model under analysis as a unit cell surrounded by identical virtual cells with atoms of the virtual systems interacting mathematically with atoms of the real system. Adding PBCs necessarily increases the total number of atoms in a system and consequently greatly extends the real-time required for simulations. Water-box or sphere size is important: a sufficient distance should be assigned to avoid exposure of the molecule to a non-aqueous region. Structures can disintegrate rapidly without the protection of a hydration shell. Electrostatic interactions using the Particle Mesh Ewald summation require a system with PBCs to be globally neutral.

If PBCs are not applied, a cubic water-box automatically re-configures to a sphere during minimization and equilibration as this conforms to the lowest energy state. The hydrated complex is then calculated as an isolated sphere in vacuo. Solvating the macromolecule in a sphere prior to a simulation reduces preparation time and computer effort. However, the system is less reflective of a physiological environment because the sphere has a boundary with the surrounding vacuum. When PBCs are applied, the cubic water-box shape is retained and consequently, for the unit cell, interaction with water boundaries is effectively removed.
The water model used was AMBER TIP3P with an explicit solvent. PBCs were applied and systems were neutralised by addition of an appropriate number of Na\textsuperscript{+} counter-ions via VMD Auto-ionize Plug-in v.1.5. A median salt concentration of 150mM was implemented using the Tool Command Language (Tcl) interface of NAMD to reflect physiological conditions and pH retained at the default setting of 7.0. Simulations for both methodologies progressed through three stages.

1.2.4 MD production

**Minimisation by steepest descent** amends unfavourable atomic geometry by reducing an energy gradient from a high initial point to achieve a gradient with the lowest energy configuration. The calculation is iterative, and because each step is an approximation of gradient reduction, a number of steps are required. For the study, Chimera models were minimised with 1 000 0.02 Å steps of steepest descent and 10 0.02 Å steps of conjugate gradient. NAMD models Systems were minimised by 50 000 x 0.01 Å steps of steepest descent and 5 000 steps of equilibration. Initial and target temperatures were 0.0 K and 310 K respectively. The default setting of zero was used for the pressure variable. The Langevin algorithm was used as the barostat. A constant volume simulation (NVT) for 100 ps was run at 310 K with the barostat off to thermalise the system. To provide good equilibration, the models were then run in a constant pressure simulation (NPT) for 100 ps at 310 K with the barostat on.

**Equilibration** produces a system with no further tendency to change in concentration over time. Initial and target temperatures were 0 K and 310 K respectively, the latter value more accurately reflecting a physiological environment. The default setting of zero was used for the pressure variable. AMBER force field ffSB14 was applied to standard residues and Gasteiger to non-standard residues for Chimera models. A setting of 5 000 steps was used. The Particle Mesh Ewald summation with a cut-off set at 8 Å was used to compute wider electrostatic effects. To reduce computer time in Chimera models, constraints were added only for the quadruplex hydrogens and H-bond distances, a total of 253 atoms. The CHARMM27 force field was used for NAMD models, with 5 000 steps of equilibration.

Parameters for Periodic Boundary Conditions were calculated in NAMD at (49.48 0.0 0.0), (0.0 44.79 0.0) and (0.0 0.0 54.75) for the three spatial co-ordinates, and models centred at (0.0 -0.86 0.66). All co-ordinates are in Ångstrom units. To reduce computer time, simulations without PBCs were first run in Chimera to identify unstable models. For NAMD models, hydrogens were constrained using the SHAKE algorithm to further reduce computer time. The algorithm is suitable for iterative time-steps ≤ 5 fs, and therefore appropriate for this study where both methodologies have a default setting of 2 fs. The constraint addition does not apply to hydrogens in waters.
**Chimera Production.** A step value was set at the default of 1 fs. Hence 1,000,000 steps equated to trajectories of 1.0 ns. Initial temperature was 0.0K and the system heated to the nominal physiological temperature of 310 K. For MD simulation, the Berendsen Barostat was retained at a default of 1.0132 bar and the Nosé thermostat set for 310K. 8 CPUs were committed to each run.

**NAMD Production,** requires more input for preparation than Chimera. In addition to a PDB file, a Protein Structure File (.psf), Topology/Parameter (toppar) Files and a configuration file (.conf) are necessary. The .psf calculates binding data and other structural information such as atomic mass and charges not written into a .pdb file. It can be generated manually or using the Autopsf_builder option in NAMD. However, for molecular modification such as methylation, .pdb and .psf files were re-written manually. A specific toppar file to allow metal ions in the molecular system (toppar_c36_jul19.tgz) was included.

For production runs, initial temperature was 0.0 K and the system heated to the nominal physiological temperature of 310 K. Simulation time was 500 ns and a time-step of 2 fs was used for plotting trajectories. The relevant configuration file for production runs, PDB files and a template command line script are detailed in List of Scripts. 96 cores were available for each run with a Linux server. More extensive simulations for adenine quadruplexes were submitted to the ARCHER2 256 core facility located within the National Supercomputer Complex.

### 1.3.1 Metrics for molecular dynamics analysis.

The behaviour of all models during molecular dynamics simulations was analysed by root mean square deviation (RMSD) and ellipsoidal RMSD (eRMSD), radius of gyration $R_g$, frequency of H-bond formation and Potential of Mean Force Torsion (Torsional PMF). Anaconda3 and Barnaba were used to determine eRMSD data and *mdtraj* to plot trajectories (64). (65). Metrics used for the analysis of coordinating cation data were RMSD, eRMSD, $R_g$ and Torsional PMF. Potential of Mean Force torsional analysis was also used compute variation in glycosidic angles.

### 1.3.2 Root Mean Square Deviation (RMSD)

RMSD analysis plots an overall reflection of dynamic activity and by implication, stability. However, the metric does not necessarily reflect a detailed description of molecular reconfiguration during a simulation. NAMD relies on the standard equation, Eq.1.1:

$$ RMSD = \sqrt{\frac{\sum_{i=1}^{N} m_i (\vec{r}_i - \vec{r}_g)^2}{\sum_{i}^{N} m_i}} $$

Eq.1.1
where \( N \) = total number of atoms, \( m_i \) is the mass of atom \( i \), \( r_i \) = sum of coordinates for atom \( i \) at a specific distance, \( r_r \) = sum of coordinates for atom \( i \) at a reference position.

Frame 1 of a simulation was set as the reference for analyses in NAMD and Chimera. However, the latter software also has an option to calculate RMSD values for individual atoms or selected groups. Accordingly, whilst plots were recorded for all entire systems, specific RMSD values were calculated where more detail was required, eg. changes in glycosidic angles. Units are Ångstrom.

1.3.3 ellipsoidal Root Mean Square Deviation (eRMSD)

A dynamic equilibrium between the sequence of bases, molecular function and tertiary structure underpins all RNAs. Incorporating these factors into a simulation adumbrates their mutual interactivity and wider influence. Trajectories and analysis of RNA dynamic activity are accurately addressed by ellipsoidal RMSD.

The standard equation (Eq.1.1) is limited by its derivation from all atomic dynamic activity, ie. there is no specificity. Recently, eRMSD has been used to analyse nucleic acid systems, in particular, RNA complexes. This metric uniquely analyses the geometry of nucleobases in a molecule and evaluates differences in their relative interaction. Measurement is continuous, symmetric and provides a more detailed analysis of MD trajectories than RMSD. Units are Ångstrom. Molecular modelling and in silico analysis of RNA is generally contingent on only a few intrinsic features: backbone dihedrals (66), hydrogen-bond networks, and stacking interactions (67). However, RNA has six torsional angles (five phosphate backbone and one sugar-base) which challenge accuracy.

A recent mathematical interpretation of base-pairing and base-stacking interactions in RNA has collated these factors and quantified them as a simple scoring function, \( \varepsilon \)SCORE (68). The methodology quantifies the comparison of a target molecule with experimentally observed RNA structures. \( \varepsilon \)SCORE is the weighted sum of base-pairing and base-stacking interactions for all pairs of bases in the structure. The scoring function underpins eRMSD for calculation of base-base deviation contingent on dynamic activity. Base-stacking in nucleic acid complexes and helices often shows a degree of overlap rather than direct \( \pi-\pi \) bonding. RMSD may not fully account for this factor, particularly purine-purine interactions.

Purines and pyrimidines are discriminated by virtue of a local co-ordinate system sited at the centre of C2, C4 and C6 of each nucleotide (figure 1.4). The x-axis is aligned through C2 for all bases; the y-axis is equidistant between C4 and N3 for pyrimidines, C6-N1 for purines. \( \mathbf{R} \) is a reference vector from which dynamic deviations are calculated and defines the eRMSD metric:
Figure 1.4: co-ordination of eRMSD base-pairing for guanine (top left) and cytosine (lower right). Orientation of y-axis determined by properties of purine or pyrimidine.

The centre of origin for the co-ordinating axes is slightly shifted in eRMSD compared to that for RMSD. In the latter, the x-axis traverses the N3 atom for pyrimidines and through the mid-point of the N1–C2 bond for purines. The y-axis lies at 90° to the x-axis, with purine-C6 and pyrimidine-C6 in the xy plane (69). Because bases are essentially planar and interactions are evaluated in both stacking and pairing vectors, the system is considered as an ellipsoidal model.

1.3.4 Radius of Gyration

Radius of gyration ($R_G$) parameterises configuration for an entire system at equilibrium and consequently determines change in molecular structure during MD simulations. It is calculated in NAMD as the sum of the root mean square distances of an atomic set from the molecular centre of mass (Eq.1.2).

$$R_G = \sqrt{\frac{\sum_{i=1}^{N} m_i r_i^2}{\sum_{i=1}^{N} m_i}}$$  

Eq.1.2

where $N$ defines the atom set, $m_i$ is the mass of the atom $i$, $r_i$ is the distance of atom $i$ from the molecular centre of mass (70). Analysis of radius of gyration reinforces data from RMSD/eRMSD by quantifying fluctuations in Cartesian dimensions of the molecule. In the course of MD simulation, variations in molecular compactness are rendered. This parameter describes the ratio of molecular surface area to that of a perfect sphere with comparable volume. Higher values indicate elevated dynamic activity at a given point in the simulation and implicit less stability. Correspondingly, lower values suggest greater
molecular integrity at a given point. The calculation is weighted for mass but not charge and hydrogens are excluded. Units are Ångstrom.

1.3.5 Frequency of hydrogen-bond formation.

The morphology and function of biomolecules is partially determined by hydrogen-bonding. Although an individual bond is empirically weaker than covalent or ionic bonding, a network of H-bonds collectively influences molecular integrity. Accordingly, interplanar hydrogen-bonding in quadruplexes contributes significantly to stability of the system. Frequency of hydrogen-bond formation analysis can elicit further insight into dynamic activity within a given complex. This factor was monitored per residue for each model.

1.3.6 N6 – H1(N6) frequency of hydrogen-bond formation.

The specific structural integrity between participating bases in the quartets of each model and inter-planar stability was assessed by calculating frequency of H-bond formation between two specified atoms.

1.3.7 N6 – H1(N6) distance

Structural integrity between participating bases in quartets was further estimated by calculating the distance between N6 – H1(N6) in participating quartets.

1.3.8 Potential of Mean Force (PMF)

Torsional analysis was used specifically to compare variations in the glycosidic angles for biplanar adenine quadruplexes in both DNA and RNA. The methodology records fluctuations in energy from a specified parameter and expresses results as a function of time-dependent distances between bases (71). Many different atomic arrangements are possible in almost any molecular system; each microstate is a particular specific configuration at any one instant with numerous factors influencing the microstate. Analysis in this work estimated the probability of a given system to be in a microstate predicated on the glycosidic angle of each residue. Given each model comprises 22 residues, global maxima and minima were concatenated for tabulation and to provide comparison.
1.4 AutoDock

AutoDock4 and AutoDock Vina (72) were selected for docking methodology contingent on high docking power (73) and acknowledged extensive citations (74). For software compatibility requiring inclusion of atom charge (q) and type (t), molecular extensions .pdb were converted to .pdbqt. Ligand flexibility was assigned to define bond-types as rotatable or non-rotatable bonds. Default docking sites were accepted and a 3-D grid overlaid for the simulation. Grid parameters were filed with a .gpf extension to specify Cartesian co-ordinates, separation of grid points and grid centre. A grid parameter file also calculates energy for virtual atoms located at specified sites in the 3-D grid for comparison with ligand atoms. Consequently, each atom type of the ligand has a corresponding grid reference, electrostatics profile and a desolvation map as waters are relocated during the simulation.

The software has four search algorithms available: the hybrid global-local Lamarkian Genetic Algorithm (LGA) was selected by default. LGA originates an assumptive set of conformations to identify the minimum binding energy level. Acting as a scoring function, binding affinity is calculated. A Docking Parameter File (DPF) is generated prior to the simulation. A DPF identifies the relevant grid map files, the ligand’s Cartesian co-ordinates and quantifies torsions. The output file displays final conformation characteristics and binding affinity for an optimal docking. Less favourable configurations are also listed.

1.5 IntaRNA

RNA-RNA interactions for the work investigating long non-coding RNAs was undertaken with the analytical algorithms embedded in IntaRNA (75). The software designers acknowledge that predicting RNA structure and interaction is more complex than simply identifying complementarity. Previous releases of RNA-RNA interaction software can filter for promising candidates as an ergonomic use of computer time. However, because IntaRNA parameterises thermodynamics in its models, physical aspects are incorporated into structure formation and hence describe more than canonical base-pairing. The algorithms predict flexible non-linear regions connecting local helices between two or more matched sequences. These can be interpreted as bulges, kinks or loops in one or both of the aligned pairs. These structures allow dynamic re-alignment between the two RNA molecules and better reflect the matched sense-antisense configuration in 3-D space.

The helical nature of the molecules can bring bases into close spatial alignment; such proximity is not always evident with a 2D “beads on a string” model. Hence, when thermodynamic parameters of a tertiary configuration are incorporated into a prediction, orthodox helices can be linked by short unpaired strands. The analysis also accounts for the molecular integrity of interacting sequences which precludes availability for
intermolecular activity beyond the matched base-pairing. The default temperature of 37°C was retained for simulations.

1.6 Scripts for NAMD and ARCHER2 simulations

1.6.1 Overview

Both software packages are only accessible via command line and neither have a dedicated GUI interface. Both are supported by Microsoft operating systems. However, Linux servers were used for all simulations and analyses apart from pilot sampling with Chimera. NAMD simulations were opened remotely using Citrix Gateway Plug-in and Interlinc tl-4 12.0. Scripts for preparing and running a molecular dynamics simulation are open resource from NAMD, although some adaptations were applied to reflect nucleic acid models and preferentially access some plug-ins. Scripts for analysis did not derive from any one source but were written from first principles applying known VMD vocabulary.

ARCHER2 is opened with MobaXterm Personal 21.5.exe. This system requires an initial security account prior to access, a generic passkey code and a personal passkey. Access protocol has seven clear stages after a terminal is opened prior to submitting a simulation (section 1.5.7).

All scripts are listed in Supplementary Material as (1.6.2) Computer scripts and ARCHER2 protocol

1.6.2.1 Script 1: Annotated command line script for NAMD preparation

1.6.2.2 Script2: Configuration file for simulation (conf.txt)

1.6.3.1 Scripts for post simulation analysis

1.6.3.2 Script 3: Single frame visualisation

1.6.3.3 Script 4: Radius of Gyration Tcl files

1.6.3.4 Script 5: Calculate waters within a given distance of specified residues

1.6.3.5 Script 6. Calculate angles or dihedrals

1.6.3.6 Script 7: Run ellipsoidal RMSD

1.6.4: ARCHER2 access protocol
1.6.3 Scripts for post simulation analysis

1.6.3.1 Overview

Post-simulation, NAMD produces numerous files including restart, velocity, coordinates and trajectory; each has a back-up file. An option to produce a protein structure file (psf) manually is available for simple calculations, but the autopsf option was implemented in this study. The key files for analysis are the ionized psf and the trajectory file (.dcd). When loaded into a VMD console, these data underpin a series of analysis protocols. Overall visualisation of a given simulation can be accessed directly in VMD. However, individual frames with enhanced detail can be better visualised in Chimera (script 1). This allows micro inspection of a molecule at any stage of its MD simulation. This facility is useful for monitoring individual bond lengths and angles at specific significant points of the molecular dynamics. The script is run in VMD Main > Tk Console and data are returned to the home folder. Radius of gyration requires three scripts: (a) gyr_radius.tcl, (b) center_of_mass.tcl and (c) rog_loop_dcd.tcl. Scripts (a) and (b) set all parameters for the calculation and script (c) returns the radius of gyration data as a text file which can be visualised with an external plotting programme.
Chapter 2. Guanine quadruplexes

2.1.1 Introduction

Specific guanine-rich hexanucleotide sequences can fold into non-canonical supramolecular structures, quadruplexes, which have significant biological implications (76). Although experimentally validated, there is currently only a limited understanding of functionality in mitochondrial DNA (77) (78). Genomic sequences that might allow guanine quadruplexes to fold are notably more prevalent in mitochondrial DNA than nuclear. However, in all cases, the complex derives from four guanine bases linked by hydrogen bonds between Watson-Crick and Hoogsteen facets into a planar quartet (detailed further in section 2.1.2). Accordingly, guanine-rich sequences favour quadruplex locations. Conventionally, three or more guanine quartets stack vertically connected by nucleotide loops (79). The electronegative central column of oxygen atoms attracts cations such as K⁺ and Na⁺ (figure 2.1). This interaction stabilises the system by neutralising the intrinsic electrostatic repulsion of carbonyl oxygens at guanine O6 (80). The effect of Li⁺ as a co-ordinating cation remains controversial (81) (19). Accordingly, some models in this study were modified to explore this point.

![Figure 2.1](PDB 1KF1): triplanar guanine quadruplex derived from d[TTAGGG] sequence in single-strand telomeric DNA. The electronegativity column of oxygens (emphasised in red) attracts metal cations, typified here by K⁺ ions (mauve). The quartets are referred to in the text as T1 (upper), centre and T2 (lower).

Stability of the system is contingent on several factors: ionic strength, the water model used for MD simulations and the force field applied (82). For the 1KF1 model, removal of one or both co-ordinating cations is structurally disruptive. In this study, biplanar and triplanar DNA models co-ordinated by K⁺, Na⁺ or Li⁺ were assessed by RMSD only.
2.1.2 Guanine quadruplexes in single-stranded telomeric DNA

Early research posited G4s were simply structural curiosities located in single-strand telomeric DNA (83). Reduction in telomere length during cell replication is offset during the cell’s natural lifespan by telomerase, until apoptosis occurs after 30-50 divisions (84). However, the activity of telomerase can continue if an occasional cell survives apoptosis, triggering runaway replication and tumorigenesis (85). The gene expressing the enzyme, human Telomerase Reverse Transcriptase (hTERT) is over-expressed in a majority of all solid tumours (86), but the process of activation is not well understood (87). Hence, inhibition of telomerase contingent on guanine quadruplexes folding in telomeric DNA has long been associated as beneficial (88). However, the mechanism of inhibition is unclear (89). It has been hypothesised that G4s might prevent telomere extension by locating in the 3’ region and compromising telomerase activity through sequestration (90). Recently, double-strand DNA (ds-DNA) is now also considered to contain G4s (91). However, contrasting with single-strand G4s (ss-G4s), these complexes are probably pathogenic either by compromising RNA transcription (92) or disrupting replication-fork progress (93). Section 6.2.2 describes possible mechanisms.

Currently, it is accepted that G4s are genomically ubiquitous (94), but distribution is localised, not random (95). Some 500,000 potential quadruplex-forming sequences have been identified from predictive algorithms (96). The actual figure is however unknown (57). Highly conserved sequences are now known in 5’ UTR regions of various species (eg H. sapiens, M. musculus, T. brucei) and particularly evident at gene promoters (97) (98). However, other species have noticeable depletion of quadruplexes in this region (C. elegans, D. melanogaster) and some appear to have no population at all (E. coli, A. thaliana). Complexes folding in DNA G4s differ from other ss-DNA complexes such as stem loops in which the ‘stem’ is constructed of purine/pyrimidine pairs and the ‘loop’ purine/purine or pyrimidine/pyrimidine pairs. A stem is therefore constructed of classical Watson-Crick pairing and the loop nucleotides are unable to pair. G4s display non-canonical binding at the Hoogsteen interface. Figure 2.2 illustrates hydrogen bonding at this interface (depicted in red), with those at the Watson-Crick interface (shown in blue).
2.1.3. Pathogenicity and intervention

Paradoxically, recent research strongly implies G4s folding in telomeres and possibly other locations compromise genomic stability via disrupting DNA replication (99). Some neurodegeneration is also associated with G4s (50). Numerous other sequences have now been identified which have the potential to fold into guanine quadruplexes (48). These allow for different lengths and composition of interplanar loops with consequent varied folding patterns. Whilst this variation prompts structural diversity and selectivity for ligands, it may also contribute to partially resolving the paradox. Several folding patterns have been described for G4s, and some may be significant in the complexes’ capacity for telomerase inhibition (9). A combined computational and experimental study concluded highly-conserved putative G4 sequences are implicated in transcriptional activity of successful embryonic development (100). Adverse aspects continue to be identified, including premature aging, cancer predisposition, and fragile X syndrome (57). Frontotemporal dementia has been recently linked to G4s (101). However, beneficial associations with respiratory function and replication of the mitochondrial genome have been noted (102). It may also be significant that helicases proliferate in G4 regions (103): quadruplex formation could depend upon ds-DNA unwinding to a transient single-strand (3).

Their distribution is localised, not random (95): sited particularly in telomeric ss-DNA and ds-DNA at replication-forks (104). It is suggested here that differential effects of G4s may be contingent on genomic location: ss-G4s sited on the telomere are associated with telomerase inhibition but some double-strand G4s (ds-G4s) are associated with neurodegeneration (92). Furthermore, it has been noted that quadruplexes can have paradoxical characteristics for the same target (105).

Given the diversity of G4s, design of effectively targeting ligands has proven challenging: planar small molecules show limited affinity and selectivity (48). Targeting specific folding
sequences in tandem with well-defined structural elements may be a promising route for intervention. Bio-informatics identify G4s populate specific sites such as immunoglobulin switch regions, promoter regions and recombination sites (22) (106). A gene regulatory role has been proposed for G4s adjacent to transcription start sites (89). However, the majority of validated G4s in PDB are telomeric. Hence research into potential G4-related tumorigenesis in promoter regions is currently nascent.

2.1.4 Molecular dynamics

1KF1 was submitted to NAMD for analysis to act as a reference for comparison with triplanar DNA and RNA models with base substitutions constructed in this work. Methylation at guanine N6 may have adverse effects on the stability of guanine quadruplexes (107). Three contributive factors have been hypothesised: the modification may distort the co-ordination site and weaken ionic interactions, and methylation could reduce local electronegativity: the consequent less effective Hoogsteen bonding would destabilise the system. Steric hindrance might also be a factor. The topology file for the simulation (top_all36_na.rtf) was revised to methylate a single guanine base at O6 in the central quartet of 1KF1 to assess any impact on quadruplex folding. O6-methylation as a disruptive factor was investigated in tandem with thionation at the same location.

2KF8 was subsequently submitted to NAMD for analysis to act as a reference for comparison with biplanar DNA and RNA models with base substitutions constructed in this work. The influence of varying the co-ordinating cation between K+, Na+ or Li+ was assessed for DNA triplanar and biplanar models only and assessed uniquely by RMSD. Given an unanticipated relevance of theoretical adenine quadruplexes to the emergence of SARS-CoV-2, only DNA and RNA adenine structures were modified with methylation at N6. The significance, if any, of co-ordination with Ca2+, Mg2+, H3O+ and NH4+ in addition to K+, Na+ or Li+ was also only assessed for DNA and RNA biplanar and triplanar models.

Single strand 22-mer DNA and RNA oligonucleotides were also constructed to assess thermodynamic stability of an unfolded chain. Necessarily, these models were not methylated and no cationic interactions were included in the simulations.
2.2. Results.

2.2.1 1KF1_non-modified

2.2.1.1 RMSD (S1)

The trajectory shows minimal RMSD variation for the duration of the simulation implying the non-modified quadruplex is dynamic but stable.

2.2.1.2 eRMSD (S2)

The data show very little deviation from a mean value of ~ 1.7 Å. There is a slight rising trend to the trajectory with marginal temporary dynamic activity between ~ 380 ns and ~ 430 ns.

2.2.1.3 Radius of gyration (S3)

The trajectory shows very little fluctuation from an average ~1.2 Å for the entire simulation implying consistent molecular compactness.

2.2.1.4 Frequency of H-bond formation (S4, S5)

The frequency of H-bond formation appears consistently regular. Although the average number of bonds at any one time is low (varying between three and four), they are maintained throughout the simulation.

2.2.1.5 N6 – H1(N6) frequency of hydrogen-bond formation

A reasonable H-bond network is predicted for most residues and exemplified by data for the residues of the central plane. Residue 9 has a significantly reduced frequency in the later stages of the simulation (figures 2.3a, 2.3b, 2.3c and 2.3d):
Figures 2.3a, 2.3b, 2.3c and 2.3d: frequency of H-bonds for residues of the central plane in GDN-N1 K^+

2.2.1.6 N6 – H1(N6) distance.

Diametrically opposite residues 9 and 21 of the central quartet show a consistent frequency of H-bond formation, whereas residues 3 and 15 located adjacent in the plane, are notably more dynamic (figures 2.4a, 2.4b, 2.4c and 2.4d):

Figures 2.4a, 2.4b, 2.4c and 2.4d: N6 – H1(N1) distances for residues of the central quartet GDN12-N1, K^+ co-ordination.
2.2.2 1KF1-O6-methylation

Guanine O6 of base 15 was arbitrarily selected for methylation, the only proviso being that the base should be one of the four central quartets. A single bond between C6 and O6 was imposed to allow an available bond for the modification (figures 2.5a and 2.5b):

Figures 2.5a and 2.5b: 1KF1 methylated at guanine O6 at t = 0ns. Left: the bases are depicted as nucleotide slabs to emphasise the symmetry of the complex. Right: methylation in the modelling software showed no preference for intramolecular orientation of the modification. Distance to the nearer K⁺ ion was 1.436 Å and to the other, 2.912 Å. Hydrogen bonds were predicted between the methyl group and both cations.

Comparison with the non-methylated system (figures 2.6a and 2.6b) discloses the noticeable degree of relocation imposed by O6-methylation. The O6- K⁺ distance in the non-methylated system is 2.7 Å compared to the values noted for figure 2.5a. Similarly, O6-K⁺ angles typify an octahedral configuration for the non-methylated system compared with variable geometries in the O6-methylated system.

Figures 2.6a and 2.6b: Distances (left) and angles (right) between K⁺ cation and guanine O6 for 1KF1. The values reinforce the assumption of stable octahedral co-ordination.
2.2.2.1 RMSD (S6)

After minimisation the trajectory shows noticeable dynamic activity with a slight rising trend to the end of the simulation. Maximum increase in RMSD occurs between ~ 50 ns and ~150 ns.

2.2.2.2 eRMSD (S7)

The system shows little activity during minimization but a noticeable rising trend thereafter. A brief period of activity occurs between ~ 70 ns and ~ 110 ns. The mean value of ~ 1.9 Å deviation is higher than for the non-methylated complex.

2.2.2.3 Radius of gyration (S8)

The molecule maintains a compact structure after ~ 250 ns. Prior to that point, the trajectory shows some activity implying slight loss of compactness. The mean value for $R_g$ is marginally higher than for non-methylated 1KF1 (~1.2 Å compared to ~1.3 Å).

2.2.2.4 H-bond formation (S9, S10)

Methylated 1KF1 appears more dynamic: although the average frequency of hydrogen bonding is slightly higher than the non-modified quadruplex, the trajectory and histogram suggest bonds are more transient.

2.2.2.5 N6 – H1(N6) frequency of hydrogen-bond formation.

The data indicates the residues of the central quartet are poorly-served for H-bonds, residue 3 in particular (figures 2.7a, 2.7b, 2.7c and 2.7d):
Figures 2.7a, 2.7b, 2.7c and 2.7d: frequency of H-bonds for residues of the central plane in GDN-N1_meth K⁺

2.2.2.6 N6 – H1(N6) distance.

All of the trajectories for N6 – H1(N6) distances in adjacent residues of the central quartet of O6-methylated triplanar guanine models are dynamic and indicative of an unstable system locally (Figures 2.8a, 2.8b, 2.8b and 2.8d):

Figures 2.8a, 2.8b, 2.8b and 2.8d: N6 – N1(H1) distances for residues of the central quartet GDN12-N1_meth, K⁺ co-ordination.
2.3.1 2KF8_non-modified

2.3.1.1 RMSD (S11)

After minimisation, the system showed a stable trajectory with \( \sim 3.0 \) Å deviation. Between \( \sim 250 \) ns and \( \sim 300 \) ns, a rising trend to \( \sim 4.5 \) Å displays a rapid drop to \(< 2.0 \) Å which is maintained for \( \sim 80 \) ns. A brief rising trend returns to a steady trajectory at \( \sim 3.0 \) Å. The dynamic activity implies a reconfiguration between two stable conditions. RMSD data for each base of the participating quartets are shown in figure 2.9:

![RMSD Graph](image)

**Figure 2.9: RMSD for all bases in the quartet pair of 2KF8 without cationic co-ordination.** This figure should be compared with RMSD for 1KF1 with and without O6-methylation (figure 2.23). Empirical data are tabulated in Supp fig RMSD SX4.

2.3.1.2 eRMSD (S12)

The trajectory is maintained at \( \sim 2 \) Å for the duration of the simulation and displays very little deviation.

2.3.1.3 Radius of gyration (S13)

The molecule remains compact for the duration of the simulation with slightly more dynamic activity in the minimisation cycle.

2.3.1.4 H-bond formation (S14, S15)

There appears to be an irregular periodicity to the trajectory with an average seven H-bonds at any given time.
2.3.1.5 N6 – H1(N6) frequency of hydrogen-bond formation.

Residues in the central plane of GDN8-N1 show a good network of H-bonds between N6 – H1(N6) although residue 9 has a less consistent frequency (figures 2.10a, 2.10b, 2.10c and 2.10d):

![Graph 1](image1)

![Graph 2](image2)

![Graph 3](image3)

![Graph 4](image4)

Figures 2.10a, 2.10b, 2.10c and 2.10d: frequency of H-bonds for residues of the central plane in GDN8-N1 no cation

2.3.1.6 N6 – H1(N6) distance

Distances between all residues in the central quartet of biplanar guanine quadruplexes without cationic co-ordination are relatively consistent in each case. Some dynamic activity is noted for residue 3 in the later stages of the simulation (Figures 2.11a, 2.11b, 2.11b and 2.11d):

![Graph 5](image5)

![Graph 6](image6)
Figures 2.11a, 2.11b, 2.11b and 2.11d: N6 – H1(N1) distances for residues of the central quartet GDN8-N1, no ionic co-ordination.

2.3.2. 2KF8 O6-methylation

Guanine O6 of base 14 was arbitrarily selected for methylation. Given the complex is biplanar, no further specificity was needed. A single bond between C6 and O6 was imposed to allow an available bond for the modification. Figures 2.12a and 2.12b illustrate the loss of structural integrity post-simulation for the methylated model.

Figures 2.12a and 2.12b: 2KF8 methylated at guanine O6. The final frame of 500ns simulation illustrates noticeable structural reconfiguration compared to the model at t = 0 ns (figure 2.5a). The quadruplex in Plan (left) shows the phosphate backbone has retained some integrity, but bases (formatted as nucleotide slabs) are notably disordered. The model viewed in Elevation (right), suggests the complex has also compacted somewhat.
2.3.2.1 RMSD (S16)

The trajectory has a higher average RMSD than non-methylated 2KF8 and with a slightly higher range between ~ 3 Å and ~ 6 Å. There is a sudden temporary reduction in the trajectory between ~ 150 ns and ~ 250 ns.

2.3.2.2 eRMSD (S17)

The trajectory is marginally more dynamic than non-methylated 2LF8 and the range slightly higher (~2.0 Å, ~2.2 Å). There is a minimal increasing trend in the eRMSD with a sharp temporary rise between ~ 320 ns and ~ 360 ns.

2.3.2.3 Radius of gyration (S18)

Although the \( R_G \) values are relatively low, the methylated molecule shows less compactness than the non-modified structure for most of the simulation, particularly during the minimisation cycle.

2.3.2.4 Frequency of H-bond formation (S19, S20)

Frequency of H-bond formation appears to be lower than for non-methylated 2KF8 and may have a slight periodicity. There are at any one time an average four or five bonds present, within a range of ~ 2 to ~ 15.

2.3.2.5 N6 – H1(N6) frequency of hydrogen-bond formation.

Residues of the central quartets in O6-methylated biplanar guanine systems show a very low frequency of H-bonding, particularly residue 15 (figures 2.13a, 2.13b, 2.13c and 2.13d):
Figures 2.13a, 2.13b, 2.13c and 2.13d: frequency of H-bonds for residues of the central plane in GDN8-N1 O6-methylation

2.3.2.6 N6 – H1(N6) distance.

Trajectories for N6 – H1(N6) distances in the central quartet of O6-methylated biplanar guanine quadruplexes are dynamic and generally show a trend towards structural disintegration (Figures 2.14a, 2.14b, 2.14b and 2.14d):

Figures 2.14a, 2.14b, 2.14b and 2.14d: N6 – N1(H1) distances for residues of the central quartet GDN8-N1, O6-methylated
1KF1 O6-methylation and 2KF8 O6-methylation are further discussed in Sections 2.5 and 2.7.4.

2.3.3 DNA single strand unfolded 22mer

A 22-mer DNA single strand was constructed to compare an unfolded sequence with the same sequence folded into a quadruplex structure. The system showed notable change in integrity when assessed by all metrics (figures 2.15a and 2.15b).

![Figures 2.15a and 2.15b. Single strand DNA 22-mer at t = 0ns (A) and single strand DNA 22-mer at t = 500ns (B).](image)

2.3.3.1 RMSD

The trajectory for a single strand DNA 22mer shows a steady increase after minimisation to ~15 Å at ~220 ns (figure 2.16). The system is very active and returns to a dynamic average RMSD of ~13.5 Å ± 1.5 Å. Given the structure unwinds continuously during the simulation, the relatively constant RMSD between 12 Å and 15 Å suggests the spatio-temporal relationship for phosphate backbone and bases is stable. Whilst the former is more dynamic, the latter is less so and in general retains its integrity.
Figure 2.16: RMSD single strand DNA 22mer. The sequence of bases is identical to that of telomeric 1KF1.

2.3.3.2 eRMSD (S21)

The trajectory maintains a generally horizontal trend after minimization at ~2 Å with a slight rising slope near the end of the simulation. Values range between ~1.5 Å and 2.5Å. The data indicate a relatively stable relationship between the phosphate backbone and bases. This interpretation strongly supports that of the RMSD result.

2.3.3.3 Radius of gyration (S22)

Compactness of the system is noticeably relaxed after minimisation from ~22 Å to ~25 Å. The system is relatively chaotic thereafter with a slight rising trend to ~27 Å. The data imply that helix unravelling is not a consistent process for both DNA and RNA structures.

2.3.3.4 Frequency of H-bond formation (S23, S24)

The trajectory indicates there is a consistent heavy population of H-bonds ranging between ~7 and 20 for the system to ~180 ns, otherwise the H-bond frequency formation is poor. The population reduces significantly from ~15 Å onwards. At the end of the simulation, approximately 20 H-bonds are apparent. Given the strand is a 22mer, this figure may represent one bond per base. The histogram confirms a low H-bond population for almost the entire simulation. A level below ten is not easily resolved by the
plotting software hence the histogram records none at scale beyond three H-bonds for any one instant.

2.3.4 RNA single strand unfolded 22mer

A similar model in single strand RNA derived from the 1KF1 sequence with uridine substitutions for thymine also showed extensive structural change when assessed by all metrics employed in this research. The system’s helical angles are less acute than the DNA equivalent (figures 2.17a and 2.17b).

![Images of RNA structures at different times](image)

**Figures 2.17a and 2.17b.** Single strand RNA 22-mer at t = 0ns (left). Single strand RNA 22-mer at t = 500ns (right).

2.3.4.1 RMSD

The unfolded single strand RNA 22-mer shows a general increase in RMSD to ~5.8 Å at ~250 ns, after which time the trajectory is dynamic but settles into a trend averaging ~4.8 Å ± 1.0 Å (figure 2.18).
Figure 2.18. RMSD trajectory for single strand RNA 22mer. The sequence is identical to the DNA model, but with uridine substituted for thymine. The difference in $y$-axes between this figure and figure 2.16 should be noted.

2.3.4.2 eRMSD (S25)

The trajectory shows a slight rising trend during minimisation which maintains an estimated eRMSD of 1.5 Å to ~420 ns and with minimal dynamic activity. There is a steady rising trend for the final ~80 ns of the simulation.

2.3.4.3 Radius of gyration (S26)

The trajectory differs significantly from Rg for the DNA equivalent. There is a rapid loss of compactness during minimisation from ~12 Å to ~15 Å which is followed by a steep reduction to ~9 Å. The system is chaotic thereafter between ~100 ns and ~200 ns before following a much less dynamic trajectory with a slight downward trend. The data reflect the steady trajectory noted for RMSD and occur in close approximation with regards to timing.

2.3.4.4 Frequency of H-bond formation (S27, S28)

The system appears to be well-populated with H-bonds for the duration of the simulation with a noticeable rising trend towards the end. Whilst this resonates with the data for a DNA equivalent model, the level of H-bonding is recorded as higher in the RNA model. There are very few occasions where no H-bonds are predicted. The histogram reinforces this interpretation of the data.
2.4 Theoretical melting temperature of double strand DNA 22-mer

A theoretical calculation for the melting temperature \( T_m \) of a \((TTAGGG)_4\) duplex was made using the Marmur-Doty (108) formula:

\[
T_m = 69.3 + 0.41(G-C) \quad [\text{Eq.2.1}]
\]

where G-C is the guanine-cytosine molar fraction \(0.5\).

\[
T_m \approx 69.5^\circ C
\]

A more recent revised formula (109) calculated:

\[
T_m \approx 71.4^\circ C
\]

The Nearest Neighbour method (110) was used to estimate the Gibbs free energy:

\[
\Delta G = -31.29 \text{ kcal mol}^{-1}
\]

If the duplex sequence is expressed as AB, then the relative concentrations at 50% dissociation is \([AB]/2\) and the free energy of association equation can be re-arranged in terms of \(T\):

\[
\Delta G = RT \log_e K_D
\]

since 

\[
K_D = ([AB]/2) \times ([AB]/2)/([AB]/2)
\]

then

\[
K_D = [AB]/2
\]

so

\[
\Delta G = RT \log_e [AB]/2 \quad [\text{Eq.2.2}]
\]

re-arranging for \(T\)

\[
T = \Delta G / (R \log_e [AB]/2)
\]

convert Joules to kcal

\[
T = \Delta G / (R \times 0.002 \log_e [AB]/2)
\]

if [AB] is 0.1mM

\[
T = \Delta G / (R \times 0.002 \log_e 0.00005)
\]

\[
T = \Delta G / (0.0382)(-9.90) \quad [\text{Eq.2.3}]
\]

The influence of \(\pi\)-stacking was included from archived stacking free energy parameters (111) and calculated as \(-25.66\text{ kcal mol}^{-1}\).

therefore

\[
\Delta G_{\text{tot}} = -56.95 \text{ kcal mol}^{-1}
\]

and from Eq.5

\[
T_m = -56.59 / -0.1646
\]

\[
T_m = 345.9K \quad (72.8^\circ C)
\]
Theoretical $T_m$ for a (TTAGGG)$_n$ duplex from three methods has a reasonable agreement with a mean of $\sim 73^\circ C$. Since only one sequence was analysed, a result predicated on concentration is reasonable assuming the G / C content is constant and concentration is the sole variable. $\Delta G$ for each of the two unwound strands was also calculated. 5' to 3' direction (sense-strand) = -26.92 kcal mol$^{-1}$ and antisense = -22.28 kcal mol$^{-1}$. Sections 4.14.5, 6.5.2.3, 7.2.3.2 discuss the significance of these data.

2.5 Potential impact of thionation and methylation at guanine O6.

2.5.1 Introduction.

Molecular dynamics data for triplanar and biplanar guanine quadruplexes when methylated at O6 show a disruptive influence of the former, but less so on the latter. Three possible contributive factors were noted earlier: reduction of local electronegativity due to distortion of the site, less effective Hoogsteen bonding would destabilise the system and steric hindrance. To assess and compare these suggestions, 1KF1 was modified with a single methylation or thionation at guanine O6 of the central quartet (figures 2.19a, 2.19b and 2.19c). Further analysis was made of single substitution O6 methylation or O6 thionation for the exterior quartets. Simulations of O6-methylated or O6-thionated models were also conducted with Na$^+$ and Li$^+$ as co-ordinating ions.

![Image](image)

**Figure 2.19a, 2.19b and 2.19c: modifications to guanine** (left). Methylation at the O6 position (centre) and thionation at O6 (right).

2.5.2 Results

A network of H-bonding and other electrostatic interactions was predicted for systems methylated or thionated at guanine O6. No direct electrostatic interaction between guanine O6-methylated and K$^+$, Na$^+$, or Li$^+$ cations was apparent. Two H-bonds were predicted for O6-thionated and fewer unfavourable interactions (figures 2.20a and 2.20b).
Figures 2.20a and 2.20b: methylation (left) and thionation (right) at guanine O6 of 1KF1. π-donor interactions shown in white, H-bonds in green. Non-polar interactions in red. K⁺ model shown, a very similar bonds network was obtained for Na⁺ and Li⁺.

The intercationic unfavourable bump interaction is predicted for both models but no direct unfavourable bump interactions for S6. Unfavourable bumps imply a steric clash but which may not necessarily prevent a other interactions. π-donor interactions are similar for both modifications but O6-thionation could allow H-bonding peripheral to the cations unlike the methylated system. Cationic rejection occurs rapidly in both methylated and thionated conditions, although a differential impact is observed for the ions. Figures 2.21a, 2.21b and 2.21c should be compared with figure 2.1.

Figures 2.21a, 2.21b and 2.21c: rejection of K⁺ co-ordinating cation (shown in mauve) from guanine quadruplex methylated at O6. Oxygens in red and methyl carbon in grey. Frames equivalent to t = 0 ns (left), t = 15 ps (centre), and t = 35 ps (right). The ion more adjacent to the methyl group (lower half of the image) relocates rapidly and maintains a dynamic position peripheral to the quadruplex exterior. The accompanying ion shifts slightly away from a median interplanar location to locate approximately in the plane of the guanine quartet for most of the remaining simulation. Bond length between the methyl group and guanine C6 is marginally reduced and the C6-O6-CH₃ angle becomes less acute.
Interatomic distances between co-ordinating cations and modified guanine O6 indicate the modifications are not equally disruptive for Li⁺, Na⁺ or K⁺ (figures 2.22a, 2.22b and 2.22c). Although notably dynamic, the least affected appears to Na⁺ for both modifications: the cation–guanine O6 interatomic distance varies most for the three ions. However, K⁺ in an O6-methylated system briefly diverts from this trend. There is a large temporary displacement between ~250 ps and 500 ps. O6-thionation appears to affect Li⁺ more than O6-methylation.

Figures 2.22a, 2.22b and 2.22c: interatomic distances between co-ordinating cations and modified guanine O6 in 1KF1. (a) O6 non-modified, (b) O6-methylated, (c) O6_thionated. Interatomic distances between cation and modified group for were typically less for Li⁺ than for K⁺ or Na⁺.

RMSD data for 1KF1 compared with O6-methylated or S6-thionated clearly show the disruptive consequence of the modification (figure 2.23)
Figure 2.23: RMSD for K⁺-co-ordinated guanine quadruplex with and without single O6-methylation or O6-thionation to residue 15 for each base of the central quartet in 1KF1 (blue), 1KF1_S6-thionated (yellow) or 1KF1_O6-methylated (mauve). Results for Na⁺ complex are similar, but generally lower values for Li⁺ model. Empirical data tabulated in Supp Mat RMSD SX1, SX2 and SX3.

Comparison of the trajectories for non-modified 1KF1 and either methylation or thionation at O6 of residue 15, show the disruptive influence of modification. Although the variation is greatest for Na⁺, the cation also shows significantly more dynamic activity in the non-modified complex. Simulations for both O6-methylated and O6-thionated models show little difference in activity for Na⁺ when compared to the non-modified structure. K⁺ and Li⁺ are influenced by both O6-methylation and O6-thionation, although Li⁺ less so. Analysis of the models with single substitutions in either of the exterior quartets (T1 and T2), indicated modification to the central quartet has greater impact of structural integrity. Residue 15 is one of the four guanines located in the central quartet of the quadruplex. Disruption appears to be greater if any base of the central quartet is either methylated or thionated when compared to single substitutions in either of T1 or T2. Data for both conditions are very similar (figure 2.24):
Figure 2.24: interatomic distances between co-ordinating K⁺ cations and single O6-methylation or O6-thionation. Residues 15 for T1 and T2, and for each base of the central quartet in 1KF1 were investigated. Data for T1 and T2 (exterior to the central quartet) are depicted in dark blue / light blue and orange / red respectively. The modified central quartet is shown in turquoise / dark green (O6-CH₃) and light green / yellow (O6-thio).

Since guanine N7 is also nucleophilic, a brief (100ns) MD simulation was run to assess potential impact of methylation. N7 of the central quartet in 1KF1 was methylated but not thionated. Models with Na⁺ and Li⁺ as co-ordinating cations were not constructed. The trajectory for cation and N7-methylation interatomic distance was dynamic but maintained a reasonably steady mean of ~6 Å up to ~100 ps at which point the complex very rapidly disintegrated. The ions appeared to not to relocate, but the methylated group was notably active. No further investigation of this system was undertaken.

2.5.3 Discussion

The MD analysis here shows methylation or thionation of the central quartet at O6 is disruptive, rapidly expelling K⁺ and Na⁺ ions early in the simulation. Except for inter-atomic distances between the cation and the exterior non-methylated quartets T1 and T2, the complex becomes unstable with loss of the co-ordinating cations. Models with Na⁺ are marginally more thermodynamically active. However, Li⁺ was not expelled. The ions are accommodated within the plane of the outer quartets and the complex maintained integrity for 1.0ns. the impact of Li⁺ in guanine quadruplexes has a contrary history (112). The ion has a reported robust inhibitory influence on quadruplexes folding, but other contemporary research does not confirm this (113). The simulations in this study suggest
Li⁺ as a co-ordinating cation in triplanar guanine quadruplexes has a stabilising influence. Whereas K⁺ and Na⁺ are expelled from the system with consequent loss of structural integrity, Li⁺ remains within the system, albeit shifted slightly to the intraplanar location of the exterior quartets. Three possible factors are suggested.

(i) Steric hindrance. The ionic radius for CH₃⁻ attached to O6 is greater than for an O6 substitution by S. Both groups should have a disruptive impact on local configuration but with more impact from O6-methylation. A marginal trend for this consideration is observed in the relevant trajectories. Moreover, unfavourable bump interactions predicted in the modelling software (Figures 15a and 15b) strongly imply suggest a steric clash for both modifications. The smaller VDW radius of Li⁺ (0.60 Å) probably allows it to remain within the G4 whilst larger K⁺ (1.33 Å) and Na⁺ (0.95) are expelled.

(ii) Lower electronegativity (EN). Pauling EN values for C and S very similar (2.5 and 2.6), but oxygen is higher (EN = 3.4). Pauling values for K⁺, Na⁺ and Li⁺ are similar (0.83, 0.93, 0.98 respectively). A reduction in local electronegativity would also contribute to reduced stability in the system, although the impact from Li⁺ should be less.

(iii) Hydration enthalpy. A minor possible factor in the final location of each co-ordinating ion. Because Li⁺ has the highest hydration enthalpy in the sequence: Li⁺ > Na⁺ > K⁺ (-520 kJ mol⁻¹ > -406 kJ mol⁻¹ > -320 kJ mol⁻¹), it should attract more of an ‘insulating’ aqueous layer than Na⁺ and K⁺. A buffer zone of water molecules might marginally improve stability.

It therefore seems likely the prime cause of instability prompted by O6-methylation derives from steric hindrance with reduced local electronegativity as a minor factor. Enhanced hydration enthalpy may also have some influence. The early observations of N7-methylation also merit some comment: since the ions were not expelled from the complex (contrasting with O6-methylation simulations), and coherence is lost so quickly, steric clashes may be involved in other ways.

2.5.4 Enzymatic repair mechanism for O6 methylation (MGMT)

2.5.4.1 Introduction

Methylation also has a high risk of miscoding due to G-T mismatching (114). Even low levels may disrupt the mis-match repair pathway (115). Such mismatching is usually repaired by O6-methylguanine DNA-methyltransferase (MGMT), but the enzyme may be less effective by poor access to a methylated central quartet.

MGMT (PDB 1EH6, MW ~ 22 kD) removes methylation at guanine O6. The reaction has 1:1 stoichiometry which implies a significant enzyme population is needed to be effective
as individual enzymes are degraded after de-methylation (116). Moreover, MGMT is not regenerated as characteristic of other enzymes. Limited or zero access to a methylated central quartet due to high molecular weight, enzyme stoichiometry and enzyme removal rather than regeneration combine to suggest MGMT is less influential in guanine quadruplexes than with single strand or double strand DNA. The disruptive impact of –CH₃ substitutions could continue even if the enzyme were present. Docking protocols between enzyme and ligand may be useful for interpreting MGMT- quadruplex interactions Target ligands were quadruplexes 1KF1, 1KF1_meth, 2KF8 and 2KF8_meth.

2.5.4.2 Methods

MGMT was downloaded as macromolecule and the target quadruplexes downloaded as ligands to Molecular Docking Server and AutoDock. Since AutoDock has no direct GUI, only free binding energies were recorded with this methodology. Molecular Docking Server options for direct visualisation of dockings, interacting residues, electrostatic energies and H-bond plotting (117) were accessed.

Docking results were calculated in Molecular Docking Server (118) and AutoDock. The AMBER MMFF94 force field was applied for energy minimisation to the target ligands (119) within the Gasteiger calculation. Partial charges were added, non-polar hydrogen atoms were merged and rotatable bonds were defined. AutoDock Tools was used to further characterise hydrogen atoms, The Kollman calculation for atom charges ratified the Gasteiger calculation and parameterised the solvation (120). The Gridbox facility in AutoDock Tools was used to generate affinity maps. A spacing of 0.375 Å was applied for the Cartesian co-ordinates defining a 25 × 35 × 25 Å grid. AutoDock Tools was further accessed to calculate van der Waals and electrostatic terms, calculate parameter set- and distance-dependent dielectric functions.

Docking simulations were generated by the Lamarckian genetic algorithm and the Solis local search method (121). Random settings were set for ligands’ initial orientation, position and torsions. The docking parameters were set at 0.2 Å for torsion steps and quaternion steps, 5° for rigid-body orientation steps, 5° for dihedral steps and 2 Å for root mean square tolerance. Defaults for the population size of 150, 250,000 for the number of energy evaluations and 540,000 for number of generations were selected. To optimise a mean value of binding affinities, ten docking simulations were run for each model.
2.5.4.3 Results

AutoDock and Docking Server binding characteristics calculated for MGMT to triplanar 1KF1 and 1KF1_meth, biplanar 2kf8 and 2KF8_meth are recorded in Table 2.1. The data strongly suggest ligand-enzyme interaction would be endothermic. No inhibition constants are predicted.

Table 2.1: Docking characteristics for biplanar and triplanar guanine quadruplexes non-methylated and O6-methylated.

<table>
<thead>
<tr>
<th>ligand</th>
<th>free energy of binding kcal/mol</th>
<th>electrostatic energy kcal/mol</th>
<th>number of interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AutoDock</td>
<td>MDS</td>
<td>MDS</td>
</tr>
<tr>
<td>2KF8</td>
<td>2.82</td>
<td>0.46</td>
<td>-0.46</td>
</tr>
<tr>
<td>2KF8_meth</td>
<td>6.11</td>
<td>1.29</td>
<td>0.74</td>
</tr>
<tr>
<td>1KF1</td>
<td>10.27</td>
<td>6.83</td>
<td>-0.30</td>
</tr>
<tr>
<td>1KF1_meth</td>
<td>11.98</td>
<td>6.86</td>
<td>6.81</td>
</tr>
</tbody>
</table>

The ligation for 1KF1 and the enzyme has higher level of interactions than that predicted for O6-methylated 1KF1 (figure 2.25). An approximate 10% of interacting residues are leucine in both O6-methylated models. Interactions predicted in Docking Server are: H-bonds, polar, hydrophobic or π-π bonds. AutoDock does not discriminate between types of bond.
Figure 2.25: docking protocols for 1KF1 (A) and O6-methylated 1KF1 (B) ligated with MGMT. A number of residues for the methylated system are located peripherally to the enzyme (circled in red) and appear to be excluded from any interaction.

H-bond plots for MGMT binding with both 1KF1 models identify regions in the O6-methylated complex with reduced interaction (figure 2.26). The plots show a particular absence of H-bonds between residues His-86 and Asn-137; only Phe-108 is predicted to interact in this region. This interaction is with N6 of an adenine in the loop structure.

Figure 2.26: H-bond plots for 1KF1 (A) and O6-methylated 1KF1 (B) ligated with MGMT. Protein residues are depicted in red. H-bonds in black: those not immediately adjacent to the enzyme are generally transient and derive from chance contacts of peripheral residues. The overall interaction pattern is broadly similar, particularly the H-bond strands generated for the earliest points in the docking progress. There are fewer interactions in the methylated system.
Biplanar 2KF8 has a similar H-bond population to triplanar 1KF1, but a notably higher number for the O6-methylated model compared to its triplanar equivalent. There appear to be little or no regions on visual inspection where the ligand is obviously isolated from the enzyme (figure 2.27).

Figure 2.27: docking protocols for 2KF8 (A) and O6-methylated 2KF8 (B) ligated with MGMT. Visual inspection supports the data from Table 2.1 of similar interaction for both models.

H-bond plots for 2KF8 and O6-methylated 2KF8 identify very few regions of difference between the two models (figure 2.28).

Figure 2.28: H-bond plots for 2KF8 (A) and O6-methylated 2KF8 (B) ligated with MGMT. The plots are very similar but contrasting with the non-methylated complex, show an absence of H-bonds in the methylated model between Phe-108 and Ala-118. The two strands of H-bonds adjacent to Leu-19 and Leu-33 (left corner of each image) are common to all four models. Visual inspection of
models does not identify any particular process involved, but a specific localised fluctuation in MGMT structure may account for the result.

2.5.4.4 Discussion

The overall interpretation of the data for MGMT interaction with biplanar and triplanar guanine quadruplexes strongly indicate O6-methylation is disruptive. Binding enthalpies for all four models are negative and therefore exothermic. Implicitly therefore, the enzyme may have some interaction with the complexes and effect repairs to an extent. However, repair of methylation on the central quartet appears less likely with these data, and tend to confirm the proposal of steric hindrance limiting access to this region. Noting that a biplanar quadruplex (ie, without a central quartet) appears less disrupted by O6-methylation than the triplanar model reinforces this interpretation.

The Protein Feature View available in Protein Data Bank identifies hydrophobic regions of the enzyme are also commensurate with regions in figures 2.20a and 2.22b showing little or no H-bonding. Given O6-methylation is also hydrophobic, the modification may further reduce local non-covalent interaction. Section 2.5.3 proposed that steric hindrance predicated on O6 methylation or thionation underpins quadruplex instability. This study suggests methylation also appears to adversely affect the repair mechanism of O6-methylguanine DNA-methyltransferase.

2.6 Occupancy of waters in guanine quadruplexes 1KF1 and 2KF8

2.6.1 Introduction

Arrays of bridging water molecules are known to be imperative in binding ligands to parallel guanine quadruplexes; hydrogen bonding mediates interaction between the two species (122). Interplanar loops are populated with waters and regarded as reinforcing stability of their conformation (123). However, such waters are superficial or peripheral to a quadruplex; little is known of waters internalised into the system. Ostensibly, molecular dynamics simulations appear to be unsuited in resolving this problem. A hydration shell neutralised with counterions and enclosing any given quadruplex model may typically contain between 15 000 and 20 000 water molecules. Consequently, manipulating the model to ascertain the number of waters located in the centre of the structure over time is highly impractical. To address this challenge, a script was written to determine by calculation occupancy of waters rather than by inspection (Section 1.6.6.4, script 5).
2.6.2 Methods

Script 5 (Section 1.6.6.4) relates to biplanar and triplanar DNA systems in which ‘resid’ defines the O6 atom for each guanine associating into planar quartets. For the biplanar models, ‘x’ residue identities were: 4 5 8 9 12 13 21 22. Residue identities were amended as appropriate for triplanar models and for quadruplexes derived from other bases. Typically, residue identities were: 2 3 4 8 9 10 14 15 16 20 21 22. Hence, all waters within 2 Å of confining atoms were characterised. Water occupancy in the central region of triplanar 1KF1 and biplanar 2KF8 was assessed. Both models were also investigated with a single methylation at O6. Figures 2.29a and 2.29b illustrate the component bases associating in quartet formation.

![Figures 2.29a and 2.29b: residues H-bonding to form quartets in 2KF8 (A) and 1KF1 (B). Predictions for waters within 2Å of guanine O6 for each residue were plotted.](image)

2.6.3 Results

Supplementary figure S2.20 indicates the non-methylated triplanar system allows a significantly higher number of waters in the interplanar spaces than the methylated equivalent. Given the greater volume available for occupancy in both triplanar models, the number of water molecules allowable is reasonably expected to be higher than for the biplanar structure. The methylated model of the latter tolerates an occupancy of either one or two water better than the non-methylated model. However, 2KF8_meth allows three, four or five waters into the central space more frequently. This model has a noticeably higher capacity for occupancy of four or five waters. The observations for both 1KF1 and 2KF8 resonate with earlier findings that steric hindrance of O6-methylation is disruptive: the methyl group is intrusive and reduces available space for water molecules. The central quartet of 1KF1_meth is methylated which is not an option in a biplanar
system. This may be causative in the greater relative disruption implied by the trajectories for triplanar 1KF1\_meth and biplanar 2KF8\_meth.

The histogram of S2.20 data shows an optimal occupancy of one to five waters for both models but with significantly less frequency for the O6-methylated model (figure 2.30).

**Figure 2.30**: number of waters over time within 2Å of guanine O6 for the three quartets of 1KF1 (left) and 1KF1 methylated at O6 (right). The former model has a much higher permanent population of waters than the methylated equivalent. It also allows a greater number of waters (twelve) than 1KF1\_meth (seven), and fewer occasions displaying a complete absence of waters.
Figure 2.31: Histogram of figure 2.30 data. 1KF1 total occurrence = 2360 waters, 1KF1_meth total occurrence = 1060 waters. Both models show an optimum of three waters allowed but the maxima differ noticeably. Values for the number of waters beyond seven in the methylated model are too low to display given the resolution available for the plotting software.

Given less available space, water occupancy for a biplanar quadruplex is proportionately lower than for the triplanar equivalent. Tolerance of O6-methylation is higher in the biplanar models than for the triplanar system (figure 2.32). A histogram of these data confirms lower occupancies than for the latter system but also less differential between the two models (figure 2.30).

Figure 2.32: number of waters over time within 2 Å of guanine O6 for biplanar quartets. Non-methylated 2KF8 appears to tolerate a higher number of waters than the methylated equivalent. Both very briefly allow five waters in the central space at ~25 ns and ~50 ns respectively.
2KF8 total occurrence = 2040 waters, 2KF8_meth total occurrence = 1935 waters. The models indicate an optimum population of two waters and a maximum of five. 2KF8_methylated shows less consistency of numbers at all levels of occupancy. This observation would resonate with other metrics indicating O6-methylation reduces the interplanar space sufficiently to permit a level of occupancy comparable with the non-methylated equivalent.

Supplementary figures S68, S69 show water occupancy data for Na⁺ or Li⁺ co-ordinated 1KF1 models and S70, S71, S72 data for 2KF8 models co-ordinated by K⁺, Na⁺ or Na⁺ respectively.

2.7 Ribose conformations and potential compromise of function

2.7.1 Introduction
The non-planarity of ribose as a sugar ring in nucleic acids is predicated on the four carbon atoms and one oxygen comprising the structure. Accordingly, the exact non-planar conformation derives from the five interior angles of the pentagonal ring. Implicitly, the most energetically favourable conformation would show the maximum possible interatomic distances between the principal components. The basic substituents of DNA and RNA sugar rings are shown in figure 2.34.
Figure 2.34: comparison of deoxyribose (A) and ribose (B) and atom numeration. The C2’ carbon of ribose has a single bonded oxygen unlike deoxyribose. The additional oxygen allows greater reactivity with consequent biological ramifications (discussed in text).

Unequal charge distribution in the molecule is contingent on these atomic components. The lone pair of O4’ could prompt dipole interactions therefore subsequent differences in non-planarity, an interaction colloquially referred to as the Rabbit Ear Effect. Interactions in DNA and RNA are susceptible to the electronegativity of the sugar-phosphate backbone (124). Reciprocity arises from re-orientation of substituents triggering changes in planarity and causing significant alterations in the backbone (125) (126). Consequently, there are ostensible links within a biological framework (127).

The displacement vector away from the plane determines terminology. Carbon facing the plane of the base is defined as an -endo configuration; facing away from the base defines an -exo configuration. However, naïve conformers are affected by substituent differentials in the sugar ring. Hence there is some latitude in an exact angular definition for both endo forms; there is a range of values for internal angles within which the endo form can arise. C2’-endo structures are ubiquitous in purines, but pyrimidines favour the C3’-endo form. The prevalence in RNA and C3’-endo conformers is contingent on the marginally greater interatomic distance between C2’ and C3’ oxygens in ribonucleosides. Fluctuations in the internal angles of the sugar component result in various conformations. Two in particular were analysed in this study: C2’-endo and C3’-endo conformations (figure 2.35).
Figure 2.35: The C2'-endo conformation is uniquely observed in β-form DNA (A) but C3'-endo is prevalent in RNA and α-form DNA (B). Distance $d$ between consecutive phosphate atoms is greater for a C2'-endo conformation as a direct result of the molecular re-alignment.

Since the change from C2'-endo to C3'-endo takes place via a series of marginally different intermediates, the progression of non-planar conformers may be presenting as a continuum of variation (128). Nevertheless, the structures are defined by energy barriers. The conversion of C2'-endo to C3'-endo is mediated by O4'-endo.

The energy barrier ranges from 2kcal mol$^{-1}$ to 5kcal mol$^{-1}$ with an overall potential energy of 1.5kcal mol$^{-1}$ (129) (130). The C3'-endo pucker produces a significantly shorter phosphate-phosphate distance in the backbone, resulting in a more compact helical conformation.

2.7.2 Methods

Data from molecular dynamics simulations for biplanar DNA and RNA methylated and non-methylated guanine quadruplex models were analysed to identify sugar planarity. The biplanar guanine quadruplex DNA and RNA models were labelled as GDN12 and GDN12_meth, GRN8 and GRN8_meth.

Mean distances between consecutive phosphate oxygens ($d$, in figure 2.35) were very similar for both systems: 5.9 Å for C3'-endo and 7.2Å for C2'-endo DNA configurations; 5.8 Å for the C3'-endo RNA configuration. The two configurations are bounded by an approximate range of dihedral angles (131); C2'endo: 0° to 36°, C3'-endo: 144° to 180°. The specific dihedral describing these configurations is C'1-C2'-C3'-C4' and calculated with the VMD command: measure dihed {a b c d} frame all where a-d are atom numbers for C1'-C2'-C3'-C4'. For DNA models these atoms are listed in the .pdb file as: 156 155 152 151, and the RNA equivalent as: 204 203 201 199. This C1'-C2'-C3'-C4' dihedral was calculated for three sugar rings of residues in methylated and non-methylated DNA and RNA biplanar guanine quadruplexes. Residues selected were: 1, 7
and 15. The latter two are partial constituents of a quartet, and the former is peripheral to the complex.

2.7.3 Results

Data returns for the sugar conformations of biplanar and triplanar guanine quadruplexes are dynamic but maintain steady trajectories. Neither system differs significantly from the apposite values: DNA models are generally C3'-endo and RNA equivalents are C2'-endo (figures 2.36a and 2.36b):

![Figures 2.36a and 2.36b: trajectories for C1'-C2'-C3'-C4' dihedrals in DNA and RNA guanine models in residue 7 co-ordinated with K⁺ (left) and no cation (right).](image)

Distribution histograms for the trajectories of residue 1 in N6-methylated and non-methylated biplanar adenine structures clearly indicate the decisive conformational change for this specific residue in the latter condition. In comparison, the N6-methylated model shows no particular conformational preference (figure 2.37):
Figure 2.37: distribution histograms from data for figures 2.36a and 2.36b data GDN8-N1 (blue) and GDN12-N1 dihedrals (pink). C2'-endo configurations are clearly defined for the DNA model but less so in the RNA model. The wider range of values for this system imply conformations are alternating frequently during the simulation.

Empirical data for conformational interconversions across guanine DNA and RNA models with relevant methylation are tabulated in Tables 2.2 and 2.3:

Table 2.2: fluctuations in deoxyribose configurations for triplanar guanine quadruplex K⁺ co-ordination

<table>
<thead>
<tr>
<th>GDN12-N1 K⁺</th>
<th>t = 0ns</th>
<th>t = 500ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo</td>
<td>C1'-endo</td>
<td>C1'-endo</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: fluctuations in deoxyribose configurations for triplanar guanine quadruplex K⁺ co-ordination

<table>
<thead>
<tr>
<th>GDN12-N1 K⁺</th>
<th>t = 0ns</th>
<th>t = 500ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo</td>
<td>C1'-endo</td>
<td>C1'-endo</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>
Table 2.3: fluctuations in ribose configurations for RNA biplanar guanine quadruplex.

<table>
<thead>
<tr>
<th></th>
<th>GRN8-N1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t = 0ns</td>
<td>t = 500ns</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>17</td>
<td>20</td>
</tr>
</tbody>
</table>

C2'-endo conformation typifies ribose in many molecules; RNA quadruplexes investigated here follows that trend. The C2'-endo conformation persists for the duration if the simulation but with some interchange to other possibilities. Interestingly C1'-endo and C3'-endo are favoured.

2.7.4 O6-methylation of biplanar guanine models.

The trajectory for dihedrals in the biplanar RNA model showed a sudden transformation from C2'-endo to C3'-endo. This transition is not observed for the same residue in the equivalent model methylated at adenine N6. Steric hindrance of O6-methylation increasing the energy barrier underpins this observation. The trajectories for biplanar DNA and RNA adenine quadruplexes are very dynamic and indicate methylation provokes an unstable system in both DNA and RNA models (figures 2.38a and 2.38b). However, the latter maintains a more dynamic C3'-endo configuration throughout the simulation than for the biplanar DNA model.
Figures 2.38a and 2.38b: trajectories for biplanar RNA adenine quadruplexes with and without N6 methylation

Table 2.5: fluctuations in deoxyribose conformations for triplanar non-methylated guanine quadruplex (black) and O6-methylated guanine quadruplex (red). Conformations at the simulation start are identical for both models. Where no definite conformation could be identified, the structure is listed as planar.

<table>
<thead>
<tr>
<th>t = 0ns</th>
<th>GDN12-N1</th>
<th>GDN12-N1 meth</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C2'-exo</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t = 500ns</th>
<th>GDN12-N1</th>
<th>GDN12-N1 meth</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C2'-exo</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>22</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>11</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The C2'-endo conformation is prevalent for both models prior to the simulation. Marginal differences are observed in the non-methylated complex whilst noting that O4'-endo and O4'-exo conformations, absent at the start, arise by the end of the simulation. However, no particular conformation is apparent for the guanine O6 methylated model after the simulation. Most of the possible conformations are identified, suggesting none is favoured. The sugar ring may fluctuate as a continuum between the possibilities during simulation.
Table 2.6: fluctuations in ribose conformations for biplanar non-methylated guanine quadruplex (black) and O6-methylated guanine quadruplex (red). Conformations at the simulation start are identical for both models.

<table>
<thead>
<tr>
<th></th>
<th>GRN8-N1</th>
<th>GRN12-N1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C1'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>t = 0ns</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GRN8-N1</th>
<th>GRN12-N1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C1'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>t = 500ns</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>20</td>
</tr>
</tbody>
</table>

Interchangeability of sugar conformations are markedly more fluent for O6-methylated systems than for non-methylated equivalents (figure 2.39):

Figure 2.39: distribution histograms from figures 2.38a and 2.38b data for GDN8-N1 (pink) and GDN8-N1_meth dihedrals (blue). C2'-endo and C3'-endo configurations are evenly represented for the duration of the simulation. Sugar ring conformations are not well-defined in the O6-methylated model but C3'-endo configurations may have a higher population.
2.7.5 Discussion of C2'-endo and C3'endo configurations

Although the sample size for this brief survey was necessarily small, consideration of some aspects is worthwhile. In particular, the rapid C2'-endo to C3'endo conversion predicted in the trajectory for figure S2.27a needs comment. The interconversion for residue 1 may be unique or occur in other bases excluded from the sample. Moreover, the change may be reversible at some unspecified point after 500ns if a simulation were run for long enough. However, the energy of the quadruplex system should be net zero eventually. This indicates the change in energy from C2'-endo to C3'endo must be offset elsewhere in the system. The interconversion in residue 1 may well be non-transient, but a complementary change in energy should be found in other sugar ring interconversions.

Conformational re-alignment from C2'-endo to C3'endo may influence nucleotide polymerisation (132) whereas C3'-endo to C2'-endo may be crucial in the catalytic capacity of RNA (133). These two observations imply that flexibility of RNA configuration is an important biological requirement. A contribution to stability of the C2'-endo conformer might derive from reduced hydration waters. A single water molecule can locate comfortably between two free phosphate oxygens of consecutive bases in the C2'-endo configuration (132). Consequently, there is an overall reduction in the enthalpy of hydration as an RNA strand construed with this configuration would require fewer water molecules. Conversely, this implies there should be an entropic cost for the reconfiguration to the C3'-endo conformer as a reduced population of waters would need to off-set the energy differential. In turn, it follows that the C3'-endo structure must be biologically imperative if the transition occurs regularly despite being thermodynamically unfavourable.

Data for C1'-C2'-C3'-C4' dihedrals in the biplanar quadruplexes investigated here, strongly indicate O6-methylation in guanine models disallows any stable configuration of the sugar ring. A recognisable and distinct switch from C2'-endo to C3'endo appears to be ruled out over the limited time scale of the simulation. The biological function of RNA as a catalyst and agent for polymerisation would therefore seem to be compromised by the entropic tariff of differential hydration between C2'-endo and C3'endo configurations.

2.8 Biplanar and triplanar DNA quadruplexes co-ordinated by K⁺, Na⁺ or Li⁺

2.8.1 Methods

To assess the effects, if any, on the integrity of guanine complexes, the co-ordinating cation was varied between K⁺, Na⁺ or Li⁺. Both 1KF1 and 2KF8 were downloaded from the PDB and opened as a PSF. The files were manually rewritten to substitute Na⁺ or Li⁺ for K⁺. The modifications were confirmed visually in Chimera and models transferred to
a Linux server for MD simulations. The only metric accessed for analysis was RMSD. A hybrid triplanar guanine-cytosine-guanine quadruplex was also constructed and co-ordinated by K\textsuperscript{+}, Na\textsuperscript{+} or Li\textsuperscript{+}.

2.8.2 Results

2.8.2.1 Models derived from 1KF1

The trajectories for Li\textsuperscript{+} or K\textsuperscript{+} co-ordinated triplanar systems show little dynamic activity when assessed as RMSD (S29). The Li\textsuperscript{+} system appears particularly stable at \(~2.5\ \text{Å}\) to 500ns. RMSD for the K\textsuperscript{+} system ranges between 3.0 Å and 3.5 Å. The Na\textsuperscript{+} trajectory is stable at \(~3.3\ \text{Å}\) to \(~220\ \text{ns}\) when the systems show noticeable dynamic activity. Similar trajectories were obtained for a guanine-cytosine-guanine quadruplex derived from 1KF1 (S30). Apart from a brief rise between \(~2.0\ \text{Å}\) and \(~3.0\ \text{Å}\) early in the simulation, Li\textsuperscript{+} as co-ordinant shows the least dynamic activity at a steady \(~2.0\ \text{Å}\). K\textsuperscript{+} maintains a steady trajectory between \(~2.5\ \text{Å}\) and \(~3.5\ \text{Å}\). The Na\textsuperscript{+} trajectory shows a dynamic rising trend to \(~120\ \text{ns}\) from \(~1.8\ \text{Å}\) to \(~4.2\ \text{Å}\) after which time, the system maintains a dynamic but stable trajectory.

2.8.2.2 Models derived from 2KF8

The trajectories for Li\textsuperscript{+}, Na\textsuperscript{+} or Li\textsuperscript{+} co-ordinated biplanar systems are broadly similar to the triplanar results (S31). Li\textsuperscript{+} shows a reasonably steady trend between \(~1.5\) and 2.5 Å with some dynamic fluctuation. The trajectory for K\textsuperscript{+} is similar, ranging between \(~2.0\ \text{Å}\) and \(~3.2\ \text{Å}\). The plot for Na\textsuperscript{+} has a steady trend at \(~2.5\ \text{Å}\) to \(~350\ \text{ns}\) at which time it follows a sharp rise to \(~4.5\ \text{Å}\) before a downward trend to the simulation end. 2KF8 modified with cytosine substitutions for all guanines of one quartet shows Na\textsuperscript{+} and K\textsuperscript{+} trajectories are notably more dynamic than the non-modified 2KF8 system (S32). There is a gradual rising trend for the former from \(~2.5\ \text{Å}\) to \(~4.4\ \text{Å}\). This contrasts with the K\textsuperscript{+} trajectory which, although ranging between \(~2.2\ \text{Å}\) and \(~3.2\ \text{Å}\), shows a sharp rise at \(~390\ \text{ns}\) before eventually levelling at \(~4.4\ \text{Å}\). This level is maintained to the simulation end with little dynamic activity. A guanine-cytosine-guanine hybrid was constructed for comparison with guanine-only equivalents. MD analysis was limited to a brief analysis.
2.9 Executive summary of guanine quadruplexes.

- all metrics applied in this study show models constructed here are stable on a biologically relevant time scale
- internal geometry of structures as well as stability depends on cationic co-ordination
- Li\(^+\) as a co-ordinant in triplanar models is demonstrated to stabilise as system as least as well as K\(^+\) or Na\(^+\). This aspect has been historically equivocal
- O6-methylation or thionation is very disruptive. K\(^+\) or Na\(^+\) are repelled rapidly and early in a simulation form the system, but Li\(^+\) relocates only to the molecular periphery.
- Steric hindrance is a likely cause but other factors cannot be discounted. Access to the central core, if methylated / thionated, would confound a correcting transferase and thereby perpetuate any disruption. Reduced local electronegativity could partially explain difference observed in cation behaviour.
- Biplanar models are significantly more stable in the absence of a cation.
- Intra-planar waters appear to mediate stability for all species, although no mechanism is directly identified here.
- Methyltransferase as a correcting factor for O6-methylation is more effective in biplanar models. The lower number of quartets \textit{ipso facto} may rationalise this observation.
- Inter-planar waters and robust \(\pi\)-bond population might enhance the stability of such models more than anticipated.
- A single strand 22-mer with an identical base sequence as required for folding into guanine quadruplexes, is not thermodynamically stable. The collapse of the model begins early in the simulation.

2.10 General discussion of guanine quadruplexes

Loss of cations in triplanar structures is shown in this study to be detrimental to the integrity of DNA-derived complexes. However, in the case of 2KF8, biplanar quadruplexes appear to be thermodynamically stable despite an absence of co-ordinating cations. Biplanar quadruplexes are non-transient to 500ns without any co-ordination. This inference is broadly supported by all the metrics used here. The observation resonates with research noted in section 2.1.1: quartet integrity in 1KF1 is less affected by cation removal than are the loop structures. The less complex biplanar system appears not to be overly affected by cationic absence predicated on fewer loops. Perhaps \(\pi\)-stacking has a greater influence in a more compact biplanar system.

Although K\(^+\) is more favoured in physiological systems, Na\(^+\) and Li\(^+\) are not unlikely as co-ordinating cations in guanine quadruplexes. As noted in Section 2.1.1, the influence, if any, of Li\(^+\) as co-ordinating cation appears to be as yet unresolved directly. The MD
simulations here for varying the co-ordinant indicate Li$^+$ is not removed from quadruplex systems, even in methylated systems. Structural integrity was lost for K$^+$ and Na$^+$ models relatively early during MD simulations, whereas Li$^+$ remained within a given quadruplex although removed to the periphery. A greater local population of water molecules contingent on the higher hydration enthalpy of Li$^+$ was proposed as marginally enhancing stability.

Across all metrics in this study, O6 methylation is disruptive except for Li$^+$ co-ordinated systems. Of the three potential causative agents noted in Section 2.4.3, the research strongly indicated steric hindrance is the prime factor. Unfavourable bump interactions were predicted for both methylated and thionated models; this observation is closely associated with steric hindrance. The data suggest reduction of electronegativity predicated on local distortion and destabilisation from less effective Hoogsteen bonding were minimal factors. However, differential hydration energy may have some influence. The rapid and permanent exclusion of K$^+$ and Na$^+$ co-ordinating cations is better accounted for by steric hindrance.

Of particular interest are the brief surveys of water occupancy in biplanar and triplanar quadruplexes and sugar ring configurations. The former showed a limited population of waters may be viable within 2 Å of the quartet structures. This observation has seemingly not been previously noted. However, it is not as yet possible to state for certain if water molecules are transient or are ionised during the course of a simulation. The numbers are low even for the optimum model (1KF1 non-methylated) and fluctuate considerably for all systems. The significance of waters in quadruplexes may lie partially in their possible interactions with ribose and deoxyribose moieties of the structure.

Periods of very low dynamic activity for residues of some DNA models were noted. Given the additional oxygen at C2’ in ribose increases the size and mass of the sugar compared to a single C2’ hydrogen in deoxyribose, the ribose structure should have a higher energy barrier for interconversion. This implicitly suggests the activation energy for interconversion is higher. Hence the data for the low sample numbers is unanticipated. Two proposals arise. Firstly, presence of the C2’ hydrogen group may predispose a specific conformation with higher energy requirements for flipping from C2’-endo to C3’-endo. This interpretation would be viable if the ground state of the DNA system is lower than for the RNA model. In which case, the activation energy would need to be higher for attaining the transition state. Secondly, an enhanced activation energy for the RNA sugar may also derive from the C2’ hydroxy forging transient or semi-transient H-bonds with adjacent atoms. H-bonding would be possible with a local nitrogen for instance, especially if it were mediated by one or more waters. Transient hydroniums may also be a possible factor.
DNA and RNA quadruplexes, biplanar and triplanar, show notable stability on a biologically relevant time scale in this study. In particular, the greater stability of biplanar complexes lacking cationic co-ordination, may be partially compensated by the core quartets themselves. The phosphate backbone and interplanar loops are affected more by removal of a co-ordinant (40). The core region may remain relatively stable (reflected in eRMSD data), whilst the loops and phosphates show noticeable dynamic activity (reflected in the RMSD data). eRMSD trajectories were assessed as being much less dynamic for both systems than their RMSD complements because the metric excludes waters and the phosphate backbone from evaluation. The structural integrity of DNA models may be limited by their descent from a duplex strand with an associated higher energy barrier. RNA systems have an inherent single strand derivation and a consequent lower kinetic impediment. A molecule will be stable if there are more attractive forces than repulsive equivalents. Repulsive forces increase the potential energy of the molecule. Hence molecules with lower energy show greater structural integrity.

The ostensible paradox of high RMSD values with apparent structural integrity in the 22mer single strand DNA and RNA systems may be resolved by inspection. Comparison with figures 2.15a and 2.15b for the DNA model and 2.17a and 2.17b for the RNA equivalent, suggests the initial high RMSD for both is contingent on the unravelling helix. The noticeable dynamic profile continues but does not increase to any great degree. This observation may imply whilst large scale unravelling has occurred, locally base-phosphate backbone interaction is not as wide-ranging. The effects, if any, of waters within and immediately adjacent to the helices is worthy of further exploration.

Given the stable yet dynamic nature of purine quadruplexes, in particular the RNA models observed in this study, the speculation arises that these systems may be continually folding and unfolding. Transient cationic ligation or small molecule binding might have a role here. Investigating cations other than K⁺, Na⁺, and Li⁺ may also provide further insights into the characteristics of quadruplex structure and interactions. Recent research supports the notion that RNA quadruplexes may only arise briefly in vivo or are very challenging to detect directly (134) (135). A fluorescent labelling methodology has recently visualised in vivo guanine quadruplexes in real-time. The technique was adapted here as a proposal for similar application to adenine quadruplexes (private communication with Prof. P Taylor). Visualisation relies on the activation of a spiro-silafluorene under UV light. Synthesis of the molecule complexed with a small molecule with affinity for a motif prevalent in biplanar adenine structures is described in Section 5.8.1.
2.11 Future work guanine quadruplexes

This survey of biplanar and triplanar guanine quadruplex form and function via the perspective of molecular dynamics analysis, provides a reference framework for a comparative interpretation of similar unorthodox structures derived uniquely from other bases. Stability of the structures derives especially from cationic co-ordination. But this observation does not apply to biplanar models which are readily disrupted by imposing an inter-planar cation. Molecular modelling of the central core in particular may add insight into cationic interactions with waters. Their role in the system is worthy of further investigation. Not only intra-planar stability may be mediated, but also modelling waters interacting with intra-planar loops could be propitious.

Intra-planar loops for all models constructed in this study are comprised of three bases, typically ...TTA... It is uncertain how loop length influences the system locally and as a whole. Given function often follows form, novel topologies of loops may add nuance to data compiled here.

Further work may establish any relationship between the number of quartets in RNA quadruplexes and individual roles in cellular responses to fluctuation in tissue-specific environments.

G4s are readily accessible to ligands and substitutions at O6 and various helicases proliferate in G4 regions. Implicitly, there may be an association to explore as well as possible histone or nucleosome interactions. Work might identify chromatin influence on quadruplex activity and accessibility of nucleosome G4s to ligands.

Future work might determine if duplex G4s responsible for incorrect transcription at replication forks might carry forward instructions prompting RNA transcripts to compile pathogenic G4s.
Chapter 3 cytosine and thymine quadruplexes

3.1. Introduction

The survey and analysis of guanine quadruplexes in Chapter 2 does not consider the complementary strand of duplex DNA. The human telomeric sequence is necessarily rich in cytosine having a general profile of $d(AATCCC)_n$. An *in vitro* supramolecular structure predicated on protonated C-C base pairing was initially reported some time ago (136). The stability of this conformation, characterised as an *i*-motif, is critically contingent on local pH (137) (138). This particular supramolecular structure is especially prevalent in cytosine-rich sequences (139). The structure is more stable in pH conditions below physiological levels and is consolidated by π-π stacking (figure 3.1).

![Figure 3.1](image.png)

**Figure 3.1:** for pH values < 7.0, cytosines can be protonated to associate into a dimer. (A) The paired bases share three H-bonds. (B) Dimers can stack vertically with consecutive planes rotated through 90°. (C) Alternate planes (light green and dark green) are linked by the phosphate-sugar backbone (light green and dark green vertical lines).

The cytosine dimer system may have other topologies (140). The low values for local pH of the *in vitro* models discussed in the earliest papers would seem to preclude an *in vivo* telomeric cytosine *i*-motif. However, recent research concludes the structure may be relatively stable at levels close to physiological pH (141). Moreover, the sequence required for potentially forming an cytosine *i*-motif appears ubiquitous in the human genome and particularly in the centromeric environment (142).

As noted in Chapter 1 and discussed in Chapter 6, guanine quadruplexes have been promising targets for anti-cancer intervention. But paradoxically, some complexes have been linked *inter alia* to neurodegeneration. It is suggested here that anti-sense cytosines may also self-associate into a quadruplex structure. Either uniquely, or in combination with guanine, these may compromise sense-strand beneficial guanine quadruplex activity. Bio-informatics identify G4s populate specific sites such as immunoglobulin switch regions, promoter regions and recombination sites (22) (106). A gene regulatory role has been proposed for G4s adjacent to transcription start sites (89). Exploring the possibility
of cytosine quadruplexes located complementary to beneficial guanine equivalents with such crucial roles, might underpin some aspects of the paradoxical pathogenicity. The proposal that anti-sense quadruplexes may also fold appears not to have been investigated. Accounting for these proposals in the C-rich anti-sense strand suggested an initial assessment of cytosine dimerisation.

3.2 Cytosine dimers.

A series of cytosine-\( X \) dimers where \( X = \) adenine, cytosine or thymine were designed and are shown in figure 3.2. A pre-requisite for a feasible structure to incorporate into a nucleic acid strand was cytosine-N1 availability (circled) for covalent bonding to the phosphate backbone.

![Figure 3.2: Cytosine – X dimers](image)

(a) cytosine-cytosine I  
(b) cytosine-cytosine II  
(c) cytosine-cytosine III  
(d) cytosine P+ cytosine  
(e) cytosine-thymine  
(f) cytosine-adenine I  
(g) cytosine-adenine II

**Figure 3.2: cytosine – X dimers.** (a), (b) \( X = \) cytosine (c) \( X = \) cytosine, dimer co-ordinated by metal cation (d) \( X = \) cytosine, protonated dimer (e) \( X = \) thymine (f), (g) \( X = \) adenine.
Structures 3.2e-3.2g were rejected because of no N1 availability. In addition, 3.2f and 3.2g were also discounted due to unfavourable H-bond alignments. 3.2d was considered feasible based on the early research into the i-motif structure. However, the caveat was noted that protonated cytosine is unlikely at physiological pH. Figure 3.2a appeared viable in principle and showed some commonality with a structure recently designed \textit{in silico} on a graphene surface (143). However, the graphene paper discusses the relevance of dimerisation to nanoparticle technology and does not extend to raising implications for a biological context. Amino hydrogens of one paired dimer have recently been suggested as capable of H-bonding to carbonyl oxygen of the other (144). Further structures were developed from the structure described in figure 3.2a to explore the possibility of forming a cytosine quartet. Both cytosines of the model were marginally re-aligned. This manipulation suggested a dimer was possible by H-bonding between H atoms on N4 of one cytosine with O2 and N3 of the other (figure 3.2b). Amino hydrogens of one paired dimer have recently been suggested as capable of H-bonding to carbonyl oxygen of the other (145) (146). Proposed here is the formation of a cytosine quartet where two such re-orientated dimers might further H-bond (figure 3.3). Significantly, N1 hydrogens are available in this model for covalent bonding to the DNA chain.

\textbf{Figure 3.3: possible alignment of a cytosine dimer pair} (from figure 3.3B). Both molecules are planar and the arrangement appears viable except that carbonyl oxygens have unfavourable angles for H-bonding with opposing amino groups.

The bonding angle O2-H1-N4 between associating cytosines is far from 180°. Within a small tolerance, this value is required for orthodox H-bonds (147) and to be sustainable, this point should be addressed.
3.3 Cytosine quartets

3.3.1 Cytosine quartet co-ordinated with non-physiologically relevant ions: Ba\(^{2+}\)

A hybrid guanine-cytosine quartet may fold in C-rich DNA and stack vertically between guanine quartets (148). The paper does not describe the molecular structure of a cytosine quartet and has no proposal of progression to a quadruplex. A cytosine quartet has been identified in a guanine quadruplex co-ordinated with Ba\(^{2+}\) mediated by water molecules (figure 3.4) (149). Although the structure is planar and appears well-served by H-bonds, barium is toxic in human physiology (150).

![Figure 3.4. Water molecules mediating Ba\(^{2+}\) co-ordinating a cytosine quartet. Adapted from Zhang, 2014.](image)

Figure 3.4. Water molecules mediating Ba\(^{2+}\) co-ordinating a cytosine quartet. Adapted from Zhang, 2014. Four self-associating cytosines, mediated by water molecules and co-ordinated with Ba\(^{2+}\). Cytosine N1 of all four bases appears to available for covalent bonding external to the quartet.

Cytosine quartets have previously only been observed in vitro DNA solutions (151). Protonation of cytosine N3 and the consequent delocalised positive charge on the quartet were suggested as a mechanism for interaction with a cation mediated by a molecule of water (151). Distances between cytosines were too large to co-ordinate a cation directly, hence waters were required to mediate the electrostatic bonding.

3.3.2 Cytosine quartets co-ordinated with non-physiologically relevant ions: Pb\(^{2+}\), Tl\(^{+}\), Rb\(^{+}\), Hg\(^{2+}\)

Quartets have recently been identified in the crystal structure of a DNA quadruplex comprising tetrads of various base combinations including cytosine-thymine pairs. (80) (152). However, a majority of these complexes are co-ordinated by non-physiologically
relevant ions such as Pb$^{2+}$, Tl$^+$, Rb$^+$ (153) Hg$^{2+}$ (154). The NMR structure of a non-hybrid uniquely cytosine quartet located between guanine quartets has been reported (155). The folding sequence derives from the Simian Virus 40 genome. No co-ordinating cation is noted, but there appears to be sufficient intramolecular space in the central channel to accommodate ions mediated by waters. The model displays single H-bonds between each cytosine O2 and adjacent amino-H, a total of four H-bonds.

3.3.3. Cytosine quartets co-ordinated with physiologically relevant ions: Li$^+$, Na$^+$ or K$^+$

A cytosine quartet derived from two re-orientated dimers (figure 3.2c: cytosine III), might further H-bond. Significantly, such a structure retains N1 hydrogen availability for non-covalent deoxyribophosphate bonding. Each cytosine has four H-bonds to adjacent molecules suggesting a degree of stability. In contrast, the cytosine quartet constructed on a graphene surface (143) can only be assembled if one of the cytosines is inverted (figure 3.5, circled in orange). Moreover, the complex does not lend itself to linkage with a DNA strand since one cytosine N1 (circled in red), is not available.

![Figure 3.5. Potential cytosine quartet formed on graphene](image)

Given these constraints, the structure was rejected as invalid. Accordingly, focus returned to the paired dimer in figure 3.3(B). This model was slightly re-aligned to allow favourable N-H…O bonding between carbonyl oxygens and amino hydrogens (figure 3.6):
Figure 3.6. Re-aligned paired cytosine dimers to allow H-bond angles (O2-H1-N4) converging on 180°. The electronegative central space is sufficient to accommodate a metal cation (M⁺). Distances between ion and each cytosine-O2 = 3.381 Å; between M⁺ and each cytosine-N3 = 3.683 Å. Six H-bonds are predicted for this structure. The model described in section 1.2.2 has four H-bonds.

Overall, these observations mitigate the plausibility of cytosine quartets outlined in figure 3.6 stacking vertically into a quadruplex. The system is predicted to be planar and has N1 hydrogen availability for non-covalent bonding to loops in a phosphate chain. There is also sufficient interplanar space for cationic co-ordination.

3.3.4. Guanine-cytosine hybrid quartets

The comment in Section 1.1 noted that cytosine in combination with guanine may associate into a quadruplex structure. Antisense cytosine-rich regions have recently been suggested as allowing a GCGC quartet to fold between a pair of guanine quartets (148). The possibility of guanine-cytosine combinations in quadruplexes was proposed as underpinning some forms of neurodegeneration (156), and also in the regulatory region of the gene PLEKHG3 strongly associated with autism (157). Some epigenetic processes might be affected by such a guanine-cytosine-guanine quadruplex (148). Developing a G-C-G-C quartet to a C-C-C-C structure is a rational next step. Such a cytosine quartet located between two guanine quartets could derive from:
which is characteristic of the single strand telomeric sequence folding into a G4, but with four G-C substitutions. Other guanines could also undergo G-C substitution. Although the sequence has loops comprising three bases, it is as yet unclear if a maximum number has been validated. Selection pressures have favoured short loops as stable and optimised folding characteristics \((158)\). However, investigation of a G-C-G hybrid was not taken up as increasing relevance of possible adenine quadruplexes developed during this thesis. Accordingly, a G-A-G hybrid quadruplex was constructed and MD analysis is described in Chapter 4, section 4.3. The system maintained stability to 500ns when co-ordinated by Li\(^+\), Na\(^+\) or K\(^+\).

### 3.4. Cytosine quadruplexes

Cytosine dimers pairing into a quartet, then three quartets stacking into a cytosine quadruplex (C4) is a logical sequential progression. As noted in a recent study, there are conflicting observations regarding quadruplex stability with or without co-ordinating ions \((40)\). For example, the stability of guanine quadruplexes co-ordinated with Li\(^+\) has been regarded as contentious \((80) (153)\). However, since 1KF1 acted as scaffold and is co-ordinated by K\(^+\), this cation together with Na\(^+\) and Li\(^+\) were included in models for MD simulations.

### 3.5 Methods

1KF1 was a scaffold for the former and 2KF8 for the latter. Cytosine was substituted for all guanines in the quartets but loops remained unaltered (refer to figure 1.2). Nine models were constructed for triplanar and biplanar quadruplexes in Chimera (Table 1.1, Chapter 2) by vertically stacking cytosine quartets linked by AAT loops using the swapna command. Triplanar models retained K\(^+\) ions co-ordination by virtue of their 1KF1 derivation. The ions were replaced in turn with Li\(^+\) or Na\(^+\) by manual reformat of the psf and pdb files. Since 2KF8 has no co-ordinating cations, the derived cytosine model endorsed this factor. However, for comparison with biplanar guanine equivalents, K\(^+\), Na\(^+\) or Li\(^+\) ions were imposed on biplanar cytosine models by the same method. Octagonal co-ordination was noted for K\(^+\) and Na\(^+\) systems. The triplanar cytosine quadruplex co-ordinated with K\(^+\) and the biplanar complex without co-ordination were further assessed with methylation at C5. Models were transferred to DSV for confirmation of structural changes and to note modified residue numbers. Monitoring triplanar models in DSV for K\(^+\) or Na\(^+\) systems confirmed co-ordinating ions have octagonal co-ordination. This observation is in close agreement with a 2016 review paper \((159)\). However, Li\(^+\) models
adopted a tetrahedral geometry. DSV models showed $\pi$-H-bond donor interactions between amino nitrogens and cytosine, resonating with in silico cytosine $\pi$-stacking interactions noted in earlier research (160).

Triplanar quadruplexes are referred to as CDN12 and biplanar as CDN8. All models were submitted to NAMD with parameters as defined in Section 1.2, Chapter 1 and metrics applied as in Section 1.3, Chapter 1.

3.6. Results

3.6.1 Triplanar cytosine quadruplex (CDN12) K$^+$ co-ordination

3.6.1.1 RMSD (S33)

The trajectory shows a strong rising trend between $t = 0$ ns and $t = \sim 60$ ns from $\sim 2.2 \text{Å}$ to $\sim 4.2 \text{Å}$ which drops suddenly to $\sim 3.2 \text{Å}$ and thereafter continues with a slight rising trend to $\sim 4.3 \text{Å}$ at the simulation end. There is some enhanced dynamic activity between $\sim 180 \text{Å}$ and $\sim 250 \text{Å}$.

3.6.1.2 eRMSD (S34)

The trajectory is marginally dynamic but maintains a steady trend for the entire simulation with eRMSD values ranging between $\sim 1.7 \text{Å}$ and $\sim 2.8 \text{Å}$.

3.6.1.3 Rg (S35)

The system shows two rapid cycles of expansion and contraction before returning to a much less dynamic trajectory to the simulation end. Maxima and minima are $0.5 \text{Å}$ and $1.5 \text{Å}$.

3.6.1.4 H-bond formation (S36, S37)

Time dependent data (S36) show an inconsistent pattern of H-bond formation. A histogram (S37) indicates the population mean ranges between seven and eleven, but values for zero or twenty are comparable.

3.6.2 Triplanar cytosine quadruplex (CDN12) Na$^+$ co-ordination

3.6.2.1 RMSD (S38)

Although the trajectory has a distinctly dynamic rising trend for most of the simulation, there are two briefly level dynamic regions between $t = 0$ ns and $t = \sim 40$ ns, and $t = \sim 460$ ns and 500 ns. A sharp peak is noticeable between $t = \sim 190$ ns and $t = \sim 230$ ns to $\sim 4.7$ Å.
3.6.2.2 eRMSD (S39)

The plot is very similar to the trajectory for the CDN12 K$^+$ co-ordinated model. Although more dynamic activity is indicated, the values maintain a steady range between ~1.6 Å and ~2.8 Å. There may be a poorly-defined cyclical trend to the data.

3.6.2.3 Rg (S40)

The system is very dynamic with continual cycles of contraction and expansion. There is a sudden and rapid reduction in Rg at ~210 ns.

3.6.2.4 H-bond formation (S41, S42)

Time dependent data (S41) show an inconsistent pattern of H-bond formation similar to data for the K$^+$ model except low single-figure values are more frequent. The histogram (S42) shows a poorly defined maximum between seven and fifteen.

3.6.3 Triplanar cytosine quadruplex (CDN12) Li$^+$ co-ordination

3.6.3.1 RMSD (S43)

The trajectory is marginally dynamic but noticeably level with values ranging between ~1.6 Å and ~2.7 Å. There is a brief peak centred on t = ~150 ns.

3.6.3.2 eRMSD (S44)

Although some dynamic activity is indicated, the trajectory maintains a fairly constant path to t = ~280 ns. Thereafter, a slight rise is noticed to the simulation end. The plot is less erratic than either the K$^+$ or Na$^+$ models, with values centred on ~2.0 Å.

3.6.3.3 Rg (S45)

The trajectory shows very little activity for the duration of the simulation. There are no obvious changes in Rg with values ranging between 0.6 Å and 0.9 Å

3.6.3.4 H-bond formation (S46, S47)

Time dependent data (S46) show a marginally less erratic frequency than for K$^+$ or Na$^+$ models but slightly greater number for zero and higher value populations. The histogram (S47) calculates a very poorly-defined maximum ranging between five and eighteen.
3.7.1 Biplanar cytosine quadruplex (CDN8) No ion

3.7.1.1 RMSD (S48)

There is noticeable dynamic activity between \( t = 0 \) ns and \( t = \sim 70 \) ns with RMSD ranging from \( \sim 2.1 \) Å to \( \sim 4.5 \) Å and a brief smaller peak centred on \( t = \sim 200 \) ns. Thereafter the trajectory is steady with minimal activity.

3.7.1.2 eRMSD (S49)

There is some dynamic activity recorded in the plot, but no obvious continued rising or descending trend is indicated. Values range between \( \sim 1.3 \) Å and 2.0 Å.

3.7.1.3 Rg (S50)

The system is dynamic but displays only minor expansions and contractions. There is a slight reduction at \( \sim 90 \) ns before a return at \( \sim 120 \) ns. Values range between 0.6 Å and 0.8 Å.

3.7.1.4 H-bond formation (S51, S52)

Frequency of H-bond formation is relatively constant for the duration of the simulation. A large majority of the population centred on a frequency of two. A marginally lower population frequency centred on four is predicted and a third stratum with a much lower population is centred on a frequency of eight. The histogram (S52) supports this observation.

3.7.2 Biplanar cytosine quadruplex (CDN8) K⁺ co-ordination

3.7.2.1 RMSD (S53)

The trajectory is noticeably dynamic and generally follows a rising trend with RMSD values ranging between \( \sim 1.6 \) Å and \( \sim 4.9 \) Å. A less dynamic region occurs between \( t = 0 \) ns and \( t = \sim 60 \) ns. There are two steep rising trends between \( t = \sim 220 \) ns and \( t = \sim 300 \) ns, and \( t = \sim 400 \) ns and 500 ns linked by a complementary descending trend.

3.7.2.2 eRMSD (S54)

The trajectory has a brief approximately level region to \( t = \sim 90 \) ns, after which there is noticeable activity. A sharp rising trend to \( \sim 2.9 \) Å is indicated at \( t = \sim 180 \) ns which levels to an erratic plot ranging between \( \sim 1.5 \) Å and \( \sim 2.5 \) Å to the simulation end.
3.7.2.3 Rg (S55)

The trajectory is dynamic with no obvious cyclical expansion or contraction. The model appears to develop more instability after ~300 ns.

3.7.2.4 H bond formation (S56, S57)

Time-dependent data (S56) show the majority of values calculated at five or less are not apparent as observed in the strata of 3.6.1.4. Maxima shown in the histogram (S57) centre on five H-bonds, but range between four and seven.

3.7.3 Biplanar cytosine quadruplex (CDN8) Na\(^+\) co-ordination

3.7.3.1 RMSD (S58)

RMSD values closely resemble the trajectory for CDN8 co-ordinated by K\(^+\) and overall show similar dynamic activity.

3.7.3.2 eRMSD (S59)

The trajectory is highly erratic for the entire simulation with eRMSD values similar to the CDN8 K\(^+\) co-ordinated model.

3.7.3.3 Rg (S60)

There is a slight downward trend in values to ~200 ns after which the system develops noticeable variation in Rg. There is no obvious cyclical character to the trajectory.

3.7.3.4 H-bond formation (S61, S62)

There appear to be minor peaks at three, five and six H-bonds (S61) although local maxima appear across most of the histogram (S62).

3.7.4 Biplanar cytosine quadruplex (CDN8) Li\(^+\) co-ordination

3.7.4.1 RMSD (S63)

There is a sharp rising trend between t = 0 ns and t = ~70 ns from ~1.5 Å to ~3.0 ns, after which the trajectory is dynamic around ~2.8 Å to 500 ns. A noticeable drop in RMSD occurs between t = ~220 ns and t = ~350 ns.

3.7.4.2 eRMSD (S64)

The trajectory is dynamic with a steady rising trend after t = ~100 ns. The range of eRMSD values is similar to the K\(^+\) and Na\(^+\) co-ordinated models.
3.7.4.3 Rg (S65)

Rg values centre on ~0.8 Å for the duration of the simulation with very little variation. There may be a minor expansion and reduction cycle.

3.7.4.4 H-bond formation (S66, S67)

There appear to be local maxima at four, five and six H-bonds (S66) with all values predicted to be less than ten (S67).

Executive summary of sections 3.6 and 3.7

- Cytosine is known to dimerise
- Two dimers are shown in this study to have the potential for self-associating non-covalently into a planar quartet
- Given a specific sequence of bases, biplanar and triplanar quadruplexes could develop from cytosine quartets
- Stability of such supramolecular structures depends on several factors and varies widely as a consequence
- Cytosines dimers participating in the central core of a theoretical quadruplex are mutually ligated with two hydrogen bonds but paired dimers have only one, a total of six per quartet. Guanine equivalents have two hydrogen bonds between each base, a total of eight per quartet.
- Cationic co-ordination marginally reduces stability in triplanar models, but inclusion of a cation in biplanar systems is very disruptive.

3.8 Occupancy of waters in cytosine quadruplexes CDN12 and CDN8

3.8.1 Introduction and methodology

A comparison of interplanar water occupancy was made with biplanar and triplanar DNA guanine quadruplexes. Script 5 (Section 1.6.6.4) was accessed to calculate the number of waters likely to occupy the interplanar region of biplanar and triplanar cytosine quadruplexes. For a closer comparison with triplanar models, biplanar equivalents were analysed with co-ordinating ions K⁺ Na⁺ or Li⁺ imposed into the systems in addition to the archetype model exclusive of ions. CDN12 was assessed with K⁺ Na⁺ or Li⁺ co-ordination to reflect its derivation from 1KF1. No residues were methylated in these analyses.

Empirical values (S68-S70) and histograms (figures 3.7 and 3.8) were plotted for each system.
Figure 3.7: Histogram of interplanar water occupancy in triplanar cytosine quadruplex models co-ordinated by K\(^+\), Na\(^+\) or Li\(^+\). Time-dependent data for this histogram are plotted in S68, S69 and S70.

Water occupancy for K\(^+\) co-ordinated models is higher overall than for the Na\(^+\) and Li\(^+\) systems, but all three indicate an average population between four and six waters. The histogram implies higher numbers of waters are slightly better tolerated than for K\(^+\) and Na\(^+\) models. This finding compares with a population approaching zero in Li\(^+\) co-ordinated triplanar cytosine quadruplexes. Figures S68, S69 and S70 reinforce this finding: time-dependent data show Li\(^+\) systems have the most consistent water occupancies of the three models.
Figure 3.8: Histogram of interplanar water occupancy in biplanar cytosine quadruplex models with no cationic co-ordination or co-ordinated by K⁺, Na⁺ or Li⁺. Time-dependent data for this histogram are plotted in S71, S72, S73 and S74.

Biplanar DNA systems with monovalent cations artificially imposed contrast with triplanar models where co-ordinants are typically present. A higher population of waters is generally tolerated better in biplanar systems without cationic co-ordination than models with co-ordinants. This observation is particularly relevant for higher occupancies. Li⁺ co-ordinated models are predicted to consistently allow a slightly higher population of waters than K⁺ or Na⁺ systems. Time dependent data (figures S72, S73 and S74) indicate a wider variation of numbers in cationic co-ordinated models.

3.9 Ribose conformations of biplanar and triplanar models

3.9.1 Introduction and methodology
C2’-endo and C3’-endo conformations (figure 2.35) of cytosine (and thymine) models were analysed for comparison with guanine and adenine systems (Section 2.7). The C3’-endo pucker produces a significantly shorter phosphate-phosphate distance in the backbone, and consequently a more compact helical conformation. Given cytosine (and thymine) models are more compact, a comparison may demonstrate the extent of this effect.
Data from molecular dynamics simulations for biplanar DNA cytosine (CDN8 and CDN12) quadruplexes. Although the former is characterised by absence of a co-ordinant, CDN8 models also had K⁺, Na⁺ or Li⁺ imposed as co-ordinating ions. Triplanar systems were co-ordinated by K⁺, Na⁺ or Li⁺. C1'-C2'-C3'-C4' dihedrals for the sugar ring of residue 7 were measured as described in section 2.7.2 for atoms numbered 156 155 152 151 in the .pdb file.

3.9.2 Results

Trajectories of triplanar DNA cytosine quadruplexes ribose C2' / C3' fluctuations are typified by data shown for residue seven with different co-ordinating ions (figures 3.9a, 3.9b and 3.9c):

**Figures 3.9a, 3.9b and 3.9c: trajectories for C1'-C2'-C3'-C4' dihedrals in DNA triplanar cytosine models in residue 7 co-ordinated with K⁺ (left), Na⁺ (centre) and Li⁺ (right).**

The majority of systems appear to favour the C2'-endo configuration, but all show notable dynamic activity. Models with Na⁺ co-ordination typically the most erratic trajectories with poor definition. Li⁺ co-ordinated systems are also dynamic but generally show clear changes between C2'-endo and C3'-endo configurations. K⁺ models appear to show brief periods of stable C2'-endo forms alternating with highly erratic and poorly-defined trajectories. Biplanar quadruplexes characterised by absence of a co-ordinant consistently show trajectories in the C3'-endo region.
Table 3.1: fluctuations in ribose configurations for triplanar cytosine quadruplexes with K⁺ co-ordina- 
tion at t = 0ns (left) and t =500ns (right). Where no definite conformation could be 
identified, the structure is listed as planar. Figures S77 and S78 show data for models co-ordinated 
with Na⁺ and Li⁺.

| C1'-endo C2'-endo C2'-exo C3'-endo C3'-exo C4'-endo C4'-exo planar |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| t = 0ns         | t = 500ns      |
| 11 | 4 | 1 | 9 | 5 | 8 | planar |
| 16 | 6 | 2 | 10 | 17 | 1 |
| 22 | 7 | 3 | 18 | 19 | |
| 21 | 13 | 14 | 15 | 20 | 21 |

However, models with K⁺, Na⁺ or Li⁺ co-ordination are notably erratic with poor definition 
for either configuration (figures 3.10a, 3.10b, S75 and S76):

Figures 3.10a and 3.10b, trajectories for C1’-C2’-C3’-C4’ dihedrals in DNA biplanar cytosine 
models in residue 7 with no ionic co-ordination (left) or co-ordinated with K⁺ (right). Trajectories 
for systems co-ordinated with Na⁺ or Li⁺ are shown in figures S75 and S76.
**Figure 3.11:** distribution histograms from figures 3.10a and 3.10b data for CDN8-N1 no ionic co-ordination (blue) and CDN12-N1 K⁺ co-ordination dihedrals (blue). C2'-endo and C3'-endo configurations are evenly represented for the duration of the simulation for both models. The wide range of conformations indicate sugars in each system fluctuate frequently during the simulation.

**Table 3.2:** fluctuations in ribose configurations for biplanar cytosine quadruplexes without ionic co-ordination (black) and with K⁺ co-ordination (red). Conformations at the simulation start were identical for both models. Figures S79 and S80 show data for models co-ordinated with Na⁺ and Li⁺.

<table>
<thead>
<tr>
<th></th>
<th>CDN8 no ionic co-ordination all residues</th>
<th>CDN8 K⁺ co-ordination all residues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t = 0ns</td>
<td>t = 500ns</td>
</tr>
<tr>
<td></td>
<td>C1'-endo C2'-endo C3'-endo C3'-endo C3'-endo O4'-endo O4'-endo O4'-endo planar</td>
<td>C1'-endo C2'-endo C3'-endo C3'-endo C3'-endo O4'-endo O4'-endo O4'-endo planar</td>
</tr>
<tr>
<td></td>
<td>1 5 15 19 20 4 6 2 7</td>
<td>1 4 3 7 10 12 6 1 8 5 11 13 15 17 4 6 15 8 9 5 11 10 11 5 7 5 18</td>
</tr>
<tr>
<td></td>
<td>15 15 19 20 10 11 12 13 14 18 21 22</td>
<td>22 23 9 12 16 14 12 14 18 19 20 21</td>
</tr>
</tbody>
</table>
3.10 Discussion

Although pyrimidines favour the C3'-endo form, the C2'-endo configuration is dominant in DNA systems. The prevalence in RNA and C3'-endo conformers is contingent on the marginally greater interatomic distance between C2' and C3' oxygens in ribonucleosides.

The concept of cytosine dimers pairing into a quartet, then three quartets stacking vertically into a cytosine quadruplex (C4), was addressed in this study. The quartets were linked with AAT loops, and physiologically relevant K⁺, Na⁺ or Li⁺ were selected as co-ordinating ions. Contrasting with guanine quadruplexes co-ordinated with Li⁺ as contentious (19) (81), a biplanar cytosine equivalent structure with no cationic co-ordination is found to be non-transient here. However, the greater dynamic activity predicted for Li⁺, K⁺ or Na⁺ co-ordinated biplanar models indicates they may be transient. All triplanar systems are characterised by greater dynamic activity and therefore strongly suggestive of reduced stability. The results indicate the stability of cytosine quadruplexes varies not only between triplanar and biplanar structures, but also the presence or absence of a co-ordinating ion. For triplanar systems, cations remain within the complex and underpin quadruplex stability. This observation contrasts notably with dynamic activity within biplanar systems where presence of co-ordinating cation reduces overall integrity of a complex.

RMSD and eRMSD data for triplanar models coincide with a sequence of higher values and more erratic trajectories in the sequence Li⁺ > K⁺ > Na⁺. Less dynamic trajectories are evidenced in biplanar models, but with lower deviation. However, all biplanar models with cationic co-ordination are notably more dynamic than a system without an ion. Radius of gyration data generally confirm this trend with some exceptions. In a triplanar system, Li⁺ and K⁺ have similar compactness and which is less than for Na⁺-co-ordinated models. However, all biplanar models with cationic co-ordination consistently show greater expansion than for a system with no cation.

Frequency of H-bond formation appears to be generally consistent amongst all models whilst noting that biplanar systems have lower values overall. This would accord with only two cytosine planes available for interaction as opposed to three in triplanar quadruplexes. As a general observation, frequency of H-bond formation in Na⁺-co-ordinated models is slightly higher than for other systems including a biplanar quadruplex without cationic co-ordination. This may be an artefact of hydrated sodium ion clusters which arise more readily than other Group 1 metals (161). If such clusters were to arise in Na⁺-co-ordinated models even transiently, H-bonding may occur at a slightly higher frequency overall.

For triplanar quadruplexes, interplanar occupancy of waters varies for K⁺ and Na⁺-co-ordinated models more than for Li⁺ system. Whilst this observation may underpin some of the enhanced frequency of H-bond formation in Na⁺-co-ordinated models, it might
plausibly apply to $K^+$ systems. This trend is not displaced by water occupancy being lower in biplanar models, although less available interplanar access must be a major factor, if not solely responsible.

Although the data for ribose configuration are generally inconsistent for all models, some trends are noted. Trajectories for triplanar Na$^+$-co-ordinated systems have no discernable patterns and this observation applies to both $K^+$ and Na$^+$-co-ordinated biplanar quadruplexes. The trajectory for a biplanar system with no cationic co-ordination is an exception to these observations. The configuration is calculated as consistently C3'-endo. There may be some periodicity to a triplanar K$^+$-co-ordinated system and an Li$^+$-co-ordinated models generally show a shift from C2'-endo C3'-endo at an approximate mid-point of the simulation.

Overall, these results imply a biplanar structure is stable to 500ns in the absence of cationic co-ordination, but models co-ordinated by Li$^+$, $K^+$ or Na$^+$ may be less stable. Triplanar cytosine quadruplexes retain their co-ordinating ions but generally exhibit more erratic trajectories across most metrics. Chapter 7 advances the issue that an antisense cytosine quadruplex may compromise the beneficial effects of a complementary sense-strand guanine equivalent. The novel suggestion of an anti-sense guanine quadruplex is also raised.

If biplanar cytosine quadruplexes do fold in single-strand DNA, their viability is credible but may be less stable than triplanar guanine systems. The latter are characterised by three H-bonds between bases contrasting with only two in cytosine-cytosine base pairing. Cationic co-ordination is also a factor for both species of theoretical cytosine quadruplexes: paradoxically, absence stabilises biplanar models, but presence reduces overall integrity of triplanar systems. However, two points arise: both show a robust network of inter-planar π-stacking. Secondly, although there is an H-bond differential between guanine and cytosine models, geometry may influence the formation. The combination may be tending towards a lower energy configuration.

Thymine substitution in quadruplex structures may further qualify the data for guanine and cytosine systems.

### 3.11 Thymine quadruplexes: Introduction

Having validated *in silico* the potential of cytosine to initially dimerise and then further self-associate into quartets, a rationale is established for investigating similar behaviour in thymine models. A theoretical quadruplex constituted by a central core of thymine planes might provide a useful comparison with guanine and cytosine equivalents. Whist thymine quartets have been identified by single crystal X-ray diffraction located within
supramolecular structures, a thymine quadruplex per se has not apparently been characterised (152).

3.11.1 Methods

As with modelling of cytosine quadruplexes, 1KF1 acted as scaffold for triplanar systems and 2KF8 for biplanar. Thymine was substituted for all guanines in the quartets using the \textit{swapna} command available in Chimera command line. Interplanar loops remained unaltered and nine models constructed. Triplanar models retained K$^+$ ions co-ordination by virtue of their 1KF1 derivation. The ions were replaced in turn with Li$^+$ or Na$^+$ by manual reformat of the psf and pdb files. In addition to a biplanar system containing no cation, K$^+$, Na$^+$ or Li$^+$ ions were imposed on cytosine models by the same method. Thymine quadruplexes were transferred to DSV for confirmation of structural changes and to note modified residue numbers. Monitoring for K$^+$ and Na$^+$ in triplanar systems confirmed co-ordinating ions have octagonal co-ordination, but a tetrahedral geometry was noted for Li$^+$ models. Triplanar quadruplexes are referred to as TDN12 and biplanar as TDN8.

3.11.2 Results

3.11.3.1 Triplanar thymine quadruplex (TDN12) K$^+$ co-ordination

3.11.3.2 RMSD (figure 3.11): the system is noticeably dynamic with a steady rising trend and collapses at ~320ns. A very steep increase in RMSD is noted just prior to loss of integrity.

\textbf{Figure 3.11}: RMSD trajectory for triplanar thymine quadruplex co-ordinated with K$^+$. 

90
3.11.3.3 eRMSD (S81)

The trajectory shows a steep rising trend to ~100 ns prior to levelling at ~2.2 Å before completion at ~320 ns. This is indicative of the model losing integrity.

3.11.3.4 Rg (S82)

Although the trajectory is approximately level varying between 1.0 Å and ~1.3 Å, it is noticeably dynamic. There may be some periodicity between compactness / relaxation states but the trajectory completes at ~320 ns.

3.11.3.5 H-bond formation (S83, S84)

H-bond frequency is consistent average of three, but varies between zero and ten prior to completion at ~320 ns.

3.11.4.1 Triplanar thymine quadruplex (TDN12) Na+ co-ordination

3.11.4.2 RMSD (S85)

A steady rising trend is observed in a very dynamic trajectory which completes at ~430 ns. Values range between ~3.0 Å and ~11.0 Å.

3.11.4.3 eRMSD (S86)

A brief rapid rise in eRMSD occurs to ~40 ns before a steady but less steep increase to ~2.5 Å, ~430 ns, at which point the trajectory completes.

3.11.4.4 Rg (S87)

An early loss of compactness occurs from simulation start to ~60 ns, after which the trajectory is very active whilst maintaining an approximately level course between 0.8 Å and ~1.3 Å. Some return to a more compact state is implied before a very steep loss immediately prior the end of the simulation at ~420 ns.

3.11.4.5 H-bond formation (S88, S89)

The plot indicates an average frequency of three H-bonds but ranges between zero and eight before completion at ~420 ns.

3.11.5.1 Triplanar thymine quadruplex (TDN12) Li+ co-ordination

3.11.5.2 RMSD (S90)
The trajectory shows an overall slight rising trend from ~0.3 Å to ~5.0 Å at 100 ns after which the system is noticeably dynamic. A sudden brief rise to ~7.0 Å occurs before maintaining a level course before completion at ~430 ns.

3.11.5.3 eRMSD (S91)

The trajectory is relatively steady showing a slight rising trend with little dynamic activity before completing at ~450 ns.

3.11.5.4 Rg (S92)

Minor changes in compactness and relaxation are indicated but the trajectory is approximately level with eRMSD values ranging between 1.0 Å and ~1.3 Å before completing at ~430 ns.

3.11.5.5 H-bond formation (S93, S94)

Frequency of H-bond formation averages two but ranges from zero to seven before completing the simulation at ~420 ns.

3.11.6.1 Biplanar thymine quadruplex (TDN8) No ion

3.11.6.2 RMSD (S95)

A very slight rising trend is observed from ~5.0 Å to ~6.0 Å at 500 ns with a brief period of minor activity between ~300 ns and ~350 ns.

3.11.6.3 eRMSD (S96)

Little dynamic activity is shown by the trajectory with eRMSD varying slightly between ~1.8 Å and ~2.1 Å.

3.11.6.4 Rg (S97)

The trajectory shows minor alternation between compactness and relaxation for the duration of the simulation. The plot is approximately level with values ranging between ~0.8 Å and ~1.2 Å.

3.11.6.5 H-bond formation (S98, S99)

Frequency of H-bond formation appears to be relatively consistent and averages two. Values ranges between zero and six.

3.11.7.1 Biplanar thymine quadruplex (TDN8) K⁺ co-ordination
3.11.7.2 RMSD (S100)

The trajectory is level to ~100 ns at ~3.0 Å after which there is an overall rising trend to ~8.0 Å at 500 ns. Minor local dynamic activity occurs between 100 ns and the simulation end.

3.11.7.3 eRMSD (S101)

The trajectory is steady at ~1.5 Å to ~120 ns at which point a rising curve completes at 2.6 Å, 500 ns. However, the system shows limited local dynamic activity.

3.11.7.4 Rg (S102)

A sharp rise in relaxation occurs at ~50 ns from ~0.7 Å to ~1.3 Å. A less dynamic trend continues to ~250 ns at which point, there is a brief loss of relaxation before a slower return to a less compacted state.

3.11.7.5 H bond formation (S103, S104)

Frequency of H-bond formation is relatively inconsistent and an average difficult to determine by inspection but the histogram indicates six. Two brief increases occur at ~200 ns and 400 ns.

3.11.8.1 Biplanar thymine quadruplex (TDN8) Na+ co-ordination

3.11.8.2 RMSD (S105)

The system is dynamic with a consistent rising trend from ~5.0 Å to ~10.0 Å at 500 ns.

3.11.8.3 eRMSD (S106)

A steep rise from ~1.4 Å to ~2.5 Å levels at ~100 ns after which the trajectory maintains an approximately level course with noticeable dynamic activity.

3.11.8.4 Rg (S107)

The trajectory is noticeably dynamic and shows a steady overall loss of compactness. A brief sudden loss of compactness occurs at ~200 ns. Values vary between ~0.7 Å and ~1.4 Å.

3.11.8.5 H-bond formation (S108, S109)

The frequency plot is inconsistent with values ranging between zero and five. The histogram indicates an average of five.
3.11.9.1 Biplanar thymine quadruplex (TDN8) Li⁺ co-ordination

3.11.9.2 RMSD (S110)

The trajectory is approximately level for the entire simulation but shows consistent local dynamic activity. Values range between ~6.0 Å and ~9.0 Å.

3.11.9.3 eRMSD (S111)

The trajectory maintains a steady rising trend from ~1.5 Å to ~2.3 Å at completion. Dynamic activity appears consistent with K⁺ and Na⁺ systems.

3.11.9.4 Rg (S112)

The system is very dynamic with a slight rising trend from ~0.5 Å to ~1.3 Å. A brief sudden loss of compactness occurs at ~280 ns.

3.11.9.5 H-bond formation (S113, S114)

Frequency of H-bonding formation is consistent and averages three. Values range between zero and seven.

3.12 Occupancy of waters in thymine quadruplexes TDN12 and TDN8

3.12.1 Introduction and methodology

Interplanar water occupancy in biplanar and triplanar DNA thymine quadruplexes was assessed with Script 5 (Section 1.6.6.4). For a closer comparison with triplanar models, biplanar equivalents were analysed with co-ordinating ions K⁺ Na⁺ or Li⁺ imposed into the systems in addition to the archetype model exclusive of ions. TDN12 was assessed with K⁺ Na⁺ or Li⁺ co-ordination. No residues were methylated in these analyses. Empirical values (S115-S117) and histograms (figures 3.12 and 3.13) were plotted for each system.
3.12.2 Results

**Figure 3.12:** Histogram of interplanar water occupancy in triplanar thymine quadruplex models co-ordinated by K⁺, Na⁺ or Li⁺. Time-dependent data for this histogram are plotted in S115, S116 and S117.

Water occupancy for K⁺ co-ordinated models is generally slightly higher than for Na⁺ and Li⁺ systems, but all three indicate a population averaging between six and seven waters. The histogram implies minimal differences between all systems. Compared with triplanar cytosine quadruplexes, lower populations of waters are better tolerated in thymine equivalents. This observation is reinforced by the prediction of a zero population are predicted to be more likely. Overall, the general trend is lower than for cytosine quadruplexes. Biplanar thymine quadruplexes have similar characteristics (figure 3.13):
Figure 3.13: Histogram of interplanar water occupancy in biplanar thymine quadruplex models with no cationic co-ordination or co-ordinated by K⁺, Na⁺ or Li⁺. Time-dependent data for this histogram are plotted in S118, S119, S120 and S121.

Absence of a co-ordinating ion in thymine models generally sanctions a slightly increased population of waters than models with co-ordinating ions. Waters in Na⁺ systems are marginally better tolerated than for other models which resonates with a similar observation of cytosine quadruplexes.

3.13 Ribose conformations of biplanar and triplanar models

3.13.1 Introduction and methodology

C2’-endo and C3’-endo conformations (Section 2.7.1 figure 2.35) of biplanar and triplanar thymine models were analysed for comparison with guanine and adenine systems (Section 2.7). Given a C3’-endo pucker produces a significantly shorter inter-phosphate backbone distance, a more compact helical conformation. Given cytosine and thymine models are more compact, a comparison may demonstrate the extent of this effect.

Data from molecular dynamics simulations for biplanar DNA thymine (TDN8 and TDN12) quadruplexes. Although the former is characterised by absence of a co-ordinant, TDN8 models also had K⁺, Na⁺ or Li⁺ imposed as co-ordinating ions. Triplanar systems were co-ordinated by K⁺, Na⁺ or Li⁺. C1’-C2’-C3’-C4’ dihedrals for the sugar ring of residue 7 were measured for atoms numbered 156 155 152 151 in the .pdb file (section 2.7.2)
3.13.2 Results

Triplanar thymine quadruplexes for K⁺, Na⁺ and Li⁺ models lose integrity before the end of a simulation configured for 500ns. Moreover, their trajectories show the glycoside dihedrals are dynamic (figures 3.14a, 3.14b and 3.14c). A possible trend for C2’-endo configuration is noted for K⁺ and Na⁺ systems before an uncertain interpretation followed by structural collapse. Li⁺ models also fluctuate significantly between a possible C2’-endo configuration and an O4’-endo intermediate.

Figures 3.14a, 3.14b and 3.14c: trajectories for C1’-C2’-C3’-C4’ dihedrals in DNA triplanar thymine models in residue 7 co-ordinated with K⁺ (left), Na⁺ (centre) and Li⁺ (right).

Concatening all ribose configurations for K⁺ models indicates a reduction of C2’-endo sugars accompanied by a marginal increase in C3’-endo configuration. However, there is also an increase of indeterminant structures (itemised as planar in Table 2.9):

Table 3.3: fluctuations in ribose configurations for triplanar thymine quadruplexes with K⁺ co-ordination at t = 0 ns (left) and t ~320 ns (right). Figures S122 and S123 show data for models co-ordinated with Na⁺ and Li⁺.
Based on analysis of a single representative residue biplanar thymine quadruplex, a model with no cationic co-ordination system shows a clear trend around C2'-endo form (figure 3.15a). However, trajectories for systems in the presence of an ion fluctuate widely (figure 3.15b and figures S124 and S125).

Figures 3.15a and 3.15b, trajectories for C1'-C2'-C3'-C4' dihedrals in DNA biplanar thymine models for residue 7 with no ionic co-ordination (left) or co-ordinated with K⁺ (right). Trajectories for systems co-ordinated with Na⁺ or Li⁺ are shown in figures S124 and S125.

Figure 3.16: distribution histograms from figures 3.15a and 3.15b data for TRN8-N1 no ionic co-ordination (pink) and TDN12-N1 K⁺ co-ordination dihedrals (blue). C2'-endo and C3'-endo configurations are evenly represented for the duration of the simulation for both models. The range of conformations for both species indicates sugars in each system fluctuate frequently during the simulation.
Concatenating ribose configurations for all residues in K⁺ biplanar models indicates disruption derived from inclusion of the ion is similar to cytosine systems (Table 2.10). There is no preference for any particular configuration. Whilst there are minor re-configurations for the ‘no ion’ analysis, the C3’-endo form remains prevalent.

**Table 3.4**: fluctuations in ribose configurations for biplanar thymine quadruplexes without ionic co-ordination (black) and with K⁺ co-ordination (red). Conformations at the simulation start were identical for both models. Figures S126 and S127 show data for models co-ordinated with Na⁺ and Li⁺.

<table>
<thead>
<tr>
<th>TDNS no ion dihedrals all residues</th>
<th>TDNS K⁺ dihedrals all residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo  C2'-endo  C2'-exo  C3'-endo  O4'-endo  O4'-exo planar</td>
<td>C1'-endo  C2'-endo  C2'-exo  C3'-endo  O4'-endo  O4'-exo planar</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
</tr>
</tbody>
</table>

3.14 Executive summary of results for thymine quadruplexes

- All metrics in this study indicate thymine quadruplexes derived from similar methodology to cytosine models are less stable. None for either species are non-transient however.
- Some enhancement of structural integrity for cytosine and thymine systems may derive from a lowest energy geometry and noticeable interplanar π-bonding networks.

3.15 Discussion of thymine quadruplexes

Thymine quadruplexes *in silico* appear to be less stable into biologically relevant times than cytosine equivalents. In particular, triplanar systems collapse prior to the run time of 500ns written into Production of the Configuration File. Whilst thymine and cytosine systems are both characterised by two H-bonds, interpretation of the lower stability for triplanar thymine models is difficult. Some contribution may derive from minor differences in inter-base distances: 3.81 Å (cytosine) and 3.84 Å (thymine). The software also predicts interplanar π-σ interactions for cytosine quadruplexes in addition to π-π stacking for both systems. It is worth noting that the melting temperatures using the
methodology in Section 2.4 calculate 36.9 °C and 57.42 °C for thymine and cytosine single strands respectively with the potential to fold into quadruplexes.

RMSD and eRMSD data for triplanar models coincide less than was noted for cytosine equivalents: all trajectories are erratic before systems lose structural integrity. Dynamic trajectories are also more in evidence for biplanar models particularly those with cations imposed. Radius of gyration data reinforce this trend as any underlying patterns are difficult to discern. All biplanar models with cationic co-ordination show greater expansion overall than for a system with no cation.

Compared with cytosine models, interplanar occupancy of waters shows a lower rate in thymine triplanar and biplanar systems but an overall similar trend. Although triplanar quadruplexes disintegrate before molecular dynamics completion, frequency of H-bonding does not appear to have the minor discrepancy relative to cytosine equivalents. Given this is noted to a small degree in biplanar models, the observation may simply reflect the collapse of triplanar thymine systems. This incomplete result may also underpin the unexpected prevalence of the C3'-endo configuration. Moreover, the incomplete simulation may reflect the population of planar orientation of sugars which implies a higher level of fluctuation. A very dynamic system complements its inherent instability and consequent indeterminant nature of configurations. There are no discernible patterns such as periodicity in theoretical thymine quadruplexes.

If biplanar thymine quadruplexes were to fold in single-strand DNA, their viability is only credible if no cations co-ordinate the system. Even so, such complexes appear to be relatively transient in a biological context. Molecular dynamics analysis here demonstrates a triplanar guanine quadruplex is considerably more stable than cytosine or thymine equivalents. However, biplanar models for guanine, cytosine and thymine species appear to be relatively non-transient in the absence of cationic co-ordination.

3.16 Future work

Ostensibly, cytosine or thymine quadruplexes may have limited physiological relevance predicated on relatively poor stability compared to guanine equivalents. If these structures were to fold genomically, perhaps their significance lies with an intermittent appearance. Although short-lived, a folding / unfolding cycle may persist into biologically relevant times and influence epigenetic processes. Such mechanisms could be amenable to molecular modelling and thermodynamic analysis. Methylation at C5 of thymine, and C5 or C6 of cytosine would probably further reduce the molecular lifetime of a particular quadruplex. However, methylation may also tag the structure for downstream modification: docking protocols with small molecules or enzymes might adumbrate further perspectives.
Chapter 4: Adenine Quadruplexes

Adenine quadruplexes

4.1 Introduction

Molecular dynamics analysis of experimentally validated guanine quadruplexes and theoretical cytosine and thymine equivalents in this thesis has determined a trend toward decreasing stability. These complexes are characterised inter alia by molecular differences in hydrogen bonding. Structural integrity in guanine systems is associated with a higher population of inter-planar waters and monovalent cationic co-ordination underpinning a robust network of H-bonding and π-stacking. Cytosine and thymine models are less well represented by such networks. Guanines associate non-canonically in planar structures by virtue of the Hoogsteeen interface. This feature is absent in cytosine and thymine equivalents. Inspection of adenines potentially self-associating clearly shows a Hoogsteeen interface is also absent. Ostensibly implicit in this observation is that adenines folding into quartets would suggest stability comparable with cytosine and thymine equivalents.

4.2 Adenine quartets

Although there may be more, three adenine quartets have been validated experimentally to date in RNA (59). A guanine- and adenine-rich crystal was found to contain an adenine quartet within a parent quadruplex which did not directly confer conformational change on the complex (162) However, the hydrogen bonding network was considered less regular at a global level. Subsequently, variant quartets were identified and similarly characterised by only four cyclic hydrogen bonds with N6 as a donor. The variants differed in acceptor sites: N1, N3 and N7. Accordingly, the structures were named $A_4$–N1, $A_4$–N3 and $A_4$–N7 (figures 4.1, 4.2 and 4.3):
Figure 4.1 planar representation of an experimentally validated adenine quartet with H-bonding between N1 and N6 (H1) of an adjacent amino group. Amino H2 hydrogens are centrally facing.

Figure 4.2 planar representation of an experimentally validated adenine quartet with H-bonding between N3 and N6 (H1) of an adjacent amino group. Amino H2 hydrogens are arranged on the periphery.
A₄–N1 has been resolved by NMR spectroscopy (163) (164); A₄–N3, and A₄–N7 have been detected by X-ray crystallography (162) (165). Computational studies have further evaluated A₄–N1, A₄–N3, and A₄–N7 models (165) (166) (167). Whilst guanine quartets (and by extension, quadruplexes) are stabilised by metal cations located centrally, adenine equivalents may be evidenced by other perspectives such as peripheral ligation (162) (168). The A₄–N3 model maintains propitious interaction between cation and the lone pair of an adjacent N1. This rationale behind this specific feature has also been noted in computational research (169). A₄–N1 and A₄–N7 quartets have exocyclic amino groups directed centrally which suggests hydrogen binding to an anion would be plausible. The observation of anionic hydrogen bonding in DNA and RNA nucleobases reinforces this notion (170). Given one hydrogen of N6 in the A₄–N3 model is external to the quartet, peripheral anionic binding is also a consideration for this system. Computational studies confirm positively charged groups may interact favourably with supramolecular nucleobase complexes (171) (172) (173). The absence of H-bond donors facing towards the central region would suggest co-ordinating ions or moieties are unlikely to bind in the inter-nucleobase space. The quartets contrast with guanine equivalents by a preference in an isolated state for non-planar geometry as has been observed in pyrimidine quartets (174). Concatenation of A₄–N1 and A₄–N7 models indicates global minima geometries are hyperbolic paraboloid or concave semi-lenticular. A₄–N3 systems are rectilinear cuboid.

A singular tautomeric imino adenine quartet characterised by N6 - N6 and N1 - N7 H-bonding has been identified (175) (Images, Supplementary figure S4.1). The model was not taken forward into molecular dynamics simulation here as the structure is considered
to occur infrequently. Although the assembly may ostensibly be biologically irrelevant, any mutation to canonical base pairing should be noted.

4.3 Adenine-guanine hybrid quartets: Introduction

As noted in section 4.2, guanines and adenines in stochastic proportion may well assemble into a hybrid quadruplex. Although investigating such a structure was initially the province of theoretical study, experimental research confirms a single adenine bonded non-covalently between two guanines in a four-purine quartet (176) (177). Biological relevance of the complex is evidenced as its functional contribution to repression of a human growth factor. Two adenine substitutions of the telomeric sequence (Chapter 1, section 1.1) exemplified by 5’ d(AGGTTA)n could fold into a quadruplex with a pair of G3A quartets intercalated between classical guanine equivalents. Adenine’s capacity for π-π stacking may enhance the system’s stability. Ab initio calculations have shown that guanine-adenine dimerisation requires minimal energy to form a co-planar structure (178). However, a pair of such dimers associating into a quartet (referred to as AGAG) with non-planarity would have little biological relevance because of this unfavourable geometry (179). Co-ordination of the hybrid quartet with Li⁺ appears to restore planarity to an extent. K⁺ and Na⁺ as co-ordinants do not resolve the issue (176). Nevertheless, research continues to identify stable substitutions in guanine quadruplexes by other molecules such as xanthine (180), cytosine (181) and adenine (182).

Theoretical work concludes that the majority of hybrids comprising one or two adenines substituted for guanine in a discreet quartet located between a pair of guanine quartets have little merit biologically. The work does not apparently address a hybrid with a quartet constituted uniquely by adenine interacted between two classical guanine quartets.

4.4 Materials and Methods

(PDB) 1KF1 was opened in Chimera and guanine bases 3,9,15 and 21 were replaced by adenine using the swapna command. K⁺ ions were retained. Consequently, the system comprised an adenine quartet located centrally between two guanine equivalents. A sequence such as that shown in Table 4.1 would fold to a guanine-adenine-guanine system:
Table 4.1: sequence for triplanar guanine-adenine-guanine hybrid complex.

\[ \text{...TTAGGGTATTGGGGTTAGGGTTAGG...} \]
\[ \text{...TTAGAGTTAGAGTTAGAGTTAGA...} \]

This sequence is the shortest possible; bases could extend it by interpolation in the TTA regions or substitution of any guanine adjacent to an adenine depicted in green. Two further structures were constructed as Na\(^+\) and Li\(^+\) were substituted for K\(^+\) as co-ordinating ions. The archetype model was labelled as GAG12.

Each complex was transferred to DSV for optimisation of geometry and lowest energies before submission to NAMD. Simulations were run for 500 ns and subsequently analysed by RMSD, eRMSD, Radius of gyration and frequency of hydrogen bond formation. Figures 4.4 and 4.5 show predicted electronic interactions between bases and co-ordinating ions prior to molecular dynamics simulation:

**Figures 4.4. Interactions between bases and co-ordinating ions for a guanine-adenine hybrid quadruplex GAG12.** Ostensibly, there is more interaction for a K\(^+\) model (A) than for Li\(^+\) (B). however, red dotted lines indicate unfavourable bump interactions which strongly imply K\(^+\) as co-ordinant is likely to disrupt the system by virtue of steric hindrance. The Li\(^+\) model has cation-donor bonds (also present in the K\(^+\) complex) but no unfavourable interactions are predicted. Interactions for the Na\(^+\) system were identical to the K\(^+\) complex. The interactions are tetrahedral for all models.

Both models are well-served by \(\pi-\pi\) stacking as indicated by mauve dashed lines in figure 4.4. This factor, plus four cation-donor bonds, indicates inherent stability. The presence of unfavourable interactions in the K\(^+\) model suggests some disruption would be likely. Bonds are formed between each guanine O6; there is none with any of the four adenines.
in the central plane. Absence of this feature for the Li\(^+\) complex implies it would be more stable. The tetrahedral co-ordination in this model is clearly visible in figure 4.5:

![Figure 4.5: tetrahedral co-ordination of Li\(^+\) in GAG12. The π-bonding network is depicted in mauve.](image)

There is a marginal loss of co-planarity in one quartet with a tendency towards semi-lenticular concave. 1KF1 is characterised by this geometry in both outer quartets. The guanine model is also slightly less well-populated by π-bonding (Images, Supplementary figure S4.2).

4.5 Results

4.5.1 Triplanar guanine quadruplex (GAG12) co-ordinated by K\(^+\).

4.5.1.1 RMSD (S4.3)

The trajectory shows marginal dynamic activity for the duration of the simulation with a steady rising trend. This reduces to an approximate level trajectory at ~300 ns but returns to a slight upward path at ~400 ns. The system retains structural integrity.

4.5.1.2 eRMSD (S4.4)

The data maintain a steady trajectory between ~2.0 Å and 2.5 Å with no appreciable dynamic activity. There is a slight reduction in eRMSD after ~150 ns which continues to the end of the simulation.
4.5.1.3 Radius of gyration (S4.5)

The model undergoes a steady relaxation of compactness for the duration of the simulation with values ranging between ~0.5 Å and 1.5 Å.

4.5.1.4 Frequency of hydrogen bond formation (S4.6, S4.7)

The hydrogen bond network appears robust and well-represented with four to eight bonds for most of the simulation. Four bonds are predicted to be almost permanently present.

4.5.2 Triplanar guanine quadruplex (GAG12) co-ordinated by Na⁺.

4.5.2.1 RMSD (S4.8)

During the minimisation cycle there is a steep rise in RMSD from 2.0 Å to ~4.2 Å at ~120 ns after which a level course is maintained. The trajectory continues to be dynamic with ranging between 2.5 Å and 3.5 Å.

4.5.2.2 eRMSD (S4.9)

There is a slight rising trend in the trajectory from ~2.0 Å to ~2.8 Å with minor dynamic activity. A very brief peak occurs at ~230 ns before returning to the original trend.

4.5.2.3 Radius of gyration (S4.10)

The trajectory is level and steady to ~100 ns with minimal variation in compactness after which a rising trend is established to 500 ns. Values range between ~0.6 Å and ~1.3 Å.

4.5.2.4 Frequency of hydrogen bond formation (S4.11, S4.12)

The model has a good hydrogen bond network with five, six or seven as most numerous. Zero bonding are predicted with marginal regularity. There is a brief overall reduction centred on ~260 ns.

4.5.3 Triplanar guanine quadruplex (GAG12) co-ordinated by Li⁺.

4.5.3.1 RMSD (S4.13)

The trajectory shows a stable system with moderate dynamic activity. There is a brief steep rising trend from the simulation start to ~80 ns after which values range between 2.5 Å and 3.5 Å.

4.5.3.2 eRMSD (S4.14)

The system shows little dynamic activity and a steady level trajectory is maintained from ~120 ns to 500 ns with values varying between ~1.9 Å and 2.3 Å.
4.5.3.3 Radius of gyration (S4.15)

The system loses some compactness slowly between 0 ns and ~120 ns but returns to a level trajectory to ~350 ns after which there is a slight rising trend. The simulation is characterised by very little variation.

4.5.3.4 Frequency of hydrogen bond formation (S4.16, S4.17)

A strong network is shown being represented by an almost permanent five bonds predicted. Two brief reductions in overall population are centred on ~190 ns and ~430 ns.

4.6 Discussion of guanine-adenine hybrids

Research prior to this study has focussed on theoretical and experimentally validated hybrid quadruplexes. The central quartet of 1KF1 has either a single or double adenine substitution where the former was referred to as G3A. A double substitution with modifications located in opposition was identified an AGAG system. Both quadruplex models showed unfavourable geometry derived from loss of co-planarity with associated low relevance in a biological context. Co-ordination with Li⁺ restored planarity to some extent, but co-ordination with K⁺ or Na⁺ had little consequence. Moreover, a K⁺ system was characterised by unfavourable bump interactions implying steric hindrance and therefore inherent instability. Bases in adenine quartets self-associate via a single hydrogen bond as opposed to two in guanine equivalents. However, this factor in reducing molecular stability may be offset by the more effective π-π stacking of the former model rather than the latter. When co-ordinated by Li⁺, the more favourable co-planarity could be an energy dividend predicated on the smaller ionic radius and a limited prospect of steric hindrance.

4.7 Adenine quadruplexes: Overview

Unpaired adenosines are known to underpin an aspect of RNA folding where the bases dock with an adjacent helix (183). This implicit self-association of adenines applied in simpler circumstances would reinforce the notion of a quadruplex with a central core uniquely constructed of adenine quartets.

A simple 1KF1-derived quadruplex constructed here with all four guanines of the central quartet substituted by adenine, showed noticeably greater coplanarity than its guanine equivalent in a brief pilot molecular dynamics simulation. Accordingly, a series of triplanar models were constructed, initially in DNA, to assess adenine models in greater detail. The scale of investigation increased significantly with the onset of SARS-CoV-2. Originally
intended to explore the impact, if any, of guanine quadruplexes in the viral genome, the work developed to include RNA systems. Furthermore, because mRNA viruses are typified by adenine-rich regions, the possibility was raised of theoretical adenine quadruplexes also folding in the genome. Moreover, in addition to triplanar complexes in DNA, biplanar systems were constructed. Biplanar adenine models were also assembled in RNA. The range of analytical metrics was increased and other co-ordinating ions were included in all biplanar models. Methylation as a disruptive factor was assessed in DNA and RNA complexes derived from A₄-N1 quartets as the structural integrity of this species was proving least dynamic hitherto.

4.8 Methods

1KF1 was the scaffold for triplanar systems and 2KF8 for biplanar models. Adenine was substituted for all guanines in the quartets using the *swapna* command in Chimera. The original ...TTA... sequence was retained for the Inter-planar connecting loops. Seventeen models were constructed for triplanar and biplanar quadruplexes in Chimera (Table 1.1, Chapter2). Triplanar models in DNA were co-ordinated by K⁺, Na⁺ or Li⁺ to align with guanine, cytosine and thymine complexes. However, for adenine systems, the range of ionic co-ordination was extended to assess the structures in greater detail. Ions imposed were: NH₄⁺ and H₃O⁺ to assess potential non-metal cationic interaction, and the divalent metal cations Ca²⁺ and Mg²⁺. The modifications were made by a manual reformat of the .psf and .pdb files. Ca²⁺ and Mg²⁺ were included for their physiological relevance, charge and ionic radius. Furthermore, Mg²⁺ has a diagonal relationship with Li⁺ in the Periodic Table. In particular, charge density for ionic lithium is closer to Mg²⁺ than other Group 1 metals: Li⁺ = 1.47 C m⁻³, Mg²⁺ = C m⁻³ (K⁺ = C m⁻³, Na⁺ = C m⁻³). Consequently, Li⁺ can be a highly polarising ion and may ligate covalently as opposed to the ionic interactions of K⁺ and Na⁺ (154). Triplanar DNA models are referred to as ADN12 structures in this study.

Biplanar RNA adenine quadruplexes are referred to as ARN8 structures. Since the models were predicated on three analogues of experimentally validated quartets (A₄-N1, A₄-N3, and A₄-N7), -N1, -N3, and -N7 are attached as suffixes. ARN8-N1, ARN8-N3 and ARN8-N7 were also methylated at N6(H1). Contingent on possible relevance to the mRNA virus SARS-CoV-2, biplanar models were also constructed in RNA. In addition to the absence of a co-ordinating ion, K⁺, Na⁺, Li⁺, NH₄⁺, H₃O⁺, Ca²⁺ and Mg²⁺ were imposed on ARN8-N1.

Models were transferred to DSV for confirmation of structural changes and to note modified base numbers. Monitoring the geometry of triplanar systems confirmed co-ordinating ions have octagonal co-ordination for K⁺ and Na⁺ models. This observation is in close agreement with guanine, cytosine and thymine complexes. Similar to biplanar structures in those bases, RNA Li⁺ biplanar adenine models showed a tetrahedral co-ordination. All models were submitted to NAMD with parameters as defined in Section
1.2, Chapter 1. Metrics for analysis as described in Section 1.3, Chapter 1, were extended to include Potential of Mean Force torsions, N6-N1 distance, N6-N1(H1) H-bond frequency and molecular energy.

4.8.1 Results overview

The analytical metrics all suggested a trend towards slightly enhanced stability for RNA models over DNA models. RNA showed marginally less dynamic activity than DNA equivalents. N6-methylation had a variable impact. Co-ordination with monovalent and divalent cations showed a commonality of early expulsion from both DNA and RNA systems. However, ions were retained within the hydration sphere at the periphery of a given quadruplex.

H-bonds between N6-N1(H1) and N6 in adjacent bases for A4-N1 quartets showed a mean distance of 2.9 Å. This value compares well with guanine N1-O6 averaging 2.8 Å. The mean distance for π-π interactions between the imidazole rings of stacked quartets was 3.9 Å. G4 π-π interactions between the pyrimidines of stacked quartets averaged 3.8 Å. Covalent bonding in ARN8-N1 was predicted to be more prevalent than ADN12-N1, in particular π-π stacking. Supplementary figure S4.6 tabulates H-bonds and π-bonds observed in both models with the differences highlighted.

RMSD data for ARN8-N3 compares well with ADN12-N3. However, the ARN12-N7 system indicated noticeable dynamic activity, contrasting strongly with ADN12-N7. Conformational change occurs at ~320 ns in this model.

The Radius of Gyration trajectory for ARN8-N7 shows a significant change in molecular structure; this contrasts with ARN8-N1 and ARN8-N3 compared with all DNA models. The loss of ARN8-N7 structural integrity at ~320 ns closely reflects a similar event in its RMSD trajectory. Compactness of the molecule was broadly similar for all co-ordinating ions except for hydronium (ΔRg = 39%) The lowest ΔRg value, for ammonium, 5.9%, contrasted with their respective ΔeRMSD values: H3O+ = 68%, NH4+ = 98%.

RMSD trajectories for N6-methylated ARN8-N1, ARN8-N3 and ARN8-N7 indicate enhanced dynamic activity compared with equivalent non-modified models. ARN8-N1 shows less activity by all metrics other than RMSD.

NH4+, H3O+, K+, Li+, Mg2+ and Na+ as co-ordinating ions in ARN8-N1 quadruplexes were rapidly and permanently expelled. However, Ca2+ relocated from its interplanar position at an early stage to a peripheral site on the phosphate backbone and the complex lost integrity at ~260 ns. The ion remained in loco for the duration of the simulation. The quadruplex maintained reasonable integrity after all ions except Ca2+ were expelled.
H-bond frequency data showed similar values for all DNA and RNA models if the difference in number of quartets is taken into account.

Other than A4-N1 derived models, all systems showed a good representation of N6-N(H1) hydrogen bonding. However, distances between adjacent quartet bases varied widely. As exceptions, ADN12-N1 and ARN8-N1 quadruplexes typically showed at least two very stable N6-N(H1) distances which were particularly characterised by vertical inter-planar π-bonding.

4.9 Results

4.9.1 Triplanar adenine quadruplex (ADN12-N1, -N3, -N7) K⁺ co-ordination

4.9.1.1 RMSD

Dynamic activity increases as ADN12-N3 < ADN12-N1 < ADN12-N7 with the latter showing the most erratic trajectory. ADN12-N1 maintains the lowest overall RMSD averaging ~3.4 Å (figures 4.6a and 4.6b):

![RMSD plots](image)

Figures 4.6a and 4.6b: ADN12-N1, ADN12-N3, ADN12-N7 RMSD values and histograms

4.9.1.2 eRMSD

ADN12-N1 shows the least eRMSD of the three systems but exhibits greater variation in values (figures 4.7a and 4.7b). All three maintain a steady trajectory centered around ~2.0 Å.
Figures 4.7a and 4.7b: ADN12-N1, ADN12-N3, ADN12-N7 eRMSD values and histograms

4.9.1.3 Radius of Gyration

Whilst all models maintain a reasonably level trajectory centered on ~1.10 Å, ADN12-N7 shows the greatest variation and range of values. ADN12-N1 and ADN12-N3 trajectories show a steady trend with a narrow range of values (figures 4.8a and 4.8b):

Figures 4.8a and 4.8b: ADN12-N1, ADN12-N3, ADN12-N7 Radius of gyration values and histograms

4.9.1.4 Frequency of H-bond formation

An overall higher rate of H-bond formation occurs in ADN12_N1 systems with the lowest values predicted for ADN12_N3 models. All complexes appear to have a stable rate of
formation with values at zero very infrequent (figures 4.9a and 4.9b):

Figures 4.9a and 4.9b: ADN12-N1 K+, ADN12-N3 K+, ADN12-N7 K+ frequency of hydrogen bond formation

The histograms of figure 4.9a data (figure 4.9b) show an optimum range of H-bond frequency between six and eight. ADN12-N1 has the most consistent and extensive range of bond numbers. The values for ADN12-N7 are the least consistent.

4.9.1.4.1 Histogram for frequency of H-bond formation per base (S4.18 - S4.37)

Predicted hydrogen bonding for each base of ADN12-N1 co-ordinated by K+ shows a stable network for most bases (exemplified by base 1, figure 4.10). In particular, the bases directly constituting the quartets show robust H-bond networks (figures S4.17 – S4.38)

Figure 4.10: Frequency of H-bond formation for base 1 ADN12_N1 K+. An average of six bonds are predicted as likely to be present at any one instant. This figure should be compared with data for the biplanar equivalent (figure 4.9.42).
4.9.1.5 Energy

Predicted energies for all three models maintain a steady range between ~880000 kJ mol$^{-1}$ and ~875000 kJ mol$^{-1}$ for the duration of the simulation (Figure 4.11a, 4.11b) The histogram of data confirms an optimum centered on ~875000 kJ mol$^{-1}$.

![Figure 4.11a and 4.11b: compilation of energies for K$^+$ co-ordinated ADN12-N1, ADN12-N3 and ADN12-N7](image)

Figure 4.11a and 4.11b: compilation of energies for K$^+$ co-ordinated ADN12-N1, ADN12-N3 and ADN12-N7

4.9.1.6 N6 – N1(H1) hydrogen bond frequency and population

4.9.1.6.1 ADN12-N1. (S4.38 - S4.59)

The majority of bases show a strong hydrogen bond network. Bases comprising the three quartets (2, 3, 4, 8, 9, 10, 14, 15, 16, 19, 20 and 21) are particularly robust with the exception of bases 2, 3 and 4 which are less well represented by bonds. Bases of the central quartet (3, 9, 15 and 21) show an overall rich population of bonds (S4.40, S4.46, S4.52, S4.58). The network is particularly strong for base 9, although bases 15 and 21 have a consistent population. Base 3 has a brief loss of bonding between ~220 ns and ~300 ns.

4.9.1.6.2 ADN12 N3 (S4.60 – S4.81)

The network of hydrogen bonding is similar overall to ADN12 N1 with minor variations. In particular, base 3 is better represented and base 9 much less so. The network for base 9 is noticeably less well populated after ~180 ns. Perhaps coincidentally, hydrogen bonding in base 22 increases at the same point.
4.9.1.6.3 ADN12 N7 (S4.82 – S4.103)

The majority of bases show highly transient hydrogen bonding: most networks collapse within ~5 ps and therefore do not register on a 500 ns scale (figure 4.12 and S4.82 – S4.103).

Figure 4.12: base-specific population of H-bonds, ADN12-N7 base 3. Supplementary figures show empty data for the majority of returns.

Bases 2,3,4,10 and 20 show reasonably strong bonding frequencies similar to ADN12-N1 and ADN12-N3.

4.9.1.7 N6 – N1 distance.

Bases comprising quartets in triplanar systems are: 2, 3, 4, 8, 9, 10, 14, 15, 16, 19, 20 and 21.

4.9.1.7.1 ADN12 N1 (K+ co-ordination) S4.104 – S4.125

Bases comprising the three quartets are generally dynamic with the majority of trajectories similar to S4.106. The central quartet comprises bases 3,9,15 and 21. The highly erratic trajectory for base 15 (S4.118) is very similar to that for base 19 (S4.122) but occurs less frequently. All planes have two N6(H1) –N1 bonds trajectories almost identical to bases 9 and 21 (S4.122 and S4.124), indicative of a stable dynamic between the relevant adjacent bases. The bases expressing this stable trajectory are almost exclusively located in a direct vertical relationship.
4.9.1.7.2 ADN12 N3 (K⁺ co-ordination) S4.126 – S4.147

Trajectories have regular similarities with data for ADN12-N1 except for bases 2,6,10 and 20 which have the very similar erratic character exemplified by figure S4.139. However, this increase in the proportion of unstable N6-N1 bonds is offset by an increase in dynamic but stable trajectories. Figures S4.138 and S4.140 typify the relevant bases: 1,5,9,13,17 and 21.

4.9.1.7.3 ADN12 N7 (K⁺ co-ordination) S4.148 – S4.169

The stable trajectory noted in ADN12 N1 and ADN12 N3 (eg. S4.104) is not predicted for any of the ADN12 N7 models. Approximately half of the trajectories are similar to the archetype of base 11 (S4.158).

4.9.2 Triplanar adenine quadruplex (ADN12 -N1, -N3, -N7) Na⁺ co-ordination

4.9.2.1 RMSD

Values for both Na⁺ co-ordinated ADN12-N1 and ADN12-N7 models are very similar which contrasts with the K⁺ system. ADN12-N3 trajectory shows a steady increase between ~100 ns and ~200 ns after which it parallels the two other plots. The histogram for this model describes a wider range of RMSD compared to the other ADN12 systems (figures 4.13a and 4.13b):

![Figures 4.13a and 4.13b: Na⁺ co-ordinated RMSD data for ADN12-N1, -N3 and -N7.]

4.9.2.2 eRMSD

The trajectories for the three systems follow very similar paths. ADN12-N1 is more dynamic and shows the widest range of values (figures 4.14a and 4.14b):
Figures 4.14a and 4.14b: Na\(^+\) co-ordinated eRMSD data for ADN12-N1, -N3 and -N7 Na+

4.9.2.3 Radius of Gyration

ADN12-N7 shows a wider range of values during the relaxation / compactness cycle than ADN12-N1 and -N7. The trajectory for ADN12-N1 indicates the model is the most stable (figures 4.15a and 4.15b):

Figures 4.15a and 4.15b: Na\(^+\) co-ordinated radius of gyration data for ADN12-N1, -N3 and -N7

4.9.2.4 Frequency of H-bond formation

The frequency of H-bond formation is predicted to be greatest for the ADN12-N1 system and least for ADN12-N7. However, all three models show a brief reduction in the frequency between ~260 ns and ~290 ns (figures 4.16a and 4.16b):
Figures 4.16a and 4.16b: ADN12-N1 Na+, ADN12-N3 Na+, ADN12-N7 Na+ frequency of hydrogen bond formation.

The optimum H-bond frequency for ADN12-N7 is centered on five, which contrasts with ADN12-N1 and ADN12-N3.

4.9.2.5 Energy

Trajectories for the predicted energies for ADN12-N1 -N3 and -N7 maintain a steady range between $\sim$880000 kJ mol$^{-1}$ and $\sim$875000 kJ mol$^{-1}$ for the duration of the simulation (Figures 4.17a and 4.17b):

Figures 4.17a and 4.17b: compilation of energies for Na$^+$ co-ordinated ADN12-N1, ADN12-N3 and ADN12-N7

Histograms of the data confirm an optimum centered on $\sim$877000 kJ mol$^{-1}$ with a similar range of values.
4.9.2.6 N6(H1) – N1 distance

4.9.2.6.1 ADN12 N1 (S170 – S174)

The majority of bases show a reduced hydrogen bond network compared to ADN12-N1 K⁺. Bases comprising the central quartets not particularly well represented except for base 21 (Figures 4.9.18a, 4.9.18b, 4.9.18c and 4.9.18d). The network for DA9 is noticeably less well populated than that for the K⁺ equivalent. Base 21 has a consistent frequency of formation which is not apparent in the majority of data.

4.9.2.6.2 ADN12 N3 (S4.174 – S4.177)

The data have close similarity for ADN12-N1 equivalent trajectories; in particular, base 21 shows a more robust network than the majority. Base 21 shows a particularly strong H-bond frequency towards the end of the simulation.

4.9.2.6.3 ADN12 N7 (S4.178 – S4.181)

The majority of data show an absence of H-bonding networks after ~5 ps. Exceptionally, base 3 maintains a very poor network for a simulation to 500 ns (figure 4.18):

![Figure 4.18: H-bonds between N6(H1) and N1 of an adjacent base in the central plane of ADN12-N7 Na⁺.](image)

4.9.2.7 N6(H1) – N1 distance

4.9.2.7.1 ADN12-N1 (S182 – S185)

Bases comprising the three quartets in Na⁺ co-ordinated ADN12-N1 models are all dynamic and erratic with no discernible common trend and typified by the bases of the
central quartet. There are no trajectories indicating a stable N6(H1)-N1 interaction as exemplified by figure S4.104.

4.9.2.7.2 ADN12-N3 (S4.185 – S4.188)

Similar to the trajectories for K+ co-ordinated ADN12-N3 systems, Na+ co-ordinated ADN12-N1 models have no overall discernible trend except for notable variation in N6(H1) and N1 distance.

4.9.2.7.3 ADN12-N7 (S4.189 – S4.192)

Similar to Na+ co-ordinated ADN12-N3 models, ADN12-N7 systems are notably dynamic with no trajectories indicating a reasonably stable inter-base relationship.

4.9.3 Triplanar adenine quadruplex (ADN12-N1, N3, N7) Li+ co-ordination

4.9.3.1 RMSD

The trajectory for ADN12-N1 shows a rising trend which is less evident in ADN12-N3 and ADN12-N7. However, all three systems maintain a relatively low variation in RMSD values (Figures 4.19 and 4.19b):

Figures 4.19a and 4.19b: ADN12-N1, ADN12-N3, ADN12-N7 Li+ RMSD values and histograms

4.9.3.2 eRMSD

ADN12-N1, ADN12-N3, ADN12-N7 Li+ co-ordinated models show little dynamic activity and similar trajectories (figures 4.20a and 4.20b):
4.9.3.3 Radius of gyration

Trajectories for the three models maintain a steady level trend indicating a stable molecular compactness with only marginal variation (figures 4.21a and 4.21b):

4.9.3.4 Frequency of H-bond formation

All three models show a consistent frequency of H-bonding with ADN12-N1 achieving a marginally higher rate. ADN12-N7 has an overall lower frequency than ADN12-N1 and ADN12-N3 systems (figures 4.22a and 4.22b):
Figures 4.22a and 4.22b: ADN12-N1 Li+, ADN12-N3 Li+, ADN12-N7 Li+ frequency of hydrogen bond formation.

A histogram of the data confirms the ADN12-N1 system maintains the highest values and ADN12-N7 the lowest.

4.9.3.5 Energy

Energies for Li⁺ co-ordinated ADN12-N1 ADN12-N3 and ADN12-N7 models maintain very similar values to K⁺ and Na⁺ systems (figures 4.23a and 4.23b):

Figures 4.23a. and 4.23b: compilation of energies for Li⁺ co-ordinated ADN12-N1, ADN12-N3 and ADN12-N7

4.9.3.6 N6(H1) – N1 distance

4.9.3.6.1 ADN12 N1 (S4.193 – S4.196)

The majority of bases show a poor hydrogen bond network compared to K⁺ co-ordinated ADN12-N1 with the exception of bases in the central quartets which have an overall better representation.
Bases are slightly less well represented by a hydrogen bond network compared to ADN12-N1 models with single H-bonds evident for ~50% of the simulation (~65% for ADN12-N1 models).

4.9.3.6.3 ADN12 N7

The majority of data show an absence of H-bonding networks after ~5 ps. Base 9 maintains some H-bonding for a simulation to 500 ns (figure 4.24):

![H-bonds between N6(H1) and N1 of an adjacent base in the central plane of ADN12-N7 Li\(^+\).](image)

Figure 4.24: H-bonds between N6(H1) and N1 of an adjacent base in the central plane of ADN12-N7 Li\(^+\).

4.9.3.7 N6(H1) – N1 distance

4.9.3.7.1 ADN12 N1 (S4.201 – S204)

Trajectories for bases comprising the central quartets in Li\(^+\) co-ordinated ADN12-N1 models are all dynamic and erratic with no discernible common trend. There are no trajectories indicating a stable N6(H1)-N1 interaction.

4.9.3.7.2 ADN12 N3 (S4.205 – S4.208)

Trajectories for ADN12-N3 models co-ordinated by Li\(^+\) are highly erratic and indicate unstable inter-base relationships for any given plane.
4.9.3.7.3 ADN12 N7 (S4.209 – S4.212)

Data returns for Li⁺ co-ordinated ADN12-N7 models are similar to ADN12-N3 and indicate significant variation in distances between bases in the same plane which range between 3 Å and 7 Å.

4.9.4 Triplanar adenine quadruplex (ADN12 -N1, -N3, -N7) Ca²⁺ and Mg²⁺ co-ordination

4.9.4.1 RMSD and 4.9.4.2 eRMSD

The trajectory for RMSD values in ADN12-N1 Ca²⁺ rise steadily to the end of the simulation. Dynamic activity is greater after ~ 320 ns. Although the eRMSD data maintain a steady trajectory to ~300 ns, after which point a similar trajectory to the RMSD is indicated. RMSD values for the Mg²⁺ model maintain a steady trend to ~310 ns at which point a steady reduction is observed. The eRMSD trajectory shows little dynamic activity after a marginal increase to ~100 ns (figures 4.25a and 4.25b):

Figures 4.25a and 4.25b: RMSD (left) and eRMSD (right) data for ADN12-N1 co-ordinated by Ca²⁺ or Mg²⁺.

4.9.4.3 Radius of gyration and 4.9.4.4 Energy

The calculations for both models show only minor variations in molecular compactness (0.5 Å to ~0.9 Å), but marginal increase in dynamic activity for the Ca²⁺ system at ~450 ns. The trajectory for the Mg²⁺ system has minimal dynamic activity. Energy trajectories for both models are also very similar and maintain a steady level (figures 4.26a and 4.26b):
Figures 4.26a and 4.26b: Radius of gyration (left) and energies (right) data for ADN12-N1 co-ordinated by Ca$^{2+}$ or Mg$^{2+}$

4.9.4.5 Frequency of H-bond formation

Both models show an average five bonds at any point, with a limited range of values. However, the data show is no similarity between the two systems on a per base analysis. Data for Ca$^{2+}$ and Mg$^{2+}$ co-ordinated models are listed in S4.390 – S4.397 and S4.398 - S4.405 respectively.

S4.390. Supplementary figure exemplifying H-bond frequency per residue. These data relate to ARN8 co-ordinated by Ca$^{2+}$. Results for all co-ordinating ions are filed between S4.18 and S4.21

4.9.5 Triplanar adenine quadruplex (ADN12 -N1, -N3, -N7) NH$_4^+$ or H$_3$O$^+$ co-ordination

4.9.5.1 RMSD and 4.9.6.2 eRMSD

RMSD trajectories for ammonium and hydronium models both show minimal dynamic activity for the duration of the simulation with the exception of NH$_4^+$ at ~100 ns. eRMSD values both systems also maintain steady trajectories trends (figures 4.27a and 4.27b):
4.9.5.3 Radius of gyration and 4.9.6.4 Energy

Molecular compactness and relaxation cycle for both models is limited and shows no obvious periodicity. Energies are calculated to be within similar ranges to ADN12-N1 systems co-ordinated with Ca$^{2+}$ and Mg$^{2+}$ (Figures 4.28a and 4.28b):

4.9.5.5 Frequency of H-bond formation per base

Both NH$_4^+$ and H$_3$O$^+$ models average three bonds at any one time. No obvious pattern on a per base analysis is indicated. Data for H$_3$O$^+$ or NH$_4^+$ co-ordinated models are listed in S4.4.06 – S4.413 and S4.414 – S4.421 respectively.
Overall, group 1 metals have a narrower range of values for all metrics, with Li⁺ showing the least. RMSD ranges between 0.4 Å and 0.65 Å for group 1 metals, 0.35 Å to 1.0 Å for Ca²⁺, Mg²⁺ and non-metal cations. eRMSD ranges between 0.5 Å and 1.1 Å for group 1 metals, 0.5 Å to 1.04 Å for Ca²⁺, Mg²⁺ and non-metal cations. Radius of gyration ranges between 1.1 Å and 1.15 Å for group 1 metals, 1.05 Å to 1.25 Å for Ca²⁺, Mg²⁺ and non-metal cations. Energies ranges between -479 000 kJ mol⁻¹ and -477 000 kJ mol⁻¹ for group 1 metals, -481 000 to -476 000 kJ mol⁻¹ for Ca²⁺, Mg²⁺ and non-metal cations (Figures 4.29a, 4.29b, 4.29c and 4.29d):

Figures 4.29a, 4.29b, 4.29c and 4.29d: RMSD (a), eRMSD (b), Radius of gyration (c) and energies (d) data for all ADN12-N1 models with ionic co-ordination.
4.10 Biplanar N6-methylated adenine quadruplexes ARN8-N1, ARN8-N3 and ARN8-N7

4.10.1 RMSD

Although ARN8-N1 generally maintains the lowest values, its trajectory shows a sudden steep rise at ~360 ns after which the trend is steady and parallels that for ARN8-N3 and ARN8-N7 (figures 4.30a and 4.30b):

Figures 4.30a and 4.30b: RMSD trajectories and histogram compilation of N6-methylated ARN8 models.

4.10.2 eRMSD

All three models show very similar trajectories for the duration of the simulation. A slight rise in eRMSD is noticeable for ARN8-N1 at ~400 ns (figures 4.31a and 4.31b):

Figures 4.31a and 4.31b: eRMSD trajectories and histogram compilation of N6-methylated ARN8 models
4.10.3 Radius of gyration

Trajectories for ARN8-N7 is more dynamic than ARN8-N1 and ARN8-N3 systems which maintain an overall steady trend. ARN8-N1 shows a rapid increase of compactness within the first 50 ns. ARN8-N3 shows a possible cycle of molecular relaxation and compactness (figures 4.32a and 4.32b):

Figures 4.32a and 4.32b: radius of gyration trajectories and histogram compilation of N6-methylated ARN8 models

4.10.4 Energy

All three models maintain consistent energies between ~-879 000 kJ mol\(^{-1}\) and ~-876 000 kJ mol\(^{-1}\) (figures 4.33a and 4.33b):

Figures 4.33a and 4.33b: trajectories and histogram compilation for energies in N6-methylated ARN8 models

4.10.5 Frequency of H-bond formation per base

Bases comprising central quartets for all models are: 1, 2, 7, 8, 15, 19, 20 and 21. Models were analysed as a per base metric because systems collapsed prior to 10ps of simulation time. The data for ARN8-N1_meth, ARN8-N3_meth and ARN8-N7_meth systems were
very similar. Histograms for the frequency of formation showed a low population of H-bonding averaging <4 bonds at any one point in the simulation. The data are available in Supplementary figures S4.213 – S4.220 (ARN8-N1_meth), S4.221 – S4.228 (ARN8-N3_meth), S4.229 – S4.236 (ARN8-N7_meth).

4.10.6 N6 – N1(H1) hydrogen bond frequency and population

The majority of bases in the pair of adenine quartets methylated at N6 for all three models show very low frequencies of H-bond formation. ARN8-N1 and ARN8-N3 are better represented than ARN8-N7. The data are itemised in Supplementary figures S4.237 – S4.245 (ARN8-N1_meth), S4.246 – S4.253 (ARN8-N3_meth), S4.254 – S4.263 (ARN8-N7_meth).

4.11 Biplanar adenine quadruplex ARN8-N1 no ionic co-ordination

4.11.1 RMSD

Trajectories for ARN8-N1 and ARN8-N3 maintain consistent values at ~1.5 Å and ~2.0 Å respectively. ARN8-N7 is also consistent to ~320 ns at which point a significant increase occurs. The system is noticeably dynamic to 500 ns (figures 4.34a and 4.34b):

![RMSD trajectories and histogram compilation of ARN8 models with no ionic co-ordination](image-url)
4.11.2 eRMSD

All three models maintain consistent trajectories to 500 ns centered on ~2.0 Å with ARN8-N1 showing marginally lower values (figures 4.35a and 4.35b):

Figures 4.35a and 4.35b: eRMSD trajectories and histogram compilation of ARN8 models with no ionic co-ordination

4.11.3 Radius of gyration

Values for all three models originate at ~0.3 Å ARN8-N1 and ARN8-N3 systems show a consistent molecular radius to 500 ns. ARN8-N7 shows sudden and rapid molecular relaxation and compaction cycle which appears to repeat immediately before the simulation end. The trajectory is very similar to the RMSD calculation (Figures 4.36a and 4.36b):

Figures 4.36a and 4.36b: Radius of gyration trajectories and histogram compilation of ARN8 models with no ionic co-ordination
4.11.4 Energy

The trajectories for energy of ARN8-N1, ARN8-N3 and ARN8-N7 are consistent for the duration of the simulation with values ranging between -876 000 kJ mol\(^{-1}\) and -878 000 kJ mol\(^{-1}\) (figures 4.37a and 4.37b):

![Figures 4.37a and 4.37b: trajectories and histogram compilation for energies of ARN8 models with no ionic co-ordination](image)

4.11.5 Frequency of H-bond formation (per base).

For bases of the two quartets (1,2,7,8,15,19,20,21), an average of four bonds are predicted as likely to be present at any one instant but no other trend is indicated. Data are listed in Supplementary figures, Hist_HB_frq per res, S4.264 – S4.271

4.12 Biplanar adenine quadruplexes ARN8-N1 with cationic co-ordination

Because of ostensible enhanced stability inherent in ARN8-N1 predicted in the work hitherto, this model was selected for further investigation by imposing monovalent and divalent ions. To compare with guanine equivalents, the influence, if any, of K\(^+\), Na\(^+\) or Li\(^+\) as co-ordinants was assessed. In addition, Ca\(^{2+}\) is isoelectronic with K\(^+\) and therefore may have a similar interaction with the models. Also selected were physiologically relevant Ca\(^{2+}\) and Mg\(^{2+}\) to compare divalency and ionic radius with the Group 1 metals. H\(_3\)O\(^+\) and NH\(_4\)\(^+\) were included as the cations are mutually isoelectronic and should have a similar interaction with adenine quadruplexes but dissimilar to metal cations.
4.12.1 ARN8-N1 K⁺, Na⁺ or Li⁺ co-ordination: RMSD and eRMSD

Trajectories for Na⁺ models show more dynamic activity for both metrics. The trajectory for the Li⁺ equivalent maintains a noticeably steady trend for the entire simulation (figures 4.38a and 4.38b):

Figures 4.38a and 4.38b: RMSD (left) and eRMSD (right) data for ARN8-N1 co-ordinated by K⁺, Na⁺ or Li⁺.

4.12.2 Radius of gyration and Energy 4.12.4

Molecular compaction varies most for K⁺ and Na⁺ systems but the Li⁺ model shows only brief relaxation immediately after the minimisation cycle of the simulation. Energies for all three models show very little variation (Figures 4.39a and 4.39b):

Figures 4.39a and 4.39b: Radius of gyration (left) and energies (right) data for ARN8-N1 co-ordinated by K⁺, Na⁺ or Li⁺.

4.12.3 Frequency of H-bond formation (per base)

An average of two bonds are predicted as likely to be present at any one instant and the range of frequencies (0 to 4) is relatively narrow as exemplified by Figure 4.40. Data for
K⁺, Na⁺ or Li⁺ co-ordinated models are listed in S4.272 – S4.279, S4.280 – S4.286 and S4.287 – S4.294 respectively.

Figure 4.40 (and S6a – S6h): Frequency of H-bond formation base 1 ARN8_N1 K⁺. This figure should be compared with data for the triplanar equivalent (figure 4.5).

4.13 Biplanar adenine quadruplexes ARN8-N1 Ca²⁺ and Mg²⁺ co-ordination

4.13.1 RMSD and eRMSD

RMSD values for ARN8-N1 Ca²⁺ rise rapidly to minimisation but after which, a steady trajectory is maintained. The simulation does not continue to completion however. The eRMSD data show a very similar trajectory. RMSD values for the Mg²⁺ model maintains a steady trend to ~460 ns at which point a rapid and steep increase is observed. The system is briefly very active which is reflected to a small extent in the eRMSD trajectory (figures 4.41a and 4.41b):

Figures 4.41a and 4.41b: RMSD (left) and eRMSD (right) data for ARN8-N1 co-ordinated by Ca²⁺ or Mg²⁺.
4.13.2 Radius of gyration and 4.13.3 Energy

Although the calculation for the Ca\(^{2+}\) model shows only minor variations in molecular compactness, the model does not run to completion and collapses at \(~450\text{ns}\). The trajectory for the Mg\(^{2+}\) system shows rapid and frequent alternations between relaxation and compaction to \(~12\text{ ns}\) after which the model is dynamic but stable (figures 4.42a and 4.42b):

![Graphs showing radius of gyration and energies for Ca\(^{2+}\) or Mg\(^{2+}\) coordination](image)

Figures 4.42a and 4.42b: Radius of gyration (left) and energies (right) data for ARN8-N1 co-ordinated by Ca\(^{2+}\) or Mg\(^{2+}\)

4.13.3 Frequency of H-bond formation (per base).

Both models show an average four bonds at any point, but also have a wide range of values. There is no commonality between the two systems on a per base analysis. Data for Ca\(^{2+}\) and Mg\(^{2+}\) co-ordinated models are listed in S4.295 – S4.301 and S4.302 -S4.309 respectively.

4.14 Biplanar adenine quadruplex ARN8-N1 NH\(_4^+\) or H\(_3O^+\) co-ordination

4.14.1 RMSD and eRMSD

Neither ammonium nor hydronium models run to completion and RMSD trajectories for both show noticeable dynamic activity to \(~200\text{ ns}\). However, eRMSD values for the NH\(_4^+\) system show a steady trend for the duration of the simulation; those for the H\(_3O^+\) model show a steady trend at \(~1.9\text{ Å}\) rise to a steady trend at \(~2.3\text{ Å}\) before collapsing (figures 4.43a and 4.43b):

![Graphs showing RMSD and eRMSD for NH\(_4^+\) or H\(_3O^+\) coordination](image)
Figures 4.43a and 4.43b: RMSD (left) and eRMSD (right) data for ARN8-N1 co-ordinated by H$_3$O$^+$ or NH$_4^+$.  

4.14.3 Radius of gyration (S65) and Energy

Molecular compactness for both models is erratic and shows no obvious periodicity before collapse prior to completion. Energies are calculated to be within similar ranges to ARN8-N1 systems co-ordinated with cations (Figures 4.44a and 4.44b):

Figures 4.44a and 4.44b: Radius of gyration (left) and energies (right) data for ARN8-N1 co-ordinated by H$_3$O$^+$ or NH$_4^+$.  

4.14.4 Frequency of H-bond formation (per base).

NH$_4^+$ models average four bonds at any one time and three for H$_3$O$^+$ systems with a reasonable range of values. No obvious pattern on a per base analysis is indicated. Data for H$_3$O$^+$ or NH$_4^+$ co-ordinated models are listed in S4.310 – S4.317 and S4.318 – S4.325 respectively.
4.14.5 RMSD, eRMSD, Radius of gyration and Energy histograms for ARN8-N1 in all coordinating ion conditions.

Overall, group 1 metals have a narrower range of values for all metrics, with Li⁺ showing the least. Histograms for Ca²⁺ and H₃O⁺ models reflect their collapse prior to the end of the simulation (Figures 4.45a, 4.45b, 4.45c and 4.45d):

For comparison with guanine, cytosine or thymine 22-mer strands, the melting temperature $T_m$ enthalpy and entropy of an adenine equivalent was calculated as in Section 2.4. Whilst enthalpic and entropic values are similar across all strands, melting temperatures resonate with data in this study of stability predicated on available H-bonding (Table 4.2):
Table 4.2: melting temperatures, enthalpy and entropy values for triplanar guanine, cytosine, thymine and adenine 22-mers

<table>
<thead>
<tr>
<th>species</th>
<th>Tm°C</th>
<th>enthalpy kJ/mol</th>
<th>entropy J/K</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDN12</td>
<td>69.3</td>
<td>178.4</td>
<td>505.46</td>
</tr>
<tr>
<td>CDN12</td>
<td>57.4</td>
<td>179.0</td>
<td>509.86</td>
</tr>
<tr>
<td>TDN12</td>
<td>39.6</td>
<td>171.5</td>
<td>515.96</td>
</tr>
<tr>
<td>ADN12</td>
<td>61.7</td>
<td>172.2</td>
<td>518.56</td>
</tr>
</tbody>
</table>

Specifically for ADN12-N1, ADN12-N3, ADN12-N7, ARN8-N1, ARN8-N3 and ARN8-N7 models, further analysis was gained by calculating Potential of Mean Force for dihedral torsions between a given base, its associated sugar and phosphate. The systems were selected to assess in more detail differences, if any, in DNA and RNA biplanar and triplanar quadruplexes derived from the three validated adenine quartets. The metric was applied to gain an overview of geometries for each base in a system.

4.15 Potential of Mean Force torsions for DNA- and RNA-derived adenine quadruplexes

4.15.1 Introduction

In a molecular system, the development of any process is defined by its reaction coordinate. Potential of Mean Force (PMF) can be interpreted as an energy profile of the reaction coordinate. The metric is pragmatic as degrees of freedom are averaged (71). Specifically for this study, collective variables (colvars) for the reaction coordinate were glycosidic and phosphate backbone torsions in DNA and RNA adenine quadruplexes. Hence, PMF torsion calculations provide insight into energy fluctuations of the reaction coordinate and expressed as a function of time-dependent angles between bases, sugars and phosphates. Many different atomic arrangements are possible in almost any molecular system with each microstate described as a particular specific configuration at any one instant. The probability of the system to be in a microstate predicated on base-sugar-phosphate angles of each base should inform temporal configurations of the
system. Of particular interest is the glycosidic bond ($\chi$) which defines the planarity of its dependent base (figure 4.46):

![Image of glycosidic bond](image)

**Figure 4.46 Glycosidic ($\chi$) and phosphate backbone torsions.** Although the sugars differ for DNA and RNA, the angles for both systems are defined as illustrated. The glycosidic angle is marked in red.

Fluctuations in energy from a specified parameter and expressed as a function of time-dependent distances between bases were analysed by Potential of Mean Force torsion (71). The seven torsion angles of the phosphate backbone were calculated for each base of all ADN12-N1, ADN12-N3, ADN12-N7, ARN8-N1, ARN8-N3 and ARN8-N7.

### 4.15.2 Methods

Solvated models neutralised with counterions were equilibrated with no constraints. The configuration file for Production (Supplementary 1.6.2.2 Script 2) was adapted with commands to access the NAMD colvars module. Script 1.6.3.7 details the colvars Config file and commands. For periodic boundary conditions to be enabled, constraints were applied to the hydration box. The system was minimized for 50 000 steps at 2 fs and Production at 24 950 000 to provide a 500 ns simulation.

PMF values on a per base analysis were calculated for DNA and RNA models constructed from the three validated quartets. Data returned as .txt files were analysed with GraphPad Prism. ADN12-N1, ADN12-N3 and ADN12-N7 trajectories for all bases were compiled into a single graph and data also expressed on a per base analysis. Data for ARN8-N1, ARN8-N3 and ARN8-N7 were similarly described. Furthermore, a global PMF was taken by summing PMF values for each base and concatenating to specifically identify $\chi$ minima for ADN12-N1, ADN12-N3 and ADN12-N7. Concatenated data for ARN8-N1, ARN8-N3 and ARN8-N7 are shown in figure 4.9.43.
4.15.3 Results

4.15.3.1 PMF Results summary

Some similarities were noticed between DNA and RNA models, but also several differences. γ and δ angles of the phosphate backbone displayed notable stability in both models. The torsional angle δ in RNA showed a large shift from an approximate median 185° to a median 60° at ~15 ns. This angle was maintained to 500 ns apart from a transient reversal at ~178 ns.

DNA ε and ζ torsions were noticeably more dynamic than RNA. δ torsions for both were very similar, showing little variation. However, in both models, the glycosidic angle (χ) changes imply an inversion of adenine along its major axis. It is unclear if the re-configuration is reversible in either condition. The torsions for glycosidic angles of each base for all models were also calculated and torsional change data for each condition were plotted.

With the exception of the ARN8_N1 χ torsion, RNA models showed very little variation. Plots for γ and ε torsions in DNA systems show considerable angular dynamic activity compared to the RNA equivalents. Modelling software predicted ribose in RNA models would adopt a C2’-endo configuration whilst deoxyribose in DNA structures conform to a C3’-endo arrangement.

4.15.3.1 Results: Concatenated data and per base analysis

Concatenation of values for the three triplanar adenine systems is shown in Table 4.3. The data indicate global minima for ADN12_N1 at -130°, +10° and +125°; for ADN12_N3 at -120°, +10° and +100°; for ADN12_N7 at -90° and +30°.
Table 4.3: Global PMF minima for all bases in ADN12 systems

Bases appear to fluctuate between these lowest energy configurations. ADN12-N7 shows significantly fewer minima than ADN12-N1 or ADN12-N3 for all bases, contrasting with all other metrics. This might be rationalised if the PMF starting point for ARN12-N7 is already at, or close to, a deep minimum and the system has insufficient energy to shift into other low energy wells. The data are expressed visually in figure 4.47

<table>
<thead>
<tr>
<th>base</th>
<th>ADN12-N1</th>
<th>ADN12-N3</th>
<th>ADN12-N7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-140  0</td>
<td>-140  80</td>
<td>-140  110</td>
</tr>
<tr>
<td>2</td>
<td>60  -140</td>
<td>-140  -50</td>
<td>40  130</td>
</tr>
<tr>
<td>3</td>
<td>30  90</td>
<td>30  90</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-140  70</td>
<td>70  150</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-140  50</td>
<td>80  60</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>70  70</td>
<td></td>
<td>60  70</td>
</tr>
<tr>
<td>7</td>
<td>-130  -140</td>
<td></td>
<td>-50  50</td>
</tr>
<tr>
<td>8</td>
<td>90  50</td>
<td>90  90</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-140  100</td>
<td>-140  100</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-140  100</td>
<td>-140  30</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-150  80</td>
<td>-130  90</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>80  80</td>
<td></td>
<td>80  80</td>
</tr>
<tr>
<td>13</td>
<td>80  80</td>
<td></td>
<td>80  80</td>
</tr>
<tr>
<td>14</td>
<td>-130  100</td>
<td>-130  -130</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-130  100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-130  100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-140  50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>70  70</td>
<td></td>
<td>70  70</td>
</tr>
<tr>
<td>19</td>
<td>-140  70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>30  110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>-130  -130</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>-130  100</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

---

Figure 4.47. Global PMF trajectories for $\chi$ values of all ADN12-N1, ADN12-N3 and ADN12-N7 bases for a 500ns simulation. For most bases, angles centered on adenine N9 totalled 356° which implies N9 is non-planar with implicit partial sp³ hybridisation.
The final configuration of ARN8-N1 shows signs of unfolding but has maintained overall structural integrity (S4.3). The A₄-N1 quartets are no longer unanimously biplanar as determined by some glycosidic angles. From a median \(~130^\circ\) prior to MD simulation, glycosidic angles were reduced to a median \(~113^\circ\). Comparison with figure 4.2.1 indicates pair-bonding has maintained a recognisable biplanar structure although with some re-configuration. The phosphate backbone also shows general integrity but with some re-configuration. The final RMSD for the backbone as calculated by DSV is in broad agreement with other methodologies: all molecule RMSD = 4.018 Å, phosphate = 3.571 Å.

Data for α and β angles were returned as NAN: not a number. Computing compliance describes NAN values as a floating point, \(ie\): undefined, unrepresentable or missing (185). In this specific calculation, it is likely precision for these angles conflicted with range and would be poorly-defined. Numerical torsional data are exemplified in supplementary figure S128.

A compilation of \(\chi, \gamma, \delta, \epsilon, \text{ and } \zeta\) torsions on a per base analysis for ADN12-N1 are shown in figures 4.48a, 4.48b and 4.48c:
χ and ζ torsions both undergo an obvious change between ~100 ns and ~200 ns. However, where the former flips from an average -160° to +170°, angles for the latter change from ~150° to -90°. ζ torsions differ to some extent from this observation as there is a significant population of negative values for the duration of the simulation. δ and γ torsions show little variation at ~120° and 60° respectively. Data for ε torsions show a slightly wider range of values averaging -120°. Concatenation of values for the three biplanar adenine systems is shown in Table 4.4

Table 4.4: Global PMF minima for all bases in ARN8 systems

<table>
<thead>
<tr>
<th>residue</th>
<th>ARN8-N1</th>
<th>ARN8-N3</th>
<th>ARN8-N7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-150</td>
<td>-130</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>80</td>
<td>-180</td>
</tr>
<tr>
<td>3</td>
<td>-150</td>
<td>-130</td>
<td>-150</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>120</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>-120</td>
<td>-180</td>
</tr>
<tr>
<td>6</td>
<td>-130</td>
<td>-100</td>
<td>-130</td>
</tr>
<tr>
<td>7</td>
<td>60</td>
<td>70</td>
<td>-180</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>-140</td>
<td>-120</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>-140</td>
<td>-150</td>
</tr>
<tr>
<td>10</td>
<td>130</td>
<td>100</td>
<td>-180</td>
</tr>
<tr>
<td>11</td>
<td>120</td>
<td>80</td>
<td>-150</td>
</tr>
<tr>
<td>12</td>
<td>40</td>
<td>80</td>
<td>-180</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>-60</td>
<td>-130</td>
</tr>
<tr>
<td>14</td>
<td>80</td>
<td>80</td>
<td>-130</td>
</tr>
<tr>
<td>15</td>
<td>-120</td>
<td>0</td>
<td>-120</td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>-120</td>
<td>-150</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>70</td>
<td>-130</td>
</tr>
<tr>
<td>18</td>
<td>-140</td>
<td>90</td>
<td>-180</td>
</tr>
<tr>
<td>19</td>
<td>30</td>
<td>0</td>
<td>-130</td>
</tr>
<tr>
<td>20</td>
<td>-60</td>
<td>40</td>
<td>-150</td>
</tr>
<tr>
<td>21</td>
<td>-60</td>
<td>-120</td>
<td>-130</td>
</tr>
<tr>
<td>22</td>
<td>60</td>
<td>-120</td>
<td>-170</td>
</tr>
</tbody>
</table>

Compiled trajectories for ARN8-N1, ARN8-N3 and ARN8-N7 are shown in figure 4.49
Figure 4.49: The glycosidic angle minimum for the ARN8-N7 system is clearly centered on -170° for the entire simulation. ARN8-N1 and -N3 models also have local minima at this value, with the former showing a deeper minimum at 0°. ARN8-N3 has a double minimum at approximately 60° and 120°.

A compilation of χ, γ, δ, ε, and ζ torsions on a per base analysis for ARN8-N1 are shown in Figures 4.50a, 4.50b and 4.50c:

![figure 4.50a](image)

![figure 4.50b](image)

![figure 4.50c](image)

Figures 4.50a, 4.50b and 4.50c compilation of per base torsions for ARN8-N1 (a), ARN8-N3 (b), ARN8-N7 (c). γ, δ, ε, and ζ torsions for biplanar ARN8-N1 models are noticeably constant with ε and ζ averaging -180° and -110° respectively. Data for δ torsions are very similar to γ and χ values at approximately 60°. However, the trajectory for χ torsions shows an obvious reconfiguration at ~300 ns which persists to the end of the simulation. The change is less than for the ADN12-N1 equivalent.

4.16 Discussion of PMF torsions

PMF data returns averaged for both triplanar and biplanar adenine quadruplexes show that the lowest energy torsional angles are about 0°, 120° and -120°. The sequence of atoms defining this dihedral is O4′-C1′-N9-C4 for purine and O4′-C1′-N1-C2 for pyrimidine.
derivatives. Hence, when $\chi = 0^\circ$ the O4'-C1' bond is eclipsed with the N9-C4 bond for purine and the N1-C2 bond for pyrimidine derivatives. Since the adenine ring is planar, the O4'-C1' bond is eclipsed with the N9-C4 bond but the N9-C8 bond is staggered between the other bonds on the C1'. This conformation implies the heterocyclic ring is non-planar with regard to adjacent bases. Given tolerable planarity, the O4-C1-N9-C4 and O4-C1-N9-C8 dihedral should sum to 180°. However, since this is not observed, other factors may be influential. A contributive factor such as a lone pair of the carbohydrate ring oxygen could affect local dipole interactions and impose a consequent tendency to a minimal energy conformation. Moreover, in a planar system, N9 would be sp² hybridised with C1-N9-C4, C1-N9-C8 and C4-N9-C8 angles totalling 360°. Since the angles are monitored as 356° (C1-N9-C4 = 123°, C1-N9-C8 = 128°, C4-N9-C8 = 105°), N9 appears to have slight sp³ hybridisation and therefore determines a marginal pyramidal structure (figure 4.51):

![Figure 4.51: the N9 (blue) environment in adenine quadruplexes shows a minor pyramidal conformation instead of an expected plane. The displacement is less noticeable in guanine equivalents.](image)

The reconfiguration for $\chi$ torsions in both DNA and RNA models may be relevant. These data resonate closely with the visual confirmation in this analysis of an N9 pyramidal structure. Such changes in purines are considered contingent in the glycosidic angle (186). Compiled PMF trajectories for RNA complexes have more clearly defined minima which reinforces the data depicted in figures 4.9.50a, 4.9.50b and 4.9.50c. Other than the reconfiguration of glycosidic angles, PMF values indicate an overall enhanced structural integrity for RNA models.

Although for molecular dynamics simulations, models were buffered by a hydration sphere, the role of waters per se was not specifically included in a given analysis. Dipolar water molecules could notionally mediate between a cation and electropositive moieties of a quadruplex. A basic script was written to gain an overview of waters in ADN12 and
ARN8. These models were accessed as analysis hitherto supported quadruplexes derived from A$_4$-N1 quartets were the most stable.

4.17 Occupancy of waters in adenine quadruplexes ADN12 and ARN8

4.17.1 Introduction and methodology

Interplanar water occupancy in biplanar RNA and triplanar DNA adenine quadruplexes was assessed in NAMD with Script 5 (Section 1.6.6.4). All models were constructed from A$_4$-N1, A$_4$-N3 and A$_4$-N7 planar quartets but RNA systems were modified by substitution of thymine by uracil. ADN12 models were assessed with co-ordination but ARN8 quadruplexes had no co-ordinant. No bases were methylated in these analyses. Empirical values ($S_{4.500}$ – $S_{4.505}$) and histograms (figures 4.9.53 and 4.9.54) were plotted for each system.

4.17.2 Results

Water occupancy for ADN12-N1 and ARN8-N1 models is generally slightly higher than either of the -N3 or -N7 systems. The DNA models indicate a larger population averaging between four and eight waters, but there is no clear prevalence for RNA models. DNA systems contrast with an apparent poor representation for zero and very low values (Figures 4.52 and 4.53). Overall, the general trend is lower for ARN8 quadruplexes, but proportionally higher than expected given the model has two participating quartets as opposed to triplanar DNA complexes. ARN8-N1 is more tolerant of waters at all populations except for zero.
For DNA and RNA models derived from A₄-N7 quartets, occupancy is the lowest across all populations and A₄-N1 / A₄-N3 equivalents except for a zero value in ARN8-N7.

Frequency of occupancy is lower than for biplanar cytosine and thymine models, but the overall population is noticeably more consistent (figures 3.8 and 3.12).

To better understand internal local geometries and develop information from Potential of Mean Force torsions data (Section 4.21), the conformations of sugar in each base were calculated. The models were selected for reasons as described in Section 4.22.

4.18 Ribose and deoxyribose conformations of biplanar and triplanar adenine models

4.18.1 Introduction and methodology

C₂’-endo and C₃’-endo conformations of ADN12-N1 and ARN8-N1 were analysed for comparison with guanine, cytosine and thymine models. In this study, cytosine and thymine models have demonstrated a more compact helical conformation derived from
the C3'-endo pucker of the relevant sugar. Ascertaining any prevalence of either conformation should reinforce trends established here for biplanar RNA structures and triplanar DNA equivalents.

Because models constructed from A4-N1 quartets have characteristically higher stability than -N3 or -N7 equivalents, ADN12-N1 and ARN8-N1 were selected for MD simulation and analysis. The former was co-ordinated by K+ but the latter without any cation imposed. For comparison with guanine equivalents, ADN12-N1 and ARN8-N1 were methylated at N6.

C1'-C2'-C3'-C4' dihedrals for the sugar ring of base 7 were monitored (section 2.7.2) for atoms numbered 156 155 152 151 in ADN12-N1 and 204 203 201 199 in ARN8-N1 the .pdb file. Mean distances monitored for distances between consecutive phosphates were 5.9Å for C3'-endo and 7.2Å for C2'-endo in ADN12-N1, 5.8Å for C3'-endo and 7.0Å for C2'-endo in ARN8-N1. The two configurations are bounded by an approximate range of dihedral angles: C2’-endo, 0° to 36°; C3’-endo, 144° to 180°. Dihedrals were measured as described in Section 2.7.2. The C1’-C2’-C3’-C4’ dihedral was also calculated for three sugar rings of bases in methylated ADN12-N1 and ARN8-N1. Bases selected were: 1, 7 and 15. The latter two are partial constituents of a quartet, and the former is peripheral to the complex.

4.18.2 Results

A noticeable difference is observed for trajectories in the analysis of adenine models, whether DNA or RNA. There is some possible commonality for adenine deoxyribose systems. Trajectories for some of these models show intervals of very little activity with bouts of temporary interconversion (figures 4.54a, 4.54b and 4.54c):

![Figures 4.54a, 4.54b and 4.54c: trajectories for C1’-C2’-C3’-C4’ dihedrals in DNA biplanar adenine models in bases 1 (left), 7 (centre) and 15 (right).]
Whilst all trajectories in DNA systems show some periodicity, spatio-temporal interconversions specifically do not have any apparent commonality. For bases 1 and 15, there are similar periods of stability between 0ns and ~150 ns, ~250 ns to ~280 ns and ~320 ns to ~370 ns. Base 7 appears to have a brief period of stability between ~130ns and ~190 ns which partially overlaps with base 1.

By contrast, trajectories for equivalent RNA models do not show periods of dynamic inactivity. The data for base 1 do indicate stable configurations prior to, and after, a major interconversion (figures 4.55a, 4.55b and 4.55c):

Figures 4.55a, 4.55b and 4.55c: trajectories for C1’-C2’-C3’-C4’ dihedrals in RNA biplanar guanine models for base 1 (left), 7 (centre) and 15 (right). There is an obvious interconversion from C2’-endo to C3’-endo for base 1 which is not apparent in the other samples.

The rapid and permanent shift from C2’ endo to C3’ endo as observed in figure 4.55a appears exclusive to the models analysed. The trajectory strongly indicated that both C2’ endo and C3’ endo configurations are stable. However, the trajectories for bases 7 (figure 4.55b) and 15 (figure 4.55c) show both configurations occur intermittently. The former may have an irregular periodicity displaying several well-defined interconversions. The range of angles in this trajectory showing a sudden transformation from C2’-endo to C3’-endo at ~220 ns, suggests this change derives from an O4’-endo configuration, intermediate between C2’-endo and C3’-endo (figure 4.56):
**Figure 4.56:** Interconversion of C2'-endo to C3'-endo progresses via the O4' endo conformation. The sugar ring has a noticeably more planar profile than either the C2'-endo or C3'-endo conformations. The distance between successive phosphates is similar to a C3'-endo structure.

**Table 4.5:** Fluctuations in deoxyribose configurations for triplanar non-methylated adenine quadruplex.

<table>
<thead>
<tr>
<th>ADN12-N1 K+</th>
<th>t = 0 ns</th>
<th>t = 500 ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The C3'-endo conformation is prevalent prior to the simulation. Marginal differences in population are noted for most configurations at t = 500 ns except for a reduction of C3'-endo and loss of O4'-endo variations. There is a slight increase in ambiguous configurations.
Table 4.6: fluctuations in ribose configurations for biplanar non-methylated adenine quadruplex.

<table>
<thead>
<tr>
<th>t = 0ns</th>
<th>ARN8-N1</th>
<th>t = 500ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo C2'-endo C3'-endo C4'-endo O5'-endo O4'-exo planar</td>
<td>C1'-endo C2'-endo C3'-endo C4'-endo O5'-endo O4'-exo planar</td>
<td></td>
</tr>
<tr>
<td>14 19 17 5 4 6</td>
<td>15 10 21 14 15 6 9</td>
<td></td>
</tr>
<tr>
<td>22 9 20 13</td>
<td>11 17 12 18</td>
<td></td>
</tr>
<tr>
<td>10 15</td>
<td>13 22</td>
<td></td>
</tr>
<tr>
<td>12 16</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Data are very similar ADN12-N1 whilst noting that C2'-endo is favoured for the entire simulation in contrast to C3'-endo the DNA system. The reduction in prevalence in both models at t = 500ns is also less for ARN8-N1 than for ADN12-N1.

4.18.3 N6-methylation of ADN12-N1 and ARN8-N1

The trajectory for dihedrals in the biplanar RNA model figure 4.55a showed a sudden transformation from C2'-endo to C3'-endo. This transition is not observed for the same base in the equivalent model methylated at adenine N6 (figure 4.55b). The energy barrier differential is based on presence or absence of N6-methylation. The trajectories for biplanar DNA and RNA adenine quadruplexes are very dynamic and indicate methylation provokes an unstable system in both DNA and RNA models (figures 4.57a and 4.57b). However, the latter maintains a more dynamic C3'-endo configuration throughout the simulation than for the biplanar DNA model.

Figures 4.57a and 4.57b: C1'-C2'-C3'-C4' dihedrals for ADN4 (left) and ADN4_meth (right). The trajectories indicate little difference between C2'-endo and C3'endo conformers for both systems. The activation energy for change in configuration is implicitly higher in these models whether O6-
methylated or not when compared with RNA equivalents. The data suggest no one particular confirmation is prevalent in the methylated model.

Distribution histograms for the trajectories of base 1 in N6-methylated and non-methylated biplanar adenine structures clearly indicate the decisive conformational change for this specific base in the latter condition. In comparison, the N6-methylated model shows no particular conformational preference (figure 4.58):

Figure 4.58: distribution histograms for ARN8-N1 (pink) and ARN8-N1_meth dihedrals (blue). C2'-endo and C3'-endo configurations are evenly represented for the duration of the simulation in the non-methylated model. Sugar ring conformations are not well-defined in the N6-methylated model but C3'-endo configurations may have a higher population.

Empirical data for conformational interconversions across adenine DNA and RNA models with relevant methylation are tabulated in Tables 4.7 and 4.8:

Table 4.7: fluctuations in deoxyribose configurations for biplanar non-methylated adenine quadruplex (black) and N6-methylated adenine quadruplex (red).

<table>
<thead>
<tr>
<th></th>
<th>ADN12-N1 K+</th>
<th>ADN12-N1_meth K+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t = 0 ns</td>
<td>t = 500 ns</td>
</tr>
<tr>
<td>C2'-endo</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>O4'-endo</td>
<td>O4'-endo</td>
<td>O4'-endo</td>
</tr>
<tr>
<td>O4'-endo</td>
<td>O4'-endo</td>
<td>O4'-endo</td>
</tr>
<tr>
<td>planar</td>
<td>planar</td>
<td>planar</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ADN12-N1 K+</th>
<th>ADN12-N1_meth K+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t = 0 ns</td>
<td>t = 500 ns</td>
</tr>
<tr>
<td>C2'-endo</td>
<td>C2'-endo</td>
<td>C2'-endo</td>
</tr>
<tr>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>C3'-endo</td>
<td>C3'-endo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>O4'-endo</td>
<td>O4'-endo</td>
<td>O4'-endo</td>
</tr>
<tr>
<td>O4'-endo</td>
<td>O4'-endo</td>
<td>O4'-endo</td>
</tr>
<tr>
<td>planar</td>
<td>planar</td>
<td>planar</td>
</tr>
</tbody>
</table>

152
Sugar ring configurations for biplanar adenine quadruplexes are broadly similar for guanine equivalents (Section 2.7.3). C2’-endo appears as the major variant throughout the simulation but less so for the methylated structure in the final condition. However, there appears to be little correlation between specific C2’-endo configurations for initial and final conditions.

Table 4.8: fluctuations in ribose configurations for biplanar non-methylated adenine quadruplex (black) and N6-methylated adenine quadruplex (red).

<table>
<thead>
<tr>
<th></th>
<th>ARN8-N1</th>
<th>ARN8-N1_meth</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = 0ns</td>
<td>C1’-endo</td>
<td>C2’-endo</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>t = 500ns</td>
<td>C1’-endo</td>
<td>C2’-endo</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>19</td>
</tr>
</tbody>
</table>

Ribose configurations’ data for a biplanar non-methylated adenine quadruplex and an N6-methylated adenine quadruplex show a close similarity to their deoxyribose equivalents. A minor difference may be the marginal correlation between the specific variants of the prevalent C3’-endo configuration. Data for C1’-C2’-C3’-C4’ dihedrals in the biplanar quadruplexes investigated here, strongly indicate N6-methylation for adenine models disallow any stable configuration of the sugar ring. Steric hindrance of increasing the energy barrier would rationalise this observation.

Most of the possible conformations are identified, suggesting none is specifically favoured. The sugar ring may fluctuate as a continuum between the possibilities during simulation.

4.19 Executive summary of Chapter 4

- specific adenine-rich sequences in DNA and RNA have the potential to fold into planar quartets
- one of these quartets has physiological relevance and a quadruplex constructed from it maintained structural integrity to 1.0 µs.
the opportunity for a quadruplex to fold does necessarily commit a complex to arise in vivo, but if so, molecular dynamics analysis indicates it could maintain reasonable integrity

DNA and RNA models are stable to a physiologically relevant timescale, contrasting with an overall understanding that DNA survives much longer than single-strand RNA.

triplanar DNA adenine models do not tolerate Ca$^{2+}$ or H$_3$O$^+$ ions as co-ordinants. Monovalent metal cations and Mg$^{2+}$ are tolerated

Li$^+$ systems are characterised by a stronger network of π stacking in guanine-adenine hybrid quadruplexes.

steric hindrance appears negligible in Li$^+$ co-ordinated guanine-adenine hybrid quadruplexes.

adenine RNA models have a greater population of interplanar π-stacking and water occupancy than DNA equivalents

the glycosidic angle for RNA systems shows less variation than in DNA models

guanine-adenine hybrid quadruplexes show molecular dynamics characteristics similar to uniquely guanine or adenine systems. Biplanar models lacking any cationic co-ordination are stable to 500ns.

concatenating global data for Potential of Mean Force indicates DNA adenine systems preferentially adopt glycosidic angles at ~120°, 0° and ~ -120°. RNA equivalents are similar but adopt less well-defined minima.

non-planarity of bases for participating quartets in adenine quadruplexes is strongly influenced by co-ordinating ion: presence of Li$^+$ maintains structural geometry and integrity.

PMF analysis indicates the N9 environment shifts to a more pyramidal geometry predicated on partial sp$^3$ hybridisation.

the glycosidic angle in DNA and RNA-derived adenine quadruplexes is more susceptible to major changes in orientation compared to the other base-sugar-phosphate angles.

radius of gyration analysis indicates RNA adenine models maintain a steadier degree of compactness compared to DNA equivalents.

RNA systems were found to support a network of hydrogen bonds with less variation than DNA models.

stability of the central quartet in triplanar adenine complexes is influenced by the local geometry: adjacent bases are less stable than those in opposition.

ADN12-N1, ADN12-N3, ARN8-N1 and ARN8-N3 models are characterised by at least two consistent H-bonds. This feature regularly occurs as a direct vertical inter-planar relationship.

models derived from A$_4$-N7 quartets are susceptible to loss of structural integrity regardless of other factors.
• models derived from A₄-N1 quartets are the most stable of the three scaffolds in all metrics and conditions except for presence of Ca²⁺, NH₄⁺ or H₃O⁺.
• all RNA systems are disrupted by inclusion of a co-ordinating ion. Ca²⁺, NH₄⁺ and H₃O⁺ models collapse prior to the end of a simulation.
• ribose N9 determined by partial sp³ hybridisation and possibly the lone pair of an adjacent O4′, slightly displace the local geometry into a pyramidal structure.
• frequency of per base hydrogen bond formation is lower for all ARN8-N1 models co-ordinated by K⁺. Many systems are predicted to have brief periods with no bonding. This feature is infrequently observed in triplanar complexes.
• RNA C2’ OH may promote stability specific to a quadruplex structure predicated on sugar configuration and an enhanced localised H-bond network. This moiety would typically determine hydrolytic susceptibility in isolated RNA.

4.20 Discussion of adenine quadruplexes

The partial in silico substitution of adenine for guanine in the central planes of 1KF1 and 2KF8 does not adversely affect the quadruplex stability in either DNA or RNA systems. This result is qualified however by the actual location of the substituted bases. Dynamic activity is marginally less where adenines are sited in opposition compared to adjacent positions. This may derive from the observation that π-stacking distances are greater and angles generally wider in the latter condition. Results in this study strongly indicate guanine quadruplexes have a higher incidence for frequency of hydrogen bond formation. The enhanced π-bonding predicated on the heterocyclic nature of adenine bases vertically stacked, may offset the inherent lower stability implicit in the reduced hydrogen bond population noted in guanine quadruplexes. Developing a quadruplex with central quartets uniquely comprised of adenines should be at least as viable as a guanine-adenine hybrid equivalent.

The plausibility of this conclusion is reinforced by both factors. All the metrics accessed in this research indicate triplanar DNA and biplanar RNA adenine quadruplexes are generally stable to 500 ns. There are however caveats. Of the three experimentally validated adenine quartets, the A₄-N1 model was favoured as a scaffold. Complexes assembled from three DNA A₄-N1 planes or two RNA analogues consistently showed less dynamic activity compared to A₄-N3 or A₄-N7 systems. Structural integrity was also shown to depend on presence or absence of ionic co-ordination. For DNA models regardless of planar derivation, K⁺, Na⁺ or Li⁺ interacting within the quadruplex, maintained stability resonant with a guanine equivalent. In complete contrast, most of the metrics imply RNA models collapse rapidly if metal cations or non-metal cations are introduced. The presence of interplanar waters may be significant: hydrogen bonding within the reduced space may
partially replace the stabilising influence of a co-ordinant. Steric hindrance could be a factor if ions were imposed.

The wide variation in bonding between N6 of one adenine and H1 of the amino an adjacent base for several models is interesting. Where bases are located as opposites in the quartet plane, the loss of an N6-H1 bond in one base may promote formation in its diametrical complement. Generally, less dynamic variation in N6-H1 distance for a given base was observed compared to the base opposite within the plane. The apparent preference for one hydrogen over the other in the N6 amino group resonates to some extent with differences noted in the effects of N6 methylation on sugar conformations (Section 4.23.3). Tables 4.7 and 4.8 for instance, show conformations dispersed across all variations. Bond angles within the amino group are predicted in the modelling software to vary between on the plane and out of the plane (figure 4.59):

![Figure 4.59: variation of amino angles for N6-methylated adenine](image)

C9-C10-meth, NH2-C10-meth and N5-C10-meth angles calculated for the amino group relative to the heterocyclic plane are very similar for out of plane configuration and therefore tend to confirm this interpretation (figure 4.60):
Furthermore, differences are noted in covalent bonding networks of ADN12-N1 and ARN8-N1. Both systems show a well-defined \(\pi\)-bond network prior to simulation but that for the latter is clearly better served (S4.4 and S4.5). Although the RNA system has lost the majority of its exclusive \(\pi\)-bonds at the simulation end, the structure shows significantly less reconfiguration than the DNA equivalent. Implicitly, an initial robust \(\pi\)-bond network preferentially promotes molecular integrity and reduces or delays loss of structure.

Local minima of PMF torsions data show frequent alternation between -120\(^\circ\), 0\(^\circ\) and 120\(^\circ\) wells and therefore portray the base-sugar-phosphate environment as flexible and dynamic on a nanosecond scale. The data strongly support the conclusion that planarity changes are associated with reversible C2’-endo to C3’-endo configurations and the glycosidic torsion angle. The C3’-endo conformer is strongly associated with a wider O3’-O5’ distance and consequently the phosphate backbone has a less acute helical angle, although both factors may be less than specific. An increased helical twist should induce a slightly enhanced interbase distance. Interestingly, this study found slightly reduced interbase distances implying that smaller torsional angles, particularly the glycosidic bond, may be an off-setting factor.

Quadruplexes constructed from ADN12-N7 or ARN8-N7 show the least structural integrity. This observation may derive from the finding that N7-N6 H-bonds were marginally longer in A2-N7 quartets. As comparison, a 40ns RMSD trajectory for an unfolded 22mer ss-RNA with the same sequence as ADN12-N1, has noticeably higher values than for an equivalent quadruplex. That the geometry may confer reasonable stability is reinforced by visual inspection of the 22mer tending to confirm this interpretation. Additionally, the preference for non-planarity of ribose in biplanar systems derives from a reconfiguration apparently contingent on partial sp\(^3\) hybridisation of N9. There may be an implicit energy dividend in this development.
A specific sequence of adenines (equivalent to the telomeric guanine-rich sequence), has the capacity to self-associate into quadruplexes. This work has demonstrated *in silico* that if such folding occurred, the complex could be non-transient within certain parameters. Hybrid guanine-adenine triplanar DNA structures compare well with uniquely guanine quadruplexes in all metrics used here. There may be further contribution to structural integrity from A-minor motifs. This supramolecular feature supports stabilisation of RNA-RNA interactions by the intercalation of an adenine pair with a G-C pair in an adjacent molecule (187). The ligation occurs within the minor groove of both species via 2'-OH hydrogen bonding.

A steady population of waters in RNA biplanar models and a lower tolerance of zero occupancy might imply inter-planar water to be a stabilising factor given the absence of any co-ordinating ion. Waters may maintain biplanar adenine quadruplexes by mediating a steady inter-planar population and also ligation within the interplanar loops. For triplanar systems co-ordinated with a cation, this may be less of a contributive factor. But for rational drug design, waters peripheral to a quadruplex could beneficially mitigate small molecule interaction.

Ca$^{2+}$/K$^+$ and H$_3$O$^+$/NH$_4^+$ are both isoelectronic pairs. However, their individual impact on the stability of adenine quadruplexes derived from A$_4$-N1 quartets differs. Ca$^{2+}$ models show incomplete trajectories, whereas K$^+$ co-ordinated systems are less dynamic. Hydronium and ammonium models have very similar trajectories. It would seem likely differences of ionic radius for the former pair, and similarities for the latter, may have an explanatory role.

### 4.21 Future work

Guanine-adenine hybrid DNA quadruplexes co-ordinated by K$^+$ and Na$^+$ show larger interaction networks than Li$^+$ systems. However, unfavourable bump interactions are noted in K$^+$ and Na$^+$ models. Hence steric hindrance is implicit. No unfavourable interactions are predicted for the Li$^+$ model. Other metal cations and non-metal cations could be imposed in the quadruplex to elucidate a clearer perspective of ionic radius and charge density as factors in stability. Biplanar RNA systems without ionic co-ordination do not share this characteristic. Further investigation might identify factors that underpin this observation.

Loop sequences have been retained for all models hitherto as concordant with the ...TTA... telomeric sequence. Base substitutions, methylation of existing bases or extending loop length have not been modelled. Moreover, the parallel topology of 1KF1 and 2KF8 loops has also been retained. Modifying these morphologies could impact on function at least
as much as the presence or absence of co-ordinating ions, choice of parent nucleic acid or number of participating quartets.

The A-minor motif proposition noted in Section 4.25, is considered potentially critical to overall compaction of the molecule. The interaction is highly favoured by Watson-Crick base pairing but adversely by non-canonical pairing. An MD investigation might elucidate mechanisms supporting the discrimination.

The minimal survey made here of waters populating the central regions of quadruplexes hints at a more complex and unanticipated interaction network. Further analysis of intra-planar waters should provide insight into the surprising stability of some quadruplexes. It was not established in this study if the population is fluent or more of a permanent feature. The latter scenario would suggest enhanced co-ordinant binding; the former would indicate transient but iterative binding.

MD simulations of both ADN12-N1 and ARN8-N1 models predicted that Ca$^{2+}$ would relocate to an adjacent phosphate niche (S4.6). Whilst it remains in the periphery of the quadruplex, binding appears significantly dynamic as the RMSD trajectory differs greatly from those with other ions ligated. The specific and unique ionic radius and charge of Ca$^{2+}$ may underpin this observation. Expulsion of all ions except Ca$^{2+}$ might also determine some of the eRMSD and RMSD trajectory differentials. Because eRMSD disregards dynamic contributions from ions or waters, if ions are expelled and the phosphate backbone remains relatively integral, interplanar activity should be more noticeable in eRMSD rather than subsumed to some extent by these other parameters in RMSD. However, there is a caveat: eRMSD only evaluates dynamic activity between C2, C4 and C6 of adjacent or paired nucleobases. Hence DNA and RNA models should show similar, or indeed identical, trajectories. Possible confounding factors suggested here include:

- RNA H-bonding between some adenine-base pairing is stronger than DNA. The mechanism for this finding is currently debated (188)
- Because H-bonding strength is intrinsically similar in DNA and RNA, stacking, sugar puckering, solvation may be responsible pro rata.
- RNA models predict a ribose-C2'-endo configuration; DNA structures show a C3'-endo configuration. Sugar-puckering conformers differentially influence the helicity of the ribophosphate chain.
- The mean interplanar distance in RNA models is 3.0 Å compared to 3.3 Å in DNA.
- π-stacking in RNA complexes is more prevalent than DNA models. Moreover, visual observation of both models clearly showed a higher proportion (~15%) of H-bonds and π-stacking in ARN8-N1 than its DNA equivalent.
In addition to base-pairing and interplanar π-bonding contributing to stability, base-stacking in the interplanar loops may also be influential. Although ADN12-N1 and ARN8-N1 both appear to have insufficient space for a cation, a proton may co-ordinate the complex.

Only a brief assessment of methylation was made here and specific to ARN8-N1, ARN8-N3 and ARN8-N7 models. Other modifications such as methylation of inter-planar loops may alter topology and possibly subsequent functions such as access for methyltransferases. The enzymatic modification of adenine to 2,6-diaminopurine may also be worthy of Molecular Dynamics enquiry as future project. Possible relevance to 2,6-diaminopurine in the evolutionary development of Homo is briefly raised in Chapter 7.

Chapters 2, 3 and 4 have described how biplanar and triplanar quadruplexes might fold in specific DNA or RNA sequences. The work confirms and extends current understanding of guanine-based complexes, but develops original proposals of equivalent structures derived from cytosine, thymine and adenine. Analysis of the first two systems does not in general support their structural integrity into serviceable biological times. However, molecular dynamics analysis for adenine quadruplexes, particularly RNA-based, demonstrate stability comparable with guanine equivalents. These latter are now known to persist on a timescale of many minutes. Hence a survey of adenine quadruplexes viewed from the combined perspectives of occurrence, persistence and stability, may be biologically propitious.

Supramolecular structures such as stem-loops or cloverleafs are now known to fold in the genomes of RNA viruses. All are also characterised by polyadenylation at the 3’ untranslated terminus. This poly(A) tail is crucial to a cascade of viral replicative activity occurring both extra- and intra-cellular during infection. Furthermore, this domain comprises a sequence capable of folding into biplanar adenine quadruplexes. SARS-CoV-2 and other mRNA viruses are characterised to some extent by this adenine-rich strand. The possibility of theoretical adenine quadruplexes folding in this region of the viral genome, and guanine quadruplexes in other locations, is addressed in Chapter 5.
Chapter 5. Purine quadruplexes in RNA viruses

5.1 Introduction

Genomic polyadenylation in eukaryota and prokaryota is a stabilising mechanism that sustains the export of mRNA from nucleus to cytoplasm and allows protection from enzymatic disposal (189). The role of this feature, a poly(A) tail, differs distinctly in viruses as evidenced by the expropriation of host biology for solipsistic viral replication (190). The poly(A) tail of SARS-CoV-2 is located downstream of the 3' untranslated region (3'-UTR) in figure 5.1:

![Figure 5.1](image)

**Figure 5.1:** Positive sense single-stranded RNA viruses are capped at the 5' terminus and further stabilised by the addition of a series of adenines at the 3' terminus. Unlike open reading frames (ORF), regions coding for the spike protein and non-specific proteins (nsp), the poly(A) tail appears not to have a direct coding function. S (structural protein), E (envelop protein), M (membrane protein), N (nucleocapsid protein). Accessory factors are proteins with poorly understood functions but have an overall impact on pathogenesis. Their locations are interspersed amongst the structural proteins. Non-specific proteins have a range of enzymatic activity.

The inherent sequence of adenines in the poly(A) tail could allow quadruplexes to fold as described in the previous chapter. If such events were to occur, their impact on viral architecture may be significant. Furthermore, although this Chapter focusses on coronavirus variants in particular, other RNA viruses are typified by polyadenylation and include Polio, Hepatitis C and HIV. Perusal of theoretical interactions between this genomic element and supramolecular purine structures may elicit prospects of chemotherapy. Accordingly, an investigation should start with a survey of the polyadenylation process in RNA viruses and its purpose in their replication.
5.2 Genomic 3’ polyadenylation

Addition of an adenine series to the 3’ terminus of a nascent mRNA strand occurs via two events (189). Initially, a short (~10 nts) tail is synthesised from polyadenylases and a hexameric sequence. This brief strand is thought to act as a template and prompts the emerging poly(A) tail to add more bases (191) (192). The hexameric sequence, AAUAAA, is located in the 3’UTR between 10 and 30 nucleotides (nts) upstream of the primary adenine in the poly(A) tail. Research now confirms the hexamer is responsible for triggering the addition of bases to the 3’ terminus (193) (189) (194). Very similar sequences are known (eg. AGUAAA, UAUAAA or AUUAAA), with an equivalent function but are far less prevalent and may not be exclusive to eukaryota (195) (196) (197). The location of the hexamer also varies slightly with viral species but is always just upstream of the poly(A) sequence. Implicitly, function is linked to location. Whilst the hexamer triggers polyadenylation, the poly(A) tail itself appears contributive to the process as well (198). Polyadenylation appears to arise where the hexameric signal is absent if a short poly(A) sequence is present. This strand can derive from host cytoplasmic poly(A)-polymerases. Moreover, absence of either polyadenylation or hexameric signal may not be entirely disruptive as one element can compensate for loss of the other (199).

Despite these validated conclusions, the process of polyadenylation has as yet no definitive mechanism. Iterative addition of adenines could arise from a form of constructive resonance: an abbreviated sequential ‘stutter’ (200). Regulation in the length of the poly(A) tail may have some impact on its own synthesis. Variation in the adenine population appears to increase early during infection, but reverts to smaller numbers in later stages (201). This interpretation resonates with the observation that efficiency of translation is proportionate to extent of polyadenylation (202). The proposal noted earlier that the hexamer and poly(A) tail are synergistic only on occasion, might be rationalised by matching their binary contributions with the varying degree of polyadenylation.

The viral poly(A) tail further differs from eukaryota (section 5.1) as it participates in replication as well as translation (203). There are minor differences between viruses for the specific mechanism of synthesis (204). Poliovirus for instance, relies on a negative-strand template for polyadenylation (205) and the hexamer signal is involved in negative-strand synthesis in addition to polyadenylation of the bamboo mosaic virus (206).

Export of mRNA is directly contingent on correct synthesis of the polyadenylated sequence and therefore has consequences for stabilising the strand and ensuring effective translation (207) (208).
5.3 Function of the viral poly(A) tail

5.3.1 Introduction

The 5’cap and the polyadenylated adjunct to the 3’UTR are crucial in viral mRNA synthesis and subsequent replication (209). These processes are regulated by adenine N6-methylation and interaction with a poly(A) binding protein (PABP). Incrementally lengthening the poly(A) sequence disproportionally increases negative strand synthesis in some species (eg. Poliovirus). Deletion of bases has been shown to significantly reduce infectivity, replication activity (210) and incomplete adenylation impacts adversely on translation in hepatitis C (211) (212).

5.3.2 Viral responses to host cell destruction of the poly(A) tail

Given the 5’cap and the poly(A) tail of RNA viruses are crucial to their integrity yet susceptible to loss of function, natural selection has favoured protective mechanisms (213). Host cell degradation of RNA is effected via specific proteins (214), or elements of host cell RNA synthesis (215). Covalent polyadenylation is probably one defence against host cell nucleases in parallel with repair mechanisms (216). Within marginal inter-species variation, poly(A) tail repair in RNA viruses is highly conserved and hence indicative of its importance.

Recovery of a deleted poly(A) tail appears to depend on the hexameric trigger (AAUAAA or a variant) and enabled by a viral polymerase (217). There also may be an adenylytransferase option for some viruses (218). Moreover, some evidence of a poly(U) cascade supports the observation that a host uridine-rich sequence can be co-opted by a virus by alternating between adenine or uridine additions (219). Canonical polyuridinylation mediated by an RNA-dependent RNA polymerase (RdRp) flags a damaged or deleted poly(A) sequence for disposal. If this pathway were sequestered by a viral activity (yet to be determined), adenine substitution for uridine would renovate an inoperative system (figure 5.2):
Figure 5.2: theoretical renovation of viral poly(A) tail targeted for destruction. (a) mature mRNA ready for nuclear export complete with poly(A) tail. (b) incomplete replication or host nuclease disperses adenines (c) host cell uridines added to 3’ UTR terminus via polyuridine polymerase (pUp) flag mRNA strand for disintegration (d) host pUp may either be adapted by an unknown viral mechanism to substitute adenines for uridines, (e) or viral RNA-dependent RNA polymerase (RdRp) may refresh polyadenylation activity.

5.3.3 Implications of viral poly(A) tail proliferation

These observations suggest genomic devaluation would ostensibly compromise viral replication but could allow persistence of population. Consequently, natural selection might favour their proliferation. Viral pathogenesis would therefore be greatly reinforced by this latitude in genomic repair. Viral replication may also be influenced by direct interaction of hairpin loops in the 3’ UTR, but translation and RNA polymerase were not obviously affected \([220]\). This enhanced opportunity for non-covalent activity might be an attractive binding site for viral or host cell RNA polymerases. Furthermore, the hairpin loop interaction appears to be promoted by the poly(A) tail. Hence, the poly(A) tail originates as a significant target for intervention. The possibility of supramolecular complexes folding in the structure, and methylation at adenine N6 (m6A,) may outline a route to chemotherapy.

5.4 Adenine N6-methylation in the polyadenylated viral genome

Methylation of N6 adenosine is a dynamic process and may be reversible when mediated by two enzymes: N6-methyladenosine methyltransferase and m6A-demethylase \([221]\). A second mechanism has been proposed which progresses the transfer via intermediates
and is considered irreversible (221). The mechanism may be specific to eukaryota. Supplementary material, Images S5.2a and S5.2b illustrate the competing proposals.

As noted in section 5.3.1, regulated poly(A) tail length may function in translation regulation, as it has been experimentally demonstrated that a longer coronavirus poly(A) tail is associated with better translation efficiency (68). Proposed here is the possibility of a role for methyltransferase inhibition in some stages of this cascade. It is worth noting that several small molecule inhibitory drugs are currently available in the treatment of some cancers (222). For instance, acalabrutinib suppression of kinase is indicated for long-term treatment in chronic lymphocytic lymphoma.

N6-methyladenosine modification of eukaryotic mRNA is ubiquitous and well-characterised, but less well understood in viruses. Wider exploration of this modification in viral genomes could signpost a better understanding of viral replication and pathogenesis. Improved understanding of a N6-methyladenosine role in viral replication may open routes to therapeutic intervention (223). Recent research anticipates a wider role in N6-methyladenosine and treatment of cancers (224).

The poly-adenine terminal of eukaryotic genomes comprise ≥ 50% m6A bases which are demethylated from the poly(A) tail prior to de-adenylation and mRNA degradation (44). The percentage in viral RNA may be slightly lower (225) although consensus is lacking: one study estimates ~15% (226) contrasting with none observed in another (227). Deletion of a sequence coding for a glycoprotein located in the 3'untranslated region (3'UTR) is responsible for this removal. In SARS-CoV-2, the 3’UTR begins after ORF16 (figure 5.1), and the base numbers 29 673 to 29 903 include the poly(A) tail.

The 5 untranslated region (5’UTR) serves as leader sequence, directly upstream from the initiation codon, and is important for the regulation of transcription. Viruses, prokaryotes and eukaryotes have differing mechanisms for this aspect of mRNA. In SARS-CoV-2, the 5’UTR is upstream of ORF1, ORF3a, ORF6, ORF7a ORF8, ORF10 and two structural proteins. Positive selection of ORF3a and ORF8 genes drives evolution of the virus SARS-CoV-2 (228). Mutation rates are low: genetic variation does not appear to have significant impact (229).

The importance of RNA m6A modification has become widely supported as research progresses (230) and, for viral genomic RNAs, has long been known as extensive (231). Showing strong similarities to SARS-CoV-2, mRNAs of influenza A virus (IAV) and adenovirus (AV) have numerous m6A residues (232) (233) (234). IAV was the first reported virus to locate m6A mRNA sites (235), the modification regulating modified RNAs via multiple stages (236). As opposed to mRNAs used only for translation, polyadenylation of mRNA in retroviruses viruses may be involved in both translation and replication (43) (237). Methylation at cytosine C5 (m5C) in mRNA appears not to be relevant in replication (238).
Because internal m6A modification can alter local RNA structures, for example, by destabilising RNA helices (239) and contributes to partially unwinding mRNA (240), it is suggested here that the coiled RNA genome of SARS-CoV-2 may be unwound by this mechanism. Interestingly, some viral dsDNA genomes code for a helicase which unwinds guanine supramolecular complexes (241). It is further proposed here that simple adenine supramolecular structures may fold in the poly(A) tail. Algorithms written to identify potential sequences and locations for adenine quadruplexes to fold are described in Section 5.6.2.2.

It has been suggested that the human metapneumovirus (HMPV) genome for example, may have acquired m6A modification to avoid immune system detection by mimicking host mRNA (227). Indeed, among more than 150 known nucleotide modifications, m6A modification is the most prevalent, being involved in RNA metabolism, protein translation, embryonic development and pathogenesis (242) (243). The modification may also be a biomarker for cellular discrimination between self- and non-self RNA (244).

N6 methylation has a general destabilising effect on duplex base-pairing but loss of m6A in cellular single-strand RNA is linked to loss of structural integrity (245). The modification has a strong association with unpaired nucleic acids (246). Other roles for m6A amongst the numerous unknown RNA modifications have been identified (239). In duplex DNA however, guanine quadruplexes are often located in hypomethylated regions; the link appears to derive from structural factors per se, not population (98). Future work might establish whether this finding persists in ss-RNA adenine quadruplexes.

In mRNA, N6-methyladenosine modification is signalled by a 5-base sequence, RRACH where \( R = A \text{ or } G \), \( H = A, C \text{ or } U \) (247). Considering the motif is so abbreviated, the SARS-CoV-2 genome of ~30 000 bases should be widely populated with m6A-preferred sites. Table 5.1 lists the 10 most prolific sequences identified in this study:
Table 5.1: N6-methylation sequences for adenine in SARS-CoV-2 mRNA. A detailed discussion of the items denoted with an asterisk is covered in Chapter 7.

<table>
<thead>
<tr>
<th>motif</th>
<th>total</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>aaaa</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>aaga</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>aacc</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>gaaca</td>
<td>35</td>
<td>*HAR1a</td>
</tr>
<tr>
<td>ggaca</td>
<td>34</td>
<td>*HAR1a</td>
</tr>
<tr>
<td>agagg</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>agagc</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>gaacc</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>ggagg</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>gaacg</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Variants of the methylation sequence are ubiquitous in the SARS-CoV-2 genome, but perhaps surprisingly, it is well populated with the AAAAA variant. Only six are located with the poly(A) tail. 37 variant sequences are found within ~60 base pairs of the poly(A) tail (for motifs agagc and ggagg), which represents a higher local proportion than for all variants at other locations. This supports the notion that m6A in the poly(A) tail is beneficial to the virus as multiple opportunities are available for the modification.

Recent research has established the relevance of m6A modification with frequent observations in transcripts (248). RNA methylating enzymes are more prevalent than demethylating agents (230). Some methyltransferases are strongly conserved and selectively methylate GAC or AAC motifs in single-stranded RNA (236). A specifically viral binding sequence has been identified as GGACU (249). In this study, Sequence Manipulation Suite (section 5.6.2.2), found 28 locations in the SARS-Cov-2 genome, with a majority in the 5’UTR coding for ORF1a and three spike protein peptides.

From studies of other viruses, it seems likely mRNA-methyltransferases derive from expression of a viral gene rather than host (250). In eukaryotes, expression is mediated by METTL3 and METTL14. Identifying a genetic equivalent for SARS-CoV-2, or elements in the downstream cascade would be a significant target for intervention. There are several unexplained gene sequences in SARS-CoV-2, although nsp9 may mediate replication by an unknown mechanism (251). However, comparisons with archived methyltransferases in PDB suggest none is specifically identified as a methyltransferase. This strongly implies N6-methyladenosine is methylated via an accessory factor (figure 5.1). Effectively, SARS-CoV-2 supplies all genetic information for its synthesis. There is still current uncertainty...
regarding the level of m6A in the poly(A) tail: one study estimated ~15% (226) contrasting with 50% in another (44).

N6-methylation was found to be only marginally disruptive in RNA models as opposed to significant disruption in adenine and guanine DNA systems (Section 4.10, Chapter 4). Therefore, it is feasible that steric hindrance and / or reduced H-bonding from increased inter-base distances may be less significant in these systems. Because N1-N6 distances for ARN8_N1 averaged 5.6 Å (Supplementary Material, N1-N6 distances, S4.383 – S4.389), limited steric hindrance from N6 methylation would seem to underpin this observation. For the ARN8-N7 model, N7-N6 H-bonds were predicted to be longer which suggests a marginal contribution to reduced stability rather than steric hindrance directly. ARN8-N3 has the largest interbase distances and histogram analysis shows the highest RMSD degree but with the narrowest range. Because steric hindrance in this case does not seem likely due to N6 methylation, weaker H-bonding is a more plausible factor (Section 4.10.6).

5.5 RNA virus infection cascade

5.5.1 Introduction

Polyadenylation of the viral 3’ UTR terminus initiates a series of integrated molecular players within the host cell upon cellular endocytosis which eventuate in an overwhelming immune response of interferon. This cytokine is crucial to an early host cell reaction and rapid wholesale influx to an infected site often manifests as inflammation. It is exclusive to vertebrates. The host antiviral default status of immune strategy relies heavily on limiting the spread of infection even after localised viral replication. General evasion of the response involves disruption of inter- and intra-cellular communication, but for coronaviruses in particular, also deception.

5.5.2 Polyadenylation in RNA viruses and the host cell interferon response

Coiled viral mRNA is unwound via m6A modification (237) (241). Steric hindrance would be a feasible mechanism in this process. The poly(A) tail acts as a template for a poly-uridine sequence. This is cleaved by two viral endonucleases derived from non-specific proteins nsp14 and nsp15 (252) into two uneven fragments, both of which are exported (226). The generalised cellular events are outline in Figure 5.3:
Figure 5.3: the poly(A) tail is crucial to RNA virus replication and suppression of the interferon early-immune response. The poly(A) tail acts as a template coding for a polyuridine chain which is then cleaved by non-host endonucleases into halves of unequal length. The shorter half is exported from the virus and subsequently regarded as a pathogen-associated molecular pattern (PAMP) because it initiates a cascade of host cell immune signalling. Viral identity can be effectively masked by an excessive interferon response. Graphic created in BioRender.

RIG-1 (Retinoic acid-inducible gene 1) and MDA5 (Melanoma Differentiation-Associated protein 5) recognise specific dsRNA sequences as viral or otherwise invasive (253). These sequences or structures are collectively referred to as pathogen-associated molecular patterns (PAMPs). The longer poly-uridine fragment folds into a stem-loop and is also identified by RIG-1 and MDA5 as a PAMP (237). The receptor / PAMP interaction activates a mitochondrial-associated signalling protein (MAVS) located inter alia on the outer membrane of host cell mitochondria (254). MAV activation upregulates production of transcription factors, particularly the interferon early-immune response cascade. Hence apoptosis is induced in virally-infected host cells. Phosphorylation may also have influence in suppressing MDA5 function since the primary ligand for RIG-I is a 5’ triphosphorylated or diphosphorylated single-stranded RNA (ssRNA) (241). The SARS-CoV-2 interferon cascade is further detailed in Supplementary material, Images, S5.1a, S5.1b and S5.1c.
The 5′ cap structure of eukaryotic ‘self’ mRNA is typically methylated at guanine N7 and ribose-2′-O positions and thus is not detected by RIG-I or MDA5. However, mRNA lacking ribose-2′-O methylation is recognized as non-self RNA. Therefore, many RNA viruses have evolved their own mRNA-cap methyltransferases to acquire ribose-2′-O methylation to avoid detection by the host innate immune system. Interferons, secreted from infected cells, are the hallmark of this innate immune response, and subsequent viral replication is compromised when viral mRNA or a poly-uridine fragment is detected. Further recognition of an invasive species is hampered by m6A methylation. The modification shields a virus from correct RIG-I activity. This proposal is reinforced by the observation that deficiency of the modification enhances the activation of the RIG-I signalling pathway, including RIG-I expression, RIG-I binding affinity, RIG-I conformational change and an interferon regulatory cascade (227).

Viral genomic sequences allowing potential guanine quadruplexes to fold are highly-conserved and are associated with some regulatory activity. Key examples include guanine quadruplexes as a countermeasure to the host’s immune system or guanine quadruplex-guided regulation of replication or transcription (241). These structures are known to fold in the hHerpes Simplex virus with a maximum abundance during viral replication within the host nucleus (46). Given the proportion of viral to host guanine quadruplex population may be several orders of magnitude (47), this suggests promising selectivity as only infected cells will have greatly increased numbers of guanine quadruplexes. Furthermore, viral guanine quadruplexes are known to regulate regulation for some viruses during lytic and latent conditions (47), hinting at the intriguing possibility of intervention at all stages in the viral life-cycle. These observations recommend further assessment of supramolecular structures in the genome of RNA viruses. Moreover, the possibility arises that non-transient adenine quadruplexes could also fold in the unwound 3′UTR and poly(A) tail of coiled mRNA. However, because this region is a template for a poly-uridine sequence crucial to the downstream cascade, such complexes folding in single-strand mRNA could compromise the template function.

5.6 Purine quadruplex assembly in RNA viruses

5.6.1 Introduction

The viral poly(A) tail appears to be not susceptible to significant mutation (255) implying potential A4-folding sequences have not been compromised. In turn, a well-conserved region suggests a non-adverse impact on the virus. Furthermore, genomic stability could improve the efficacy of a vaccine.

Viral A4s are only implicit because they may not persist, or perhaps fold uniquely at a defined stage in the replication cycle, eg. to stabilise the poly(A) tail template prior to, or
during, cleavage by an endonuclease. Because coronavirus mRNA is translated directly from positive strand viral RNA, the replication process is cytoplasmic (256). As such, a host cell nucleus cannot instruct the addition of adenines to cap the viral mRNA and consequently the virus itself must code for polyadenylation. There are two steps in translation: after initial production of an RNA polymerase, structural, non-structural proteins and the poly(A) tail derive from phase II. Viral assembly then continues in the rough endoplasmic reticulum.

Cytosine methylated at C5 is ubiquitous in cytosine-guanine dinucleotides of the vertebrate genome (257). The modification is generally disruptive. Sequences of cytosine-guanine dinucleotides up 1000 base-pairs in length are characterised by a high prevalence of guanine and cytosine composition notably without C5-methylation. Since guanine quadruplexes in these sequences, CpG islands, of telomeric dsDNA inhibit disruptive methylation (98), the possibility arises adenine quadruplexes may have a similar function in SARS-Cov-2 mRNA. However, there are several reasons adenine complexes have not been observed in vivo:

- folding may be a short-term event occurring at a specific stage in the replication cycle.
- if adverse, mechanisms exist for their removal
- there has been no previous proposal or investigation
- they do not exist

5.6.2 Relevance of biplanar quadruplexes to SARS-CoV-2

5.6.2.1 Introduction

To investigate the potential of purine and pyrimidine quadruplexes folding in RNA viruses, the SARS-CoV-2 complete genome for this study was taken from the National Centre for Biotechnology Information GenBank (NCBI Reference Sequence: NC_045512.2). Biplanar systems were preferentially selected because of their stability as indicated by the RNA models described in Chapters 2, 3 and 4. The SARS-CoV-2 reference genome is ~30 000 bases comprising ~60% adenine / uracil, ~40% guanine / cytosine, and terminating in a poly(A) tail comprising ~35 adenines.

5.6.2.2 Materials and Methods

Algorithms of nucleotide sequences that have the potential to form simple biplanar guanine quadruplexes, cytosine quadruplexes, adenine quadruplexes and uridine quadruplexes were written:
where bases within square brackets can be cytosine, guanine, adenine or thymine (uridine in RNA models). The algorithm would provide a biplanar system with four pairs of identical bases linked by three interplanar loops with up to four bases in each.

An algorithm for biplanar quadruplexes derived from alternating adenine and guanine was written:

- \( \text{ag[cagt]} \text{ag[cagt]} \text{ag[cagt]} \text{ag[cagt]} \)

additionally, to assess the prevalence if any, of planar adenine quartet or a planar guanine quartet in the viral genome, algorithms which could fold accordingly, were also written:

- \( \text{a[cagt]} \text{a[cagt]} \text{a[cagt]} \text{a[cagt]} \)
- \( \text{g[cagt]} \text{g[cagt]} \text{g[cagt]} \text{g[cagt]} \)

A recent study applied more complicated patterns (258). Algorithms designed to identify triplanar guanine quadruplexes, were:


DNA Pattern Finder of the Bioinformatics Sequence Manipulation Suite was applied to scan the genome for the selected algorithms (259). The results were plotted using a scatter graph with x and y values corresponding to the start and end base numbers respectively.

5.6.2.3 Results

No patterns for cytosine, thymine or uridine planar quartets or biplanar quadruplexes were identified, but a preponderance of adenine quartets and quadruplexes was noted (Table 5.2):
Table 5.2: The number of locations identified for potential supramolecular complexes and purine and pyrimidine quartets in SARS-CoV-2

<table>
<thead>
<tr>
<th>structure</th>
<th>number of locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>adenine quartet</td>
<td>271</td>
</tr>
<tr>
<td>guanine quartet</td>
<td>71</td>
</tr>
<tr>
<td>thymine quartet</td>
<td>0</td>
</tr>
<tr>
<td>uridine quartet</td>
<td>0</td>
</tr>
<tr>
<td>biplanar adenine quadruplex</td>
<td>37</td>
</tr>
<tr>
<td>biplanar guanine quadruplex</td>
<td>4</td>
</tr>
<tr>
<td>biplanar adenine-guanine quadruplex</td>
<td>9</td>
</tr>
</tbody>
</table>


The Bioinformatics Sequence Manipulation Suite (BSMS) located a significant majority of potential adenine quadruplex-folding sequences (PA4s) in the poly(A) tail as anticipated (labelled as multiple in figure 5.4). Detailed BSMS information with notes for adenine and adenine-guanine algorithms is filed in Supplementary material, Images, S5.3a and S5.3b. The software also identified 14 PA4s, of which sequences 11069, 11072, 11177 are in a region coding for structural proteins involved in viral replication (260) (261).
Figure 5.4: predicted locations of potential bilaterally adenine quadruplex-folding sequences in the SARS-CoV-2 reference genome identified by plotting start base number against end base number.

The sequence at 25755-25765 occurs in the ORF3a region carrying a nucleotide variation which may have an impact on the clinical outcome. The study comparing this variation with clinical outcomes is an accelerated article preview(229) and notes that genomic mutation is predicated on expression of ORF3a and ORF8 genes.

Figure 5.5: graphic representation of data in figure 5.4. PA4s for bilaterally adenine quadruplexes folding show some traits of grouping. Two sequences are located immediately prior to the poly(A) region.
The specific locations and protein function of predicted adenine quadruplexes in SARS-CoV-2 are itemised in Table 5.3. Complexes potentially folding in the poly(A) tail are not included.

**Table 5.3 description of potential adenine quadruplex locations**

<table>
<thead>
<tr>
<th>location start</th>
<th>gene</th>
<th>protein / function</th>
</tr>
</thead>
<tbody>
<tr>
<td>8465</td>
<td>ORF1b</td>
<td>nsp3 9-subunit complex crucial to replication</td>
</tr>
<tr>
<td>11069</td>
<td>ORF1b</td>
<td>nsp6 facilitates assembly of replicase proteins</td>
</tr>
<tr>
<td>11072</td>
<td>ORF1b</td>
<td>nsp6</td>
</tr>
<tr>
<td>11177</td>
<td>ORF1b</td>
<td>nsp6</td>
</tr>
<tr>
<td>14145</td>
<td>ORF1ab</td>
<td>nsp12 RNA-dependent RNA polymerase</td>
</tr>
<tr>
<td>21567</td>
<td>S</td>
<td>spike protein: receptor recognition</td>
</tr>
<tr>
<td>21571</td>
<td>S</td>
<td>spike protein</td>
</tr>
<tr>
<td>22001</td>
<td>S</td>
<td>spike protein</td>
</tr>
<tr>
<td>25755</td>
<td>ORF3a</td>
<td>ORF3a protein: ion channel activity / replication / pathogenesis?</td>
</tr>
<tr>
<td>26297</td>
<td>ORF4</td>
<td>envelop protein: viroporin</td>
</tr>
<tr>
<td>27592</td>
<td>ORF7a</td>
<td>ORF7a protein: transmembrane protein</td>
</tr>
<tr>
<td>27595</td>
<td>ORF7a</td>
<td>ORF7a protein</td>
</tr>
<tr>
<td>28989</td>
<td>ORF9</td>
<td>nucleocapsid protein: antiviral antagonist</td>
</tr>
<tr>
<td>29387</td>
<td>ORF9</td>
<td>nucleocapsid protein</td>
</tr>
</tbody>
</table>

Two PA4s are located in a region coding for nucleocapsid phosphorylation, three for surface glycoprotein coding and five for ORF1ab polyproteins. PA4s located in the accessory factors’ region have no known association with protein coding.

Locations for PA4s in SARS-CoV-2 variants MW185823.1 and MW185825.1 were plotted (figure 5.6). The data are similar to the reference genome except for a lower population in the overlap between ORF1a and ORF1b. There are also three locations for the variants in the early ORF1a region that do not appear in the reference genome. Variants and reference models match with the highest populations located in the structural protein region, accessory factors and 3’ UTR. The poly(A) tail has an identical structure for all three models and consequently supports the same population of A4s.
Although the number of locations is similar for both variants, with the majority in regions of overlap, MW185823.1 has at least two more PA4 sequences.

5.7 Guanine and adenine-guanine hybrid quadruplexes in SARS-CoV-2

5.7.1 Introduction

Prior to the rapid onset of the 2020 pandemic, research into coronavirus guanine quadruplexes received limited attention. Although potential folding sequences were recognised, no supramolecular structures per se were identified (47). Evidence for G4s in SARS-CoV-2 developed from research contemporary with the pandemic and focused on one of four sequences (262). A stable G4 located in the nucleocapsid phosphorylation protein was verified in vivo. A single G4 was subsequently identified by the same team in nsp10 and two in the spike protein. Potential sequences for a further three complexes were noted in ORF1a (258). Locations, proposed or definite, are shown in Figure 5.7:
Figure 5.7: identified biplanar RNA G4s (blue) and putative G4s (green) in SARS-CoV-2.

The search algorithm for the earlier study was: $gg[xxx]gg[xxxx]gg[x]gg$, and for the subsequent study: $gg[xxxx]gg[xxxxxxxx]gg[xxxx]gg$. The latter pattern is unusual as the central loop would comprise eight nucleotides.

5.7.2 Results

Four sequences allowing biplanar RNA guanine quadruplexes to fold were identified in this study. These are plotted in figure 5.8 with the currently identified complexes described in figure 5.7 excluded.

Figure 5.8: predicted locations of potential biplanar guanine quadruplex-folding sequences in the SARS-CoV-2 reference genome. Sequences’ starts identified at 20111, 22102, 24977, 25085.
All potential biplanar guanine quadruplexes identified in this study arise in the region for coding structural proteins and are in addition to the complexes noted in figure 5.7.

Data for biplanar RNA adenine-guanine hybrid quadruplexes show broadly similar locations to SARS-CoV-2 variant MW185823.1 with the obvious exclusion from the poly(A) tail (figure 5.9):

![Image](image.png)

**Figure 5.9:** predicted locations for adenine-guanine hybrid quadruplexes show similarities with the SARS-CoV-2 reference genome and variant MW185823.1. Sequences’ starts identified at 4107, 7851, 8004, 15011, 15992, 16879, 17347, 19992, 24836.

5.7.3 Discussion

3-D biplanar adenine quadruplexes in the poly(A) tail would fold from a strand which is essentially 2-D. Consequently, surface area availability for any molecular interaction is reduced by such folding. Whereas N6 methylation in the poly(A) tail was previously considered absent (263), more recent research concludes the modification is prevalent in the post-3’ UTR polyadenylation process (44). Whilst this may simply represent a difference between eukaryota and viruses, the disparity is striking. Both m6A in or adjacent to the 3’UTR region, and guanine N7 methylation in the 5’ cap, resemble significant stabilising factors in viruses (264). Access for demethylating enzymes to the central region of quadruplexes is decisively reduced compared to the unfolded nucleic acid strand. Presumably, in the process of polyadenylation, enzymatic demethylation
would also be less effective. Methylation of an adenine-guanine hybrid quadruplex at either adenine N6 or guanine N7 should share similar constraints.

The nucleocapsid protein is considered a probable promoter of G4 folding and repair particularly when co-operating with nsp15 (13). Interestingly, this region was the first to be identified as supporting a G4 sequence. A simple algorithm applied here to identify potential folding sequences determined at least two. G4s in the nucleocapsid gene may be present intermittently as the protein is capable of dismantling it in HIV-1 (265). The interaction allows transcription to proceed.

It has been proposed that RNA quadruplexes arising in nsp3 of SARS-CoV-2 might bind unspecified proteins to avoid the host cell immune response (266).

5.8 Direct visualisation of purine quadruplexes

5.8.1 Introduction

The fluorophore silicon-rhodamine (SiR, figure 5.10) has been covalently linked to a guanine quadruplex-specific pyridostatin derivative PyPDS (figure 5.11) (39). The synthesis of SiR-PyPDS was designed to assess and visualise activity of guanine quadruplexes in vivo without global interference in formation or dynamics of interaction. The methodology successfully achieved this aim.

![Figure 5.10: Silicon-rhodamine (SiR)](image)

A 6-carbon chain linked SiR to PyPDS. Binding of the active form of SiR-PyPDS to a DNA guanine quadruplex is visualised by single molecule fluorescent imaging. The inactive analogue has a very low affinity for quadruplexes registering few false positives.
Figure 5.1: Active (A) and inactive (B) forms of SiR-PyPDS. The silo-fluorophore in the active form fluoresces regularly on ligation with a quadruplex. The inactive form fluoresces occasionally without binding.

This methodology was adapted to explore ligation with adenine quadruplexes and possible subsequent direct visualisation. A small molecule, C5 (figure 5.12), was identified as a suitable ligand since it binds as dose-dependent to a SARS-CoV-2 attenuator RNA hairpin located on the boundary between ORF1a and ORF1b (267). C5 binds to a UUA motif. Inter-planar loops of ARN8-N1 also comprise this sequence of trinucleotides.

Figure 5.12: C5, small planar molecule 3-[[5-(methylamino)-3,4-dihydroquinazolin-2-yl]amino]benzoic acid. $K_D = 11.0nM$.

5.8.2 Proposed method

SiR and C5 are both carboxylic acids and can therefore be linked with a diamine such as 1,6-diaminohexane (figure 5.13A). To avoid complicated purification, one amine group of 1,6-diaminohexane should be temporarily protected to ensure only a single binding reaction with C5 takes place. tert-Butyloxycarbonyl (N-BoC) would suit this purpose (figure 5.13B):
Figure 5.13 1,6 diaminohexane (A) and N-BoC (B). the tert-butyloxycarbonyl compound acts as a protective shell to avoid secondary coupling with the alternative amino group.

In practice, this step could be omitted by replacing both components with N-BoC-cadaverine as this compound has tert-butyloxycarbonyl group linked with an monoaminohexane (figure 5.14).

Figure 5.14: C5 and N-BoC-cadaverine prior to coupling (A) and with an amide link (B)

The available amino group of N-BoC-cadaverine-1,6-diaminohexane group should couple to the C5 carboxylic acid group. If the N-Boc screen were removed with trifluoroacetic acid to reveal the other amine group (C), the same coupling reaction with the carboxylic acid of SiR (D), should yield the target C5-spirofluorene (figures 5.15 and 5.16).

Figure 5.15 Trifluoroacetic acid (TFA) removes N-BoC to reveal a terminal amino group available for coupling

Figure 5.16: C5-spirofluorene (SiR-C5)
The SiR-C5 complex was submitted for docking protocols with ARN8-N1 and analysed in AutoDock and MDS. Five dockings were run for each methodology.

5.8.3 Results

Binding affinities were predicted to be exothermic by both methodologies and therefore thermodynamically favourable. The inhibition constant for each docking was reasonable and RMSD between the optimum configuration and less favourable analogues was minimal (Table 5.3):

**Table 5.3 C5-spirofluorene docking with ARN8-N1.** ADT: Autodock tools, MDS: molecular docking server, $K_i$: inhibition constant, TIE: total intermolecular energy.

<table>
<thead>
<tr>
<th>docking</th>
<th>binding affinity kcal/mol</th>
<th>$K_i$</th>
<th>RMSD (Ång) from optimum</th>
<th>TIE kcal/mol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADT</td>
<td>MDS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-5.5</td>
<td>-4.03</td>
<td>1.11mM</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>-5.3</td>
<td>-4.00</td>
<td>1.16mM</td>
<td>1.014</td>
</tr>
<tr>
<td>3</td>
<td>-5.1</td>
<td>-5.23</td>
<td>1.23mM</td>
<td>3.527</td>
</tr>
<tr>
<td>4</td>
<td>-5.1</td>
<td>-4.19</td>
<td>1.73mM</td>
<td>2.024</td>
</tr>
<tr>
<td>5</td>
<td>-4.7</td>
<td>-4.78</td>
<td>1.47mM</td>
<td>3.157</td>
</tr>
</tbody>
</table>

Reflecting the original experimental synthesis, modelling a 6-carbon aliphatic linker between C5 and SiR achieved an optimum balance for binding energy and number of carbons (Table 5.5). This model has an optimum number of bonds (10) for an acceptable binding affinity (-7.8 kcal mol$^{-1}$) and $K_i$ (1.91 μM).
Table 5.4: binding energies and number of carbons in aliphatic linker.

<table>
<thead>
<tr>
<th>carbons</th>
<th>optimum binding energy kcal mol(^{-1})</th>
<th>Ki</th>
<th>ligand moiety</th>
<th>ligated residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>3C</td>
<td>-7.86</td>
<td>1.74 µM</td>
<td>C37, C38, C33, C34, Spire moiety, Aliphatic linker, C5 moiety H2</td>
<td>A22, A22, A12, A22, A21, A21, A20, A8, A22</td>
</tr>
<tr>
<td>4C</td>
<td>-7.34</td>
<td>4.14 µM</td>
<td>C34, C35, C16</td>
<td>A12, A21, A21, A7, A8, A21, A22</td>
</tr>
<tr>
<td>5C</td>
<td>-8.02</td>
<td>1.31 µM</td>
<td>C40, C35, C35, Aliphatic linker O2, C9-C10-C11-C14-C13 ring</td>
<td>A12, A21, A20, A21, A21, A21</td>
</tr>
<tr>
<td>6C</td>
<td>-7.8</td>
<td>1.91 µM</td>
<td>C16, C20, C1-C2-N1-C4-N2-C3 ring, C5 moiety spiro-moiety</td>
<td>A20 6C, A20 5C, A21 6C, A21 6C, A21 6C, A7, A8, A20, A21, A22</td>
</tr>
<tr>
<td>8C</td>
<td>-6.89</td>
<td>8.90 µM</td>
<td>C37, C38, C23-C24-C25-C26-C27-C28 ring H8</td>
<td>A12, A21, A7, A8</td>
</tr>
<tr>
<td>10C</td>
<td>-6.39</td>
<td>20.79 µM</td>
<td>O4, C5 6C-ring H2, H13</td>
<td>A12, A22, A22, A22</td>
</tr>
</tbody>
</table>

5.8.4 Discussion of direct visualisation

Negative values for total intermolecular energy between C5-spirofluorene and ARN8-N1 reinforce the finding that the target and query molecules have a high probability of a favourable thermodynamic ligation. Interestingly, a noticeable reconfiguration for C5-spiro-fluorene is evident on docking with ARN8-N1. The closely proximate interaction
around an interplanar UUA loop implies a robust ligation. A well-defined visualisation of the ligation is filed as video S5.4 in Supplementary Material, Images. The majority of reconfiguration is derived from a change in the planar structure of C5. The fluorene silicon atom is depicted in light grey.

Figure 5.17: screenshot of C5-spiro-fluorene docking with ARN8-N1. The ligand is predicted to reconfigure in the optimum AutoDock calculation.

Although the in silico modelling, synthesis and docking would benefit from direct experiment, the data are encouraging. Spiro-fluorene labelling a molecule characterised by a high affinity for RNA supramolecular structures supports the notion of direct in vivo visualisation.

5.9 Executive summary

- polyadenylation in RNA viruses could allow biplanar adenine quadruplexes to fold
- a conserved hexameric 3'UTR sequence prompts polyadenylation
- compromising correct synthesis of the poly(A) tail is a possible chemotherapeutic target
- synthesis is regulated by m6A and a poly(A) binding protein
- poly(A) tail repair mechanisms are highly conserved in RNA viruses and by implication, their significance. The feature is also only marginally susceptible to mutation.
• algorithms to identify potential sequences capable of folding into quadruplexes determined locations for a small number of guanine, adenine and hybrid structures.
• no sequences capable of folding into pyrimidine quadruplexes were established
• purine quadruplexes may be intermittent events occurring at specific stages of the viral replication cycle.
• a verified synthesis for visualising guanine quadruplexes in vivo was adapted and proposed here to similarly identify adenine events.
• although an indigenous helicase is now known, other non-host proteins fold mRNA into supramolecular structures. This suggests the complexes are not necessarily adverse.
• several aspects of the molecular cascade for viral polyadenylation and specifically purine quadruplex folding appear promising targets for intervention.

5.10 Discussion

The possibility arises that non-transient A4s could fold in the poly(A) tail of coiled SARS-CoV-2 and other RNA viruses. However, because this region is a template for a polyuridine sequence crucial to the downstream cascade, A4s in single-strand mRNA could compromise the template function. Consequently, the replication cycle may be influenced by this proposal.

Whilst direct evidence for RNA G4s in viruses is currently still putative, there is a strong consensus for the likelihood of their presence (268). There is a broad agreement of RNA G4 potential contribution to viral pathogenesis, but opinions of specifics differ (269) (270). Viral persistence in host cells and avoidance of detection may depend on G4s inhibiting viral proteins, repression of transcription and influence viral replication (271) (272) (273). In SARS-CoV-2, potential G4-folding sequences interact with nsp13, a helicase responsible inter alia for catalysis of G4 unfolding. The complex appears to act as a brake on self-destruction (103). Quadruplex topology and variation in quartet bases may also underpin G4 unfolding (268). It is possible therefore that the points raised in Section 4.3 discussing guanine-adenine hybrid quadruplexes are relevant in this perspective.

A concurrent view of G4s as adverse in the viral replication cycle, points to natural selection. A depletion of G4s from the viral genome might be advantageous (274). Non-transient G4s compromise of replication or transcription would be obviated and viral G4s
may avoid a host recognition cascade by cellular binding proteins. Two consequences may follow the prevalence of G4s and their corresponding folding sequences. A low (or zero) population would facilitate rapid infectivity, but a greater ubiquity would require a more covert strategy of invasion. Such a tactic could resolve persistence of infection and favours a hypothesis of SARS-CoV-2 behaviour (268).

The genomic impact of viral G4s extending to purine hybrids and adenine quadruplex analogues seems plausible in the light of data reported for this thesis. All the observations referenced here suggest allocating an adverse impact for indigenous supramolecular structures to RNA viruses. A feasible corollary would be further investigation as suitable targets for chemotherapy. Chapter 6 proposes small molecule interventions to address pathogenesis credibly associated with DNA and RNA quadruplexes.

5.11 Future perspectives

Some aspects of this study strongly indicate m6A may underpin rational design of enhanced live or attenuated candidates for a vaccine. Theoretical work could overlap with experimental approaches and reinforce such developments.

Greater definition of the actual mechanisms of polyadenylation would be helpful in addition to understanding the variable length of poly(A) tail in any given RNA virus. The hexameric sequence necessary for polyadenylation may work in tandem with the poly(A) tail in a form of self-catalysis. Varying the contribution from either source would test the validity of this notion by establishing the adenine population in the feature.

Since ds-DNA G4s are often located in hypomethylated regions, research might establish if this finding persists in ss-RNA A4s. Future work may focus on experimentally confirming the data presented here and further exploring non-canonical supramolecular structures in cellular and viral contexts. With particular reference to SARS-CoV-2 in silico docking protocols predict G4-targetting carboxy pyridostatin also has a high binding affinity for A4s. Current research elsewhere is exploring the antiviral potential of this molecule.

This Chapter has only considered RNA viruses; a general audit of quadruplexes in other viral species, bacterial genomes and mitochondria appears to be poorly-represented in the literature. Guanine quadruplexes have been intensely investigated, but not exhaustively.
6.1 Introduction

As discussed in Section 2.1.2, G4s arising in duplex DNA are often described as pathogenic in contrast to equivalents folding in single-strand telomeric DNA. Although understanding is far from complete, this paradox extends to RNA complexes. Support for adverse conditions that encourage the development of neurodegeneration has been noted (275) (156). Moreover, potential RNA quad-forming sequences are associated with a specific long non-coding RNA known to be upregulated in colon cancer (276). In contrast, RNA quadruplexes have beneficial associations in telomeric regulation and critical roles in gene expression (277) (278). Furthermore, polyadenylation to 3’ UTR mRNA is essential and directly associated with RNA quads (279). This finding resonates with proposals described in Section 5.7. Consequently, reconciling the apparent pathogenicity of RNA quads with their prevalence in diverse RNAs and their stability is challenging. Hybrid DNA/RNA complexes may underpin the persistence of potentially damaging structures (280). However, there may sufficient differentials in the chemistry of DNA and RNA to resolve the paradox.

Section 2.6 discussed the higher population of waters in the RNA models analysed compared pro rata with DNA systems. In combination with peripheral waters attracted by the additional 2’-hydroxyl groups of ribose, ligation with waters was shown to promote stability of RNA models over DNA equivalents. The presence of 2’-hydroxyl groups located with regular spacing might allow other interactions to enhance structural integrity. Ribose hydroxyls also impose constraints on the overall quadruplex topology. Sections 2.7, 3.9, 3.13 and 4.23 noted the preference for C3’-endo conformations favoured stability for biplanar RNA systems without cationic co-ordination. Uridine bases as opposed to thymidine further endorse stability (281). Given hydration contributes to the demethylation of thymine and consequent conversion to uracil (282), the disproportionate higher occupancy of waters in biplanar RNA systems could minimise uridine methylation.

DNA quadruplexes derive stability from cationic co-ordination paired with the central quartet contributing a greater share than the two outer quartets. Stability in RNA systems is predicated inter alia on an equitable contribution from all quartets (283). Ostensibly, this reflects the structural integrity of RNA models, particularly biplanar quadruplexes, demonstrated by the molecular dynamics data detailed throughout Chapters 2,3 and 4.

These differentials of chemistry signpost possible routes not only to selectivity between DNA and RNA models, but perhaps also to discrimination within a group. The disparity may subsequently extend to selection for viral mRNA. Strategic small molecule
interventions for adverse quadruplexes whilst excluding beneficial systems would be propitious. Rational design of therapeutics is outlined here with elucidation of three theoretical approaches, after a brief survey of three quadruplex-based interactions proposed as mechanisms for cellular compromise.

6.2: Concept of quadruplexes as molecular obstacles

6.2.1 Ribosome pausing.

During translation, mRNA traverses a dedicated ribosomal channel after unfolding the nucleic acid strand. Effectively, the ribosome acts as a helicase (284), but that function can be paused idiopathically (285). mRNA stem-loops have been proposed as candidates for the disruption (286). Since these supramolecular structures act as a physical obstruction, it is feasible that quadruplexes might also compromise ribosomal activity. A temporary quadruplex docking in the aminoacyl site for RNA binding (287) may trigger a pause in protein synthesis. The charged surface availability of a quadruplex compared to an unwound strand implies enhanced electrostatic interaction. The obvious caveat of helicase efficiency (288) dispersing quadruplex occlusion could be countered by noting that even a transient interruption may prompt accumulation of incomplete oligopeptides (289).

In the specific instance of SARS-CoV-2, coronaviruses are characterised by an abbreviated molecular cascade that regulates endogenous protein expression (290). A marked escalation of upregulation is mediated by a pseudoknot comprising three intertwined RNA hairpin loops. The structure is considered unique to SARS-CoV-2. Correspondingly, the outcome of compromising the cascade triggers a downregulation of viral protein expression. Inhibition of the pseudoknot suggests promising therapeutics. The possibility of biplanar RNA quadruplexes also mediating the cascade would similarly identify them as suitable targets for intervention.

6.2.2 Lagging strand cytosine quadruplexes

Telomeres mediate cell-division and are reduced in length during each replication cycle. This problem is resolved during the cell’s natural lifespan by telomerase until apoptosis occurs after 30-50 divisions. However, enzymatic activity can continue if an occasional cell survives apoptosis, triggering runaway replication and tumorigenesis (291). Telomeric maintenance mechanisms are closely associated with underpinning genetic instability (292). Single strand telomeric G4es are associated with telomerase inhibition. However, a guanine complex folding in the leading strand of telomeric duplex DNA will automatically disrupt the regular structure of the lagging strand. The complement distorts to
accommodate the new morphology. Disruption implies genetic problems, but a complex sited in telomeric single strand DNA, has no complementary strand to disrupt. Implicitly therefore, telomeric location dictates character: systems in duplex DNA are associated with neurodegeneration (92), whilst single strand equivalents are beneficial. This latter scenario allows telomerase suppression to proceed with consequent negative impact on tumorigenesis. Unfortunately, the issue becomes less binary when noting that some quadruplexes appear to have opposite effects on the same target (105). G4s located at or near replication forks have adverse associations (293) and hence plausibly interrupt routine progress. In contrast, beneficial suppression of oncogenic transcription has been directly linked to a guanine complex also located adjacent to a replication fork (294).

Section 3.4 described the logical development of cytosine quadruplexes folding in duplex telomeric DNA. The lagging strand *de facto* has a sequence that would allow a cytosine quadruplex to fold (Section 3.1). Therefore, a possible situation arises where a leading strand guanine complex and lagging strand cytosine equivalent are exactly complementary.

To repair terminal DNA lost during cell replication, telomerase binds to single strand telomeric DNA. The enzyme comprises two major components, telomerase reverse transcriptase (TERT) and telomerase RNA (TERC), together with several sub-units. Acting as a template, TERC interpolates the sequence 5’-TTAGGG to the chromosome 3’ terminus (section 2.1.2). The complex relocates downstream by six bases and repeats the addition process. The lagging strand is upgraded to duplex strand by a DNA polymerase. A leading strand guanine complex and a complementary lagging strand cytosine quadruplex may be transient or intermittent (Sections 5.7.3 and 5.9). But the presence of either, separately or together, during TERT telomeric reinstatement could impair the process (figure 6.1):
Figure 6.1: the RNA molecule TERC provides template reconstruction for a telomere reduced by cell division. Presence of guanine or cytosine quadruplexes in either the leading or lagging strands could disrupt the process of telomeric elongation.

If either scenario arises, at least three inferences are possible:

- **Pathogenicity of double strand G4s** may derive from the presence of complementary cytosine equivalents because both strands are corrupted and replication is compromised.

- **Cytosine complexes** are uniquely responsible for pathogenicity. Beneficial effects may still be associated with single strand G4s even though the strand is corrupted but no cytosine complexes is present. If correct, this implies routes to destabilise double strand G4s may be unproductive.

- **Pathogenicity of an anti-sense cytosine complexes** complementary to sense-strand guanine complexes may be more significant if located near a replication fork or promoter sequence. Moreover, a sense-strand (AATCC)n sequence would logically allow an anti-sense G4 to fold.

Telomerase reinstates the reduction of chromosome termini derived from cell division, but routine telomeric maintenance devolves to Shelterin.
6.2.3 Conjectural quadruplex-induced telomeric dislocation of Shelterin

Comprising six subunits, shelterin suppresses inappropriate repair responses and balances telomerase activity (295) (296). Two dimerised subunits, TRF-1 and TRF-2, collaborate as a default mechanism in limiting telomerase extent (297). The TRF-1 complex co-operates with a third subunit, POT-1 in a dynamic protective manoeuvre. The TRF complex can adapt to altered circumstances and facilitate enzymatic function by sequestration of helicases (298). Reduced levels of the TRF complex have negative consequences for the telomere (299). A series of kinetic interactions occur between a TRF-1 or TRF-2 monomer binding at a TTAGGGTTA sequence in the duplex telomeric region. POT-1 binds to the single strand telomere at a TTAGGGTTAG sequence (figure 6.2):

![Figure 6.2: step 1 of shelterin capping end of telomere.](image)

**Figure 6.2: step 1 of shelterin capping end of telomere.** POT-1 sub-unit of the shelterin complex binds to the 3’ single strand overlap near a telomeric terminus. One of two TRF-1 or TRF-2 monomers binds to upstream duplex telomere.

The unattached TRF-2 monomer, in seeking the duplex telomeric sequence, by default folds the region back on itself. Since POT-1 is still attached to the single strand, a loop is completed (figure 6.3):
Figure 6.3: step 2 of shelterin capping end of telomere. The unattached TRF monomer binds to a specific sequence in the duplex telomere just upstream of the 3’ overlap start (A). Consequently, a loop is created with both monomers and POT-1 bound to the telomeric duplex. The chromosome is effectively capped (B).

Both the duplex and the 3’ overlap have sequences conducive to allowing G4s and/or C4s to fold. Even if arising intermittently, presence of either or both could compromise routine shelterin interaction.

6.2.4 Discussion of Section 6.2

The scenarios outlined here are entirely conjectural; the premise that a quadruplex could fold does not mean it will. The appropriate sequences are available, but a signal or mechanism for folding remains anonymous. Moreover, there are noticeable contextual differences between each proposal. Varying levels of repair mechanism advocacy seem likely for the proposals. Nevertheless, these uncertain outcomes do have some commonality. The structures as envisaged are functionally adverse, and ligation of small molecules would be a prior requirement for any form of intervention.

6.3 Challenges of interpreting purine quadruplex contribution to aetiology.

6.3.1 Introduction

Somatic cells show very low levels of telomerase (300) contrasting with over-expression in ~85% of all tumours (301). The possibility of anti-cancer intervention has prompted several therapeutic developments, although some significant disadvantages were noted. For instance, stem cells require the enzyme to survive and stress may relocate telomerase
expression to mitochondria. (302). Furthermore, short telomeres are preferentially targeted, implying chemotherapy is limited to advanced tumours.

Several studies suggest enhancing G4 stability as promising chemotherapy (303). Current research reports telomeric oligonucleotide structures are in development as potential candidates (85). For instance, the DNA of melanoma cells is adversely affected by a duplex G4 homologue folded from an 11-mer characteristic of telomeric DNA (304). The folded double-strand complex was found to be more stable than a comparable single-strand nucleotide sequence but appeared to show less antiproliferative activity. Three reasons are suggested for this observation: the oligonucleotide shows enhanced binding to a protective telomeric protein complex compared to the folded form, subsequently triggering DNA damage. Moreover, because enzymatic activity of DNase-1 can convert quadruplexes into single-strand oligonucleotides, the relative concentration of the two forms varies, making comparison of efficacy challenging. Furthermore, steric hindrance may also be a factor as noted in Sections 2.1.4 and 2.5.1; some telomeric regions could be more available to the oligonucleotide.

Distinguishing between duplex DNA and RNA quadruplexes in drug design has been challenging. However, research addressing the issue of selectivity has progressed with the development inter alia of ICZC-3 (293). Murine studies indicate this novel molecule displays significant preference for G4s in duplex c-Myc promoter regions. Persistent expression of the gene upregulates downstream genes associated with cell proliferation. Whilst stabilising G4s in c-Myc promoters and suppressing transcription is ostensibly propitious, this strategy has not achieved the anticipated success to date (305). The biochemistry underpinning G4 selectivity for melanoma cells and anti-cancer activity is poorly understood. It has been proposed that telomerase per se may be involved (85). Defining the interaction of G4-type structures with telomerase may well reinforce current therapies and improve prognoses. However, progress in development to clinical trials is hampered by the marginal understanding of possible mechanisms.

Challenges also remain in developing modalities to apply different therapies working in parallel to tackle solid tumours. Resonating to some extent with section 6.2.2, telomeric oligonucleotide structures targeting hTERC, telomerase expression and activity might be additional future chemotherapeutic options (306).

G4s are now considered to populate most guanine-rich regions, having been routinely described as unique to single-strand telomeric DNA (76). Transient separation into a single strand, for instance at replication forks (Section 6.2.2) or prior to DNA repair (Section 6.2.3), may be a pre-requisite of G4 folding. G4-putative sequences may inhibit methylation in locations specifically associated with gene expression (307). Some high-throughput sequencing suggest there may be ~700 000 human-genomic potential G-rich regions available (308).
Section 4.26 summarised results that indicate RNA A4s were less susceptible to dynamic activity than DNA equivalents. Marginally reduced inter-planar distances and enhanced non-covalent bonding probably underpin this observation. Moreover, the ribose 2’-hydroxyl group may have direct influence on the enhanced stability of RNA models due attracting waters into the system. Additionally, steric constraints on the quadruplex topology may derive from 2’-OH and reduced lability. Furthermore, the C3’-endo configuration is imposed exclusively upon RNA sugars and the consequent general influence on reducing large fluctuations in the ribophosphate angles. Interplanar distances can be diminished as a consequence. Possible pathogenic aspects of RNA quadruplexes continue to be reported, including tumorigenesis, neurodegeneration and cardiac dysfunction (28). Section 5.7.3 noted a viral strategy to evade detection centred on endogenous RNA quadruplexes.

Contrasting significantly with G4s, stability of A4s does not appear to be contingent on coordinating cations. Interplanar loop length is known to influence G4 stability and may be factor in A4 integrity. Furthermore, loop direction in G4s has a range of possibilities but remains to be evaluated for in silico A4s. N6-methylation destabilised G4s and all DNA A4 models to a greater or lesser extent probably predicated on steric hindrance. Antithetically, m6A-signalling sequences in the mRNA poly(A) tail were found to be ubiquitous (section 5.4). Implicitly therefore, the modification is pragmatic. Methylation of loop bases was not explored in this study.

6.3.2. Vaccination

There are numerous research efforts globally to exploit viral susceptibilities and mass-screening to identify current chemotherapies serendipitously targeting them (309) (310) (311). Of lead candidates that reach trials, the overall probability of success is 33.4% (309). Given the global imperative of vaccine development, scaling up to supply widespread demand currently appears possible in less than the average timeline of 10.7 years (312). Achieving the stated target of ~18 months suggests that later stages of development might commence whilst previous ones are still in assessment. This could have implications for safety (313). Market entry hitherto is years, not months (314). If a non-Phase III-approved candidate had even very low adverse effects, popular opinion may greatly reduce widespread acceptance of a properly approved vaccine. Accordingly, limiting infection is a reasonable strategy and potentially achievable by several routes such as vaccination running concurrently with inoculation or repurposing current drugs. Other non-chemotherapeutic interventions include quarantine, widespread reliable testing and effective contact tracing. Viral mutation may need to be accounted for, there might also be unknown long-term outcomes and length of significant immunity has yet to be
established. Furthermore, vaccination may not be suitable in specific circumstances such as allergic reaction or pregnancy.

Recent research concludes antibody prevalence declines noticeably with time, implying herd immunity could be difficult to achieve (315). This finding also suggests the probability of increased re-infection. Some current vaccine preparations are not protein-based, but derive from inactivated or attenuated viral RNA. Synthesis of protein-based vaccines is slow and may be much later reaching approval than RNA-based vaccines (316) and there are several uncertainties in the mechanisms of viral infectivity (311) (317).

Vaccination remains a primary route to confronting infection. However, regular booster inoculation and emergence of new strains have consequences for modified vaccines. Effective immunity may be further confounded by scaling to widespread availability (318). A reasonable option therefore is to mitigate the virus’ effects (319) with rational design de novo of small molecule intervention. m6A modification may underpin rational design of enhanced live or attenuated candidates for a vaccine.

6.3.3: Targeting purine quadruplexes as chemotherapy

Whilst direct evidence for RNA G4s in viruses is currently still putative, there is a strong consensus for the likelihood of their presence (268). There is a broad agreement of RNA G4 potential contribution to viral pathogenesis, but opinions of specifics differ (269) (270). Viral persistence in host cells and avoidance of detection may depend on G4s inhibiting viral proteins, repression of transcription and influence viral replication (271) (272) (273). In SARS-CoV-2, potential G4-folding sequences interact with nsp13, a helicase responsible inter alia for catalysis of G4 unfolding. The complex appears to act as a brake on self-destruction (103). Quadruplex topology and variation in quartet bases may also underpin G4 unfolding (268). It is possible therefore that the points raised in Section 4.3 discussing guanine-adenine hybrid quadruplexes are relevant in this perspective.

A concurrent view of G4s as adverse in the viral replication cycle, points to natural selection. A depletion of G4s from the viral genome might be advantageous (274). Non-transient G4s compromise of replication or transcription would be obviated and viral G4s may avoid a host recognition cascade by cellular binding proteins. Two consequences may follow the prevalence of G4s and their corresponding folding sequences. A low (or zero) population would facilitate rapid infectivity, but a greater ubiquity would require a more covert strategy of invasion. Such a tactic could resolve persistence of infection and favours a hypothesis of SARS-CoV-2 behaviour (268).

Microbial nucleic acids are important activators of the innate immune response by triggering pattern recognition receptors. Telomeric G4s in non-viral DNA have been investigated for some time. Viral genomic sequences allowing potential G4s to fold are
highly-conserved and are associated with some regulatory activity. Key examples include G4s as a countermeasure to the host’s immune system or G4-guided regulation of replication or transcription. The genomic impact of viral G4s extending to purine hybrids and adenine quadruplex analogues seems plausible in the light of data reported for this thesis. The complexes may also contribute to strategies evading host cell immune responses (Section 5.5.2). Proteins binding to RNA quadruplexes are critical in dismantling RNA quadruplexes and subsequent pathogenic downregulation (Section 5.10). The folding and unfolding cycle is closely balanced by the dynamic interplay of stabilising ligands or binding proteins and helicases (Section 3.16). Revision of the balance can prompt pathogenesis. Validation of adenine quadruplexes folding in the viral genome may suggest concepts for neoteric chemotherapy. All the observations referenced here suggest allocating an adverse impact for indigenous supramolecular structures to RNA viruses. The corollary would be further investigation as suitable targets for chemotherapy and subsequent ligand designs.

A range of stabilising ligands have been synthesised such as telomestatin and quarfloxin, but to date, targeting has not been significantly improved (156). Distinguishing between helical- and quadruplex-DNA underpins low selectivity in the human genome. Drug development is often delayed or halted by unanticipated side-effects derived from poor selectivity. Recent work proposes modified loop structures may not necessarily compromise G4 activity and might be exploited for target signaling and / or drug delivery (320).

6.4 Purine quadruplex ligands

6.4.1 Telomestatin

Telomestatin is a planar macrocyclic compound with a central core of nitrogens arranged in an electronegative ring (figure 6.4). The molecule is highly selective for G4 structures compared to duplex DNA (321) and has a stabilising influence by capping the complex. In addition, access for disruptive agents is greatly hindered (321).
Telomestatin shows a robust network of $\pi-\pi$ interaction with aromatic heterocyclics superficial to the quadruplex (322). Intercalation and binding association with the interplanar loops have also been proposed (323). The various binding proposals were explored here with ADT and MDS. Both methodologies predicted the molecule is no longer planar when docked 1KF1 and reconfigures to an ellipsoidal arc (figure 6.5).

Similar results were predicted for ADN12-N1 docking. Co-ordinating cations were K$^+$, Na$^+$ or Li$^+$. The reconfiguration of telomestatin from a planar molecule to the proximate ligation around an interplanar loop closely resembles docking for C5-spiro-fluorene described in Section 5.8.4. Although figure 6.5 has a triplanar DNA guanine quadruplex, and Supplementary figure S5.4 shows a biplanar RNA system, the binding sites for both are very similar. The interplanar loop for the former is TTA and the latter, UUA.
6.4.2 Adapted porphyrins

6.4.2.1. Introduction

There have been some biomedical issues regarding formulation of metallo-porphyrins generating cytotoxic reactive oxygen species and accurate targeting. Recent nanotechnology appears to be addressing these challenges (324). Based on its specificity for G4s over duplex DNA, telomestatin should be a suitable scaffold. Modifications should encompass the Lipinski guidelines: a molecular weight <500, low lipophilicity (logP<sub>O/w</sub> typically between 2 and 5), good solubility and oral bioavailability. Blood Brain Barrier (BBB) permeance would be advantageous and also resistance to effluence by the phospho-glycoprotein (P-gp) mechanism. G4 inhibition of telomerase in cortical cells could reduce neural neoplasms. However, designing drugs capable of crossing the BBB has proven very challenging, and many are rendered less effective by the cellular efflux action of P-gp (325). Although telomestatin has a high affinity for telomeric G4s, and may enhance quadruplex stability, it has characteristics that contravene three Lipinski requirements. It has very poor solubility and gastrointestinal absorption. It is a substrate for P-gp and not BBB permeant. Nevertheless, telomestatin was selected as a scaffold despite the higher affinity of pyridostatin for G4s (section 5.8) and carboxy-pyridostatin characterised by selectivity of RNA models over DNA equivalents (326). Although telomestatin is a relatively small molecule, a smaller structure would facilitate docking protocols and monitoring.

6.4.2.2 Methods

A pharmacophore within telomestatin was tentatively identified as the planar nature of a nitrogen octet. This structure has a very electronegative exterior and associated lipophobic character which may underpin its non-BBB permeance. The relatively low lipophilicity and co-ordination with Fe<sup>2+</sup> supports this notion. A preliminary design removed the cation and reduced the structure to a planar cyclic pyrrole quartet with no overall charge as an initial reference model. ADMET data were a slight improvement on telomestatin, but a ChemSpider (327) similarity check compared it closely with porphyrin, differing mainly by lack of a co-ordinating cation. The design was named Adapted Porphyrin 1 (Figure 6.6).
Figure 6.6: incremental modifications to porphyrin intended to develop improved ADMET data. (A): porphyrin, (B): AP1, (C): AP2, (D): AP3, (E): AP4, (F): AP5

The nitrogen quartet and Fe$^{2+}$ of the porphyrin scaffold were removed to increase lipophilicity (AP1). Although Lipinski rules were satisfied, this proved to be a retrograde step. Accordingly, the N-quartet was restored and lipophilicity increased by a single methylation of each pyrrole ring (AP2). AP3 had increased lipophilicity by extending the methyl groups, but further methylation in AP4 implied P-gp efflux. AP5 increased electronegativity by substitution of the N quartet by O and retained the lipophilicity of AP3, but lost BBB permeance. The model was also a substrate for P-gp. None of the designs included a co-ordinating cation because the novel structures required assessment in their simplest form; the effect of adding an ion would be better interpreted by separate study.

A compound similar to AP2 has been synthesised for trials in photodynamic therapy. There is no reference to the molecule as a quadruplex ligand. Chemspider lists no entries for AP3 similarity. Metrics applied for analysis were: Lipinski Rules, lipophilicity, bioavailability, P-gp efflux, solubility, blood-brain barrier permeance for telomestatin, porphyrin and analogues. MD simulations to 500 ns were run for AP2 or AP3 docked with 1KF1 co-ordinated by K$^+$, Na$^+$ or Li$^+$ and analysed by RMSD and Radius of gyration.
6.4.2.3 Results

ADMET data for porphyrin were marginally favourable. However, the unmodified molecule as a candidate for was rejected for several reasons. Oral bio-availability was low (0.55) due to the degree of unsaturation and synthetic accessibility was 6.5 (1 = very easy, 10 = very difficult). Values for AP2 and AP3 were: bioavailability (AP2) 0.74, (AP3) 0.72, synthetic accessibility (AP2) 8.1, (AP3) 7.9. Further ADMET data are shown in Table 6.1.

Table 6.1. ADMET data for telomestatin, porphyrin and porphyrin analogues. Properties regarded as favourable are: moderate or good solubility, high gastro-intestinal availability, no substrate interaction with P-gp and compliance with Lipinski Rule of Five. Porphyrin, AP2 and AP3 satisfy all three constraints.

<table>
<thead>
<tr>
<th>model</th>
<th>MW</th>
<th>logP&lt;sub&gt;oct&lt;/sub&gt;</th>
<th>solubility</th>
<th>G-I</th>
<th>BBB</th>
<th>P-gp</th>
<th>Lipinski</th>
</tr>
</thead>
<tbody>
<tr>
<td>telomestatin</td>
<td>582.5</td>
<td>2.2</td>
<td>very poor</td>
<td>low</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>porphyrin</td>
<td>308.34</td>
<td>2.25</td>
<td>soluble</td>
<td>high</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>AP1</td>
<td>304.38</td>
<td>4.63</td>
<td>mod/sol</td>
<td>low</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>AP2</td>
<td>362.44</td>
<td>3.27</td>
<td>soluble</td>
<td>high</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>AP3</td>
<td>420.55</td>
<td>4.65</td>
<td>mod/sol</td>
<td>high</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>AP4</td>
<td>484.72</td>
<td>5.4</td>
<td>mod/poor</td>
<td>high</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>AP5</td>
<td>440.61</td>
<td>4.37</td>
<td>mod/sol</td>
<td>high</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Telomestatin, AP2 and AP3 were taken forward for ADT docking protocols with 1KF1 to compare binding data. AP2 and AP3 docked to 1KF1 K<sup>+</sup>, Na<sup>+</sup> or Li<sup>+</sup> were subsequently run in MD simulations and analysed with RMSD and radius of gyration.

6.4.2.4. Telomestatin, AP2 and AP3 docking with 1KF1

ADT calculated three binding interactions for each of the ligands when docked with 1KF1. All were exothermic. The optimum values are listed in Table 6.2 together with pharmacokinetic data predicted by MDS.
Table 6.2: binding data predicted by ADT and MDS for telomestatin, AP2 and AP3 docked to 1KF1. The values are optimum predictions for each model when co-ordinated by K⁺, Na⁺ or Li⁺.

<table>
<thead>
<tr>
<th></th>
<th>affinity kcal/mol</th>
<th>Ki μM</th>
<th>VDW+H-bond+solvation energy kcal/mol</th>
<th>electrostatic energy kcal/mol</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K⁺</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>telomestatin</td>
<td>-5.04</td>
<td>203.13</td>
<td>-7.18</td>
<td>-0.17</td>
</tr>
<tr>
<td>AP2</td>
<td>-6.35</td>
<td>22.18</td>
<td>-6.28</td>
<td>-0.07</td>
</tr>
<tr>
<td>AP3</td>
<td>-6.08</td>
<td>34.83</td>
<td>-7.15</td>
<td>-0.10</td>
</tr>
<tr>
<td>Na⁺</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>telomestatin</td>
<td>-5.11</td>
<td>242.02</td>
<td>-7.79</td>
<td>-0.19</td>
</tr>
<tr>
<td>AP2</td>
<td>-6.67</td>
<td>30.21</td>
<td>-7.02</td>
<td>-0.09</td>
</tr>
<tr>
<td>AP3</td>
<td>-6.33</td>
<td>32.61</td>
<td>-7.43</td>
<td>-0.14</td>
</tr>
<tr>
<td>Li⁺</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>telomestatin</td>
<td>-5.40</td>
<td>109.42</td>
<td>-7.00</td>
<td>-0.53</td>
</tr>
<tr>
<td>AP2</td>
<td>-6.34</td>
<td>22.68</td>
<td>-6.28</td>
<td>-0.06</td>
</tr>
<tr>
<td>AP3</td>
<td>-6.00</td>
<td>39.93</td>
<td>-7.09</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

Both methodologies find AP2 and AP3 are exactly parallel with the three tetrads and located directly on a line through the central core of 1KF1. Transferring the docking data to Chimera, Figure 6.7 shows the predicted alignment in MDS of AP2 with a VDW surface imposed on the quadruplex. AP2 the distance between the planar centre of AP2 to the planar centre of the outer guanine quartet (3.6 Å) is significantly less than that calculated for telomestatin (~9 Å). AP3 binds at the same site, in the same orientation as AP2 and has a very similar distance to 1KF1.

Figure 6.7: Chimera interpretation of AP2 docked to 1KF1 with a VDW surface imposed on the quadruplex.
RMSD returns for AP2 docked with 1KF1 K⁺ indicate the complex has good structural integrity to 500 ns. There is some dynamic activity to ~50 ns which probably reflects minimisation prior to production. RMSD maintains a steady trend at ~2.8 Å to the simulation end with minor variations in activity (figure 6.8).

**Figure 6.8: RMSD for AP2 docked with 1KF1 K⁺ co-ordination.** Data for Na⁺ and Li⁺ co-ordinated systems were very similar.

Dynamic activity early in the RMSD data resonates to some extent with NAMD calculations for radius of gyration. The trajectory shows a brief near-symmetrical cycle of relaxation and expansion prior to maintaining steady values centred on ~13 Å (figure 6.9).

**Figure 6.9: Radius of gyration data for AP2 docked with 1KF1 K⁺ co-ordination.** Values for Na⁺ and Li⁺ co-ordinated systems were similar.
6.4.2.5 Discussion

Taken together, an overview of the MD results indicates a stable interaction between AP2 or AP3 and a triplane guanine quadruplex. ADT predicted three exothermic conformations for AP2 ligated to 1KF1, two of which align parallel to one particular outer quartet of 1KF1. But a third docking conformation locates AP2 to the other outer quartet. Since the ligand cannot physically travel through the G4, implicitly the software finds in that site also favourable for docking AP2. If a second porphyrin analogue were ligated, the complex would be effectively capped on both outer faces. Consequently, attempts were made to construct an AP2-complex-AP2 association. Because ADT parameters accept only one ligand and a macromolecule, a submission cannot progress; adding a second is interpreted as an artefact.

However, two AP2 molecules linked with a flexible aliphatic chain (AP2-AP2) were constructed and accepted by ADT. The ligands bind parallel to both the outer quartets as implied by initial ADT results with an affinity of -8.2 kcal mol\(^{-1}\). Pharmacokinetics data are less promising (Table 6.3).

Table 6.3: ADMET data for AP2-AP2.

<table>
<thead>
<tr>
<th></th>
<th>MW</th>
<th>logP(_{o/w})</th>
<th>solubility</th>
<th>G-I</th>
<th>BBB</th>
<th>P-Gp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1077.54</td>
<td>14.24</td>
<td>insoluble</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

High molecular weight may be less significant in this instance as larger molecules can still transit neoplasmic capillaries due to their thinner endothelium (328). But the high lipophilicity and insolubility of the linked AP2 molecules are problematic. A linked structure appears unlikely to cap a quadruplex. The possibility remains open that non-transient ligation by two unlinked porphyrin analogues, possibly intermittent, could secure the opposing outer quartets.

6.4.2.6 N-meso methylporphyrin

The folding direction of G4 loops discriminates between single- and double-strand DNA (Introduction, Thesis Overview). Quadruplexes forming in the latter are potentially pathogenic. Several molecules are known to recognise the parallel looping of ds-DNA. One in particular, N-meso methylporphyrin (NMM) was taken forward here as a scaffold for in silico ligand designs. A sequence known to form parallel looping in triplanar G4s (PAGQ), was constructed for docking with modified NMM derivatives (329).
6.2.4.7 Methods

Because molecular weight, solubility and BBB-permeance of NMM were unfavourable, some modifications were made in a series of analogues: N-MM2 – N-MM6 (figure 6.10).

Removing double bonds to make the structure more flexible proved to be a retrograde step by reducing solubility (N-MM3). N-MM4 simplified the design but was recognized as a P-Gp substrate. However, it retained an electronegative centre and hydrophobic exterior. Adjusting these two parameters for NMM5 and NMM6 were also retrograde modifications. Table 6.4 details ADMET data for the N-MM series.

Figure 6.10: N-meso methylporphyrin (N-MM) and analogues N-MM2 – N-MM6
6.4.2.8 Results

Table 6.4: ADMET data for N-meso methylporphyrin and analogues. Properties regarded as favourable are: moderate or good solubility, high gastro-intestinal availability, no substrate interaction with P-Gp and compliance with Lipinski Rule of Five. Only N-MM2 satisfies all three constraints.

<table>
<thead>
<tr>
<th>Model</th>
<th>MW</th>
<th>logP_{o/w}</th>
<th>Solubility</th>
<th>G-I</th>
<th>BBB</th>
<th>P-Gp</th>
<th>Lipinski</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-MM1</td>
<td>582.73</td>
<td>4.55</td>
<td>Poor</td>
<td>High</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N-MM2</td>
<td>480.69</td>
<td>5.30</td>
<td>Soluble</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N-MM3</td>
<td>474.72</td>
<td>5.16</td>
<td>Poor</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N-MM4</td>
<td>472.79</td>
<td>4.45</td>
<td>Mod.</td>
<td>High</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N-MM5</td>
<td>504.84</td>
<td>3.87</td>
<td>Mod.</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>N-MM6</td>
<td>476.82</td>
<td>5.34</td>
<td>Mod.</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

N-MM2 had favourable lead candidate characteristics and was taken forward to docking protocols with PADQ in ADT and MDS. Although N-MM4 was recognized as a substrate for P-Gp, it was also submitted to ADT and MDS. ADT predicted poor binding affinity for N-MM2 (0.39 kcal mol\(^{-1}\)) and N-MM4 (0.46 kcal mol\(^{-1}\)). Ligation in both methodologies was limited to the periphery of the macromolecule and did not locate parallel to either outer quartet of the quadruplex. Both potential candidates noticeably lost planar conformation during the docking process, MDS predicted slightly higher binding affinities (N-MM2, -0.6 kcal mol\(^{-1}\) and N-MM4, -0.2 kcal mol\(^{-1}\)) but unlike ADT, the ligands maintained approximate planar configurations.

6.4.2.9 Discussion

Poor pharmacokinetics and insufficiently specific selectivity do not ostensibly recommend NMM or its analogues as suitable compounds for quadruplex ligation. Pharmacokinetic and binding properties of some telomestatin and porphyrin analogues are encouraging. Such ligands might stabilise beneficial single-strand telomeric quadruplexes associated with telomerase suppression. Section 2.10 referred briefly to a proposal that the length of interplanar loops may influence quadruplex stability. The notion advocates a potential route to destabilise non-beneficial duplex-derived quadruplexes.
6.5: Conjectural approach to destabilise pathogenic purine quadruplexes

6.5.1: Introduction

Nucleotide sequences with the potential to fold into guanine quadruplexes have been observed in single- and double-strand DNA models \(^9^3\) \(^3^3^0\). Whereas single-strand telomeric sequences could readily fold into a quadruplex linked by nucleotide loops, double-strand sequences showed a much-reduced capability. Moreover, the number of nucleotides in the loops was adversely influential on ds-G4s; only a small number would be tolerated. More than three or four nucleotides in the loop sequence would disallow folding. Although this thesis has identified locations allowing potential adenine quadruplexes to fold, it is not reinforced by experiment.

The finding that the length of interplanar loops may dictate selectivity for DNA systems and reduce their structural integrity is an interesting polemic. It suggests a method of interpolating extra nucleotides into a loop sequence might not only destabilise pathogenic G4s, but may prevent their formation \textit{ab initio}.

6.5.2: Peptide nucleic acids

6.5.2.1 Introduction

Synthetic nucleotide analogues constructed around an aminoethyl-glycine unit (PNAs) were considered as a vector for adding specific nucleotides into a G4 interplanar loop. Several properties recommend PNAs including high binding strength and selectivity (figure 6.11). Since PNAs are flexible and uncharged, they are not routinely recognised by proteases and consequently resist degradation \(^3^3^1\).

![Figure 6.11: basic unit of peptide nucleic acids.](image)

A peptide polymer substitutes for the deoxyribose phosphate backbone.
Interplanar loops comprising a maximum of three bases can fold in duplex DNA; even a slightly higher population appears to veto a quadruplex assembling (330). This observation could reinforce selectivity. Although introducing any of the four bases into an interplanar loop would increase its length, thymine was chosen because its smaller size might be conducive to BBB-permeance. Various small molecules were designed with thymidine monophosphate as a side chain to a peptide nucleic acid (PNA) pharmacophore. 1KF1 was selected as a target macromolecule because of its suitability as a reference model. PNAAs are disadvantaged by modest to poor cell membrane transit. Molecular modification needs to address low lipophilicity in particular. Favourable characteristics for lead candidates are as noted in Table 6.5. A series of six compounds were constructed (figure 6.12) and subsequently ligated with 1KF1 in ADT and MDS.

6.5.2.2 Methods

Compound 1 was a simple addition of thymidine monophosphate to a PNA unit as a baseline. Given the overall high distribution of electronegative charges, the ADMET data returned were unfavourable other than molecular weight and as a substrate for P-gp. (Table 6.6).

![Figure 6.11: serial modifications to a thymine-PNA complex.](image)
Compounds 2 and 3 were similar designs to assess if morphology might be a factor towards more favourable characteristics. Both were retrograde steps. Compound 3 also replaced the thymine component with an adapted PNA structure of similar conformation.

All of the electronegative atoms of the PNA structure were removed for Compound 4 leaving an aliphatic region to the design which should improve lipophilicity. Because the data were more promising despite almost identical morphology with compound 3, it was inferred configuration could be relevant to a BBB-permeant design. Compound 5 developed this motif, but retaining high GI-absorption and improving solubility also reduced lipophilicity. Compound 5 was predicted to be a substrate for P-gp efflux.

Compound 6 consolidated the aliphatic component and re-introduced the thymine component. Although lipophilicity was lower than anticipated, marginal BBB-permeance was predicted. However, the molecule would also be removed by the P-gp efflux mechanism. GI-absorption was also low. All of the designs complied with Lipinski, but were generally disappointing (Table 6.5).

Table 6.5: ADMET data for PNA-thymine complexes. Properties regarded as favourable are listed in Table 6.4

<table>
<thead>
<tr>
<th>compound</th>
<th>MW</th>
<th>logP&lt;sub&gt;a/w&lt;/sub&gt;</th>
<th>solubility</th>
<th>G-I</th>
<th>BBB</th>
<th>P-Gp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>462.35</td>
<td>-1.73</td>
<td>very</td>
<td>low</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>466.38</td>
<td>-2.44</td>
<td>very</td>
<td>low</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>368.23</td>
<td>0.59</td>
<td>very</td>
<td>low</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>408.51</td>
<td>3.49</td>
<td>poor</td>
<td>high</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>488.43</td>
<td>2.79</td>
<td>mod</td>
<td>low</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>488.43</td>
<td>2.79</td>
<td>mod</td>
<td>low</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Although compound 6 had latitude for improvement, it was selected as a structure that might interpolate thymidine monophosphate into an interplanar G4 loop if consecutive nucleotides in the loop could be separated.

6.5.2.3 Separating loop nucleotides: type II restriction endonucleases

Restriction endonucleases are a bacterial enzymatic defence against viruses and generally regarded as a simple but effective immune system. Viral DNA or RNA entering the bacterium is degraded by cutting at specific nucleotide sequences. This specificity suggests a tentative route to separating a G4 loop at a precise point. Numerous nucleotide sequences have been identified as scission points (332).
Each enzyme has an indigenous corresponding methylase which protects host DNA containing identical sequences from self-attack. Methylated bases within the recognition sites of the host DNA are thereby marked as non-viral (333). The enzymes are dimeric and therefore recognise sense- and anti-sense strands (334). A three stage mechanism has been suggested for binding (335). The enzyme’s open conformation is typified by a C-shaped profile. After random attachment to a DNA strand, the enzyme progresses until the relevant nucleotide sequence is encountered. There may be a specific recognition of the first and last base-pairs which confirms the site and triggers H-bonding. Finally, more H-bonds form as the central nucleotides are recognized. Common to most endonucleases are co-ordinating cations, typically Mg$^{2+}$, located on the periphery of the active site. The lone pair of oxygen of H-OH attacks phosphorus forming a doubly negative intermediate species. Subsequent proton transfer from an oxygen of the intermediate to another prompts loss of the cytosine groups and reconstitution of the phosphate (figure 6.12).

![Diagram](image)

**Figure 6.12:** proposed mechanism for compound 6 interpolation of a single base exemplified by guanine (adapted from P. Taylor, 2022).

It is proposed here that an intermediate state might be stabilised long enough by Mg$^{2+}$ to allow cleavage of the cytosine phosphodiester bond. Both separated cytosine and the guanine-phosphate group are hydroxylated and thus receptive to attachment with extraneous nucleotides and/or continuation of the nucleotide sequence.

If the monophosphate group of compound 6 were sufficiently adjacent, the disrupted loop might allow linkage with the protonated guanine monophosphate and hence interpolate an extraneous thymidine motif. Deprotonation of the detached hydroxyl-cytosine would
then also be available for binding to 3′C of the interpolated thymidine. The loop is consequently extended by the extraneous nucleotide. This interpolation could conceivably occur at any or all of the G4 loops and hence destabilise the complex. If a DNA sequence with G4-forming potential were targeted, scission and interpolation might prevent the structure from folding *ab initio*.

Unfortunately, there is currently no endonuclease specific for the 1KF1 loop sequence of ...AAT.... However, the non-palindromic sequence (GGGGCC)_n is identified with neurodegenerative diseases when folded into a quadruplex (92), and was chosen here to model its interaction with an extraneous thymidine monophosphate interpolated by compound 6. The phosphate group of compound 6 was accepted by 3′ and 5′ carbons of adjacent cytosine and guanine and hence a (CCGGGG) sequence was extended to (CCTGGGGG) in DSV and Chimera. The ‘clean geometry’ facility of DSV adjusted bond angles and distances very slightly.

The restriction endonuclease BsaJI recognises the sequence C*NNGG, where *N* is any nucleotide and the cleavage point is indicated by the asterisk (336). Unfortunately, BsaJI is not archived in PDB and therefore could not be submitted to any docking software. However, BglI and BcnI were identified as having recognition sequences very similar to the target (CCGGGG). Both are available in PDB as 1D2i and 2ODH respectively. BglI recognises GCCNN*N*NGC and BcnI recognises CC*N*NG where *N* = C or G. Where a cleavage point is asymmetric, the exposed DNA ends are cohesive and readily available for binding. (337). Transient base-pairing can occur with this type of cleavage and might allow interpolation of extra nucleotides (338). Asymmetrically cleaved cohesive ends also promote re-establishment of a continuous new strand.

For K⁺ co-ordinated G4s, melting temperature (Section 2.4) decreases by 2° for each single nucleotide increase in the length of interplanar loops (339). The data are less well-defined for Na⁺ models. However, the inverse relationship between melting temperature and nucleotide population of interplanar loops is compelling. Whilst the impact of shorter loops varies between systems, G4s with loops in excess of six nucleotides are always more prone to instability. This finding recommends the incremental addition of extra bases to adversely affect non-beneficial quadruplexes. Loop direction may promote selectivity. The data only consider K⁺ and Na⁺ as co-ordinants, Li⁺ was not modelled in this context.
6.6 Incidence of cancer in patients with long-term lithium medication

Li⁺ has been included throughout this thesis to attempt reconciliation of conflicting perspectives (Section 2.1.1). Guanine sequences folding into supramolecular structures may be inhibited in the presence of the ion but the paradoxic effect has also been noted (112) (81). Given these uncertainties and the incorporation of Li⁺ in most quadruplex models here, a brief survey of the ion’s physiological context is reasonable.

Lithium continues to be a successful prime therapeutic intervention in bipolar disorder after many decades even though a precise mechanism of action is unclear. The carbonate and citrate have high gastro-intestinal absorption and renal uptake; side effects are mainly associated with the urinary system. Cortical microcysts are regularly observed, some of which have papillary projections usually linked to pre-malignancy. Lithium-induced inhibition of a renal kinase has been suggested as underpinning microcyst formation (340).

A Standardised Incidence Ratio of renal cancer in a longitudinal study found a significant increase in lithium-treated patients compared with a control group of 7.51: 13.69 (341). However a re-evaluation of this research using a different statistical strategy (340), and an earlier study (342), found no significant association of lithium with renal neoplasia. Increased incidence of thyroid or renal cancer was also not confirmed in an epidemiological study, but noted these neoplasms are typically characterised by low occurrence and mortality (343).

Lithium, when no longer co-ordinating a disintegrating adenine or guanine system, may not be easily available for gastro-intestinal uptake after extra-cellular transport. If so, a direct link to oncogenesis may not be easily observable.

6.7 Executive summary of Chapter 6

- The paradox of DNA quadruplex as adverse or beneficial extends to RNA complexes
- RNA quadruplexes have been shown in this thesis to maintain a higher occupation of waters compared with DNA models
- The contribution to systemic stability by the core of quartets differs between DNA quadruplexes and RNA equivalents. The outer quartets support the former models more than the central group, whereas the distribution is equitable in RNA systems
- Discrimination between DNA and RNA quadruplexes may derive from these observations plus other differences in local chemistry: absence or presence of a cation, waters binding peripherally to a given complex and more the regular geometry of RNA structures
predicated on the ribose hydroxyl groups

- Quadruplexes folding in nucleic acids during ribosomal translation could compromise the process
- The lagging strand of telomeric duplex DNA has sequences that could allow cytosine quadruplexes to arise. This eventuality might prove disruptive.
- Whilst vaccination remains a prime route to controlling viral infection, concurrent chemotherapy should be propitious.
- Rational drug design would need to address contradictory stabilising and destabilising intervention in purine quadruplexes
- Planar macrocyclics such as telomestatin show excellent binding affinities but poor selectivity. Some porphyrins modified here show enhanced ADMET properties; selectivity may be improved for triplanar guanine quadruplexes
- Increased length of interplanar loops is inversely correlated to instability; a conjectural exercise to interpolate additional bases into a loop outlines some possibilities. Modified peptide nucleic acids and restriction enzymes might progress synergistically to effect a selective and destabilising outcome
- Although a validated mechanism for ribosome pausing is currently unidentified, transient quadruplex occlusion in a ribosomal site for protein synthesis could produce a backlog of incomplete polypeptides

Whilst supramolecular structures arise in both DNA and RNA, assumption of quadruplexes as counterparts is no more than a preliminary overview. Single-strand RNA endorses various molecular designs unavailable in DNA. Hairpin loops and i-motifs as imperatives were noted in Section 5.3.3 and Section 3.1; further instances are discussed in Chapter 7.

6.8 Future perspectives.

- Given its diagonal Periodic Table relationship with Li⁺, Mg²⁺ as a stabilising factor in DNA purine quadruplexes might be a consideration in future work

- Porphyrin analogues are in current use *inter alia* as drug therapy. If docking and MD studies were applied, a comparison with AP2 and AP3 data might be synergistic. Similarly, AP2 or AP3 substitution for the targets of current porphyrin analogues should be explored. The impact of presence or absence of a co-ordinating ion significant for these circumstances

- If it were possible to simulate docking of two separate AP2 molecules ligated with a quadruplex, pharmacokinetics data may be favourable because the characteristics of a
single AP2 are encouraging (Table 6.1). Two such ligands located on opposite sides of 1KF1 would be separated by a notional ~15 Å. Any direct interaction between the two would be very unlikely at this distance. Implicitly therefore, the overall pharmacokinetics could still be estimated on an individual basis. At present, docking and molecular dynamics simulations of the linked structure are not available with current software. Exploring the viability of two planar molecules bound parallel to the outer quartets in silico lies in the future.

- Porphyrin analogues are in current use as drug therapy but not targeting purine quadruplexes. A comparison of their docking and MD data with the porphyrins modified in this Chapter might be pragmatic.

- ADT data indicate two AP2 molecules linked by flexible aliphatic chain may be interacting with the inter-tetrad loops. Telomestatin interacting with interplanar loops of a purine complex may partially explain its radical reconfiguration when bound to 1KF1 (section 6.4.1).

- The deprotonation of Cyt-OH (figure 6.12) could hydroxylate thymidine monophosphate of compound 6. Interpolation ex silico would probably not progress in this case: software accessed here automatically protonated cytidine monophosphate 3’O.

- Cytosine detached from the original sequence by enzymatic scission could re-associate before any interpolation. If the unattached cytosine 3’O accepted methylation, interpolation could proceed.

- Detachment of the aliphatic heterocyclic group of compound 6 might be achieved by protonation of thymidine C1 if there were an increase in local basicity. It is not clear however, if removal of the heterocyclic aliphatic group might be necessary for the nucleotide interpolation to disrupt loop formation or disintegration.
Chapter 7. Purine quadruplexes in long non-coding RNAs.

7.1 Introduction.

7.1.1 Long non-coding RNAs in the context of Natural Selection

The legacy of evolutionary functional innovation is often characterised by abrupt mutations in highly-conserved locations historically predicated on negative selection. Some 200 highly-accelerated regions (HARs), mainly situated upstream of telomeres, were identified in two 2006 studies [344]. HAR sequences were also closely associated _inter alia_ with nucleic acid metabolism regulation [345]. Comparison of nucleotide substitution rates for all HARs across several species were found to be unexpectedly high specifically for _Homo sapiens_: HAR1 to HARS show particularly advanced mutation rates. Because these sequences in vertebrates were highly conserved, attention focussed on potential human-specific evolutionary advantages. The statistically aberrant incidence of single nucleotide substitutions in these mRNA fragments for _Homo_ relative to other vertebrata, might be an important factor in Hominid speciation. One specific region of HAR1 (HAR1a) contains a 118nt hairpin sequence with 18 base pair strengthening mutations. The expected rate would be 0.27 single base mutations [346]. This value is observed in all other higher vertebrates. Because the sequence does not apparently code for proteins, it typifies long non-coding RNAs (lncRNAs). However, a very small number of lncRNAs are now not only associated with polypeptide binding and an oncogenesis cascade, but also contain G4s [276] [347]. Some studies report the lncRNA transcript HAR1 binds with one or more (as yet) unknown molecules vital to embryonic cortical development in _Homo_ [348].

Given their proximate relationship to a genomic region characterised by mutation, HAR1a and HAR1b are historically very stable in non-human species. The comprehensive transformation epitomised by _Homo_ has been estimated by base-pair changes to have begun within the last 5 Mya [344]. This date of _Homo_ / _Pan_ speciation resonates reasonably well with other estimates of between 5 and 7 Mya [349] [350]. The last known common ancestor with HAR1 dates to at least 310 Mya [346].

HAR1a mediates embryonic neurological progress: expression is high in the developing neocortex. Because this region is notably more developed in _Homo_ than _Pan_ for instance, this observation is significant. Furthermore, expression of HAR1a is absent in other regions of the forebrain [351]. Taken together, these observations could underpin the emergent pre-eminence of _Homo sapiens_ as developing exceptionally cognitive species. Some neurological dysfunction and tumorigenesis have been associated with HAR1 loci [352].
The single nucleotide mutations contrasting *Homo* with other species are almost exclusively adenine or thymine/uridine to cytosine or guanine. These particular point mutations augment non-covalent interaction from two hydrogen bonds to three. Consequently, base-pairing and overall stability are enhanced. Guanine-cytosine biased conversions can occur during chromosomal recombination as repair mechanisms may preferentially incorporate guanine or cytosine to adenine or thymine/uridine (353). However, this process does not appear to completely account for the high guanine-cytosine content of HARs. An increase in the local H-bond population would enhance secondary structure. Positive selection for a G-C bias would be one conspicuous consideration (354). The presence of purine quadruplexes located within a HAR could be another.

7.1.2 Purine quadruplexes in HAR1

Supramolecular structures may also preferentially influence the notable stability of *Homo* HAR1a compared to *Pan*. Data described in Section 2.9 indicated RNA quadruplexes are more thermodynamically stable than DNA equivalents. Despite lack of a co-ordinating cation, an additional ribose 2′-hydroxyl allowing enhanced H-bonding in RNA systems. Although both are ubiquitous in the genome, there is a preponderance for the DNA complexes in telomeres (Section 2.1.3). HAR1 is located just upstream of the telomere on chromosome 20. The gene is guanine-rich and numerous potential biplanar and triplanar G4-forming sequences are prevalent both upstream and downstream of the 118nt hairpin sequence in **HAR1a** (figure 7.1). It has been proposed that localised G4s may promote structural integrity of HAR1 (351).
Figure 7.1: potential sequences for biplanar purine quadruplex folding in HAR1. Guanine models (PG4S) are slightly more prevalent than adenine systems (PA4S). guanine-adenine hybrid quadruplexes are the least numerous, however a model located at ~850:880 may be of interest.

Several sequences are located downstream of the hairpin fragment. Misincorporation of adenine for guanine in quadruplexes (Section 4.6) has been deemed possible and probably biologically relevant (355). HAR1 harbours several sequences that could fold into simple biplanar adenine-guanine hybrid systems. One in particular is partially located immediately upstream of the 118nt fragment (located at ~850:880 in figure 7.1). The sequence appears to be uniquely human as there is no directly similar feature observed in Pan.

In addition to numerous potential guanine and adenine quadruplexes, HAR1 terminates in a short poly(A) tail. The AAAUAA hexamer associated with prompting terminal polyadenylation is present at the appropriate site just upstream of the tail (Section 5.2). The nuclear mechanism accounts for the 3’polyadenylation of HAR1, and possibly the structure is a legacy of historical integration of viral nucleic acid into the human genome predicated on retroviral activity (356). These events are considered to have been extensive and may still be contemporary (357). Resonating with this notion is a unique uridine / cytosine sequence in HAR1 located on Chromosome 20 63.102.407 > 63.102.630 with a total absence of guanine and adenine.
7.1.3 Materials and Methods

Models of biplanar guanine and adenine quadruplexes including a hybrid were constructed in Chimera and DSV from the relevant HAR sequences:

biplanar G4 (2642 > 2667): GGgagGGuccucaGGgcaccuuGG

biplanar A4 (2686 > 2715): AAgcuaAAugguuggAAaaaaucAAaa

biplanar guanine-adenine hybrid (851 > 879) uGAaaGAgacguuacAGcaacgugucAG

None of the systems included a co-ordinating cation and models were prepared for Molecular Dynamics simulation as described in Section 1.1. Metrics used for analysis were RMSD, eRMSD, Radius of Gyration, Frequency of H-bonding, occupancy of waters and ribose configurations.

7.1.4 Guanine-adenine biplanar hybrid results

HAR1 was found to be well-populated by potential sequences for biplanar G4 folding but with an apparent non-random distribution. A small number of sequences for triplanar guanine systems were identified such as GGGucugaGGGaggaaGGGauGGGa (3056 > 3080). A sequence at 2686 > 2715 could also form a triplanar adenine quadruplex. No other sequence was found with this potential. Molecular dynamics simulations were run to 500 ns.

7.1.4.1 RMSD

A steep rise in the trajectory from ~1.0 A at t = 0 peaks at ~70 ns prior to a sudden reduction in RMSD. The trajectory maintains a steady trend to the end of the simulation although with some dynamic activity (Figure 7.2).
7.1.4.2 eRMSD

Values for eRMSD show a steady trend with little dynamic activity for the duration of the simulation. The trajectory shows a minor peak at ~ 70 ns which resonates with the RMSD data (figure 7.3).
7.1.4.3 Radius of Gyration

There is a trend of molecular relaxation to ~180 ns followed by brief compression prior to maintaining limited dynamic activity to 500 ns. A minor peak in relaxation at ~60 ns matches similar data in RMED and eRMSD. The trajectory maximises at ~180 ns at which point the RMSD data show a noticeable reduction in dynamic activity (figure 7.4):

![Figure 7.4: HAR1 guanine-adenine biplanar hybrid Radius of Gyration.](image)

7.1.4.4 Hydrogen bond frequency of formation

The guanine-adenine biplanar hybrid quadruplex maintains a steady level of H bonding between four and six for the duration of the simulation with three brief peaks at ~60 ns, 280 ns and ~450 ns (figure 7.5). The lowest population at ~170 ns reflects a peak in molecular relaxation indicated by the Radius of Gyration trajectory.
7.1.5 Guanine and adenine biplanar HAR1 quadruplexes

7.1.5.1 RMSD

Trajectories for both systems maintain similar trends of limited dynamic activity centred on ~1.8 Å. Guanine shows a minor peak at ~150 ns echoed by the adenine model at ~360 ns (figure 7.6):
7.1.5.2 eRMSD

Data for both models are very similar and imply base-pairing has limited dynamic activity with the central region of the two systems (figure 7.7):
7.1.5.3 Radius of Gyration

Both models show marginally limited cycles of compression and relaxation. Whereas the adenine system maintains a steady trend, the guanine equivalent has a slight rising trend to the end of the simulation (figure 7.8):

![Figure 7.8: Radius of Gyration data for guanine and adenine quadruplexes derived from HAR1 sequence.](image)

7.1.5.6 H-bond frequency of formation

Both models show a hydrogen bond frequency formation between four and six for the duration of the simulation. There are however numerous occasions for both systems where the frequency peaks between ten and twelve (figure 7.9).
Figure 7.9: H-bond frequency of formation for guanine and adenine quadruplexes derived from HAR1 sequence.

7.1.6 Ribose configurations for HAR1 G4, A4 and hybrid systems.

C2’-endo configurations are preferentially adopted for guanine and adenine systems for the duration of the simulation but the hybrid model configures C3’-endo by 500 ns (Table 7.1). The stabilising influence of a 2’-hydroxyl pair in ribose extends from uniquely guanine or adenine quadruplexes to the hybrid model (Section 6.3.1). The influence of steric constraints derived from these groups on local topology may also apply to the hybrid system.
Table 7.1: ribose configurations for G4 A4 and guanine-adenine biplanar hybrid quadruplex.

<table>
<thead>
<tr>
<th>t = 0ns</th>
<th>HAR1 G4</th>
<th>HAR1 A4</th>
<th>HAR1 G_A</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C2'-exo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>19</td>
<td>6</td>
<td>8</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t = 500ns</th>
<th>HAR1 G4</th>
<th>HAR1 A4</th>
<th>HAR1 G_A</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1'-endo</td>
<td>C2'-endo</td>
<td>C2'-exo</td>
<td>C3'-endo</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>18</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>17</td>
<td>19</td>
<td>21</td>
<td>22</td>
</tr>
</tbody>
</table>

7.1.7 Interplanar occupancy of waters for HAR1 G4, A4 and hybrid models

The hybrid system maintains a marginally higher population of interplanar waters than G4 or A4 models. The occupancy for G4s is lower than for adenine equivalents (Figure 7.10). The data are similar to results for biplanar quadruplexes derived from the telomeric sequence (Sections 2.9 and 4.22.2).

Figure 7.10: Interplanar occupancy of waters for G4, A4 and hybrid models derived from HAR1 sequences
7.1.8 Discussion

Biplanar purine quadruplexes derived from genomic sequences located within HAR1 compared with 2KF8 equivalents have very similar in silico properties. The number of locations identified in this study is reflected pro rata by reports of guanine quadruplex prevalence in lncRNAs (358). Of particular interest is a sequence immediately prior to the 118nt HAR1 progression which would allow a biplanar guanine-adenine hybrid to fold. Given guanine quadruplexes regularly fold in genomically significant locations, and promote lncRNA binding with interacting proteins (359), a hybrid model may feasibly have similar influence. The sequence does not arise in the Pan HAR1.

The metrics used here to analyse molecular dynamics data all indicate a reasonable structural integrity to 500 ns. None of the models was methylated at guanine O6 or adenine N6 for a complete comparison with 2KF8 systems (Sections 2.5.4.3 and 4.25) because of the differentials in interplanar loops.

Whilst the number of identified lncRNAs has increased significantly (360), HAR1 and its associated 118nt region, HAR1a, are worthy of a specific survey. The lncRNA carries mutations with potential evolutionary advantages for Homo. Furthermore, RNA-RNA interactions explored here inform a possible cellular enterprise of ionic calcium.

7.2 RNA-RNA interaction: HAR1A, HAR1 and small nucleolar RNAs

7.2.1 Introduction

The primary sequence of lncRNAs is poorly conserved compared to that of protein-coding genes. However, function is likely to be conserved within short sequence motifs and molecular structure (361) (362) (363). For instance, some 300 bp upstream of the motif is an unusual 240 bp sequence with regular iterative alternation of cytosine and uracil. No function is currently attributed to the feature. The secondary structure of the HAR1 region in Homo and Pan has been mapped and analysed (351) (364) (365), with potential binding sites identified (348). These sequences were found to have vastly different structures and therefore molecular evolution may be responsible for a human-specific function of HAR1A. Furthermore, the 18 nt single base mutations between Pan and Homo all increase the H-bond population and correspondingly enhance stability of HAR1 secondary structure. This significant trait may underpin the evident contrast of respective embryonic cortical development.

HAR1A is primarily localised in the nucleus (366), and its current mechanisms in health and disease are unknown. Exploring potential interactions between the gene and small nucleolar RNAs (snoRNAs) is feasible. Predominantly localised in the nucleolus, snoRNAs
are a class of RNAs between 60-300nts in length. There are three classes of snoRNA depending on their structural motifs: H/ACA snoRNAs (snoras), C/D box snoRNAs (snords), and small Cajal body specific RNAs (scarnas) (367). SnoRNAs are ubiquitous but the highest tissue-specific expression for Homo has been identified in the brain (368) (369). Although extensive research has identified the role of this group as guide RNAs, in gene expression and RNA modification. However, other functions continue to be identified (370) (371), notably a possible mechanism for IncRNAs regulation of snoras.

Due to the relative novelty of snora research specific to cellular function, information is limited. LncRNAs have been identified with snora termini (sno-lncRNAs) and are linked to RNA transcription and lipid metabolism (372) (373) (374). In addition, snords interacting with lncRNA may be inherent in diffuse intrinsic pontine glioma (375). Given their tissue and species specificity, investigation informing the relationship between snoras and lncRNAs would be plausible. With this aspiration, the structures of HAR1A and the HAR1 region were mapped to assess any interaction and complementarity between the two RNAs. Elucidating mechanisms of action, significant differences between the three classes of snoRNA were noted.

7.2.2 Methods

7.2.2.1 Predictive methodology of RNA-RNA interactions

Complementarity for matching a given snoRNA sequence to any location in the gene was investigated with IntaRNA (75). The software design acknowledges that predicting RNA structure and interaction is more complex than exclusive identification of complementarity alone. Previous releases of RNA-RNA interaction software can filter for promising candidates as an ergonomic use of computer time. However, because IntaRNA parameterises thermodynamics in its models, physical aspects are incorporated into structure formation and hence describe more than simple base-pairing. The algorithms predict flexible non-linear regions connecting local helices between two or more matched sequences. These can be interpreted as bulges, kinks or loops in one or both of the aligned pairs. These structures allow dynamic re-alignment between the two RNA molecules and better reflect the matched sense-antisense configuration in 3-D space. Loops and kinks in RNA-RNA interactions from recent research (376) are exemplified in Supplementary Figure 3 (snord115-48 binding to the serotonin receptor 5HT2C).

The helical nature of the molecules can bring bases into close spatial alignment; such proximity is not always evident with a 2D “beads on a string” model. Hence, when thermodynamic parameters of a tertiary configuration are incorporated into a prediction,
canonical helices can be linked by short unpaired strands. The analysis also accounts for the molecular integrity of interacting sequences which precludes availability for intermolecular activity beyond the matched base-pairing.

7.2.2.2 Probability methodology

Choice of a threshold value directly influences accuracy of a model. Moreover, the extent of predicted locations populated by the variable (prevalence), is also contingent on this selection. The canonical value of \( p \leq 0.05 \) does not necessarily maintain prevalence or yield the highest accuracy (377). These aspects are particularly relevant to prevalence at the upper and lower boundaries of a data set. Selection of \( p \leq 0.05 \) is therefore arguably arbitrary: both statistically significant or non-significant results may derive from the same test. Such ambiguity has repercussions for data sets where \( n \leq 10 \) (378).

Although historically \( p \leq 0.05 \) is often accepted a default (379), lower values prevail in later work: \( p \leq 0.03 \) has been accepted as a threshold in a theoretical study of duplex DNA in random knotting activity (380). An alternative approach calculates pattern discovery in DNA by determination of an error threshold contingent on sample length. The algorithm evaluates probabilities in terms of an error ratio rather than a \( p \) value. Nevertheless, the research highlights the exponential effect of a unit increase in pattern length on the error ratio (381). Equation 1 resonates very closely with this observation. A unit increase in the exponential denominator translates as a greatly amplified outcome for the same numerator.

\[
p = \frac{(y-n+1)(x-n+1)}{4^n}
\]

Eq.7.1

Using equation 1 (derivation in Supplementary Material), the probability of complementarity for matching a non-specified snoRNA sequence to any location in the gene was calculated. The number of nts in the snoRNA (\( x \)), and the length of the query sequence (\( n \)) are variables. The number of nts in the target gene is constant (\( y \)). The mean value of \( x \) for the three snoRNA species analysed was 140.

Since the data analysis considers comparison of a large population in specified groups, the probability of Type I errors, Family-Wise Error Rate (FWER), becomes relevant. This was calculated using equation 7.2:

\[
\alpha_{FW} = 1 - (1- \alpha)^c
\]

Eq.7.2

where \( \alpha \) is an appropriate significance level and \( c \) the number of comparisons.
7.2.2.3 Mapping HAR1A and HAR1

7.2.2.3.1 Introduction

HARs are located almost exclusively in long non-coding RNA: strands generally comprising at least 200 nucleotides (382). Although protein-coding function is absent (383), IncRNAs are now emerging as having significant biological roles. Regulation of transcription and translation have been noted (384). Even though the almost total lack of coding ability in HAR sequences tends to confound inclusive understanding of their attested role, IncRNAs may have direct mutual interaction. Intronic cytosolic microRNAs are buffered by IncRNAs: a recent study proposes such interaction may mediate inhibition of target binding (385). HAR1a is nuclear (386) but not cytosolic. Nevertheless, the possibility of interaction with other small RNAs is advanced here. In particular, interaction with small nucleolar RNA (snoRNA) is explored.

Numerous snoRNAs are known and were originally associated uniquely with ribosomal biosynthesis. Later research extends this function to ribose-O2’ methylation in other nuclear factors (388). This epigenomic mechanism is ubiquitous and often guided by a protein-snoRNA complex (389). However, full understanding of snoRNAs’ roles is incomplete. In silico molecular investigations can be very time-expensive.

In common with numerous biomolecules, the structure and function of IncRNAs are closely related. IncRNAs are species-specific and have poor sequence conservation, therefore their functions and molecular interactions can be inferred from tri-dimensional domains formed by specific folding. Due to the large size of IncRNAs and their heterogeneity, mapping their structure has proven challenging. Previously, the secondary structure of the HAR1 region has been mapped (351) (390). Molecular models of Homo and Pan HAR1 secondary structure were constructed for morphological and topological comparison. A model of HAR1 tertiary structure was also constructed for molecular dynamics analysis. Prior to this study, the tertiary structure of HAR1a was ostensibly limited to a sequence (PDB 5UNE), characterised by Mg2+ ligation. 5UNE incorporates the initial 47 nucleotides of the118nt IncRNA fragment. The region has been recognised as an important structural difference between Homo and Pan (348). Subsequent consideration suggests the adenine-rich motif has crucial role in binding currently anonymous molecules contributing to embryonic brain development. For this current study, the sequence was interpreted as a 3D model and ligation characteristics of other divalent metal cations with 5UNE and HAR1 were compared with data for binding to the ADT optimal docking site.
7.2.2.3.2 Methods and materials

HAR1A sequences for Homo and Pan were accessed from Ensembl Release 104 (391). The gene (*Homo*) is located at: Chromosome 20 63.102.142 > 63.102.259 (accession NC_036899), and HAR1A (*Pan*) at: Chromosome 20 60.485.365 > 60.485.482 (accession NR_003244.2).

Using an RNA Webserver modelling algorithm (392), the secondary structure of HAR1A in *Homo* and *Pan* was mapped in order to identify any differences between the two structures. Based on the structure of HAR1A, greater detail of specific motifs was considered expedient. Applying a modelling algorithm, the tertiary structures of the HAR1 region in *Homo* and *Pan* were mapped. Sequences were submitted to RNA-Fold to obtain predicted 2D structures (393). Centroid-Fold was used for secondary structure method prediction since fewer errors are associated with this method (394). Minimum free energy and partition function were selected and ten interpretations requested.

The 118nt sequences were copied into RNA Composer (395) via batch mode. The 3D filter structure elements depository in CentroidFold was selected and the X-ray determined structures-only option chosen with the resolution threshold set to 3.0. 3D predictions for both species were downloaded. Molecular models were opened in Discovery Studio Visualiser (Client) v16.1.0.15350 (DSV Client) and UCSF Chimera 1.12_build_41623 (396). Two modelling methodologies were used for practical purposes, *ie.* manipulation and confirmation of data returns.

7.2.2.3.3 Results

Minimum free energy for HAR1a_*Homo* was estimated at -33.5 kcal mol\(^{-1}\) and for HAR1a_*Pan*, -20.20 kcal mol\(^{-1}\).

Differentials between *HAR1A_Pollard* in *Homo* and *Pan* may be indicative of function. Inter-species structural differences deriving primarily from their base-pairing sequences were found. The overall configuration of *HAR1A_Pollard* is similar in both *Homo* and *Pan* (Figure 7.11). However, there are notable differences in their local architecture. These moieties may be short motifs and binding sites with consequent *Homo*-specific functionality.
Figure 7.11. Secondary structures of HAR1A for Homo and Pan. The modelling algorithm depicts regions of high probability in red, orange and yellow. Regions of reasonable probability are green; lower confidence is shown in blue. Microdomains relevant to this study are identified in Figure 7.19.

RNAfold calculations for the HARI secondary structures of Homo and Pan indicate reasonable similarity for the two species (figure 7.12). The calculation indicates much higher overall probabilities of base-pairing in the Homo model than for Pan. This observation reflects the predictions for HAR1A (figure 7.12):
However, laboratory analysis concluded *Homo* HAR1a adopts a cloverleaf configuration but *Pan* a more linear hairpin secondary structure (351). The enhanced stability of *Homo* HAR1a was attributed to this differential. A possible binding site in the SUNE fragment was proposed in a recent X-ray crystallography study (348).

From the perspective of molecular biology, 3D structure can be more informative than 2D. Complete 3D models for HAR1 are not apparent in the literature. For this current study, tertiary configurations were obtained via RNA Composer for HAR1a of *Homo* and *Pan*. These models harbour some characteristics of the 2D predictions described by both earlier research evaluations.

An overall similarity in the tertiary modelling of the HAR1 fragment for both species was noted (Figure 7.13). The cloverleaf structure, and the *Pan* sequence as an extended and unstable hairpin noted earlier, are evident in the mapped tertiary structures (figures 7.12). The main difference is an extended unpaired base loop at H4, as this is larger in *Homo* than in *Pan*. H4 (*Homo*) is significantly less open than the equivalent in *Pan*. Two of the stabilising mutations are located in the H4 (*Homo*) loop. Implicitly, these results indicate that HAR1 H4 could be a binding site for various ligands with electropositive surface availability.
Figure 7.13. 3D structures of HAR1 for *Homo* (A) and *Pan* (B). Helix H2 has closely similar configuration in both species. Predictions for Helix H1 are also very similar. However, the H4 loop in *Homo* is notably more compact than for *Pan*. This feature may underpin the evident angular perturbation in the overall linearity differential between models. SUNE is located between helices C/H4 and H3.

### 7.2.2.3.4 Discussion of HAR1A and HAR1 mapping

Molecular modelling and prediction of tertiary structure is not experimental validation and therefore has limits on interpretation. For instance, information about base-pairing, bond angles and distances are contingent on comparison with archives of laboratory-validated values. Predictions are therefore optimised. However, enhanced insights into the complexities of 3D RNA structures improves understanding of biological function. Very large molecules can be manipulated and analysed relatively quickly. If a model or ligation seems less than optimal, the simulation can be halted with very little time lost or expense incurred. Experimental techniques such as NMR and X-ray crystallography are limited at a genomic scale. Accordingly, identifying sequences *in silico* has complemented biophysical characterisation. Computational comparisons of HAR1 for *Homo* and *Pan* suggest some differences which may have implications in evolutionary biology. Whilst molecular configurations for both show close similarity in the helices, the median region may be open to various interpretations. 2D and 3D predictions for the helixC-H4 loop in both species predict >50% probability and a more open loop in *Pan*. Free energy of binding for a snoRNA to this sequence is more favourable for *Homo*.
7.2.2.3.4 Snora binding sites defined by probability

Following the mapping of 3D structures of HAR1A\_Pollard and HAR1, the threshold value for the number of nts of a snoRNA complementary to HAR1A\_Pollard was calculated. FWER for the three species were: snora: 0.1610, snord: 0.1136 and scarna: 0.0361. In this analysis, $\alpha = 0.005$. To counteract the FWER, a Bonferroni correction was applied which calculated $p = 2.0 \times 10^{-4}$. Query strands modelled with a range of 13 to 20 nts were tested in equation 1 for consequent $p$ values (Table 7.2):

<table>
<thead>
<tr>
<th>$n$</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$3.1 \times 10^{-3}$</td>
<td>$7.8 \times 10^{-4}$</td>
<td>$1.9 \times 10^{-4}$</td>
<td>$4.7 \times 10^{-5}$</td>
<td>$1.2 \times 10^{-5}$</td>
<td>$2.9 \times 10^{-6}$</td>
<td>$7.1 \times 10^{-7}$</td>
<td>$1.8 \times 10^{-7}$</td>
</tr>
</tbody>
</table>

Given the range of $y$-axis values precluded clarity of visualisation, these empirical data were plotted with a square root transformation \((397)\) \((398)\) as shown in figure 7.14:

![Figure 7.14: square root transform of $p$ values for query strands comprising 13 – 20 nts](image)

The selection of a threshold value for $n$ was contingent on two factors. Firstly, figure 5 indicates $n = 15$ is the lowest value to satisfy $p \geq 2.0 \times 10^{-4}$. Secondly, when $n$ values are plotted against the number of predicted complementary strands for forty snords (figure 7.15), $n = 15$ is the first point where the data appear to be approaching an asymptote.
Figure 7.15: predicted number of complements for nts in query strands comprising 12 to 20 nts

Given the proportion of type I errors increases with the volume of data analysed, a statistical correction of multiple data returns in this study was considered appropriate. The Bonferroni method was chosen because of recommendation for large data sets and maximum avoidance of false positives (399).

Prior to a Bonferroni correction, 11.6% of snoras, 11.2% of snords and 27.6% of scarnas have complementarity for $\alpha = 0.005$ from a total of 568 samples. When corrections are applied, these figures reduce to 0%, 7.3% and 0%. Data for snords with $p \geq 2.0 \times 10^{-4}$ and $\alpha = 0.005$ are shown in Table 7.3:
Table 7.3: Bonferroni-corrected data for snords with $p \leq 2.0 \times 10^{-4}$, $\alpha = 0.005$

<table>
<thead>
<tr>
<th>snord</th>
<th>nts (x)</th>
<th>mer (n)</th>
<th>p value</th>
<th>NCBI accession</th>
<th>chromosome</th>
<th>MFE kcal/mol</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b</td>
<td>85</td>
<td>27</td>
<td>$1.8 \times 10^{-4}$</td>
<td>NR_004396.1</td>
<td>17q25.1</td>
<td>-5.39</td>
<td></td>
</tr>
<tr>
<td>35a</td>
<td>88</td>
<td>22</td>
<td>$1.8 \times 10^{-4}$</td>
<td>NR_000181.8</td>
<td>19q12.33</td>
<td>-9.02</td>
<td></td>
</tr>
<tr>
<td>35b</td>
<td>88</td>
<td>83</td>
<td>$1.8 \times 10^{-4}$</td>
<td>NR_004285.1</td>
<td>19q13.33</td>
<td>-7.13</td>
<td></td>
</tr>
<tr>
<td>36b</td>
<td>71</td>
<td>27</td>
<td>$1.3 \times 10^{-4}$</td>
<td>NR_000171.7</td>
<td>9q34.2</td>
<td>-3.27</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>74</td>
<td>27</td>
<td>$1.5 \times 10^{-4}$</td>
<td>NR_000152.5</td>
<td>1p34.1</td>
<td>-5.21</td>
<td></td>
</tr>
<tr>
<td>58c</td>
<td>90</td>
<td>15</td>
<td>$1.9 \times 10^{-4}$</td>
<td>NR_003701.1</td>
<td>18q21.1</td>
<td>-5.14</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>67</td>
<td>15</td>
<td>$1.3 \times 10^{-4}$</td>
<td>NR_001294.1</td>
<td>15q11.2</td>
<td>-1.65</td>
<td>P-W syndrome; 2xG-U doublet</td>
</tr>
<tr>
<td>87</td>
<td>76</td>
<td>24</td>
<td>$1.5 \times 10^{-4}$</td>
<td>NR_002598.1</td>
<td>8q13.1</td>
<td>-9.12</td>
<td>2xG-U doublet</td>
</tr>
<tr>
<td>113-6</td>
<td>74</td>
<td>15</td>
<td>$1.5 \times 10^{-4}$</td>
<td>NR_003234.1</td>
<td>15q32.1</td>
<td>-4.88</td>
<td>P-W syndrome; 2xG-U doublet</td>
</tr>
<tr>
<td>114-2</td>
<td>77</td>
<td>19</td>
<td>$1.5 \times 10^{-4}$</td>
<td>NR_003194.1</td>
<td>14q32.0</td>
<td>-6.03</td>
<td>P-W syndrome</td>
</tr>
<tr>
<td></td>
<td>-30</td>
<td>71</td>
<td>$1.3 \times 10^{-4}$</td>
<td>NR_003223.1</td>
<td>14q32.1</td>
<td>-7.44</td>
<td>P-W syndrome</td>
</tr>
<tr>
<td>115-36</td>
<td>82</td>
<td>15</td>
<td>$1.7 \times 10^{-4}$</td>
<td>NR_003351.1</td>
<td>15q11.2</td>
<td>-3.55</td>
<td>P-W syndrome; 2xG-U doublet</td>
</tr>
<tr>
<td>-38</td>
<td>82</td>
<td>16</td>
<td>$1.7 \times 10^{-4}$</td>
<td>NR_003353.1</td>
<td>15q11.2</td>
<td>-6.26</td>
<td>P-W syndrome</td>
</tr>
<tr>
<td>-45</td>
<td>53</td>
<td>17</td>
<td>$9.6 \times 10^{-5}$</td>
<td>NR_003498.1</td>
<td>15q11.2</td>
<td>-9.08</td>
<td>G-U doublet</td>
</tr>
<tr>
<td>-47</td>
<td>33</td>
<td>22</td>
<td>$4.4 \times 10^{-5}$</td>
<td>NR_003491.1</td>
<td>15q11.2</td>
<td>-0.21</td>
<td></td>
</tr>
<tr>
<td>-48</td>
<td>76</td>
<td>32</td>
<td>$1.5 \times 10^{-4}$</td>
<td>NR_003362.1</td>
<td>15q11.2</td>
<td>-3.96</td>
<td>P-W syndrome; ser-rep 2c</td>
</tr>
<tr>
<td>130</td>
<td>85</td>
<td>15</td>
<td>$1.8 \times 10^{-4}$</td>
<td>NR_047032.1</td>
<td>15q11.2</td>
<td>-8.42</td>
<td></td>
</tr>
<tr>
<td>137</td>
<td>79</td>
<td>17</td>
<td>$1.6 \times 10^{-4}$</td>
<td>NR_132576.1</td>
<td>9p23</td>
<td>-7.49</td>
<td></td>
</tr>
<tr>
<td>147</td>
<td>85</td>
<td>16</td>
<td>$1.8 \times 10^{-4}$</td>
<td>NR_145810.1</td>
<td>11p15.1</td>
<td>-3.49</td>
<td></td>
</tr>
<tr>
<td>157</td>
<td>68</td>
<td>19</td>
<td>$1.4 \times 10^{-4}$</td>
<td>NR_145781.1</td>
<td>19q13.42</td>
<td>-7.39</td>
<td></td>
</tr>
</tbody>
</table>

Data for Table 7.3 prior to correction and excluded for analysis are shown in Table 7.4.

Table 7.4: snords with $p \leq 2.0 \times 10^{-4}$, $\alpha = 0.005$ prior to data correction

<table>
<thead>
<tr>
<th>snord</th>
<th>nts (x)</th>
<th>mer (n)</th>
<th>p value</th>
<th>NCBI accession</th>
<th>chromosome</th>
<th>MFE kcal/mol</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>148</td>
<td>19</td>
<td>$3.4 \times 10^{-4}$</td>
<td>NR_002604.2</td>
<td>17p13.1</td>
<td>-3.07</td>
<td>G-U doublet</td>
</tr>
<tr>
<td>90</td>
<td>107</td>
<td>11</td>
<td>$2.4 \times 10^{-4}$</td>
<td>NR_003071.1</td>
<td>9q33.2</td>
<td>-1.69</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>142</td>
<td>27</td>
<td>$3.3 \times 10^{-4}$</td>
<td>NR_004402.1</td>
<td>11p15.4</td>
<td>-6.64</td>
<td></td>
</tr>
<tr>
<td>116-10</td>
<td>104</td>
<td>17</td>
<td>$2.3 \times 10^{-4}$</td>
<td>NR_003325.1</td>
<td>15q11.2</td>
<td>-5.28</td>
<td>Poly(U)</td>
</tr>
<tr>
<td>-22</td>
<td>104</td>
<td>55</td>
<td>$2.3 \times 10^{-4}$</td>
<td>NR_003356.2</td>
<td>15q11.2</td>
<td>-5.18</td>
<td></td>
</tr>
<tr>
<td>-25</td>
<td>98</td>
<td>32</td>
<td>$2.1 \times 10^{-4}$</td>
<td>NR_003462.1</td>
<td>15q11.2</td>
<td>-9.36</td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>137</td>
<td>22</td>
<td>$3.2 \times 10^{-4}$</td>
<td>NR_003294.2</td>
<td>17p13.1</td>
<td>-5.92</td>
<td>Poly(A) site; G-U doublet</td>
</tr>
<tr>
<td>126</td>
<td>99</td>
<td>26</td>
<td>$2.1 \times 10^{-4}$</td>
<td>NR_002692.1</td>
<td>14q11.2</td>
<td>-6.03</td>
<td>G-U doublet</td>
</tr>
<tr>
<td>142</td>
<td>195</td>
<td>17</td>
<td>$4.7 \times 10^{-4}$</td>
<td>NR_132758.2</td>
<td>10p15.2</td>
<td>-3.65</td>
<td>Caja?</td>
</tr>
<tr>
<td>145</td>
<td>106</td>
<td>26</td>
<td>$2.2 \times 10^{-4}$</td>
<td>NR_145816.1</td>
<td>1p34.1</td>
<td>-3.04</td>
<td></td>
</tr>
</tbody>
</table>

Results with minimum free energy (MFE) $< ~10$ kcal mol$^{-1}$ and $> 0$ kcal mol$^{-1}$ were excluded from further analysis as values outside of this range are considered less reliable (400). Such RNA-RNA complements with MFE $< ~10$ kcal mol$^{-1}$ would be unduly exothermic and those with MFE $> 0$ kcal mol$^{-1}$ endothermic. Both conditions are thermodynamically unfavourable.
45% of the snords itemised in Table 3 have commonality with pathogenic structures and functions. Of these twenty results, 25% contained guanine-uridine doublets and 35% were associated with Prader-Willi (P-W) syndrome. G-U doublets were noted in 22% of snoras, 38% of snords and 57% of scarnas prior to Bonferroni corrections. However, only a small minority of snords were characterised by $p \geq 2.0 \times 10^{-4}$ post-correction for association with P-W syndrome or G-U doublets. The potential significance of this latter observation is noted in Future Perspectives. A relatively high proportion of complementary interactions appears in variants of snord113 to snord115.

7.2.2.3.5 snoRNA complementarity in HAR1A

Primary analysis of snoRNAs with HAR1A_Pollard was made after determining the threshold of 15nts as suitable for snoRNA complementarity. snoRNA sequences were downloaded from comprehensive databases and complementarity with HAR1A_Pollard investigated using IntaRNA. Based on binding potential and significance values, initial data were filtered to benchmark plausible candidates. IntaRNA reckons HAR1A_Pollard as the target and the interacting snora as query. Target sequences run 5’ to 3’ and query sequences as anti-sense. Where numerically possible, five nts are included both upstream and downstream of the interaction.

Snora29 has a 36-mer anti-sense strand with good complementarity to HAR1 nt2394>2423. The strand also has high complementarity within an 83nt sequence for snord35B. Although both interactions do not occur simultaneously, the complementary alignments for snora29 and snord35b with HAR1A are shown in Figure 7.16.

![Figure 7.16: snora29 and snord35b alignments with HAR1A_Pollard.](image)

The majority of mutual complementarity aligns with a central core section of HAR1. Gaps in a sequence indicate a loop or kink in the interacting strand.

Figure 7.17a illustrates complete complementary pairing, whilst Figure 7.17b has a single adenine kink opposite a single guanine kink.
Figures 7.17a and 7.17b: predicted complementarity for HAR1a_Pollard with snord116-10 and snord116-28. Interestingly, snord116-10 is uniquely characterised by polyuridinylation between nt29 and nt44. In RNA viruses, this structure provides a template for polyadenylation and is critical in the replication cycle. Its relevance in HAR1a is currently unknown.

The energy term derives from the difference between the initial condition of two non-interacting (separate) RNA strands and the interacting state. If the bound condition has higher energy than the initial condition, the system is more thermodynamically stable. Consequently, the interacting condition is exothermic. Effectively, this metric predicts the degree of Gibbs free energy required to restore the bound system to separated non-interacting strands. H-bonding involves hybridisation of orbitals which reduces local energy of the bound pair. Hence hybridisation energy is also exothermic.

Free energy values in the range ~0 kcal mol\(^{-1}\) to -10 kcal mol\(^{-1}\) are consensually regarded by the software designers as valid; values outside these markers imply an unstable system. An interaction will either not easily occur or the bound state will rapidly collapse. Hybridisation energies are acceptable for values below -10 kcal mol\(^{-1}\).

Figure 7.18 shows a 9nt loop in the leading strand and a 7nt loop in the lagging strand. These features are predicted by the algorithm contingent on the hybridisation free energy of the interacting strands and the free energy required to separate the interacting sites in both target and query. If these thermodynamic parameters are not met, even where canonical H-bonding would be expected, no interaction is predicted. For instance, in this calculation, G-C and G-U opposites are located upstream and downstream of the interacting sequences but are not predicted to H-bond.
The two helices and loops have slightly different individual free energies.

Microdomains in HAR1_Pollard relevant to this study are identified in Figure 7.19:
Figure 7.19: microdomains in HAR1A_Pollard. (1) cytidine / uridine iterative sequence, (2) polyadenylated region providing interaction site for poly(U) sequence of snord116-10, (3) overlap region for snora29 and snord35b (4) HAR1a 118nt sequence. (1) is a true loop, (2) is an open loop located at the gene’s terminus, (3) is a pair of sequences linked by four very short loops in opposing strands, (4) is a stem loop associated with three small loops.

7.2.3 Overview of pathogenic IncRNAs interacting with HAR1A

HAR1A is a human-specific sequence, containing the evolutionary accelerated HAR1 region (401). The IncRNA is specifically localised in the nucleus. Although HAR1A has been found to have aberrant expression primarily in cancer and neurological disorders, the mechanism of action is currently unknown. Function of IncRNAs can be related to structure because of their derivation from a poor sequence conservation.

Several of the statistically-validated snoRNAs with predicted interaction with HAR1A are yet to be characterised. These include snord1B, snord36B, snord58C, snord114-30, snord130, snord137, snord147 and snord157. Databases carry minimal information on these. However, a significant fraction of candidates identified here have ostensible pathogenic associations.
SnoRNAs predicted to interact with HAR1A have also been studied in neurological disorders such as Alzheimer’s disease. A murine model correlated a cluster of differentially correlated snoRNAs (including snord55 and snord87), with Alzheimer’s disease (401). Further research is required to understand their role in this illness.

Several candidates with predicted HAR1A interactions (e.g., snord64, snord113, snord114 and snord115), are correlated with P-W syndrome (402) and as reflected in Table 7.3. The condition derives from an absence of genes from chromosome 15. Infantile hypotonia, developmental delay, childhood obesity and behavioural problems have been noted. LncRNA U-UBE3A-ATS, located in chromosome 15q11-q13, contains *inter alia* snord64 and snord115. Clusters of both were found to have very low expression in P-W syndrome patient-derived induced pluripotent stem cells, compared to a normal control (403), and are therefore a part of the deletion leading to the P-W phenotype (404).

**Snord35A and snord35B** have little information in the literature, but probably guide 2’O-ribose methylation of other small nuclear RNAs (371); snord35A is expressed in the choroid plexus. Snord35B ostensibly has very high complementarity with HAR1 but this may be an artefact of the software. The algorithm might have correctly identified a suboptimal match, e.g., the two helices linked by the UGCACCC loop and the thermodynamic profile suggests further interactions adjacent to the region. If so, the energy for any suboptimal matches should be very similar to -7.1 kcal mol⁻¹ (Supplementary figure 5). This value was not found. Data input and formatting was checked and re-submitted several times with identical results. It therefore remains possible the 84-mer sequence is wholly or partially complementary to HAR1: a series of helices with small bulges and linked by a 7nt loop. The stable core may be the central group linked by this loop, or the downstream groups linked by AACC and CCC kinks. The paired fragments upstream and downstream of these stable cores may be interacting transiently or intermittently. The region’s 81% complementarity might also be considered as a series of very closely-located “islands” of bonding. Both interpretations imply a dynamic but stable pairing over some distance. Snord35A was found to be specifically overexpressed in pancreatic cancer, reducing the expression of C-Met (405) and overexpressed in breast cancer brain metastases (406). Deletion of snord35A was reported to suppress clonogenic potential of leukaemia cells and leukaemogenesis (407). Snord35A was part of a signature released from macrophages during inflammation (408). Snord35B was determined as contributive to an rRNA methylation complex signature, and analysed in relation to B-cell precursor acute lymphoblastic leukaemia (409).

Snord35B has been implicated in COVID-19. Significant upregulation for post SARS-Cov-2 infection in lung adenocarcinoma cells has been reported (410). Moreover, recent
research concludes that snord113-6 participation in methylation, integrin signalling and mRNA degradation in human primary fibroblasts (411).

**Snord55** was found to be significantly decreased in tumour-educated platelets of early non-small cell lung cancer patient samples (412).

**Snora100** has a polyuridine sequence that exactly complements the 3′ poly(A) region of HAR1.

**Snord114-2** was part of a snoRNA signature that was downregulated in omental metastasis tissue where a lncRNA MIAT had increased expression, in comparison to ovarian cancer tissue. However, no significant difference was recorded for snord144-2 (413).

**Snord115** clusters are reported to be exclusively expressed in brain tissue (414). There are currently 48 variants known. It is present in neurons but absent in the choroid plexus, also having sequence complementarity with the serotonin receptor 5-HT(2C)R (415). The conserved 18-nt antisense element 5′ AUGCUAAUAGGAUUACG 3′ has perfect complementarity to 5-HT(2C)R. IntaRNA predicts a 15-mer complementarity to HAR1 where G-U pairing is included (figure 7.19): Snord115 has an 18nt complementarity to serotonin receptor 5-HT(2C)R expressed in neurons whereas 5-HT(2C)R pre-RNA is expressed in the choroid plexus. However, the formation of a duplex comprising snord115 and 5-HT(2C)R pre-RNA as an *in vivo* gain-of function has been reported (416).

![Figure 7.20: conserved bases in HAR1, snord115 and 5ht(2C)R. Leading strand bases in common shown in red, lagging strand complements in mauve. In terms of H-bonding, leading strand differences are neutral.](image)

**Snord116-10** has a similar sequence to snora100. The implications of this characteristic appear not to be noted in the literature. Both feature a poly(U) strand which the algorithm predicts to bind with the polyadenylated 3′ region of HAR1.
7.2.4 Discussion of snora interactions with HAR1A

The legacy of evolutionary functional innovation is often characterised by abrupt mutations in highly-conserved locations historically predicated on negative selection. Highly accelerated regions (HAR) sequences are closely associated *inter alia* with nucleic acid metabolism regulation. Comparison of nucleotide substitution rates for all HARs across several species were found to be unexpectedly high specifically for *Homo sapiens*. Because these sequences in vertebrates were highly conserved, attention focussed on potential human-specific evolutionary advantages. The statistically aberrant incidence of single nucleotide substitutions in these mRNA fragments for *Homo* relative to other vertebrata, might be an important factor in Hominid speciation.

Given their proximate relationship to a genomic region characterised by mutation, *HAR1a* and *HAR1b* are historically very stable in non-human species. *HAR1a* mediates embryonic neurological progress: expression is high in the developing neocortex. Because this region is notably more developed in *Homo* than *Pan* for instance, this observation is significant. Furthermore, expression of *HAR1a* is absent in other regions of the forebrain. Taken together, these observations could underpin the emergent pre-eminence of *Homo sapiens* as developing towards an exceptionally cognitive species. Some neurological dysfunction and tumorigenesis have recently been associated with HAR1A loci.

Current information is minimal for interactions between HAR1A and small nucleolar RNAs. Secondary structure models of HAR1A and tertiary structure models for *HAR1a* assessed in this thesis for potential complementarity with snoras revealed a small number of statistically significant results. Whereas these are reported here in purely computational terms, tangible validation can only derive from direct experimentation.

7.2.3.1 5UNE

Given its proposed significance for pre-frontal cortical development (Sections 7.2.1 and 7.2.2.3), a brief survey of (PDB) 5UNE may support inquiry into potential snora-HAR1 interaction. The sequence was opened in RNA Composer to provide a 3D configuration for comparison with the Protein Data Bank model. This structure is helical with two minor bulges and ligates Mg\(^{2+}\). The ion binds to phosphate oxygens of uracil 10, cytosine 30 and guanine 31 (figure 7.21):
Binding energies are moderately exothermic and therefore thermodynamically favourable. Ions appear to preferentially bind with a phosphate group oxygen of HAR1, and guanine O6 of 5UNE. There are no H-bonds, but each ion except for Mg$^{2+}$ has two electrostatic bonds of < 3 Å. Mg$^{2+}$ has three electrostatic bonds of < 3Å. Binding data for Mg$^{2+}$, Zn$^{2+}$, Ca$^{2+}$ or Fe$^{2+}$ with 5UNE or HAR1 are shown in Table 7.5:

Table 7.5: Autodock optimum binding affinities for divalent metal cation ligation with 5UNE or HAR1.

<table>
<thead>
<tr>
<th></th>
<th>Mg$^{2+}$</th>
<th>Zn$^{2+}$</th>
<th>Ca$^{2+}$</th>
<th>Fe$^{2+}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUNE</td>
<td>HAR1</td>
<td>SUNE</td>
<td>HAR1</td>
</tr>
<tr>
<td>binding energy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kcal/mol</td>
<td>-1.6</td>
<td>-1.1</td>
<td>-1.5</td>
<td>-1.0</td>
</tr>
<tr>
<td>ligating atoms</td>
<td>U10-OP1</td>
<td>U19-OP1</td>
<td>G8-O6</td>
<td>U18-OP1</td>
</tr>
<tr>
<td></td>
<td>C30-OP2</td>
<td>G32-O6</td>
<td>U7-O6</td>
<td>U19-OP2</td>
</tr>
<tr>
<td></td>
<td>G31-OP2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distance (Ang)</td>
<td>2.322</td>
<td>2.561</td>
<td>2.597</td>
<td>2.587</td>
</tr>
<tr>
<td></td>
<td>2.451</td>
<td>2.574</td>
<td>2.952</td>
<td>2.333</td>
</tr>
<tr>
<td></td>
<td>2.334</td>
<td>2.811</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AutoDock predictions calculate Ca$^{2+}$ and Fe$^{2+}$ have greater dynamic activity than Zn$^{2+}$, while Mg$^{2+}$ has the least. The majority of interactions are between a phosphate oxygen
and the ion. No H-bonds were noted. ADT identifies an optimal binding site for snord35b and 5UNE mediated by Mg$^{2+}$ (figure 7.2):

![AutoDock calculation for snord35b ligation with the HAR1 fragment located between helices C/H4 and H3 (5UNE). A single ADT calculation found an optimal binding energy of 1.63 kcal mol$^{-1}$. Since this estimation is marginally endothermic, some input of energy may be required to complete the docking. However, ADT can only calculate a global value: given the target and query molecules are large, the local binding data may be exothermic and therefore thermodynamically viable. This interpretation is reinforced by the initial ADT selection of a binding site which expected a close interaction between target and query mediated by Mg$^{2+}$.](image)

Figure 7.22: AutoDock calculation for snord35b ligation with the HAR1 fragment located between helices C/H4 and H3 (5UNE). A single ADT calculation found an optimal binding energy of 1.63 kcal mol$^{-1}$. Since this estimation is marginally endothermic, some input of energy may be required to complete the docking. However, ADT can only calculate a global value: given the target and query molecules are large, the local binding data may be exothermic and therefore thermodynamically viable. This interpretation is reinforced by the initial ADT selection of a binding site which expected a close interaction between target and query mediated by Mg$^{2+}$.

7.2.3.1.1 Discussion of 5UNE

The tertiary structure of 5UNE derived from RNA Composer compared well with the PDB model. The data for divalent cationic non-covalent ligation are very similar and ADT calculates parallels in location: all are peripheral to a minor groove in the duplex system.

The eRMSD analysis implies the base-pairing quartets remain in a reasonably stable geometry, whilst bases in the loops are more dynamically active. Radius of Gyration data show the system maintains a marginally variable compactness. Because RMSD provides an overall interpretation of activity, individual contributions from the other metrics are somewhat masked. Minimal quartet-pairing and phosphate backbone activity is subsumed at ~50ns by loop base activity. Visual inspection of the simulation tends to confirm this interpretation.
7.2.3.2 Degeneracy in molecular biology

Most cellular systems at biochemical level display degeneracy: that is, unalike molecules can have a similar function in one context, but different interactions in another. For instance, because initiation of transcription may occur at various 5' locations, a given gene sequence does not necessarily have immutable definition \(^{(417)}\). Furthermore, alternative pre-RNA splicing patterns derived ultimately from a single transcript can induce differential degenerate isoforms \(^{(418)}\). Degeneracy has been considered a *sine qua non* of evolution because a majority of mutations would be adverse in non-degenerate circumstances \(^{(419)}\).

The integrity of RNA responds to selection pressure; primary sequences and secondary structures are characterised and sustained by thermostability. An experimental metric of RNA function is thermostability \(^{(420)}\). Functionally significant regions of non-coding RNA have greater stability than random control fragments \(^{(421)}\). Regions of high \(T_m\), *ie.* thermal stability, may therefore comprise important functional elements which are conserved despite robust selection pressure.

Variation of RNA melting temperature \((T_m)\) is strongly associated with regions of high and low thermostability. For this study, melting temperatures were calculated *in silico* for potential G4-folding sequences (PG4S) in hHAR1 using Equation 7.3. \(T_m\) was also calculated for fragments with similar guanine /cytosine content but without G4-folding potential (non-PG4S). These data are shown in table 7.6. Degenerated nucleotides are excluded from base-stacking \(T_m\) calculations but are included for basic \(T_m\). Salt concentration was 50mM at physiological pH.

\[
T_m = 64.9 + 41 \times (yG+zC-16.4)/(wA+xU+yG+zC) \quad \text{equation 7.3}
\]

where \(w,x,y,z\) are the number of the bases A,U,G,C in the sequence respectively \(^{(109)}\).
Table 7.6: Guanine / cytosine ratios and calculated melting temperatures for potential and non-potential G4-folding sequences.

<table>
<thead>
<tr>
<th>PG4S</th>
<th>non-PG4S</th>
</tr>
</thead>
<tbody>
<tr>
<td>start</td>
<td>length</td>
</tr>
<tr>
<td>106</td>
<td>26</td>
</tr>
<tr>
<td>150</td>
<td>24</td>
</tr>
<tr>
<td>1095</td>
<td>26</td>
</tr>
<tr>
<td>1174</td>
<td>23</td>
</tr>
<tr>
<td>1425</td>
<td>22</td>
</tr>
<tr>
<td>1558</td>
<td>23</td>
</tr>
<tr>
<td>1752</td>
<td>20</td>
</tr>
<tr>
<td>2131</td>
<td>20</td>
</tr>
<tr>
<td>2482</td>
<td>18</td>
</tr>
<tr>
<td>2601</td>
<td>24</td>
</tr>
<tr>
<td>mean</td>
<td>22</td>
</tr>
</tbody>
</table>

G / C ratios are very similar for both systems and G / C enthalpy changes are generally slightly higher compared to A / T. However, melting temperatures suggest PG4S fragments are more stable than the enthalpy change differentials allow. Histograms for PG4S and non-PG4S agree well with the mean values of Table 7.6 (figures 7.23a and 7.23b):

Figures 7.23a and 7.23b: histograms and melting temperatures of potential G4 sequences and non-potential G4 sequences.
7.2.3.3 Guanine-uridine pairs in tandem.

Whilst base-pairing is mainly responsible for folding of nucleic acids, binding small nucleolar strands, even intermittently, might play a part. In particular, this analysis suggests a specific characteristic of some snoRNAs could allow divalent metal cations to bind. Guanine-uridine pairing is regarded as thermodynamically favourable and biologically essential (422). It is noteworthy that there are appears to be neither a repair mechanism for a G/U “mismatch” nor a tagging enzyme. Moreover, divalent metal cations show strong affinity for G-U pairing in tandem (423). Some research concludes parallel doublets of G/U pairs may also serve as binding sites for catalysis (424). G-U doublets located in the major groove of double-stranded nucleic acids afford an effective site for metal cation binding (425). Surface electronegativity is notably higher in the G-U environment compared to canonical pairing (Figure 7.24).

![Figure 7.24](image.png)

**Figure 7.24.** Duplex RNA strands with guanine-uridine pairs (A) show higher surface electronegativity compared to guanine-cytosine pairs (B). Electronegative regions depicted in red. (Adapted from Varani et al, 2000)

The enhanced metal cation binding is exemplified in figure 7.25 where divalent calcium shows more hydrogen-bonding in a guanine-uridine doublet compared with tandem guanine-cytosine pairing.
Figure 7.25. Ca\(^{2+}\) binds with 6 hydrogen bonds in *Homo* duplex RNA G-U pairs; the ion has four H-bonds in G-C pairs. Such interaction may influence molecular folding. Several snoRNAs with very low \(p\) values for HAR1a interaction in this study were noted to harbour G-U pairs in tandem (Table 7.3).

Binding divalent metal cations would reduce local electronegativity. This effect may offset or neutralise any repulsive force sufficiently to allow a polypeptide with local electronegative surfaces to ligate with the snoRNA. Perhaps transiently, perhaps a more durable interaction. Ligation with cations and / or macromolecules could influence snoRNA folding and subsequent function such as potential interaction with HAR1A for example.

### 7.24 Executive summary of Chapter 7

- Studies of highly-accelerated genomic regions (HARs) are attracting attention because of association with statically aberrant mutation rates between *Homo* and other vertebrates.
- HARs have particular relevance contingent on proposals of evolutionary advantages.
- Stability of HAR1A (*Homo*) is enhanced by specific mutations.
- This thesis finds sequences within HAR1A (*Homo*) that would allow purine quadruplexes to fold. These structures could also reinforce local stability.
- There is a significant sequential difference between *Homo* and *Pan* which tends to support the concept of evolutionary advantage.
- Secondary and tertiary modelling of HAR1A and HAR1 *Homo* and *Pan* in this work reveals structural differences that may promote ligation with polypeptides, RNAs or other oligonucleotides.
• Molecular docking protocols for ligation between a HAR1A fragment and a small nucleolar RNA indicate binding would be mediated by Mg\textsuperscript{2+}. The ion is known to be present peripherally to HAR1

• Interactions of HAR1A (expressed in the nucleus) with small nucleolar RNAs were identified in a small number of analyses.

• Such interactions are plausibly influential in routine cellular processes

• Further analysis of specific small nucleolar RNAs finds several instances of non-canonical base-pairing between guanine and uridine. Where two such instances occur directly adjacent to each other, local electropositivity is reduced. Divalent metal cation binding is effectively greater than for RNAs without this base-pairing.

• It is proposed here that cationic ligation would re-configure a small nucleolar RNA with (as yet) unknown consequences for cellular processes.

• Inspection of the SARS-CoV-2 genome also finds instances of sequences allowing purine quadruplexes to fold.

• These sequences are mainly located in coding regions for structural proteins in the virus, but some occur in a polyadenylated feature highly relevant to viral replication.

• The poly(A) tail is common to all RNA viruses

• Biplanar purine quadruplexes in HAR1A and SARS-CoV-2 are found to be stable in this study, but their impact is indeterminate.

Future perspectives

The HAR1A (*Homo*) sequence allowing a guanine-adenine hybrid to fold is located immediately upstream of *HAR1* is absent in *Pan HAR1*. This difference may contribute to the stability of the 118nt sequence in *Homo* with consequent influence on its integrity. The notion commends further enquiry.

The single-base mutation rate between *Homo* and *Pan* in *HAR1* could be different. A given base superficially changed from say, C to G may have followed a route such as C>A>G. in this case, although the number of actual mutations is greater, the end result is still C>G. But if the route were C>G>C, the number of mutations is greater, but there is no apparent change. The expected rate of *HAR1* single-base mutation is 0.27 between *Homo* and other vertebrates, but the observed rate is 18. An even higher rate of change would reinforce the premise for special consideration in *Homo*. Statistical analysis could explore the dichotomy.
Spatial differences between *Homo* and *Pan* in the H4 loop would plausibly affect binding of small molecules. Downstream investigation of this possibility is uncharted.

Adenine in both DNA and RNA can be amino-modified at C2’ if adenylosuccinate synthetase were present. The enzyme does seem to be ubiquitous in human cell biology. Future MD analysis might factor in this modification to assess any impact on quadruplex stability.

Docking protocols of snord35b ligated with a *HAR1* fragment provide some insight into the character of binding, particularly possible mediation by Mg$^{2+}$. Follow-up work should analyse divalent metal cations with physiological relevance. Although ligation is probably covalent, other factors (*eg*. enzymes, steric hindrance) might induce transience. Whilst metrics such as RMSD and Radius of Gyration would inform an overall view, mapping covalent and non-covalent networks would provide more detail.

No suggestion was made here regarding classes of molecules other than snoRNAs that might interact with HAR1A.

m6A in lncRNAs has been identified as a molecular switch in transcription for a specific lncRNA. Computational analysis of the modification if applied to other lncRNAs may further inform this observation.

Guanine-uridine pairing is regarded as thermodynamically favourable and biologically essential. With specific relevance to this study, Ca$^{2+}$ shows strong affinity for G-U pairing in tandem. The feature affords an effective site for metal cation binding derived from enhanced surface electronegativity. Ca$^{2+}$ ligation could influence snoRNA folding and subsequent function. This raises a possible productive line of inquiry by the consideration of guanine-uridine tandem pairs. Consolidation of published research and data returned in this study support proposals of divalent cation binding could reconfigure some snoRNAs. Altered form may alter function. Further work could explore issues such as anionic binding and physiologically relevant cations other than Ca$^{2+}$. Suitable candidates may be Mn$^{2+}$, Mg$^{2+}$ and Zn$^{2+}$.

Inclusion of the A-minor motif in IncRNA models may have some leverage in some RNA-RNA interactions by enhancing stability. Some of the enhanced stability for *HAR1 Homo* might derive from an approximately adjacent RNA containing an A-minor motif.

The iterative cytidine-uridine pattern in HAR1A appears not to have received comment. The unusual feature may be a stutter in transcription, an ancient fragment of invasive RNA, possibly viral, or an inefficient repair mechanism. Modelling the region might elucidate some insight: either it has a function, or is redundant.
Single-base mutations in human *HAR1 Homo* are all strengthening by virtue of increasing H-bonding by 50%. This strongly implies a consequent advantage in terms of selection pressure. However, there may be other strengthening mutations specific to *HAR1 Homo* as yet undetected. 2,6-diaminopurine has been recognised as a beneficial modification to adenine in a pre-biotic context (426). Furthermore, substitution of 2,6-diaminopurine for adenine in RNA is considered a stabilising modification (427). If the substitution were to occur in *HAR1 Homo*, the development would be better seen as a post-transcriptional modification rather than a mutation. In which case, an enzymatic mechanism would seem likely. Such systems are well-characterised in a DNA context involving 2,6-diaminopurine triphosphate and polymerase (428). A change of *HAR1* (*Homo*) adenine to 2,6-diaminopurine can therefore be seen as a strengthening transition. Were the modification to occur in *HAR1*, it would resonate closely with the advantageous single-base substitutions noted earlier. Molecular modelling and dynamics should provide a background to further work such as NMR spectroscopy. Both are suitable methodologies for identifying (or excluding) the molecule.
Conclusion

Molecular Dynamics analysis is not experimental validation and therefore has a boundary on interpretation. RMSD for instance, provides a general overview of dynamic activity but has limited information about base-pairing. Because values are contingent on the quantitative scale of data, a very large sample could show a biased trajectory. Furthermore, because the calculation is predicated on squared data, outliers may be disproportionately represented. However, enhanced in silico insights into the complexities of 3D RNA structures can improve understanding of biological function. Experimental techniques such as NMR and X-ray crystallography are limited at a genomic scale. Accordingly, identifying sequences computationally has complemented biophysical characterisation. The two approaches support and reinforce each other.

Whilst 2D representation software visually allocates a spectrum of probabilities, 3D programmes do not as yet have this amenity. To some extent, this is offset by IntaRNA algorithms factoring a degree of uncertainty into 3D calculations. Hence, a summary view of local thermodynamic characteristics would favour the likelihood of significant complementarity.

All molecular modelling packages accessed in this study have an option to find the lowest energy configuration. Reliability of models is therefore greatly enhanced by an iterative process which eventually yields the most efficient geometry. However, atomic distances and angles are not flexible in final lowest energy models. IntaRNA and molecular dynamics calculations compensate for this aspect providing a limited range of probable geometries.

Guanine quadruplexes were originally considered as putative features exclusive to single-strand telomeric DNA. Consensually, they are now regarded as non-randomly ubiquitous and have been validated in vivo by direct visualisation. Moreover, Introduction (i) signposted that their influence covers a spectrum from genomes of eukaryota and prokaryota to mitochondria and viruses. That such phylogenetically remote archetypes display extensive commonality supports the notion of an archaic evolutionary selection for guanine quadruplex motifs. Another layer of interpretation is added if, as this thesis proposes, genomic adenine sequences can also fold into quadruplexes.

Although context and understanding of supramolecular structures is nuanced, the overall impact of quadruplexes, hairpin loops (Section 5.3.3) and A-minor motifs (Section 4.26) ranges from viral anatomy to human physiology. Some of these features are highly-conserved and are implicitly imperative in molecular biology. Nucleic acids are perceptibly dynamic systems and the stability of such structures is contingent on many factors, only some of which have been tested here. For instance, the structural integrity of biplanar RNA adenine and guanine quadruplexes contrasts with DNA equivalents by the absence or presence of a co-ordinating cation (Section 1.1). The orthodox Watson-Crick base pairing in duplex DNA presents a marginally higher kinetic and thermodynamic barrier to
quadruplex folding than apposite to single strand RNA with consequent enhanced stability (Section 2.9).

Whilst guanine quadruplexes have been very well characterised to date, debate continues regarding their apparent paradoxical functions (Section 2.1.3) not only in duplex genomic DNA, but recently long non-coding RNAs (Section 7.1.1). Resolving the complete nature of adenine equivalents, if they exist, may be as genomically extensive and paradoxical. In RNA viruses, crucial post-3’ UTR polyadenylation could be adversely affected by adenine quadruplex folding. This observation resonates with guanine quadruplexes as arguably deleterious to the replication cycle of RNA viruses (Section 5.10). However, in some non-viral contexts, two observations imply the complexes may have neutral effect or even beneficial influence: the absence of a known repair mechanism, and data in this study indicating the model is non-transient (Section 4.25). Natural Selection may favour purine quadruplexes in certain biological contexts, but not in others.

Thesis retrospective

This thesis has explored various supramolecular purine- and pyrimidine-derived biplanar and triplanar quadruplexes modelled in DNA and RNA. In the more stable complexes, a trend has emerged of systems folding and unfolding, dynamically and intermittently, and not without eventual physiological consequences. This binary activity may be regulated by helicases dismantling systems and, in some cases, N6-methylation labelling them for repair. However, implicit in this industrious cellular environment, is an unidentified rationale: the repetitive folding-unfolding iteration requires periodic energy transactions. Macro-physiological and micro-cellular benefits are contingent on the enterprise of such transactions. However, with regard to quadruplexes, the benefits are seldom clear.

Cellular consonance relies on many actors and numerous restoratives for disharmony. Some purine quadruplexes participate in roles that are paradoxically adverse and beneficial. Their contribution appears dependent on morphology, topology and timing. But uncertainties recur. Having proposed in this study how and where these supramolecular structures might arise and behave, why a folding-unfolding manoeuvre would eventuate commends further inquiry.
Software Acknowledgements

I am obliged to Dassault Systèmes BIOVIA, Discovery Studio Modelling Environment, Release 2017, San Diego for licence waiver.

Grateful thanks also to ARCHER2 and the Engineering & Physical Sciences Research Council for computing facilities and funding.

References


020-0236-x.


98. Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M.,


238. Liu,R.J., Long,T., Li,J., Li,H. and Wang,E.D. (2017) Structural basis for substrate binding and 268


following first peak of the pandemic: REACT2 study in 100,000 adults. medRxiv, 10.1101/2020.08.12.20173690.


336. Alphabetized List of Recognition Specificities | NEB.


276


392. Lorenz, R., Bernhart, S.H., Hoener, C., Siederdissen, Z., Tafer, H., Flamm, C., Höner Zu


