Characterization of Nucleic Acids and Nucleic Acid Binding Proteins Using NMR

Marina Evich
Georgia State University

Follow this and additional works at: https://scholarworks.gsu.edu/chemistry_diss
CHARACTERIZATION OF NUCLEIC ACIDS AND NUCLEIC ACID BINDING PROTEINS

USING NMR

by

MARINA EVICH

Under the Direction of Markus W. Germann, PhD

ABSTRACT

Biological systems are governed by molecular interactions. At the core of all biological systems is the processing of nucleic acids, and preserving the fidelity of the genome is crucial to those systems. This manuscript follows different biological systems interrogated by nuclear magnetic resonance: the structural impact of single ribonucleotide damage in DNA, the formation of oligomeric protein-DNA complexes of an ETS family transcription factor, how methylation affects the charge state of the amino acid arginine and how the base pair opening rates of an oligonucleotide is affected by inosine.
INDEX WORDS: Nucleic acids, Nuclear magnetic resonance, Single ribonucleotide damage,
Nucleic acid structure, Transcription factors, ETS family protein, PU.1 protein,
Biological interactions, Arginine methylation, Base pair opening, Inosine nucleobase
CHARACTERIZATION OF NUCLEIC ACIDS AND NUCLEIC ACID BINDING PROTEINS
USING NMR

by

MARINA EVICH

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy
in the College of Arts and Sciences
Georgia State University
2017
CHARACTERIZATION OF NUCLEIC ACIDS AND NUCLEIC ACID BINDING PROTEINS USING NMR

by

MARINA EVICH

Committee Chair: Markus Germann
Committee: David Wilson
Kathryn Grant

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
December 2017
DEDICATION

This work is dedicated to my mother Svetlana, my sisters Viktoria and Janet and to my husband Charlie. Without your love and support, this would not be possible.
ACKNOWLEDGEMENTS

All of the work done for these projects was carried out under the supervision of Dr. Markus W. Germann. Through Dr. Germann’s patience, direction and insight, I have gained valuable knowledge and experience. I have enjoyed working and learning in this laboratory, and for that, I cannot express my gratitude enough. Thank you Dr. Germann for investing time in me and teaching skills that I will have for the rest of my life. I would like to thank my committee members Dr. W. David Wilson and Dr. Kathryn Grant and Dr. Donald Hamelberg for the insight and all of your help during my studies at Georgia State University. Additionally, I could not have gone so far without the help of my lab mentor and fellow graduate student/post doctorate Dr. Alexander M. Spring-Connell. I appreciate all of the time and effort you have devoted in teaching me everything I needed to succeed. Additionally, I would like to thank all of my former and fellow lab mates: Dr. Jin Zhang, Ekaterina Stroeva, Christopher Hamilton, Qiushi (Shauna) Li, Sarah Nguyen, Beatrice Edjah, Jessica Siemer, Zachary Ferris and Christien Fadler and my collaborators: Dr. Gregory Poon and his graduate student Shingo Esaki; with your assistance and input, I have gained valuable knowledge and friendship. Thank you for making my laboratory experience a joyful one. And finally I would like to thank Brains and Behavior and the Georgia State University Neuroscience programs for funding my research.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ V

LIST OF TABLES .................................................................................................................. XII

LIST OF FIGURES ................................................................................................................. XIII

1  INTRODUCTION ................................................................................................................. 1

1.1  Nucleic Acids: the Basis of Life ...................................................................................... 1

1.2  Bio-macromolecule Interactions .................................................................................... 2

1.3  Chapter Outline ............................................................................................................... 4

1.4  References ...................................................................................................................... 4

2  SINGLE RIBONUCLEOTIDE DAMAGE: HOW A SINGLE ATOM CAN
AFFECT STRUCTURE AND DYNAMICS OF DNA ......................................................... 6

2.1  Preface .......................................................................................................................... 6

2.2  Abstract ........................................................................................................................ 6

2.3  Introduction .................................................................................................................... 7

2.3.1  Structure and physical characteristics of DNA and RNA molecules ................. 8

2.3.2  Ribonucleotide excision repair pathway ................................................................. 9

2.3.3  Consequences of single ribonucleotide damage ..................................................... 10

2.4  Materials and methods ................................................................................................ 12

2.4.1  Oligonucleotide sample preparation ....................................................................... 12

2.4.2  Thermal stability studies .......................................................................................... 13
2.4.3 NMR experiments ................................................................. 13
2.4.4 Structure calculation and analysis ............................................. 15
2.4.5 Computational and dynamics studies ......................................... 17
2.5 Results ...................................................................................... 18
  2.5.1 Thermodynamics and base pairing ......................................... 18
  2.5.2 Base stacking and NMR chemical shifts .................................. 19
  2.5.3 T1 relaxation measurements .................................................. 19
  2.5.4 Sugar conformations .............................................................. 20
  2.5.5 Backbone torsion angles and NMR chemical shifts .................. 21
2.6 Structure analysis ...................................................................... 23
  2.6.1 Analysis of the rG_AxC structure ........................................... 23
  2.6.2 Analysis of the rG_CxG structure ......................................... 24
  2.6.3 Analysis of the rG_GxC structure .......................................... 24
2.7 Probing for dynamics .................................................................. 25
  2.7.1 MDTar analysis of rG_AxC duplex ........................................ 26
  2.7.2 MDTar analysis of rG_CxG duplex ....................................... 26
  2.7.3 MDTar analysis of rG_GxC duplex ........................................ 27
  2.7.4 Free MD analysis of rG_AxC duplex ....................................... 27
  2.7.5 Free MD analysis of rG_CxG duplex ..................................... 28
  2.7.6 Free MD analysis of rG_GxC duplex ....................................... 28
2.8 Discussion ...................................................................................................................... 28
  2.8.1 Structure and implication for diseases ................................................................. 28
  2.8.2 Potential evolutionary benefits ........................................................................ 29
2.9 Conclusion .................................................................................................................... 30
2.10 References .................................................................................................................. 70
2.11 Supplementary information ....................................................................................... 76
  2.11.1 Preparation of Pf1 bacteriophage as partial alignment media ....................... 76
  2.11.2 Protocol for NMR structure generation .............................................................. 77
  2.11.3 Protocol for free molecular dynamics calculations ............................................. 81
  2.11.4 Calculation of syn/anti populations ................................................................. 82
3 PROBING COMPLEX FORMATION OF PU.1 BY NMR .................................................. 84
  3.1 Preface ...................................................................................................................... 84
  3.2 Abstract ................................................................................................................... 84
  3.3 Introduction ............................................................................................................. 84
    3.3.1 Role of PU.1 in biological systems ................................................................. 85
    3.3.2 Structure and dynamics of PU.1 ............................................................... 86
    3.3.3 Evidence of multimeric complex formation in PU.1 and other ETS proteins
           87
  3.4 Materials and methods ........................................................................................... 87
    3.4.1 Protein and DNA NMR sample preparation .................................................. 87
3.4.2 \( ^1\text{H} \) NMR experiments ................................................. 89
3.4.3 Diffusion-ordered spectroscopy experiments .................................. 89
3.4.4 \( ^1\text{H}-^{15}\text{N} \) HSQC of PU.1 ........................................... 90
3.4.5 Simulating \( ^1\text{H} \) methyl exchange ........................................ 90
3.5 Results .................................................................................. 91
3.5.1 \( ^1\text{H} \) NMR of PU.1 ................................................................. 91
3.5.2 Simulated exchange peaks ....................................................... 91
3.5.3 \( ^1\text{H} \) NMR of DNA substrates ............................................. 92
3.5.4 Diffusion-ordered spectroscopy ................................................. 92
3.5.5 \( ^1\text{H}-^{15}\text{N} \) HSQC ................................................................. 94
3.6 Discussion ............................................................................. 96
3.6.1 Evidence for dimerization of PU.1 in the presence of site-specific DNA ... 96
3.6.2 Complex formation of PU.1 bound to nonspecific DNA is unique from the site-specific complex .......................................................... 98
3.6.3 An extended DNA interface is necessary for sequential dimerization ....... 99
3.7 Conclusion .............................................................................. 99
3.8 References ............................................................................. 116
3.9 Supplementary information ....................................................... 118
3.9.1 Pulse field gradient calibrations for DOSY .................................... 118
3.9.2 Estimating hydrodynamic radii from diffusion measurements .......... 119
3.9.3  Simulating methyl peak exchanges .......................................................... 121

4  NMR METHODOLOGY APPLIED TO UNIQUE SYSTEMS: ARGinine PKa AND BASE PAIR OPENING OF INOSINE V. GUANOSINE ............................. 123

4.1  Preface .............................................................................................................. 123

4.2  Introduction .................................................................................................... 123

4.2.1  Using NMR to measure the pKa of the side-chain of methylated arginine

124

4.2.2  Using NMR to gauge local base pair dynamics of an inosine for guanosine substitution in a DNA hairpin...................................................................................... 125

4.3  Materials and methods.................................................................................... 127

4.3.1  Arginine pKα sample preparation............................................................... 127

4.3.2  Inosine for guanosine base substitution for base pair opening rates ...... 128

4.4  Results ............................................................................................................ 129

4.4.1  Methylated arginine pKα study results...................................................... 129

4.4.2  Base pair opening results for a DNA hairpin containing an inosine substitution 132

4.5  Discussion and conclusion ............................................................................. 133

4.5.1  The effect of methylation on the pKα of arginine ....................................... 133

4.5.2  The effect of an inosine substitution in place of a guanosine on local base pair dynamics 134

4.6  References ..................................................................................................... 135
4.7 Supplementary information ........................................................................................................ 138

4.7.1 The Henderson-Hasselbalch equation modified for NMR $pK_a$ titrations and sample calculations .......................................................................................................................................................... 138

4.7.2 Degradation products observed for arginine with backbone mimics at N- and C- termini ................................................................................................................................................................................. 139

4.7.3 Base pair opening lifetimes determination using inversion recovery NMR 140

5 CONCLUSIONS ........................................................................................................................................... 158

5.1 Summaries and Future Outlook ........................................................................................................ 158

5.2 References ........................................................................................................................................... 160
LIST OF TABLES

Table 2.1 Oligonucleotide Sequences and Thermal Stabilities ............................................. 33
Table 2.2 $^1$H and $^{31}$P NMR chemical shifts (ppm) for the AxC, CxG and GxC duplexes .......... 34
Table 2.3 Total number of restraints, Corma validation scores and Amber energy penalties for the final NMR structures ........................................................................................................ 35
Table 2.4 $^3$J$_{H3'-P}$ Coupling Values ..................................................................................... 36
Table 2.5 Backbone torsion angles for the top PDQPRO conformers and final rMD ............... 37
Table 3.1 Translational diffusion coefficients of PU.1, complexes and DNA substrates. ........ 101
Table 4.1 Extrapolated pK$_a$ values for the modified arginine variants and control compounds imidazole and trifluoroethanol (TFE). Values are based on the $^1$H NMR chemical shift data as a function of pH (1). ........................................................................................................... 142
Table 4.2 Base pair opening lifetimes for inosine and guanosine substitutions in a DNA hairpin. Rates were calculated in collaboration with Qiushi (Shauna) Li. ............................................. 143
LIST OF FIGURES

Figure 2.1 Monomeric units of nucleic acids include deoxyribonucleoside monophosphate (A) and ribonucleoside monophosphate (B), where Base represents the canonical nucleobases adenine, guanine, cytosine and thymine or uracil for DNA or RNA, respectively. .............................. 38

Figure 2.2 Transesterification in alkaline aqueous solutions from a highly reactive hydroxyl group resulting in spontaneous backbone cleavage. The figure is adapted from Sponer, J. et al. (70)......................................................................................................................... 39

Figure 2.3 Ribonucleotide excision repair, RER, in normal cells results in error-free repair. However, alternative pathways, such as Top1-mediated cleavage can result in mutagenic repair or DSB. Figure adapted from Huang, S.N., et al. (39).......................................................................................................................... 40

Figure 2.4 Thermal melting stability was gauged by T_m plots tracking absorbance (260 nm) as a function of temperature (K) for dG_AxC (blue trace) and rG_AxC (red trace) at 9.6 and 9.7 μM (A). A plot of 1/T_m as a function of ln of total strand concentration (B)................................................................. 41

Figure 2.5 Imino 1H NMR of rG_CxG (A) and dG_CxG (B) at various temperatures............. 42

Figure 2.6 Imino protons at 280 K are indicative of base pairing. All base pairs are accounted for in the AxC (A), CxG (B) and GxG (C) duplexes where the rGMP-containing duplex is on top and all DNA control is on bottom for each panel. ......................................................................................... 43

Figure 2.7 Chemical shift differences between control and rGMP-containing duplexes for AxC (A-C), CxG (D-F) and GxG (G-I) for base (white) and sugar (black) 1H (A, D, G), 31P (B, E, H) and imino 1H (C, F, I). ...................................................................................................................................................... 44

Figure 2.8 Base to H1’ pathways from 250 ms NOESY spectra are shown for rG_AxC (A), rG_CxG (B) and rG_GxG (C)................................................................................................................................................. 45
Figure 2.9 $T_1$ relaxation of base (H6/H8) and sugar (H1’) protons were measured for rG_AxC (A). A graphical representation of the elevated $T_1$ relaxation rates (> 2.2 s) of base and sugar protons (red) (B). Data not measured shown in grey................................................................. 47

Figure 2.10 Coupling constants (Hz) were measured from COSYDF with low flip angle at 294 K. The selection shown is for the rG_AxC C15 H1’-H2” and C15 H1’-H2’ crosspeak. Measured values were used to estimate fraction south using the Altona Graphical Method (44,71). ............ 48

Figure 2.11 Sugar conformations for riboguanosine-DNA (A-C) and all DNA controls (D-F) for AxC (A, D), CxG (B, E) and GxC (C, F). ............................................................................................................................. 49

Figure 2.12 $^{31}$P NMR spectra for AxC,(A), CxG (B) and GxC (C) between riboguanosine-containing duplexes (top) and control duplexes (bottom). ................................................................. 51

Figure 2.13 Final bundles of solved rMD structures, with the riboguanosine shown in red for rG AxC (A), CxG (B) and GxC (C). ........................................................................................................ 52

Figure 2.14 Helical analysis for the rG (black circle) and dG (grey circle) AxC (A), CxG (B) and GxC (C) duplexes. Standard A and B type helical values are depicted by large and small dashed lines, respectively. The dG_AxC (A) control structure was solved by Dr. Alexander Spring-Connell ................................................................. 53

Figure 2.15 Plots of the minor groove widths for the final 10 minimized, restrained structures for AxC (A), CxG (B) and GxC (C). Data for the rGMP-containing duplexes shown in red and the control structures are shown in blue. ................................................................. 57

Figure 2.16 Dynamic torsion angles for core nucleotides (5’-…A5-rG6-C7-T8…-3’) for rG AxC (A), (5’-…C4-rG5-G6-T7…-3’) for rG CxG (B) and (5’-…G4-rG5-C6-C7…-3’) for rG GxC (C) from an ensemble of 10,000 snapshots of the 10.0 ns rMD simulation........................................................................ 58
Figure 2.17 Dynamic torsion angles for core nucleotides A) (5’-…A5-rG6-C7-T8…-3’) for rG AxC, B) (5’-…C4-rG5-G6-T7…-3’) for rG CxG and C) (5’-…G4-rG5-C6-C7…-3’) for rG GxC from an ensemble of 5,000 snapshots of the 5.0 ns MDtar simulation.

Figure 2.18 Final rMD (A-C) and MDtar (D-F) core nucleotides for each duplex.

Figure 2.19 The formation and duration of hydrogen bonds (donor and acceptor distances) during the 5.0 ns MDtar simulations for rG_AxC (A), rG_CxG (C-D) and rG_GxC (F-G) duplexes hydrogen bonds formed (B, E, H).

Figure 2.20 Backbone torsion angles and phase angles for the core nucleotides of AxC (A), CxG (B) and GxC (C) as a function of time from an ensemble of 4,000 total structures from the unrestrained 400 ns MD simulation.

Figure 2.21 Subtle backbone perturbations affect base stacking of dinucleotide steps (A) shown for the AxC duplex. The minor alteration is shown as a decrease in dihedral angle between two pyrimidines in the control and rGMP-duplex (B).

Figure 3.1 ¹H NMR spectra of PU.1 (250 μM) titrated with 16 base pair high-affinity DNA (A) and methyl region (B).

Figure 3.2 Methyl peaks were simulated to probe for chemical exchange to explain 2:1 methyl peak broadening in the 16 base pair site-specific titration.

Figure 3.3 NOESY spectra for 10 base pair high-affinity DNA imino protons. Resonances too broad to yield observable NOESY intensities were confirmed through a temperature series. Resonances which suffered from overlap were assigned by 1D NOE experiments.

Figure 3.4 Imino ¹H spectra of the 16 base pair high-affinity DNA substrate (A) titrated with PU.1 (bottom to top) (B) at 293 K.
Figure 3.5 Imino $^1$H spectra of the 10 base pair high-affinity DNA substrate (A) titrated with PU.1 (bottom to top) (B) at 293 K. ............................................................................................................. 106

Figure 3.6 Imino $^1$H spectra of the 16 base pair nonspecific DNA substrate (A) titrated with PU.1 (bottom to top) (B) at 293 K. ............................................................................................................. 107

Figure 3.7 Diffusion coefficients of PU.1 titrated with DNA substrates in 100% D$_2$O at 25°C determined by DOSY (5). ......................................................................................................................... 108

Figure 3.8 Thermal denaturation of PU.1 monitored by $^1$H-$^{15}$N HSQC. Thermal denaturation is reversible, as protein crosspeaks were restored after lowering temperature to 25°C. ............ 109

Figure 3.9 HSQC titrations of PU.1 with DNA substrates (5)................................................................. 110

Figure 3.10 Dimerization interface from HSQC titrations (A,B) and complexed PU.1 (C,D). 111

Figure 3.11 Gradients calibrations were benchmarked from HOD $^1$H resonance peaks using a 100% D$_2$O sample at 298 K. Uncalibrated GCC diffusion curve (A) compared to new GCC diffusion curve (B) ......................................................................................................................... 112

Figure 3.12 DOSY parameters were optimized with sucrose (A) and DNA (B). .................. 113

Figure 3.13 1D slices of the DOSY at varied gradient strengths (A) for unbound PU.1. Diffusion plots for select peaks (color coded) (B) for: water (blue) and two PU.1 resonances (red, green). ......................................................................................................................... 114

Figure 3.14 The structure of PU.1 from the crystal structure 1PUE (17). ................................. 115

Figure 4.1 Arginine and methylated arginine variants. N- and C- termini contain protection groups to mimic a peptide environment ......................................................................................................................... 144

Figure 4.2 Base pairing with guanosine and cytosine (A) found in the guanosine-containing DNA hairpin (B), in contrast to the inosine and cytosine base pair (C) found in the inosine-containing DNA hairpin (D) ......................................................................................................................... 145
Figure 4.3 ¹H NMR spectrum of sDMA at pH 13.14. The delta protons (3.2 ppm) were monitored during titrations. Methyl peak (3 ppm) from the guanidino methylation. The methyl peak (2 ppm) is from the N-terminus of the residue.

Figure 4.4 600 MHz ¹H NMR of the delta protons of 0.5 mM MMA at 25°C from pH 13.6-11.1 (top to bottom).

Figure 4.5 NMR pH titration of aDMA (black), sDMA (red), MMA (green) and control arginine (blue) (1).

Figure 4.6 Titration curves for (A-B) aDMA, (C-D) sDMA and (E-F) MMA. Experimental data are represented by data points, closed for delta proton chemical shifts and open for methyl proton chemical shifts.

Figure 4.7 Titration curves for (A) control compounds and (B) the control arginine.

Figure 4.8 Titration curves for the control arginine for natural abundance (A) ¹³Cζ and (B) ¹³Cγ chemical shifts for a 50 mM control sample.

Figure 4.9 Electrostatic potential maps of guanidinium head groups calculated using Spartan 10. The constructs shown above depict representative shortened, model structures of the four arginine variants: (A) unmethylated, (B) MMA, (C) aDMA and (D) sDMA.

Figure 4.10 ¹H NMR spectra of a degraded sDMA at pH 11.62.

Figure 4.11 Imino ¹H NMR of guanosine-containing hairpin at 293 (A) and 278 K (C) and inosine-containing hairpin at 293 (B) and 278 K (D). Spectra were collected in collaboration with Qiushi (Shauna) Li.

Figure 4.12 Resonance intensity as a function of delay from T₁ inversion recovery experiments for guanosine-containing hairpin. Data for 0 mM catalyst at 278 (A) and 293 K (B) and 2 mM
catalyst at 278 (C) and 293 K (D). Data was calculated in collaboration with Qiushi (Shauna) Li.

Figure 4.13 Resonance intensity as a function of delay from $T_1$ inversion recovery experiments for inosine-containing hairpin. Data for 0 mM catalyst at 278 (A) and 293 K (B) and 5 mM catalyst at 278 (C) and 293 K (D). Data was calculated in collaboration with Qiushi (Shauna) Li.

Figure 4.14 Base pair opening lifetimes for inosine-containing (red) and guanosine-containing (green) hairpins for residue 4 (A) and 14 (B) at 278 K. Data was calculated in collaboration with Qiushi (Shauna) Li.
1 INTRODUCTION

Molecular structure and interactions are crucial for biological processes. Sequence specific and typically high affinity interactions must undergo careful regulation, as aberrant interactions are often linked to a number of diseases including cancer (1).

1.1 Nucleic Acids: the Basis of Life

Nucleic acids, the ubiquitous biomacromolecules, contain the genetic information for all known cellular life. Biological processes are regulated on a molecular level, encoded by deoxyribonucleic acid (DNA). DNA is replicated to make additional copies of itself, transcribed into RNA, and RNA is in turn translated into proteins. Maintaining the integrity of the genome is crucial in biological systems, the disruption of which leads to diseases. DNA/RNA hybrids often times exist in some of these fundamental processes, such as DNA replication (2) and transcription (3). While genetic information is stored in the chemical makeup of the nucleobases, the global and local conformation and structural plasticity contribute to recognition and binding of proteins and drugs.

Although the structure of DNA was unlocked almost 65 years ago by Watson and Crick (4), DNA and RNA structures are still constantly being solved. Unique perturbations caused by lesions, sequence effects or complexation with drugs or proteins reveal that nucleic acid structure is dependent on a number of variables. From gross structural perturbations to minute subtleties, these effects drive recognition, binding and repair. A fundamental understanding of DNA structure and dynamics is crucial for our understanding of some of these biological processes.

Structurally, nucleic acids consist of a phosphodiester backbone, with a five-membered (deoxy)ribose sugar ring and phosphodiester bonds. The backbone consists of seven torsion angles: α, β, γ, δ, ε and ζ. The eighth (deoxy)ribonucleotide torsion angle, χ, is found at the
glycosidic bond, where the nucleobase is attached to the (deoxy)ribose C1’ atom. These nucleobases are typically found in standard Watson-Crick base pairs, with imino and amino protons involved in hydrogen bonds: adenosine (A) base paired with thymine (T) or uracil (U) and guanine (G) base paired with cytosine (C).

Nucleic acids exist in various conformations depending on their fundamental components or environment. Double-stranded RNA, consisting of ribonucleotides, typically exist in A-form conformation, while double-stranded DNA, with units of deoxyribonucleotides and more conformational flexibility, exist in B-form under biological conditions and other forms depending on hydration, sequence, or salt concentrations. Globally, these molecules take on a primarily right-hand helix, with most X-ray and NMR structures existing in canonical A- or B-form; however, left-handed helices, namely the Z-form have also been shown to exist in vitro. The nitrogenous bases of nucleic acids dictate the local geometry of the helices through energetic differences in base stacking, yet it is the conformation of the furanose sugar which is an essential predictor of the global conformation. The sugar is not planar and exists in a range of conformations from the C3’-endo or North (N) conformation to C2’-endo or South (S), depending on the variation of the phase angle found in A-form and B-form, respectively.

1.2 Bio-macromolecule Interactions

Interactions between biomolecules drive essentially all biological processes. These interactions can occur at the surface, catalyzing reactions between proteins and ligands including nucleic acids, lipids, and other proteins. Similarly, interior residues can be responsible for protein activity and catalysis and retaining folded structure. Sequence specific interactions between proteins and DNA can consist of direct and indirect readout, where interactions include contacts between the protein amino acids and the nucleobases or between proteins and the sugar-
phosphate backbone; these essentially describe “base” or “shape”, respectively (5). Non-specific interactions occurring with proteins and the backbone also exist and contribute to binding affinity. Often times, binding or recognition events are dependent on not one, but a number of such mechanisms (5).

Molecular binding events are favorable when coupled to a negative Gibbs free energy, consisting of enthalpic and entropic components. The enthalpic aspect is related to substrate specificity and affinity, including hydrogen bonding, electrostatic and van der Waals interactions. Interactions, especially found on the surface and in catalytic active sites, tend to be governed by electrostatics; unsurprisingly, these surfaces are populated by many charged and polar residues, with primarily hydrophobic residues found in interiors (6). Electrostatic interactions can drive protein-protein interactions, as the protein dimerization of PU.1, the subject of Chapter 3, which can be disrupted when charged residues $^{195}DKDK^{198}$ are mutated to structurally similar but uncharged residues $^{195}NINI^{198}$ (7). Alternatively, the entropic contribution to interactions is often driven by solvation effects, such as displacement or rearrangement of water or ions. Thermodynamically, the binding of proteins to major or minor grooves of DNA shows no difference in Gibbs free energies; differences are primarily enthalpic/entropic. Major groove binding of DNA by proteins is associated with negative enthalpies, whereas minor groove binding is associated with positive enthalpies and higher entropic contributions. Similarly, positive enthalpies are also associated with small minor groove binding molecules (8).

Major groove binding proteins, such as gene regulatory proteins, often share common structural motifs, many of which contain α helices or β sheets. These include helix-turn-helix, which is the first DNA-binding motif to be recognized and is a structural motif found in ETS family proteins, such as PU.1 (9), zinc fingers, which are zinc-coordinated motifs usually
consisting of an α helix and β sheet, the leucine zipper motif, which contain two α helices held by hydrophobic side chains such as leucine, and more (10).

1.3 Chapter Outline

The following chapters describe unique projects detailing structures and interactions of different bio-macromolecules. Chapter 2 explores the structural impact of one of the most common types of DNA damage: single ribonucleotides embedded in DNA. This single atom difference imparts genomic instability and is linked to a number of diseases, while structurally, the manifestation is quite subtle. Interestingly, single ribonucleotides locally distort DNA in a sequence dependent manner, complicating impact predictions, as two different sequences with a single riboguanosine in a purine – X – pyrimidine context exhibit unique structural and dynamic perturbations (11,12). The following chapter, Chapter 3, describes the characterization of a transcription factor, PU.1, and the formation of substrate specific complexes. NMR evidence reveals a 2:1, protein:DNA, complex which forms in the presence of sequence and length specific DNA at excess protein concentrations, potentially unveiling an auto-inhibitory mechanism at the protein-DNA level (7). Finally, Chapter 4 details two independent studies, where NMR was used as a powerful tool revealing important characteristics of biomolecules. In the first study, the pKₐ of the amino acid side chain arginine is revisited and the impact of methylation is considered using NMR (13). The second study uses NMR to reveal subtle dynamics of inosine relative to guanosine in a DNA hairpin.

1.4 References

2 SINGLE RIBONUCLEOTIDE DAMAGE: HOW A SINGLE ATOM CAN AFFECT STRUCTURE AND DYNAMICS OF DNA

2.1 Preface

Project conception was a collaborative effort by Dr. Markus W. Germann and Dr. Francesca Storici at the School of Biological Sciences at Georgia Institute of Technology. The all DNA control structure for the AxC duplex was solved by Dr. Alexander M. Spring-Connell.

2.2 Abstract

Single ribonucleotides represent the most common non-standard nucleotide embedded in genomic DNA, yet little is known on the structural impact of such a lesion. The presence of isolated ribonucleotides promote genomic instability in addition to affecting the physical properties of the DNA. To probe for structural and dynamic effects of such lesions in various sequence contexts, AxC, CxG and GxC, where x=rG or dG, the structures of three single ribonucleotide containing DNA duplexes and the corresponding DNA controls were studied. A single ribonucleotide subtly and locally perturbs the structure asymmetrically 3’ of the lesion on both the riboguanosine-containing and complementary strand of the duplex, primarily in the sugar and phosphodiester backbone. The ribose sugars and 3’ downstream deoxyriboses are also affected, and exist in predominately N-type conformation, while backbone torsion angles epsilon and/or zeta of the ribonucleotide or upstream deoxyribonucleotide are affected. The structural manifestation of this lesion is also sequence dependent; a single riboguanosine flanked by different nucleotides results in the formation of different hydrogen bonds and perturbed backbones. Although this structural perturbation is overall subtle, the base stacking is locally altered, which can sensitize the DNA to other more deleterious types of DNA damage.
2.3 Introduction

The monomeric units of DNA and RNA, deoxyribonucleoside monophosphates (dNMP) and ribonucleoside monophosphates (rNMP), respectively, are chemically quite similar, Figure 2.1. Chemical differences between the two are limited to a single atom difference on the (deoxy)ribose moiety and the presence or absence of a methyl group on the thymine or uracil nucleobase. With such a subtle difference on the sugar ring, it is unsurprising that ribonucleotides, the units of RNA, can become embedded in genomic DNA; however, the full extent of such a lesion was only recently studied within the last decade. Ribonucleotides represent the most common nonstandard nucleotide in DNA, with more than one million ribonucleotide monophosphates (rNMPs) predicted in the genome of mammalian cells (1,2), yet relatively little is known on the impact of such lesions. Various mechanisms for ribonucleotide incorporation have been proposed, from inherent systems such as the incorporation of about 10 nucleotide RNA primers during eukaryotic genome replication initiation (3) to the accidental incorporation of ribonucleotides by error-prone DNA polymerases (4). In eukaryotic cells, replication of the nuclear genome is primarily carried out by three DNA polymerases (Pols) α, δ and ε (3,5). Replicative DNA polymerases have long been thought to properly select for the appropriate substrate, favoring dNTPs over rNTPs by regulating sugar discrimination via a steric gate, such as a conserved Tyr or Phe (6,7). However, due to low polymerase fidelity coupled with higher cellular rNTP concentrations, all three replicative DNA polymerases incorporate more than 10,000 rNMPs during each replication cycle in yeast (4,8). Additionally, Ribose-seq mapping of the budding yeast genome has revealed widespread but not random incorporation both in nuclear and mitochondrial DNA with unique hot spots (9).
2.3.1 Structure and physical characteristics of DNA and RNA molecules

Nucleic acids have the ability to take on a number of conformations. The structure and mechanical properties of DNA are important for many cellular functions, such as DNA binding to proteins, replication and repair (10-13). Understanding the dynamics and physical properties of DNA is therefore crucial to a better understanding of the biological processes. Double strand DNA typically has a B-form conformation in biological conditions and double strand RNA has an A-form conformation, with DNA-RNA hybrids forming unique structures with characteristics of both B- and A-forms. Less studied are chimeric hybrids of DNA containing stretches or single ribonucleotides.

X-ray crystallography structures of chimeric DNA/RNA molecules containing two to four consecutive ribonucleotides have been solved in the past decade; similar to DNA-RNA hybrids, these structures form duplexes with regions of hallmark characteristic A- form DNA (14-19). Several NMR solution structures have also been published with chimeric RNA/DNA oligonucleotides (containing stretches of rNMPS) with similar results (20,21).

Fewer structural studies exist of isolated ribonucleotides in DNA. One study has suggested isolated, or single ribonucleotides contribute to global conformational changes from B- to A- form DNA, while more recent NMR work has suggested an overall B-form structure is maintained, with localized perturbations. The X-ray crystal structure of a decamer containing a single riboguanosine was shown to have a global conformation change from B-DNA to A-DNA, with all sugars converted to the C3’-endo pucker (22). More recently, an isolated riboguanosine in the well-studied self-complementary Dickerson-Dodecamer sequence (total of two riboguanosines in the duplex) (23) was studied by NMR and was found to primarily retain B-form structure, with localized regions of symmetric A- form characteristics. The discrepancy in
the X-ray and NMR results could be due to the sequences selected. The selected sequence for the X-ray crystallography study was G/C-rich; G/C-rich oligonucleotides have a tendency to crystallize in the A-form (24,25) and although the authors were successful in crystallizing their all G/C control in the B-form, their single riboguanosine containing duplex, which likely contained A-form like regions, was induced to the A-form during crystallization.

Other studies have included circular dichroism (26) and atomic force microscopy (27) results done by our lab and our collaborators (28). This work confirms that single and isolated ribonucleotides are well tolerated in DNA duplexes and duplexes maintain an overall B- form conformation, unlike RNA-DNA hybrids. The presence of scattered, single riboguanosines has been shown to affect elasticity of 30 base pair DNA sequences (28). Interestingly, although both sequences were the same size and contained the same base composition, single riboguanosines affected each sequence differently, with one an increase in the stretch modulus of one sequence, and decrease in the stretch modulus of the other sequence, with respect to the all DNA controls, suggesting an important sequence dependence on the physical properties of ribonucleotide-containing DNA.

### 2.3.2 Ribonucleotide excision repair pathway

The enzymes specifically implicated in removal of rNMPs are RNase H1/HI and H2/HII, with different cleavage patterns and substrate specificities in prokaryotic/eukaryotic systems [9]. In prokaryotic cells, RNase HI and HII potentially work in cooperation to remove single misincorporated ribonucleotides (29). RNase H1/HI requires a substrate containing a stretch of at least four consecutive RNA intrusions (30). In mammalian cells, the removal of single ribonucleotides is predominately facilitated by RNase H2, the central enzyme in ribonucleotide excision repair (RER), and the primary source of RNase H activity (31). Eukaryotic RNase H2 is
more complex than its prokaryotic counterpart (which consists of a single protein); eukaryotic RNase H2 contains three subunits, all of which are required for activity: RNASEH2A, RNASEH2B and RNASEH2C (31). RNASEH2A is the catalytic domain, and shares sequence homology with the prokaryotic RNase HII, while the other two subunits are believed to modulate interactions with other assembly proteins (31). During removal of single ribonucleotides, RNase H2, a minor groove binding protein, recognizes the ribonucleotide and initiates RER. The enzyme nicks the strand 5' of ribonucleotide. The RNASEH2B subunit contains a PIP box, which interacts with the proliferating cell nuclear antigen (PCNA), involved in recruiting additionally assembly proteins (32). Flap endonuclease (FEN1) further excises the lesion and DNA polymerase δ and DNA ligase I complete the strand displacement, synthesis and ligation in the mechanism. Partial redundancies, such as the substitution of Exo1 for FEN1 and Pol ε for Pol δ, have also been found in the mechanism (33).

2.3.3 Consequences of single ribonucleotide damage

Unrepaired single ribonucleotides in genomic DNA promote genomic instability and have been linked to diseases. Complete RNase H2 deficiency is lethal for embryonic mice; biallelic null mutations for any of the *Rnaseh2* genes results in early embryonic death (1). In humans, mutations in any of the RNase H2 subunits, causing partial loss of RNase H2 enzyme function, results in the autosomal-recessive disorder Aicardi-Goutières syndrome (AGS) and has been linked to familial chilblain lupus and systemic lupus erythematosus (SLE) (34,35). SLE and AGS can mimic a congenital viral infection and are related to the immune recognition of nucleic acids, which includes an induction of antiviral type 1 interferons (IFNs). Briefly, these contribute to a reduction of B cell tolerance and in increase in autoantibody production which can lead to destructive cell tissue and inflammation. Often times, UV irradiation or a viral infection can
trigger this response. Gunther, et al., has linked embedded rNMPs to UV-induced cyclobutane pyrimidine dimer (CPD) formation and enhanced type I IFN signaling, both in vivo and in vitro (35).

Other enzymes have also been implicated in AGS. Some patients with AGS have mutations in enzymes such as the DNA exonuclease TREX1 and dNTP triphosphohydrolase SAMHD1, and more (26,35,36). Subunits of mammalian RNase H2 have been targets of mutations to gain insight into the role of the mammalian RNase H2 enzyme, concluding that the removal of ribonucleotides is essential for preserving genome integrity (2). Similar studies done on yeast cells found rnh201 negative strains move slowly through S phase with elevated dNTP pools and increase 2-5 base-pair deletions (37).

Defective repair in the primary pathway (RER) leads to an accumulation of single ribonucleotides and is linked to disease; however, the exact role of single ribonucleotides is still unknown. An accumulation of rNMPs in the genome in AGS and SLE mimic congenital viral infections (34, 35). The presence of a hydroxyl group at the C2’ position of a ribonucleotide is reactive and can affect the chemical stability of the genome, sensitizing the DNA to backbone cleavage, via a transesterification process, Figure 2.2. Another possibility is the role of secondary repair mechanisms. The primary repair mechanism RER is robust and fairly efficient, resulting in error-free repair. However, if the primary repair pathway has been compromised or the efficiency drastically altered, secondary pathways, some of which are error-prone, will remove single ribonucleotides, Figure 3.2. Topoisomerase I (Top1) has ribonuclease activity and Top1-mediated cleavage of unrepaired ribonucleotides can result in error-free repair or can result in mutations or double-strand breaks (DSBs), which are toxic to cells and can lead to further mutations, genetic rearrangements and cancers (38,39). Finally, there is a possibility, that even
with secondary repair pathways, unrepaired rNMPs may still exist in the genome, which can affect binding or contribute to more deleterious types of DNA damage. These lesions affect the physical properties such as flexibility, thermal stability, and structure (28,40). Distorted structure of DNA can affect the binding efficiency of housekeeping enzymes. Even a subtle, local physical perturbation in DNA has been linked with minor changes in stacking of the bases, potentially priming the DNA for more deleterious types of DNA damage, such as the formation of UV-photoproducts, creating a link in the formation of skin lesions in patients with SLE and AGS with the accumulation of single rNMPs (35).

2.4 Materials and methods

2.4.1 Oligonucleotide sample preparation

DNA and rGMP-containing sequences in this chapter are short 9 base pair oligonucleotides, Table 2.1, from sequences used in the AFM study (28). Of the two 30 base pair duplexes, Sequence 2 contains a fragment of *Saccharomyces cerevisiae* chromosome III and Sequence 1 is a fully synthetic construct, containing other flanking combinations not found in Sequence 2. The all DNA control strands were purchased from Integrated DNA Technologies and synthesized in-house using an Applied Biosystems PCR-MATE 391 DNA Synthesizer and reagents and solid phosphoramidites purchased from Glen Research. DNA was synthesized with the “Trityl-On” setting and deprotected and cleaved from the solid support in concentrated ammonium hydroxide at 60°C overnight. Following deprotection and cleavage, the DNA was further purified via the “Trityl” group using Glen-Pak Purification Cartridges following standard protocols on the technical bulletin (Glen Research) and the “Trityl” group was cleaved with 2% trifluoroacetic acid. The riboguanosine-containing strands were purchased from Thermo Scientific Dharmacon. DNA only strands were desalted using HI-TRAP desalting high-

performance liquid chromatography (HPLC) columns on an ATKA FPLC and purity was checked on 15% polyacrylamide, 8 M urea denaturing gels. Strands containing the riboguanosine were deprotected for 30 minutes at 60 °C in 2’-deprotection buffer (100 mM acetic acid, adjusted to pH 3.8 with TEMED) and then desalted using HPLC HI-TRAP columns. Purity was confirmed using denaturing polyacrylamide gels. Strand concentrations were calculated using extinction coefficients derived from the sum of the mononucleotides at an absorbance of 260 nm at 80 °C on a Cary UV-Vis Spectrophotometer using the following extinction coefficients for A, C, G and T, respectively: 15500, 6980, 11200 and 9360 M⁻¹cm⁻¹.

### 2.4.2 Thermal stability studies

Thermal stabilities of the duplexes at 2-20 μM strand concentration were assessed by analyzing thermal melting (Tₘ) values on a Cary UV-Vis Spectrophotometer in 100 mM sodium chloride, 10 mM sodium phosphate, 0.5 mM EDTA buffer at pH 6.6. UV absorbance was measured at 260 nm from 15 °C to 85 °C at a rate of 0.5 °C/minute. Tₘ values were obtained from a six parameter fit of the UV melting curves assuming a two-state transition: helix to coil (41). The enthalpies, entropies and Tₘ values (30 μM) were determined from a linear fit of the van’t Hoff plot and thermodynamic variables for the biomolecular association of non-self-complementary duplexes (28,42). Thermodynamic stability of duplexes was also qualitatively assessed by monitoring the imino protons (1-1 Jump and Return ¹H NMR) at various temperatures from 275 to 333 K (40).

### 2.4.3 NMR experiments

All NMR samples were prepared by combining single strands at 1:1 ratios to make 1 mM DNA or riboguanosine-containing DNA duplexes in 100 mM sodium chloride, 10 mM sodium phosphate and 0.1 mM ethylenediaminetetraacetic acid (EDTA) buffer at pH* 6.6-6.7 in 99.99%
D$_2$O or pH 7.0-7.1 in H$_2$O/D$_2$O (90%/10%) as previously described (28). Strand stoichiometry was confirmed from $^1$H NMR at high temperature (333 K) by integrating relevant base H6/H8 peaks using a presaturation solvent suppression NMR pulse program. For residual dipolar coupling (RDC) restraint calculations, NMR samples were aliquoted into two parts to make two samples for each sequence: 0.73 mM duplex for phage samples and 0.53 mM duplex without phage. Prepared Pf1 bacterial phage was purchased from Asla and solvent exchanged to make 56 mg / mL stock; of this stock, appropriate amounts of Pf1 phage and buffer were added to one of the aliquots of NMR sample for each sequence and a $^2$H deuterium splitting of ~14 Hz was observed, corresponding to ~15 mg / mL of phage per 0.73 mM duplex sample. Additional details can be found in the Supplementary Information.

All spectra were obtained on a Bruker Avance 600 MHz system using a 5 mm QXI probe and Avance III HD 850 MHz NMR using a TCI cryoprobe. Proton spectra were referenced to an internal 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) standard and $^{31}$P spectra were referenced to an external capillary of 85% H$_3$PO$_4$ in D$_2$O. Nuclear Overhauser effect spectroscopy (NOESY) spectra were recorded for 150 and 250 ms mixing times ($\tau_M$) with 4.0 s relaxation delays for the AxC and 5.0 s relaxation delays for the CxC and GxC duplexes to ensure complete relaxation at 294 K. Data was zero filled (4K x 4K) and processed with a shifted sine bell curve (SSB=2). T$_1$ relaxation values were measured from the T$_1$ inversion recovery experiment. All spectra were assigned and integrated using SPARKY 3.33 UCSF with Gaussian or Sum over Box integration when necessary (43). Low flip angle COSY experiments were run to estimate the pseudorotational angles of the deoxyribose and ribose sugars (44). Phosphorous resonances were assigned from $^1$H-$^{31}$P correlated 2D spectra (HPCOR) (45). Epsilon torsion angles were calculated from $^{3}$J$_{H3'-P}$ coupling constants obtained from constant
time NOESY (CT-NOESY) experiments, by selectively exciting H3’ protons with a 12 ms REBURP pulse with and without 31P decoupling (46). Both 1D and 2D 1-1 Jump and Return and excitation sculpted NOESY experiments (τM 50 and 150 ms) and TOCSY experiments (τM 60 ms) were used to observe and assign exchangeable protons for the water NMR samples (90% H2O/10% D2O) with 1.5 and 2.5 s relaxation delays. Residual dipolar couplings were measured from hetero-nuclear F2 coupled natural abundance 13C-1H (HSQC) spectra for the sugar range and the base range in both the presence and absence of alignment media, Pf1 phage.

2.4.4 Structure calculation and analysis

Starting structures were initially built using Assisted Model Building with Energy Refinement (AMBER) (44) 9.0’s NUCGEN and XLeap using the parm99 forcefield (27). The starting structure was initially built as a perfect B-type DNA based on the results of the NMR NOESY, imino and 31P NMR data, which suggest that the structure is mostly standard B-form, with isolated regions of perturbed structure. For the riboguanosine- containing structures, the nucleotide was changed from a DG to RG manually. The molecule was then neutralized with 16 Na+ ions and solvated 8.0 Å from the solute to the edge of the octahedral water box with ~ 2,500-3,000 TIP3P water molecules. Electrostatic interactions were accounted for using Particle Mesh Ewald (PME). Simulations were run at constant pressure, using SHAKE for bonds involving hydrogens in Sander (AMBER 9) with a 1 fs time step and trajectories were recorded every ps. Solvated systems were minimized in two stages prior to implementing restraints. More details can be found in the Supplemental Information.

Restrains used consisted of both standard and experimentally derived restraints. Standard restraints consisted of backbone torsion angles: α, β, γ and ζ and Watson-Crick hydrogen bonded base pair angles and distances. Sugar pucksers (containing δ torsion angles), ε torsion angles and
syn/anti constraints to restrain the 5’ terminal base pair (where necessary) were experimental determined. All of these restraints represent the non-distance restraints and were converted to AMBER format using internal scripts. Each minimized starting structure was subjected to an initial AMBER cycle (minimization, 100 ps molecular dynamics run at 300 K, minimization) with only these non-distance restraints (47). These dynamic runs provided the starting structures for implementing distance restraints. Quantitative nonexchangeable inter-proton distances, based on NOESY crosspeak volumes, were obtained from an iterative RANDMARDI procedure (Corma, Mardigras and Amber cycles) using a correlation time of 3.5 ns that agreed best with the data (48,49) (50); crosspeaks integrated using Gaussian were given a 20% error, and Sum over Box integrations were assigned higher errors (40%). This procedure is done independently for all 150 and 250 ms NOESY data sets, then, subsequently the 150 and 250 ms distance files were averaged and underwent further iterative cycles. Structures after each Amber cycle were assessed based on Corma R^X scores, Amber energy penalties and qualitatively using Visual Molecular Dynamics 1.8.7 (VMD) (51).

Finally, RDC restraints were implemented by first minimizing the alignment tensor only, then a full system minimization implementing all restraints followed by a 50-100 ps production run and additional minimization to yield Mardigras structures for 10.0 ns restrained molecular dynamics simulations (rMD) at 300 K. The rMD trajectory was recorded at intervals of 1 per ps and 10 structures from the final 100 ps (one snapshot every 10 ps) were individually minimized to yield the final bundle of rMD structures. From each ensemble of 10 rMD structures, a representative structure was selected for low Corma R^X scores (< 6%) and low Amber restraint violations. Bundles of the 10 final minimized rMD structures were deposited in the PDB: rG_AxC (PDB 5KGV), rG_CxG (PDB 5KI7), dG_CxG (PDB 5KI4), rG_GxC (PDB 5K1B) and
dG_GxC (PDB 5KI5). The control DNA duplex for dG_AxC (PDB 2N5P) was solved by Dr. Alexander M. Spring-Connell. The root-mean-square deviation (RMSD) was calculated using and helical parameters were assessed using VMD and Curves+.

### 2.4.5 Computational and dynamics studies

The Mardigras calculated restrained NMR structures were also subjected to a 5.0 ns molecular dynamics simulation with time-averaged restraints (MDtar) (52) and from that data, a pool of high probability structures was accessed using Probability Distribution by Quadratic PROgramming (PDQPRO) (53) as previously described (54). Additional details are found in the Supplemental Information. The same NMR restraint lists were used from the rMD simulation, with the exception of RDC and sugar restraints. A 20 ps window (τ= 20 ps) was used for MDtar restraints. The MDtar trajectory generated a pool of 5,000 structures for each riboguanosine containing duplex. PDQPRO then was iteratively used to find the most probable conformers on the basis of experimental NOESY rates. For the first iteration all 5,000 conformers were included. Structures with probabilities < 5% were discarded and the new pool of 8, 8 and 6 (rG_AxC, rG_CxG and rG_GxC, respectively) structures were input for a second PDQPRO iteration to yield 3, 3 and 2 results with >15% probabilities (PDB codes for rG_AxC, rG_CxG and rG_GxC are 5KIE, 5KIF and 5KIH, respectively).

Additional dynamics were probed using extended unrestrained molecular dynamics (MD) simulations on Amber 14 on NVIDIA GPUs for 400 ns, using a 2 fs timestep (55). Sequences were built using Nucleic Acid Builder (NAB) with additional GC basepairs on each terminal end to prevent fraying effects. Simulations were run using parmBSC0 forcefield with ε/ζ reparameterizations (56). Additional information can be found in the Supplementary Information.
2.5 Results

2.5.1 Thermodynamics and base pairing

The presence of a riboguanosine did not significantly affect the thermodynamic stability compared to the deoxyriboguanosine control sequences. All Tₘ values, reported for 30 μM strand concentration, were within 1-3 degrees (± 0.5), suggesting a modest effect on the oligonucleotide thermodynamic stability, Table 2.1 and Figure 2.4. Similarly, a temperature series from 275 to 315 K on the imino protons qualitatively correlated with the Tₘ values, Figure 2.5. The presence of a single ribonucleotide at the core of an oligonucleotide was not destabilizing to the DNA duplex and imino proton peak intensities qualitatively appear identical between riboguanosine and control duplexes. At 280 K, all base pairs were accounted for, Figure 2.6, with the exception of the terminal 5’ base pair of the GxC duplex and its control, due to the presence of a terminal AT base pair (Table 2.1). In all cases a sharp peak for the riboguanosine/cytosine base pair (position 6 of rG_AxC and 5 of rG_CxG and rG_GxC) was observed. Additionally, the chemical shifts of most of the imino protons were similar, between riboguanosine and control duplexes, suggesting a minimal or localized perturbation in the helical structure, Figure 2.7.

Syn/anti populations were observed at 5’ terminal ends in both control and ribo-containing AxC and CxG duplexes supported from by the increased intra-residue base-H1’ NOE intensity, Figure 2.8. Populations of < 20% syn gycosidic bond formation of those adenosines (A1) were confirmed by integration of the A1 H8 to H1’ NOEs at 75 ms, Supplemental Information. The syn/anti populations at the terminal ends are far from the riboguanosine and have no anticipated effect on the results of core structure. The 5’-A1 base to A1 sugar and A1 base to T2 sugar integrations were not used, due to the presence of syn populations, and instead,
the 5’ terminal bases were “locked” in the dominant anti-conformation during NMR structure calculations. No evidence for syn populations was detected for the riboguanosine or flanking nucleotides.

2.5.2 Base stacking and NMR chemical shifts

A single RNA intrusion has a minor, localized effect on the chemical shifts of the base and sugar protons, Figure 2.7 and Table 2.2. As expected, for all of the $^1$H chemical shifts, the largest differences from the controls were for the riboguanosine H2’ proton at position 6 for AxC and 5 for CxG and GxG, with smaller deviations flanking the ribonucleotide. Complete base to H1’ sugar NOESY pathways were assigned, indicative of a right-handed B- form DNA, Figure 2.8. A close examination of the NOE cross peak intensities reveal unique, localized markers. For the rG_AxC and rG_CxG duplexes, cross peaks between the ribonucleotide sugar H1’ and downstream +1 base, rG6H1’-C7H6 and rG5H1’-G6, respectively, have higher NOE intensities compared to rGH1’-rGH8 peaks, Figure 2.8. This increased intensity corresponds to a very slight decrease in the rGH1’ to downstream base distance by ~0.2 Å as compared to the controls. In contrast, in the rG_GxG base to H1’ NOESY pathway, Figure 2.8, the intensity of the rG5H1’-C6H6 crosspeak is lower. The lower intensity corresponds to a 1 Å increase in the distance of rG5H1’-C6H6 (4.7 Å) versus the control dG5H1’-C6H6 (3.7 Å). A complete list of assigned NMR chemical shifts can be found in the Table 2.2.

2.5.3 $T_1$ relaxation measurements

$T_1$ relaxation values were measured from the $T_1$ inversion recovery experiment; riboguanosine sugar H1’ and base H8 protons and those of surrounding residues were found to be up to 2 times the length of normal deoxyriboguanosine $T_1$ values, Figure 2.9. This is supported by values measured from RNA and chimeric or hybrid DNA-RNA duplex studies, as
the environment of sugar and base protons in a ribose sugar differ to a deoxyribose sugar (57). The riboguanosine H8 base proton, for example, was found to be 3.1 s, compared to other H8 protons at 2.0 s. Similarly, the H1’ of riboguanosine was 3.5 s, compared to other H1’ protons at 2.0 s.

2.5.4 Sugar conformations

Deoxyribose and ribose sugar conformations, described as fraction south (fS), were estimated from the vicinal sugar couplings (3JH1’-H2’1, 3JH1’-H2’2, Σ H1’, Σ H2’1 and Σ H2’2) as previously described from COSY spectra, Figure 2.10 and 2.11 (44). Sugars that could not be resolved due to overlap in the H1’-H2’1/H2’2 regions were qualitatively estimated by the intensity and coupling values of the COSY H3’/H4’ crosspeaks and compared to simulated COSY crosspeaks for north and south sugars, simulated using Bruker Topspin (40). The sugars for the riboguanosine-containing duplexes had high fS for most residues that are distant from the riboguanosine, indicative of normal B-form DNA, Figure 2.11, while the ribose sugars were in the range of 40-60 % fS. Interestingly, a lower fS was also observed in deoxyribonucleotides 1-2 residues downstream of the lesion and opposite the ribonucleotide, for the rG_AxC and rG_CxG duplexes, indicating an asymmetric localized perturbation 3’ of the ribonucleotide. In addition to these traits in the rG_GxC duplex, the upstream sugar (G4) is also estimated to have a lower fS, resulting in a more symmetric perturbation (Figure 2.11), which is more consistent with the findings of DeRose, et al., where pseudorotational angles of flanking deoxyribose sugars were symmetrically increased. A direct comparison, however, is difficult, as the self-complementary sequence selected in that study may play a role in the nature of the perturbation (23). Control structure sugars were mostly in the normal DNA, south conformation, with the exception of some terminal sugars and core sugars, which have some north characteristics, Figure 2.11.
2.5.5 Backbone torsion angles and NMR chemical shifts

The $^{31}$P resonances for control DNA resonate in a narrow range, consistent with B-form DNA. The addition of just a single atom (in the riboguanosine-containing duplexes) radically changes the local $^{31}$P spectra, indicative of a perturbed backbone localized around the lesion, Figure 2.7 and 2.12. The extent of perturbation is not the same for all of the riboguanosine-containing duplexes, suggesting that the flanking sequence context is important, as indicated in the AFM studies (28). Two of the sequences, rG_AxC and rG_CxG, show a similar trend, with minor upfield shifts in $^{31}$P chemical shift of the riboguanosine compared to the control (<0.25 ppm), followed by a large downfield shift (0.80 and 1.28 ppm, respectively) of the downstream +1 phosphate and an upfield shift of +2 phosphorous residue (0.27-0.29 ppm), Figure 2.7 (28). This is consistent with the base stacking observed in the NOESY base to H1’ pathway for the rG_AxC and rG_CxG duplexes previously described, where cross peaks between the ribonucleotide sugar H1’ and downstream +1 base, have higher NOE intensities compared to rGH1’-rGH8 peaks. This is an indication of altered backbone torsion angles around the riboguanosine and following base, which correlates with the shifted residues in the $^{31}$P spectra, C7 and G6 for rG_AxC and rG_CxG, respectively.

The other sequence, rG_GxC, exhibited a different trend. Here, the riboguanosine $^{31}$P resonance exhibited the largest deviation, a downfield shift of ~0.4 ppm. The $^{31}$P chemical shifts of residues 6 and 7 could not be determined with certainty, due to overlap; however, despite ambiguity, neither of these residues experienced $^{31}$P chemical shifts larger than the riboguanosine. Again, this different trend correlates to the NOESY pathway rG_GxC base to H1’ NOESY pathway, Figure 2.8, where the intensity of the rG5H1’-C6H6 crosspeak is lower, further supporting the flanking sequence effect on the riboguanosine perturbation. Furthermore,
in each riboguansine-containing duplex, the shifted $^{31}$P signal (visualized as a difference in $^{31}$P chemical shift compared to controls) correlated with northern sugar conformations, illustrating the localization of the structural perturbation.

Backbone ε torsion angles were measured from $^{3}J_{H3'\cdot P}$ coupling constants from the CT-NOESY spectra, Table 2.4. As anticipated from the NOESY pathways, pseudorotational analysis and $^{31}$P NMR spectra, the ε torsion angles not flanking the riboguanosine were similar to the control and were in the normal trans range for B-DNA. The rG_AxC and rG_CxG duplexes again showed similarities in $^{3}J_{H3'\cdot P}$: the $-1$ residue, (A5 and C4, respectively) had normal to slightly larger coupling constants, the ribonucleotides (g6 and g5) had small couplings (1.5 Hz; ε ~170°) and the downstream +1 deoxyribonucleotide (C7 and G6) had larger coupling constants (3.6-5.7 Hz; ε ~190-200°) compared to the corresponding control residues. For both of these ribonucleotide-containing duplexes, there appears to be a pattern where the riboguanosine $^{3}J_{H3'\cdot P}$ and the +1 downstream deoxyribonucleotide $^{3}J_{H3'\cdot P}$ coupling constants are at opposite extremes, which correlates with the downfield shifted +1 and upfield shifted +2 deoxyribonucleotide $^{31}$P residues (Figure 2.12 and 2.7). Again, the rG_GxC was different; the upstream (-1) residue has a large coupling, compared to the control, but showed a normal riboguanosine coupling constant, indicating a perturbation 5' of the riboguanosine. In the previously solved solution structure of the Dickerson-Dodecamer duplex containing riboguanosines, the $^{3}J_{H3'\cdot P}$ of the riboguanosine was reported to increase to 7 Hz. In contrast, for two of our duplexes, the coupling decreased to 1.5 Hz while rG_GxC remained at a normal range (3.6 Hz). This demonstrates that sequence contexts is important and is not readily predicted by just considering purine/pyrimidine neighbors.
2.6 Structure analysis

A bundle of the final 10 restrained structures for each sequence was calculated after 10 ns of restrained molecular dynamics (rMD), Figure 2.13. All of the riboguanosine-containing structures are highly restrained, with > 400 restraints, ~23, 26 and 24 restraints/nucleotide for the rG AxC, CxG and GxC sequences, respectively, and in good experimental agreement; all Corma \( R^X \) values are below 6%, Table 2.3. The final bundle of structures, individually minimized after the 10.0 ns rMD run, yields a heavy atom RMSD of < 0.5 Å. For each sequence, the best minimized structure from the final ensemble of structures was selected on the basis of having the lowest \( R^X \) values and Amber energy penalties, Table 2.3.

Overall, the presence of a riboguanosine does not severely distort the DNA with the AxC and CxG duplexes experiencing slightly larger deviations, as seen in the helical analysis, Figure 2.14. Additionally, the minor groove widths of the final structures were slightly increased for the rG_AxC and CxG duplexes, with little to no change in the GxC duplexes, Figure 2.15, consistent with previously published work, where the presence of riboguanosines had no appreciable impact on the minor groove widths (23).

2.6.1 Analysis of the rG_AxC structure

An analysis of the 10.0 ns rMD trajectory for the rG AxC duplex, Figure 2.16, revealed that the \( \alpha/\gamma/\epsilon/\zeta \) backbone torsion angles were primarily in the canonical g-\( \gamma \)+/BI conformation, with the exception of a perturbed, decreased \( \zeta \) torsion angle for the riboguanosine (\( \epsilon-\zeta = BI \)) near shifted \( 31P \) resonances. In the rG_AxC duplex, a hydrogen bond formed for the duration of the simulation between the C2’-OH of the riboguanosine and following nucleotide backbone (O2P) for >98% of the rMD simulations, correlating with the perturbed, but static torsion angles.
2.6.2 Analysis of the rG_CxG structure

Similar to rG_AxC, the rG_CxG duplex, Figure 2.16, showed a perturbed, decreased ζ torsion angle at the riboguanosine near shifted $^{31}$P resonances; however, riboguanosine α/γ and ε torsion angles were also perturbed. Flanking deoxyribonucleotide α/γ/ε,ζ backbone torsion angles were in canonical g-/g+/BI conformation. The riboguanosine α/γ torsion angles flipped after 5 ns to the t/t conformation, which has been observed for RNA and RNA/DNA hybrids (58) and the riboguanosine ε/ζ torsion angles are in the BII conformation. The formation of a hydrogen bond was observed between the C2’-OH of the riboguanosine and following base (N7 of dG6) for >98% of the rMD simulations. The differences in ε/ζ torsion angles were greatest in the rG_CxG duplex, which incidentally had the largest shifted out $^{31}$P resonance (at the following phosphate).

2.6.3 Analysis of the rG_GxC structure

The α/γ torsion angles also flip ~180° after 5 ns for the rG_GxC riboguanosine, similar to rG_CxG, with riboguanosine α/γ torsion angles in the t/t and g+/t conformations. Additionally, increased sampling for the rG_GxC occurred in the ε/ζ torsion angles of the upstream dG4 residue (ε-ζ = BII), in contrast of the other duplexes, where ε/ζ torsion angles were perturbed at the riboguanosine backbone. The ε/ζ torsion angles of the riboguanosine in rG_GxC are in the canonical BI conformation. Here, the dynamic flipping between ε/ζ torsion angles correlate with the formation of two hydrogen bonds (riboguanosine C2’-OH and downstream nucleotide backbone O5’ and sugar O4’), Figure 2.16 C. The ε/ζ torsion angles of the upstream nucleotide of the rG_GxC sequence correlate well with the shifted riboguanosine $^{31}$P resonance.

DeRose, et al., reported an increase in both ε/ζ torsion angles for their rG (in CxA context) (23); this compares to the increase in both ε/ζ torsion angles for the rG_GxC sequence
of the present study and contrasts to the increased $\epsilon$, decreased $\zeta$ torsion angles of rG_AxC and rG_CxG, reinforcing the notion that the perturbation is sequence dependent. Although overall structures have low error scores (Corma R$^5$) and low energy penalties (44), evidence of dynamics exist from sugar puckering and perturbed backbone torsion angles in the vicinity of the riboguanosine.

2.7 Probing for dynamics

Surprisingly, the rMD trajectories contained dramatic 180° changes at the $\alpha/\gamma$ torsion angles for the rG in the CxG and GxC duplexes. To determine if this flip is an artefact or a consequence of the biased contraints, where the riboguanosine $\alpha$, $\beta$, $\gamma$ and $\zeta$ backbone torsion angles were deliberately unrestrained, while all other torsion angles were broadly restrained, we further investigated the backbone dynamics using molecular dynamics with time-averaged restraints (MDtar), as rMD does not reproduce conformational dynamics (41). Furthermore, the algorithm PDQPRO was utilized to determine the most probably conformers from the pool of 5,000 MDtar structures (for each riboguanosine-containing duplex) based on the NOESY relaxation rates. The pool of probable conformers was analyzed for two iterations of PDQPRO: the first iteration on all 5,000 possible structures, followed by a second iteration on a pool of conformers with higher probability (>5%).

To further probe the extent of the dynamics and structural perturbations from a single ribonucleotide, unrestrained MD simulations were run for 400 ns on duplexes containing additional GC base pairs at the ends. Fraying effects, especially at AT terminal base pairs have been observed by both our group and others (59). To counter these effects, two GC base pairs have been added to each 9 base pair duplex, to yield 13 base pair constructs containing the NMR sequences at the core.
2.7.1 MDtar analysis of rG_AxC duplex

The results of the 5.0 ns MDtar trajectory for the rG_AxC duplex revealed similarities with the 10.0 ns rMD trajectory, Figure 2.16 and 2.17. From this trajectory, the most perturbed backbone torsion angle was found at the riboguanosine ζ torsion angle, corresponding to the most shifted phosphate residue (C7) from the 31P NMR. For the majority of the simulation (99% of the trajectory), a single hydrogen bond formed between the hydroxyl group and the backbone O2P atom of the downstream nucleotide.

The second PDQPRO iteration resulted in seven probable conformation structures. Of those, the top three conformers represent a probability of 67% and shared similarities with the final rMD structure, Table 2.5, where the riboguanosine ζ torsion angle is most perturbed. Although the riboguanosine ζ torsion angle is lowered, the conformation is still in the BI conformation. Final MDtar derived structures are found in Figure 2.18.

2.7.2 MDtar analysis of rG_CxG duplex

The results of the MDtar simulation for the rG_CxG duplex revealed dynamics in the backbone, Figure 2.17. The riboguanosine α/γ torsion angles sampled the t/t conformation at 2 ns, as in the rMD simulation. In contrast to the results of the rMD simulation, the ε/ζ torsion angles alternated between two states in MDtar, which corresponded to the formation of two different hydrogen bonds: the riboguanosine C2’-OH and the following base N7 for a majority 73% of the simulation and to a lesser extent to the following nucleotide backbone O5’ for 18% of the simulation. The hydrogen bond with the N7 atom of the guanosine base was also observed in the rMD trajectory and final rMD structure.

The PDQPRO analysis resulted in eight highly probable conformers from the rG_CxG pool, the top three of which represent a total probability of 53%. Similar to the rG_AxC
structures, the most perturbed backbone torsion angles of the rG_CxG structures are again found in the riboguanosine ε (increased) and ζ (decreased), with the riboguanosine backbone existing in the BII conformation (ε-ζ), Table 2.5 and Figure 2.17. In the three top PDQPRO conformations, the α/γ torsion angles, including those in the riboguanosine, exists in the canonical g-/g+ conformation.

2.7.3 MDTar analysis of rG_GxC duplex

The results of the MDTar simulation for the rG_GxC duplex (Figure 2.17) showed similar results as for the rMD simulation, with similar hydrogen bonds forming between the C2’-OH and following nucleotide O5’ and O4’; however, the α/γ torsion angles displayed differences, which were in the canonical g-/g+ conformations.

PDQPRO results revealed six top structures for the rG_GxC duplex with two of those conformers representing 99% optimal probability, Table 2.5. Unlike the rG_AxC and CxG duplexes, the most perturbed backbone torsion angle of the rG_GxC trajectory is upstream of the riboguanosine, at the -1 (dG4) ζ torsion angle, which is slightly decreased, but still in the canonical BI conformation. These MDTar structures are found in Figure 2.18.

2.7.4 Free MD analysis of rG_AxC duplex

The rG_AxC duplex was assessed after 400 ns and compared to the all DNA control. Similar to the results of the MDTar simulations, the free MD α/γ torsion angles also revealed slight perturbations in the backbone at the rG α/γ torsion angles and downstream +1 (dC7) γ torsion angle, Figure 2.20. Additionally, subtle backbone changes are observed in the ε/ζ torsion angles of the upstream -1 (dA5) nucleotide and riboguanosine (rG6). The dynamics in the upstream dA5 nucleotide ζ torsion angle are not observed in rMD or MDTar simulations. The sugar conformation of the riboguanosine moiety sampled both north and south conformations,
which is reasonable with the experimental results, suggesting that the sugar has a $f_S$ of $\sim 50\%$.

However, the downstream +1 (dC7) deoxyribose, which has a higher north conformation experimentally, was shown to have a predominately south conformation.

### 2.7.5 Free MD analysis of rG_CxG duplex

The rG_CxG free MD simulation revealed dynamics at the $\alpha/\gamma$ of the riboguanosine (rG5) and downstream nucleotide (dG6); slightly perturbed torsion angles were briefly sampled in the free MD, which is comparable to the results of the MDtar and rMD simulations, albeit to a lesser extent, Figure 2.20. Different conformations were also sampled at the $\epsilon/\zeta$ torsion angles of the upstream (dC4) nucleotide and ribonucleotide (rG5). The phase angle of the riboguanosine for rG_CxG free MD results also revealed a dynamic ribose moiety, with $40\%-60\%$ $f_S$ sampling.

### 2.7.6 Free MD analysis of rG_GxC duplex

After 400 ns of unrestrained MD simulation time, the rG_GxC duplex revealed very subtle differences at primarily the $\epsilon/\zeta$ torsion angles of the upstream (dG4) nucleotide and (rG5) ribonucleotide compared to the control duplex simulation, Figure 2.20. However, the largest differences here are observed for the ribonucleotide, in contrast to the results of the MDtar and rMD trajectories, where the upstream (dG4) nucleotide was most dynamic. Additionally, the free MD simulation primarily sampled the north conformation for the ribose moiety, while the experimental data suggests the conformation is $\sim 50\%$ $f_S$.

### 2.8 Discussion

#### 2.8.1 Structure and implication for diseases

Many biological processes, which bind, cleave or read DNA require or alter structural or mechanical properties of DNA. These properties are affected by sequence and the presence or absence of lesions. The presence of a single, isolated ribonucleotide subtly perturbs the structure,
primarily in the phosphodiester and sugar backbone locally. Although this perturbation is subtle and localized, it can still alter physical properties and affect biological processes. Isolated ribonucleotides affect the flexibility of DNA, as shown by our collaborators in an AFM study (28). Interestingly, two different 30 base pair duplexes containing the same nucleotides in different sequences revealed that isolated ribonucleotides can make the overall oligonucleotide more or less flexible, depending on sequence. This correlates with the findings of the NMR and molecular dynamics experiments, where the location and extent of structural perturbation and dynamics depends on sequence, even when comparing purine-rG-pyrimidine duplexes, such as the different results found in the AxC and GxC duplexes.

Although a global B-form conformation is still retained, subtle differences can result in serious consequences. Patients with Aicardi-Goutières and systemic lupus erythematosus suffer from UV light sensitivity, resulting in skin lesions. Günther et al. correlated UV light sensitivity with photodamage in rNMP-containing DNA (35). UV light induces especially deleterious forms of DNA damage such as UV photoproducts, which include cyclobutane pyrimidine dimers (CPD), such as T-T, C-T and T-C dimers. Single ribonucleotides subtly alter the backbone and sugar conformation of DNA, and this minor alteration affects base stacking, priming dinucleotide pyrimidines for the formation of CPDs, as shown by a decrease in dihedral angle between two pyrimidines in a ribonucleotide containing duplex compared to control DNA, Figure 2.21.

2.8.2 Potential evolutionary benefits

Although single ribonucleotides represent the most common nonstandard nucleotides in the genome, they are tolerated quite well, as shown by the structural work, where overall B-form DNA is still maintained. Additionally, the presence of a single ribonucleotide has only a minor effect on the thermal stability of the duplex, as shown by the slight T_m change and imino ¹H
spectra, where the base pairs appear sharp and comparable in peak intensity to the control DNA. The fact that this lesion is so common, in addition to the minor impact suggests potential evolutionary benefits. Mutation studies on yeast replicative DNA Polymerase (Pol) ε found M644L resulted in the polymerase to stably incorporate three-fold fewer rNMPs than wild-type (37). Furthermore, an M644G mutation in yeast Pol ε stably incorporated an eleven-fold increase of rNMPs, further suggesting ribonucleotide incorporation in the genome is a well-tolerated at even higher concentrations. Unlike some bacteria, which have DNA methylase and endonucleases to direct and nick newly synthesized strands for MMR, eukaryotic cells lack this system. Because the majority of isolated ribonucleotides become embedded during eukaryotic DNA synthesis, single ribonucleotides can potentially act as flags for repair mechanisms. Several studies have suggested that RNase H2-dependent processing of ribonucleotides generates nicks, directing mismatch repair (MMR) to the nascent strand (60-62).

2.9 Conclusion

Ribonucleotides are incorporated into DNA at a high rate through both normal metabolic processes and environmental factors. Despite their common occurrences, they are tolerated well and lead to minor structural perturbations localized near the ribonucleotides. These non-self-complementary duplexes contain a single lesion, unlike the two isolated riboguanosines in a previous study (23). This lesion manifests perturbations primarily confined to the sugars and the phosphodiester backbone. Furthermore, the extent and pattern of distortion depends on the sequence and also asymmetrically extends to next-nearest neighboring nucleotides, in agreement with a previous finding showing a sequence dependence on the elasticity of DNA containing scattered riboguanosines (28).
An asymmetric perturbation was observed, localized 3’ of the riboguanosine for two duplexes, rG_AxC and rG_CxG, in contrast to a more symmetric perturbation in both sugar and backbone for rG_GxC. This suggests that the symmetric perturbations previously reported is not a uniform structural characteristic of ribonucleotides in DNA, as previously reported, but rather a consequence of the sequence context (DeRose, et al. report a riboguanosine, flanked by a C and A), in addition to the self-complementary nature of the duplex (23). The extent and nature of perturbation due to a single riboguanosine is different, even for two purine-rG-pyrimidine (rG AxC and GxC) duplexes.

Sugar conformation is an important and potential structural recognition site of repair enzymes (63,64). The sugars of rG_AxC and rG_CxG have north-like characteristics asymmetrically 3’ of the riboguanosine on the riboguanosine-containing and complementary strands, whereas the sugars of the rG_GxC duplex has a more symmetric distribution of north-like sugars flanking the riboguanosine. Nevertheless, the symmetry is not as pronounced as seen in the self-complementary sequence by DeRose, et al. (23). The crystal structure of RNase H2 in complex with a single ribonucleotide containing DNA duplex revealed that recognition is facilitated by multiple RNase H2 interactions formed with the 2’-OH group of the ribonucleotide (65). While the shift of the sugar conformation to an N downstream of the ribose may be too subtle a feature for RNase H2, it could potentially still modulate interactions with other proteins. Restriction enzymes may cleave DNA-RNA hybrids albeit with decreased activity (66), and proteins can still recognize single ribonucleotide containing duplexes with lower affinity (67). Future work will need to focus on additional proteins to evaluate the impact of ribonucleotide lesions on efficiency.
The perturbation of the spectrum indicates a distortion in the phosphate backbone due to BI/BII transitions of the ε/ζ torsion angles. The perturbed $^3\text{P}$ residues correlate with perturbed experimentally derived ε torsion angles and $^3\text{J}_{\text{H3'-P}}$ coupling constants and perturbed α/γ and ε/ζ torsion angles in the solved structures, correlating with the sugar conformations and formation of hydrogen bonds. The formation of these hydrogen bonds correlates with the flipping of the backbone torsion angles in the MDtar and the perturbed torsion angles in the rMD simulations.

Using restrained molecular dynamics (MD) simulations and MD simulations with time-averaged restraints (MDtar), dynamics in the systems around the riboguanosine of each duplex correlates with shifted resonances in the $^3\text{P}$ NMR spectra. In addition, the predicted formation of hydrogen bonds, which is dependent on the sequence context, influences the structure and dynamics of the lesion and flanking nucleotides. The application of the ribose-seq protocol to diseased cells may provide specific signatures of ribonucleotide incorporation associated with particular human diseases (9,68). Such hot spots may allow identification of genomic alterations clinically relevant for treatment, in which the ribonucleotide sites and their sequence context could serve as targets for screening and/or the design of drugs of specific human pathologies (40).
Table 2.1 Oligonucleotide Sequences and Thermal Stabilities

All DNA control and single riboguanosine-containing DNA were designed to be non-self-complementary containing the same base composition with a single riboguanosine embedded in the core of the duplex in the context of various flanking nucleotides. Thermal stabilities of riboguanosine-containing duplexes with DNA controls calculated at $C_\text{T} 30 \mu M \pm \text{S.E.}$ (40).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Name</th>
<th>$T_m$ (K)</th>
<th>Enthalpy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'-ATGGAGCTC</td>
<td>dG_AxC</td>
<td>315.3 ± 0.4</td>
<td>-57.5 ± 1.0</td>
</tr>
<tr>
<td>TACCTCGAG - 5'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5'-ATGGAgCTC</td>
<td>rG_AxC</td>
<td>312.8 ± 0.5</td>
<td>-61.9 ± 3.7</td>
</tr>
<tr>
<td>TACCTCGAG - 5'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5'-ATCCGGTAG</td>
<td>dG_CxG</td>
<td>314.3 ± 0.4</td>
<td>-55.9 ± 0.6</td>
</tr>
<tr>
<td>TAGGCCATC - 5'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5'-ATCCgGTAG</td>
<td>rG_CxG</td>
<td>314.0 ± 0.4</td>
<td>-48.7 ± 6.0</td>
</tr>
<tr>
<td>TAGGCCATC - 5'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5'-TTAGG CCTG</td>
<td>dG_GxC</td>
<td>317.0 ± 0.4</td>
<td>-47.7 ± 4.3</td>
</tr>
<tr>
<td>AATCCGGAC - 5'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5'-TTAggCCTG</td>
<td>rG_GxC</td>
<td>314.8 ± 0.4</td>
<td>-43.8 ± 6.6</td>
</tr>
<tr>
<td>AATCCGGAC - 5'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.2 $^1$H and $^{31}$P NMR chemical shifts (ppm) for the AxC, CxG and GxC duplexes

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>Super*</th>
<th>$^{1}$H</th>
<th>$^{1}$H</th>
<th>$^{1}$H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
</tr>
<tr>
<td></td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
</tr>
<tr>
<td></td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
<td>$^{1}$H</td>
<td>$^{31}$P</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
<th>$^{1}$H</th>
<th>$^{31}$P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3 Total number of restraints, Corma validation scores and Amber energy penalties for the final NMR structures

The final bundle of restrained structures of each duplex was highly restrained, with > 20 restraints per residue and overall low Amber energy penalties and Corma scores, < 6% for rG_AxC (PDB 5KGV), rG_CxG (PDB 5KI7), dG_CxG (PDB 5KI4), rG_GxC (PDB 5KIB) and dG_GxC (PDB 5KI5) (40). The control AxC, dG_AxC, structure was solved by Dr. Alexander M. Spring-Connell (PDB 2N5P) (69).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>rG_AxC</th>
<th>rG_CxG</th>
<th>dG_CxG</th>
<th>rG_GxC</th>
<th>dG_GxC</th>
<th>force constant (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>quantitative distance restraints</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no. of nonexchangeable (total)</td>
<td>157</td>
<td>219</td>
<td>186</td>
<td>172</td>
<td>111</td>
<td>30</td>
</tr>
<tr>
<td>no. of intraresidue</td>
<td>104</td>
<td>146</td>
<td>117</td>
<td>110</td>
<td>86</td>
<td>30</td>
</tr>
<tr>
<td>no. of interresidue</td>
<td>53</td>
<td>73</td>
<td>69</td>
<td>58</td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>qualitative (lock into anti for A1)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>average well width (Å)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>endocyclic torsion angle restraints</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no. of deoxyribose/ribose</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>50</td>
</tr>
<tr>
<td>Watson-Crick restraints</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no. distance</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>no. flat angle</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>backbone torsion angle restraints</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no. of broad restraints α, β, γ, ζ</td>
<td>60</td>
<td>60</td>
<td>66</td>
<td>60</td>
<td>66</td>
<td>50</td>
</tr>
<tr>
<td>no. ε (CT NOEEST)</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>15</td>
<td>14</td>
<td>50</td>
</tr>
<tr>
<td>residual dipolar coupling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no. RDC restraints (total)</td>
<td>45</td>
<td>41</td>
<td>50</td>
<td>47</td>
<td>46</td>
<td>1.0 (dwt)</td>
</tr>
<tr>
<td>total number of restraints</td>
<td>413</td>
<td>472</td>
<td>453</td>
<td>426</td>
<td>373</td>
<td></td>
</tr>
<tr>
<td>no. of restraints per residue</td>
<td>22.9</td>
<td>26.2</td>
<td>25.2</td>
<td>23.7</td>
<td>20.7</td>
<td></td>
</tr>
</tbody>
</table>

| CORMA R² (%)                                   |        |        |        |        |        |                  |
| 7m=150 ms                                      | intra, inter, total | intra, inter, total | intra, inter, total | intra, inter, total | intra, inter, total | intra, inter, total |
|                                               | 4.62, 5.10, 4.75 | 3.50, 5.86, 4.38 | 4.10, 5.15, 4.43 | 4.36, 5.14, 4.54 | 4.10, 5.15, 4.43 |
| 7m=250 ms                                      | 4.63, 5.26, 4.83 | 3.45, 5.48, 4.16 | 3.20, 5.36, 3.94 | 4.03, 5.39, 4.35 |                  |

| final Amber parameters (best)                  |        |        |        |        |        |                  |
| total distance penalty (kal/mol)               | 22.12  | 25.54  | 16.20  | 20.96  | 26.95  |                  |
| total angle penalty (kal/mol)                  | 0.48   | 0.25   | 0.52   | 0.31   | 0.31   |                  |
| total torsion angle penalty (kal/mol)          | 6.45   | 4.07   | 5.67   | 5.75   | 5.251  |                  |
| RDC alignment constraint (kal/mol)             | 6.23   | 5.19   | 15.91  | 10.89  | 6.07   |                  |
Table 2.4 $^3J_{\text{H3'P}}$ Coupling Values

Coupling values (Hz) were measured from CT-NOESY experiments for the riboguanosine-containing and all DNA control structures. Some of the values, denoted by a dash, could not be measured due to extensive overlap of the H3’ protons. The 3’ terminal residues have no phosphate group (40).

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>$^3J_{\text{H3'P}}$ Control</th>
<th>rGMP-DNA</th>
<th>Nucleotide</th>
<th>$^3J_{\text{H3'P}}$ Control</th>
<th>rGMP-DNA</th>
<th>Nucleotide</th>
<th>$^3J_{\text{H3'P}}$ Control</th>
<th>rGMP-DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>4.6</td>
<td>4.7</td>
<td>A1</td>
<td>4.0</td>
<td>4.2</td>
<td>T1</td>
<td>5.2</td>
<td>5.4</td>
</tr>
<tr>
<td>T2</td>
<td>4.6</td>
<td>-</td>
<td>T2</td>
<td>3.3</td>
<td>3.5</td>
<td>T2</td>
<td>3.9</td>
<td>4.4</td>
</tr>
<tr>
<td>G3</td>
<td>4.0</td>
<td>4.7</td>
<td>C3</td>
<td>-</td>
<td>-</td>
<td>A3</td>
<td>3.5</td>
<td>4.0</td>
</tr>
<tr>
<td>G4</td>
<td>3.8</td>
<td>4.4</td>
<td>C4</td>
<td>-</td>
<td>3.9</td>
<td>G4</td>
<td>4.5</td>
<td>5.4</td>
</tr>
<tr>
<td>A5</td>
<td>3.3</td>
<td>4.8</td>
<td>G5/rG5</td>
<td>3.6</td>
<td>1.6</td>
<td>G5/rG5</td>
<td>-</td>
<td>3.6</td>
</tr>
<tr>
<td>G6/rG6</td>
<td>3.8</td>
<td>1.5</td>
<td>G6</td>
<td>3.0</td>
<td>3.6</td>
<td>C6</td>
<td>-</td>
<td>5.4</td>
</tr>
<tr>
<td>C7</td>
<td>4.5</td>
<td>5.5</td>
<td>T7</td>
<td>-</td>
<td>-</td>
<td>C7</td>
<td>3.8</td>
<td>4.3</td>
</tr>
<tr>
<td>T8</td>
<td>4.2</td>
<td>5.4</td>
<td>A8</td>
<td>3.2</td>
<td>3.4</td>
<td>T8</td>
<td>4.4</td>
<td>4.7</td>
</tr>
<tr>
<td>G9</td>
<td>-</td>
<td>-</td>
<td>G9</td>
<td>-</td>
<td>-</td>
<td>G9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>c10</td>
<td>4.3</td>
<td>4.3</td>
<td>C10</td>
<td>5.0</td>
<td>5.0</td>
<td>C10</td>
<td>5.4</td>
<td>5.4</td>
</tr>
<tr>
<td>A11</td>
<td>3.3</td>
<td>3.6</td>
<td>T11</td>
<td>3.9</td>
<td>3.8</td>
<td>A11</td>
<td>3.7</td>
<td>3.9</td>
</tr>
<tr>
<td>G12</td>
<td>3.7</td>
<td>3.8</td>
<td>A12</td>
<td>2.7</td>
<td>3.1</td>
<td>G12</td>
<td>4.2</td>
<td>4.8</td>
</tr>
<tr>
<td>C13</td>
<td>4.2</td>
<td>4.4</td>
<td>C13</td>
<td>3.8</td>
<td>3.9</td>
<td>C13</td>
<td>3.6</td>
<td>4.2</td>
</tr>
<tr>
<td>T14</td>
<td>3.1</td>
<td>3.2</td>
<td>C14</td>
<td>4.4</td>
<td>4.7</td>
<td>C14</td>
<td>3.7</td>
<td>4.3</td>
</tr>
<tr>
<td>C15</td>
<td>4.3</td>
<td>-</td>
<td>G15</td>
<td>3.9</td>
<td>4.0</td>
<td>C15</td>
<td>3.7</td>
<td>4.2</td>
</tr>
<tr>
<td>C16</td>
<td>4.4</td>
<td>-</td>
<td>G16</td>
<td>3.5</td>
<td>3.5</td>
<td>T16</td>
<td>3.9</td>
<td>-</td>
</tr>
<tr>
<td>A17</td>
<td>3.7</td>
<td>4.0</td>
<td>A17</td>
<td>3.3</td>
<td>3.4</td>
<td>A17</td>
<td>3.3</td>
<td>3.7</td>
</tr>
<tr>
<td>T18</td>
<td>-</td>
<td>-</td>
<td>T18</td>
<td>-</td>
<td>-</td>
<td>A18</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 2.5 Backbone torsion angles for the top PDQPRO conformers and final rMD

Backbone torsion angles for the top, most probable PDQPRO conformers is from an ensemble of 5,000 MDtar structures and compared to the final minimized rMD structure for the riboguanosine-containing duplexes (40).

<table>
<thead>
<tr>
<th>Structure probability nucleotide</th>
<th>Backbone Torsion Angles</th>
<th>Backbone Torsion Angles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>γ</td>
</tr>
<tr>
<td>2587 0.2371 A5</td>
<td>299.8</td>
<td>40.6</td>
</tr>
<tr>
<td>rg6</td>
<td>276.2</td>
<td>39.7</td>
</tr>
<tr>
<td>C7</td>
<td>289.4</td>
<td>67.3</td>
</tr>
<tr>
<td>T8</td>
<td>291.0</td>
<td>62.8</td>
</tr>
<tr>
<td>4877 0.2037 A5</td>
<td>286.0</td>
<td>58.3</td>
</tr>
<tr>
<td>rg6</td>
<td>285.2</td>
<td>41.5</td>
</tr>
<tr>
<td>C7</td>
<td>289.8</td>
<td>62.9</td>
</tr>
<tr>
<td>T8</td>
<td>280.0</td>
<td>58.7</td>
</tr>
<tr>
<td>3119 0.2273 A5</td>
<td>287.7</td>
<td>41.0</td>
</tr>
<tr>
<td>rg6</td>
<td>299.3</td>
<td>40.3</td>
</tr>
<tr>
<td>C7</td>
<td>295.0</td>
<td>50.2</td>
</tr>
<tr>
<td>T8</td>
<td>289.3</td>
<td>60.3</td>
</tr>
<tr>
<td>total</td>
<td>0.6681</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structure normalized nucleotide</th>
<th>Backbone Torsion Angles</th>
<th>Backbone Torsion Angles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>γ</td>
</tr>
<tr>
<td>1493 0.1761 C4</td>
<td>280.9</td>
<td>49.2</td>
</tr>
<tr>
<td>rg5</td>
<td>293.4</td>
<td>51.4</td>
</tr>
<tr>
<td>C6</td>
<td>268.4</td>
<td>49.1</td>
</tr>
<tr>
<td>T7</td>
<td>292.0</td>
<td>66.4</td>
</tr>
<tr>
<td>2634 0.1747 C4</td>
<td>277.9</td>
<td>50.4</td>
</tr>
<tr>
<td>rg5</td>
<td>306.2</td>
<td>55.2</td>
</tr>
<tr>
<td>C6</td>
<td>273.1</td>
<td>47.5</td>
</tr>
<tr>
<td>T7</td>
<td>293.1</td>
<td>55.9</td>
</tr>
<tr>
<td>3991 0.1812 C4</td>
<td>288.6</td>
<td>54.5</td>
</tr>
<tr>
<td>rg5</td>
<td>286.3</td>
<td>55.2</td>
</tr>
<tr>
<td>G6</td>
<td>288.9</td>
<td>38.6</td>
</tr>
<tr>
<td>T7</td>
<td>295.6</td>
<td>56.9</td>
</tr>
<tr>
<td>total</td>
<td>0.532</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structure normalized nucleotide</th>
<th>Backbone Torsion Angles</th>
<th>Backbone Torsion Angles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>γ</td>
</tr>
<tr>
<td>3764 0.6932 G4</td>
<td>287.2</td>
<td>46.0</td>
</tr>
<tr>
<td>rg5</td>
<td>291.0</td>
<td>53.2</td>
</tr>
<tr>
<td>C6</td>
<td>298.4</td>
<td>61.3</td>
</tr>
<tr>
<td>C7</td>
<td>294.2</td>
<td>50.7</td>
</tr>
<tr>
<td>4901 0.2988 G4</td>
<td>293.2</td>
<td>48.2</td>
</tr>
<tr>
<td>rg5</td>
<td>300.9</td>
<td>59.4</td>
</tr>
<tr>
<td>C6</td>
<td>284.6</td>
<td>64.4</td>
</tr>
<tr>
<td>C7</td>
<td>286.6</td>
<td>57.5</td>
</tr>
<tr>
<td>total</td>
<td>0.992</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.1 Monomeric units of nucleic acids include deoxyribonucleoside monophosphate (A) and ribonucleoside monophosphate (B), where Base represents the canonical nucleobases adenine, guanine, cytosine and thymine or uracil for DNA or RNA, respectively.
Figure 2.2 Transesterification in alkaline aqueous solutions from a highly reactive hydroxyl group resulting in spontaneous backbone cleavage. The figure is adapted from Sponer, J. et al. (70).
Figure 2.3 Ribonucleotide excision repair, RER, in normal cells results in error-free repair. However, alternative pathways, such as Top1-mediated cleavage can result in mutagenic repair or DSB. Figure adapted from Huang, S.N., et al. (39).

The tyrosine of Top1 can form a covalent bond with the phosphate, which can then relegate; however, in the presence of a hydroxyl group at the H2’ position, this process can result in cyclic phosphate (2’,3’-CP) (insert). A second round of Top1 results in DSB.
Figure 2.4 Thermal melting stability was gauged by $T_m$ plots tracking absorbance (260 nm) as a function of temperature (K) for $dG_{AxC}$ (blue trace) and $rG_{AxC}$ (red trace) at 9.6 and 9.7 μM (A). A plot of $1/T_m$ as a function of ln of total strand concentration (B).

(B) The data for $rG_{AxC}$ reveals a straight line, fit with the Van’t Hoff equation for non self-complementary duplexes. Data was collected at several concentrations for AxC, CxG and GxC duplexes (data not shown). The unstacked bases of single stranded DNA (and rGMP-containing DNA) absorb more light than double stranded; this effect is hyperchromicity. The point of inflection, the $T_m$, represents the scenario where 50 % of molecules are denatured, and 50 % are native. The $T_m$ value can be calculated by a 6 parameter fit. This variable is concentration dependent for double-stranded DNA. A plot of $1/T_m$ as a function of ln of total strand concentration (B) for $rG_{AxC}$ reveals a straight line, which can be fit with the Van’t Hoff equation for non self-complementary duplexes:

$$
\frac{1}{T_m} = R/\Delta H^0 \ln \left(C_t\right) + (\Delta S^0 - R \ln \left(4\right)) / \Delta H^0,
$$

where enthalpy, $\Delta H^0$, and entropy, $\Delta S^0$, can be calculated and used to determine other physical parameters.
Figure 2.5 Imino $^1$H NMR of rG_CxG (A) and dG_CxG (B) at various temperatures.

Temperature was increased to 320 K for CxG, AxC and GxC duplexes (data not shown). Imino protons spectra are indicative of base pairing and sharp imino protons indicate stable base pairs. Qualitatively, a similarity between the resonance intensities of control and rGMP-containing DNA reveal that a single riboguanosine in DNA is well tolerated and does not grossly perturb structure or base pairing, correlating with $T_m$ data.
Figure 2.6 Imino protons at 280 K are indicative of base pairing. All base pairs are accounted for in the AxC (A), CxG (B) and GxC (C) duplexes where the rGMP-containing duplex is on top and all DNA control is on bottom for each panel.

The 5’ terminal T1 imino proton of GxC (C) is missing in both the rG and dG duplexes. This proton can be observed as a broad resonance at 13.2 ppm for each duplex at 275 K (data not shown); this base pair undergoes exchange with bulk water.
Figure 2.7 Chemical shift differences between control and rGMP-containing duplexes for AxC (A-C), CxC (D-F) and GxC (G-I) for base (white) and sugar (black) $^1$H (A, D, G), $^{31}$P (B, E, H) and imino $^1$H (C, F, I).

Note, due to ambiguity, in assigning the GxC control $^{31}$P resonance between residues 6 and 7, values can be 0.11-0.20 and 0.23-0.33, respectively (H).
**Figure 2.8** Base to H1’ pathways from 250 ms NOESY spectra are shown for rG_AxC (A), rG_CxG (B) and rG_GxC (C).

Complete pathways are observed, suggesting the presence of a single riboguanosine does not grossly perturb the structure. Additionally, the pattern of intra and inter residue crosspeak intensities suggest primarily B-type DNA. However, slightly localized differences are observed. Two crosspeaks (circled) correspond to rG6H1’-C7H6 and rG6H1’-rG6H8 correlations (A). The two crosspeaks (circled) in panel B, corresponding to rG5H1’-G6H8 and rG5H1’-rG5H8 of rG_CxG. The two crosspeaks (circled) for the rG_GxC pathway (C) correspond to rG5H1’-C6H8 and rG5H1’-rG5H8 (40).
Figure 2.9 $T_1$ relaxation of base (H6/H8) and sugar (H1') protons were measured for rG_AxC (A). A graphical representation of the elevated $T_1$ relaxation rates (> 2.2 s) of base and sugar protons (red) (B). Data not measured shown in grey.

The pulse program consists of a 180° pulse, a delay, then a 90° pulse, with varied delay times. Peak intensities are plot as a function of delay, $\tau$, and fit using the equation:

$$I(\tau)/I(0) = 1 - 2 \times \exp(-\tau/T_1).$$

Typical base (H6/H8) and sugar (H1’) protons for DNA are in the 2.0 s range; adenosine H2 has the longest nucleobase relaxation of 4.0 s. (B) A graphical representation of the elevated $T_1$ relaxation rates (> 2.2 s) of base protons (red letters) or sugar H1’ protons (red pentagons) demonstrate the localization of the effect. $T_1$ rates of < 2.2 s are shown in black. Data not acquired in grey.
Figure 2.10 Coupling constants (Hz) were measured from COSYDF with low flip angle at 294 K. The selection shown is for the rG_AxC C15 H1'-H2'' and C15 H1'-H2' crosspeak. Measured values were used to estimate fraction south using the Altona Graphical Method (44,71).
Figure 2.11 Sugar conformations for riboguanosine-DNA (A-C) and all DNA controls (D-F) for AxC (A, D), CxG (B, E) and GxC (C, F).

The sugar conformations in fS: green > 90%, light green 80-90%, yellow 60-80%, orange 40-60% and red <40%. The $^{31}$P $|\Delta\delta|$ (ppm) from the controls are depicted by color coded “P”: 0-0.05 green, 0.06-0.11 light green, 0.12-0.2 yellow (outlined for visibility), 0.21-0.49 orange and >0.5 red (A-C). Data not collected, due to overlap is shown in grey. Grey sugars outlined in colored boxes denote data qualitatively assessed from cross peak intensities and estimated coupling values from the H3’/H4’ region only when data in the H1’/H2’1/H2’2 region could not be obtained; green to yellow/orange boxes represent sugars with south to north characteristics, respectively.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th></th>
<th>B</th>
<th></th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P</td>
<td>18</td>
<td>P</td>
<td>18</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td>P</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td>P</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>5′-</td>
<td>A-T</td>
<td>5′-</td>
<td>A-T</td>
<td>5′-</td>
<td>T-A</td>
</tr>
<tr>
<td>11</td>
<td>T-A</td>
<td>12</td>
<td>T-A</td>
<td>13</td>
<td>T-A</td>
</tr>
<tr>
<td>12</td>
<td>G-C</td>
<td>13</td>
<td>C-G</td>
<td>14</td>
<td>A-T</td>
</tr>
<tr>
<td>13</td>
<td>G-C</td>
<td>14</td>
<td>C-G</td>
<td>14</td>
<td>C-G</td>
</tr>
<tr>
<td>14</td>
<td>A-T</td>
<td>15</td>
<td>G-C</td>
<td>15</td>
<td>G-C</td>
</tr>
<tr>
<td>15</td>
<td>G-C</td>
<td>16</td>
<td>G-C</td>
<td>16</td>
<td>C-G</td>
</tr>
<tr>
<td>16</td>
<td>C-G</td>
<td>17</td>
<td>T-A</td>
<td>17</td>
<td>C-G</td>
</tr>
<tr>
<td>17</td>
<td>T-A</td>
<td>18</td>
<td>A-T</td>
<td>18</td>
<td>T-A</td>
</tr>
<tr>
<td>18</td>
<td>C-G</td>
<td>19</td>
<td>G-C</td>
<td>19</td>
<td>G-C</td>
</tr>
<tr>
<td>19</td>
<td>G-C</td>
<td>20</td>
<td>G-C</td>
<td>20</td>
<td>C-G</td>
</tr>
<tr>
<td>20</td>
<td>C-G</td>
<td>21</td>
<td>T-A</td>
<td>21</td>
<td>T-A</td>
</tr>
<tr>
<td>21</td>
<td>T-A</td>
<td>22</td>
<td>A-T</td>
<td>22</td>
<td>T-A</td>
</tr>
<tr>
<td>22</td>
<td>C-G</td>
<td>23</td>
<td>G-C</td>
<td>23</td>
<td>G-C</td>
</tr>
<tr>
<td>23</td>
<td>G-C</td>
<td>24</td>
<td>G-C</td>
<td>24</td>
<td>C-G</td>
</tr>
<tr>
<td>24</td>
<td>C-G</td>
<td>25</td>
<td>T-A</td>
<td>25</td>
<td>T-A</td>
</tr>
<tr>
<td>25</td>
<td>T-A</td>
<td>26</td>
<td>A-T</td>
<td>26</td>
<td>T-A</td>
</tr>
<tr>
<td>26</td>
<td>C-G</td>
<td>27</td>
<td>G-C</td>
<td>27</td>
<td>G-C</td>
</tr>
<tr>
<td>27</td>
<td>G-C</td>
<td>28</td>
<td>G-C</td>
<td>28</td>
<td>C-G</td>
</tr>
<tr>
<td>28</td>
<td>C-G</td>
<td>29</td>
<td>T-A</td>
<td>29</td>
<td>T-A</td>
</tr>
<tr>
<td>29</td>
<td>T-A</td>
<td>30</td>
<td>A-T</td>
<td>30</td>
<td>T-A</td>
</tr>
<tr>
<td>30</td>
<td>C-G</td>
<td>31</td>
<td>G-C</td>
<td>31</td>
<td>G-C</td>
</tr>
<tr>
<td>31</td>
<td>G-C</td>
<td>32</td>
<td>G-C</td>
<td>32</td>
<td>C-G</td>
</tr>
<tr>
<td>32</td>
<td>C-G</td>
<td>33</td>
<td>T-A</td>
<td>33</td>
<td>T-A</td>
</tr>
<tr>
<td>33</td>
<td>T-A</td>
<td>34</td>
<td>A-T</td>
<td>34</td>
<td>T-A</td>
</tr>
<tr>
<td>34</td>
<td>C-G</td>
<td>35</td>
<td>G-C</td>
<td>35</td>
<td>G-C</td>
</tr>
<tr>
<td>35</td>
<td>G-C</td>
<td>36</td>
<td>G-C</td>
<td>36</td>
<td>C-G</td>
</tr>
<tr>
<td>36</td>
<td>C-G</td>
<td>37</td>
<td>T-A</td>
<td>37</td>
<td>T-A</td>
</tr>
<tr>
<td>37</td>
<td>T-A</td>
<td>38</td>
<td>A-T</td>
<td>38</td>
<td>T-A</td>
</tr>
<tr>
<td>38</td>
<td>C-G</td>
<td>39</td>
<td>G-C</td>
<td>39</td>
<td>G-C</td>
</tr>
<tr>
<td>39</td>
<td>G-C</td>
<td>40</td>
<td>G-C</td>
<td>40</td>
<td>C-G</td>
</tr>
</tbody>
</table>
Figure 2.12 $^{31}$P NMR spectra for AxC (A), CxG (B) and GxC (C) between riboguanosine-containing duplexes (top) and control duplexes (bottom).

Only select residues are labeled; complete $^{31}$P assignments can be found in the supplemental data.
Figure 2.13 Final bundles of solved rMD structures, with the riboguanosine shown in red for rG AxC (A), CxG (B) and GxC (C).
Figure 2.14 Helical analysis for the rG (black circle) and dG (grey circle) AxC (A), CxG (B) and GxC (C) duplexes. Standard A and B type helical values are depicted by large and small dashed lines, respectively. The dG_AxC (A) control structure was solved by Dr. Alexander Spring-Connell.
Figure 2.15 Plots of the minor groove widths for the final 10 minimized, restrained structures for AxC (A), CxG (B) and GxC (C). Data for the rGMP-containing duplexes shown in red and the control structures are shown in blue.

The minor groove widths are defined as $P_i - P_{i+4}$ minus phosphate radius ($5.8$ Å). The $dG_AxC$ control structure (A) was solved by Dr. Alexander Spring-Connell.
Figure 2.16 Dynamic torsion angles for core nucleotides (5’-...A5-rG6-C7-T8...-3’) for rG AxC (A), (5’-...C4-rG5-G6-T7...-3’) for rG CxG (B) and (5’-...G4-rG5-C6-C7...-3’) for rG GxC (C) from an ensemble of 10,000 snapshots of the 10.0 ns rMD simulation.

Torsion angles are plotted in degrees (0 to 360, ordinate) as a function of time in ns (0 to 5, abscissa). Riboguanosine torsion angles shown in red.
Figure 2.17 Dynamic torsion angles for core nucleotides A) (5'-...A5-rG6-C7-T8...-3') for rG AxC, B) (5'-...C4-rG5-G6-T7...-3') for rG CxG and C) (5'-...G4-rG5-C6-C7...-3') for rG GxC from an ensemble of 5,000 snapshots of the 5.0 ns MDtar simulation.

Torsion angles are plotted in degrees (0 to 360, ordinate) as a function of time in ns (0 to 5, abscissa).
Figure 2.18 Final rMD (A-C) and MDtar (D-F) core nucleotides for each duplex.

rMD structures rG_AxC (PDB 5KGV) and dG_AxC (white) (PDB 2N5P, solved by Dr. Alexander Spring-Connell) (A). rMD structures rG_CxG (PDB 5KI7) and dG_CxG (white) (PDB 5KI4) (B). rMD structures rG_GxC (PDB 5KIB) and dG_GxC (white) (PDB 5KI5) (C). Bundle of 3 MDtar structures for rG_AxC (PDB 5KIE) (D). Bundle of MDtar structures for rG_CxG (PDB 5KIF) (E). Bundle of 2 MDtar structures for rG_GxC (PDB 5KIH) (F). Hydrogen bonds shown in green.
Figure 2.19 The formation and duration of hydrogen bonds (donor and acceptor distances) during the 5.0 ns MDtar simulations for rG_AxC (A), rG_CxG (C-D) and rG_GxC (F-G) duplexes hydrogen bonds formed (B, E, H).

A hydrogen bond formed and remained for 99% of the simulation between the rG6 hydroxyl group and C7 O2P atom (A,B), for 73% of rG5 hydroxyl group and G6 N7 atom (C,E) and 18% with G6 O5' (D,E) and formed the most dynamic hydrogen bond between rG5 hydroxyl and 44% of the simulation with C6 O4' (F,H) and 36% of the simulation with C6 O5' (G,H). Note, in the rG_GxC duplex, a hydrogen bond with the riboguanosine hydroxyl proton is not observed in the final restrained, minimized structure.
Figure 2.20 Backbone torsion angles and phase angles for the core nucleotides of AxC (A), CxG (B) and GxC (C) as a function of time from an ensemble of 4,000 total structures from the unrestrained 400 ns MD simulation.

In each panel, the nucleotides from the riboguanosine-containing duplex are shown in red (top) and all DNA control duplex is shown in blue (bottom). Angles are plotted in degrees (0 to 360, ordinate) as a function of time in ns (0 to 400, abscissa).
Figure 2.21 Subtle backbone perturbations affect base stacking of dinucleotide steps (A) shown for the AxC duplex. The minor alteration is shown as a decrease in dihedral angle between two pyrimidines in the control and rGMP-duplex (B).
2.10 References


2.11 Supplementary information

2.11.1 Preparation of Pf1 bacteriophage as partial alignment media

Alignment of DNA can be achieved with filamentous phage, yielding dipolar couplings which in turn can be used for residual dipolar coupling (RDC) restraints. Pf1 bacteriophage was purchased from Asla Biotech and prepared by a series of buffer exchange cycles with 20 mM sodium phosphate, 20 mM sodium chloride in D$_2$O. The buffer was exchanged at least three times. Each exchange involves filling a centrifuge tube with about $\frac{3}{4}$ volume of the new exchange buffer at the appropriate pH*. Note, the pH of the phage cannot be directly adjusted. The sample was centrifuged for 1 h at 90,000 rpm (TLA 120.2 Beckman Rotor); care was taken to avoid over compacting the pellet. Then, the supernatant was replaced with fresh NMR buffer and rocked gently overnight at room temperature. A long glass pipette was used as the phage is highly viscous and sheared phage will not align the media. After the third exchange, the NMR buffer added was high quality D$_2$O (99.996%).

The concentration of the Pf1 phage cannot be measured by UV, as only intact phage aligns the media. Thus, the concentration of functional phage in the stock solution was determined from the quadrupolar splitting of the $^2$H D$_2$O signal in the presence of phage. The measured splitting was compared to the concentration of phage, found on Asla’s website: http://asla-biotech.com/products/pf1-phage-for-nmr-analysis. After determining the concentration of active phage in the stock, an aliquot of phage was added to each DNA NMR sample. For RDC measurements, a phage concentration of 15-25 mg/mL, corresponding to 13-28 Hz quadrupolar splitting, was used.


2.11.2 Protocol for NMR structure generation

Following experimental NMR collection, a series of restraints were built using Amber 9 (27); these restraints include: quantitative NOESY distance restraints derived from NOESY crosspeak integrations and calculated with Mardigras, broad sugar restraints derived from fraction south ($f_S$) values determined from 3-bond scalar coupling values ($^3J_{1H-1H}$) from COSY spectra, epsilon backbone torsion angles derived from $^3J_{P\cdot H3}$ coupling values from constant time NOESY (CT-NOESY) spectra, and residual dipolar coupling (RDC) values from measured $^1J_{1H-13C}$ couplings from $^1H-^{13}C$ HSQC spectra in the presence and absence of phage as well as qualitative canonical base pair distance/angle restraints, broad backbone restraints and qualitative distance restraints from H$_2$O NOESY crosspeaks.

Quantitative distance restraints were determined from integration of NOESY crosspeaks at three mixing times: 250, 150 and 75 ms were measured and integrated using Sparky UCSF. These values were converted to distance restraints using an iterative Mardigras cycle. Sugar restraints were estimated from the intra-sugar $^3J_{1H-1H}$ measurements and “sum-of” ($\Sigma$) coupling values from COSY spectra and estimated using the graphical method (44,71) for $f_S$:

\[
\Sigma 1' = ^3J_{H1'\cdot H2'} + ^3J_{H1'\cdot H2''},
\]

\[
\Sigma 2' = ^3J_{H1'\cdot H2'} + ^3J_{H2'\cdot H2''} + ^3J_{H2'\cdot H3'},
\]

\[
\Sigma 2'' = ^3J_{H1'\cdot H2''} + ^3J_{H2'\cdot H2''} + ^3J_{H2''\cdot H3'},
\]

\[
\Sigma 3' = ^3J_{H2'\cdot H3'} + ^3J_{H2''\cdot H3'} + ^3J_{H3'\cdot H4'}.
\]

The pseudorotational angles was converted to torsion angles using the Amber program makeANG_RST, with Amber force constants of 50 kcal/mol·degree. The epsilon ($\varepsilon$) torsion angles was determined by measuring the $^3J_{P\cdot H3}$ coupling values from a CT-NOESY. This pulse program consists of two interleaved NOESY experiments, with selective excitation on the H3’
protons with (attenuated) and without (reference) $^{31}\text{P}$ decoupling. Then, using the following equations $^3J_{\text{H}_3'}\text{-P}$ can be determined:

$$\text{Attenuated} / \text{reference} = \cos [ \pi ({}^3J_{\text{H}_3'\text{-P}} + D_{\text{H}_3'\text{-P}}) T_{\text{eff}} ],$$

where attenuated and reference are the peak intensities; $T_{\text{eff}} = T + 2 \times \tau_{180} \times 0.9$; $T$ is the constant time evolution period, $\tau_{180}$ is the duration of the excitation (RE-BURP) pulse, and 0.9 is scaling factor to take into account the $^{31}\text{P}$ evolution during the RE-BURP pulse. At lower field strengths ($\leq 600 \text{ MHz}$), $D_{\text{H}_3'\text{-P}}$ (dipolar coupling) can be neglected. Then, using a modified Karplus equation, the torsion angle can be calculated:

$$^3J_{\text{H}_3'\text{-P}} = 15.3 \cos 2 \phi - 6.1 \cos \phi + 1.6,$$

resulting in a $\epsilon'$ torsion angle defined as $\text{H}_3'\text{-C}_3'\text{-O}_3\text{-P}$. To convert to the literature standard of $\text{C}_4'\text{-C}_3'\text{-O}\text{-P}$:

$$\epsilon \text{ CCOP} = \epsilon' \text{ HCOP} - 120^\circ.$$ 

From this calculation, there are four possible solutions; however, the $\epsilon$ value in DNA is constrained to only values between $-150$ and $-190^\circ$ in standard B DNA. Values between $-130$ and $-220^\circ$ were considered. The Amber program makeANG_RST was used to generate restraints with force constants of 50 kcal/mol-degree. Residual dipolar coupling values can be measured during partial alignment, which occurs at high magnetic fields or in alignment media such as Pf1 bacteriophage. This negatively charged and rod-like particle can partially align the DNA in a magnetic field. The RDC values measured reflect the CH bond vector, which can be used to define the helical axis orientation in DNA. These are calculated from $^1J_{\text{H}-\text{13C}}$ values from F2-coupled $^1\text{H}-\text{13C}$ HSQC experiments using the following equation:

$$\text{RDC} = ^1J \text{ in the presence of phage} - ^1J \text{ without phage},$$
and created in Amber. Standard broad restraints for backbones (72) to maintain a right-handed conformation: \( \alpha \) -90 to -30\(^\circ\), \( \beta \) 135 to 215\(^\circ\), \( \gamma \) 30 to 90\(^\circ\), \( \varepsilon \) 140 to 300\(^\circ\), and \( \zeta \) 150 to 315\(^\circ\) can be used. Amber force constant penalties of 50 kcal/(mol·degree) were used for each. Additionally, base pair restraints (distance and angle) were used from ideal structures, as all base pairs were accounted for by NMR imino \(^1\)H spectra: G:C base pairs: G-O6:C-N4 2.81 – 3.01 Å, G-N1:C-N3 2.85 – 3.05 Å, G-N2:C-O2 2.76 – 2.96 Å; A:T base pairs: A-N1:T-N3 2.72 – 2.92 Å, A-N6:T-O4 2.85 - 3.05 Å. Parabolic boundaries beyond the distance restraint well widths were 0.5 Å and angle restraints were 170 to 190\(^\circ\). Amber force constant penalties of 10 kcal/(mol·degree) and 25 kcal/(mol·Å) were used for angle and distance restraints, respectively.

After creation of restraints, a model of the system was constructed. Starting structures were generated as B-form using the NUCGEN tool of Amber. For the riboguanosine containing duplexes, pdb files were opened in vi editor and the DG residue was manually changed to an RG residue, deleting the H2’’ residue manually. Then, using xleap (Amber), libraries and parameters for force-field were implemented. The DNA was neutralized with sodium ions and then solvated, using an octahedral box with TIP3P water 10.0 Å from the edge of the solute molecule. Xleap was used to generate the solvated/ion-containing topology file with the embedded forcefield parameters (.prmtop), a starting coordinate/restart file (.crd) and a starting structure file (.pdb). The starting structure was first minimized while the DNA was held rigid to minimize solvent and counter-ions. Then the entire system was minimized. Finally, the broad backbone torsion angles, epsilon torsion angles, sugar and base pair restraints were implemented for a single Amber cycle consisting of energy minimization steps and a molecular dynamics run with constant pressure with fast heat exchange and 1 fs time step for 100 ps, followed by a second minimization cycle.
The cycle was repeated with the addition of quantitative distance restraints calculated using Mardigras and validated using Corma for multiple cycles. Mardigras calculates proton-proton distance constraints for nucleic acid structures from crosspeak intensities using 2D NOESY intensity/integral files and Amber calculated structures that are refined as the cycles proceed. The algorithm calculates a relaxation rate matrix from the experimental NOESY intensity data and the Amber structure file. Mardigras accounts for spin diffusion and therefore permits the use of NOESY data recorded at longer mixing times thus allowing the use of longer mixing times in the NOESY experiments and for more and stronger NOESY crosspeaks. Through an iterative procedure (RANDMARDI) introducing random noise and relative errors, deviations between the calculated and experimental NOE intensities are minimized and distances are calculated from the final cross relaxation rates. The Mardigras package produces relatively narrow distance bounds which in conjunction with Amber results in high quality structures from NOESY data. The structures are then directly validated against the experimental NOESY data using Corma which calculates a theoretical NOESY intensity file using a structure and compares that to the experimental data producing a quality factor $R^\alpha$. Inter-proton NOESY integrations were brought into Mardigras in small batches. The Mardigras output must be converted to Amber format using the Mardigras script $m2ahomo$ and 2 Å for upper and lower parabolic widths and 20-30 kcal/mol-Å force constants. Thymine methyl integrations, defined as "M7" for atoms H71, H72 and H73, were treated as pseudoatoms and distances calculated iteratively to the pseudoatom at the geometric center of the three methyl protons.

After adding the Mardigras distances to the Amber restraint file, and following an Amber cycle, the resulting structure was validated against the experimental data. The pdb file, stripped of water and counter ions, was reformatted using the $corma.in$ script and validated using Corma.
Corma creates a theoretical intensity file based on the structure, field strength, mixing and correlation time. The validation, in terms of a Corma \( R^x \) score, defined as the sixth-root residual index:

\[
R^x = \sum_i |V_{\text{exp}}^{1/6}(i) - V_{\text{calc}}^{1/6}(i)|/\sum_i V_{\text{exp}}^{1/6}(i),
\]

where \( V_{\text{exp}}(i) \) and \( V_{\text{calc}}(i) \) are the measured and theoretical volumes of cross peak \( i \), respectively.

The cycle was repeated until all restraints were implemented and yielded final structures with low \( R^x \) scores (below 6%) and low Amber energy penalties.

Amber and Mardigras/Corma cycles were run to generate structures for each NOESY mixing time measured. After structural comparisons and validation, the Mardigras `avgbnds` script was used to average the distances from each Mardigras distance constraint output file (.bnds) and used to calculate an average distance constraint file for final average structure calculations. RDC restraints were implemented on the final average structure and subjected to a 10 ns rMD run at 300 K. From the final rMD trajectories, a sampling of the last 100 ps, with a snapshot every 10 ps, was used to generate a final bundle of structures. Each snapshot was individually minimized and compared and validated against experimental NOESY data for each mixing time via Corma \( R^x \) scores.

### 2.11.3 Protocol for free molecular dynamics calculations

Unrestrained MD simulations were run on NVIDIA GPUs using Amber 12-14 (55,73). Structures were generated as perfect B-form DNA using Amber’s Nucleic Acid Builder (NAB). The riboguanosine was manually changed in the pdb file using vi editor. For the free MD simulations, additional GC base pairs were added to each termini to prevent fraying effects previously observed at AT ends. Using Amber’s tleap, forcefields parmBSC0 and/or BSC1 and additional reparameterization files was implemented, the duplexes were neutralized and
solvated with TIP3P water molecules. A gentle equilibration protocol was implemented, modified from a procedure by the Orozco lab (74) and protocol by Ross Walker (75). The system was carefully minimized and slowly heated to avoid “vacuum” bubbles. The solvent and ions were first minimized, holding the DNA rigid, followed by a full system minimization. The system was slowly heated from 0 to 100 K with weak 25 kcal/mol·Å restraints and Langevin temperature calibration scheme for 10-20 ps using constant volume. Next, the system was heated 100 to 300 K with constant pressure for 80-100 ps. Then, a short 100 ps, 300 K dynamics run with constant pressure was run while slowly reducing DNA constraints. Finally, an equilibration run for at least 1 ns was implemented prior to the final production run of 0.2-1 μs. All production runs were implemented with Amber’s pmemd.cuda for GPU calculations with Particle Mesh Ewald (76) method for electrostatic effects. Production runs were NPT (isothermal, isobaric ensemble) at 300 K, 1 atm with SHAKE (77) with an integration step of 2 fs. Trajectory snapshots are written every 10-100 ps, depending on simulation size.

2.11.4 Calculation of syn/anti populations

The nucleobases can exist in either the syn or anti conformations. In the case of the AxC and CxG duplexes, the 5’ terminal adenosine nucleobase displayed characteristics of having a syn population in both the control and riboguanosine containing structures. The 5’ terminal adenosine base (H8) to sugar (H1’) crosspeak was observed to be qualitatively more intense in each of the NOESY Base-to-H1’ pathways, Figure 2.8, when compared to other equivalent crosspeaks. The significant increase in intensity is due to different conformation yielding different distances between the base and H1’ resonances. Canonical purine syn and anti-distances are 2.2 Å and 3.7 Å, respectively. Thus, some syn populations will disproportionately skew the intensity of the crosspeak, as peak volume is proportional to \( r^{-6} \), where \( r \) is the distance...
inter-proton distance. To confirm syn populations, the volume of the adenosine H8 to H1’
crosspeak was used to calculate percent syn population using internal cytosine H5-H6
integrations as references in each short 75 ms mixing time NOESY. Theoretical syn and anti
NOE integrations were calculated for each set of data using the following equation:

$$\frac{\text{NOE}_{\text{anti}}}{\text{NOE}_{\text{ref}}} = \frac{r_{\text{ref}}^5}{r_{\text{anti}}^6},$$

$$\frac{\text{NOE}_{\text{syn}}}{\text{NOE}_{\text{ref}}} = \frac{r_{\text{ref}}^5}{r_{\text{syn}}^6},$$

where $\text{NOE}_{\text{ref}}$ is the value for the cytosine H5-H6 integration and $r_{\text{ref}}$ is the H5-H6 distance: 2.46 Å, and $r_{\text{anti}}$ and $r_{\text{syn}}$ distances are 3.7 and 2.2 Å, respectively. Assuming that only two populations
exist, syn and anti, each theoretically full anti and syn NOE integration was used to calculate the
percent anti population from the observed adenosine H8-H1’ integration, $\text{NOE}_{\text{obs}}$:

$$X \ (\text{NOE}_{\text{anti}}) + (1-X) \ (\text{NOE}_{\text{syn}}) = \text{NOE}_{\text{obs}},$$

$X$ is the population of the anti-conformation and $1-X$ is the syn population. The AxC and CxG
duplexes had between 15-20% syn populations. Because of this, the 5’ terminal adenosine to
intra or inter residue NOESY integrations provided unreliable distance calculations are were not
used for these duplexes.
3 PROBING COMPLEX FORMATION OF PU.1 BY NMR

3.1 Preface

The project was initiated by Dr. Gregory K. Poon. Project conception for probing complex formation of PU.1 by Diffusion Ordered Spectroscopy (DOSY) was a collaborative effort by Dr. Markus W. Germann and Dr. Gregory K. Poon. Protein was expressed and purified by Shingo Esaki.

3.2 Abstract

The DNA-binding domain of ETS (E26 transformation specific) transcription factors are known to bind sequence-specific sites as monomers over a single helical turn. PU.1, a member of the ETS family, has been shown to sequentially dimerize using titration calorimetry at excess protein concentrations (μM) in the presence of sequence-specific DNA. To confirm the dimerization of PU.1 and further explore complex formation, spectroscopic studies were carried out on PU.1 and a series of DNA substrates using NMR. Sequential dimerization is observed at site-specific DNA containing more than 10 base pairs of DNA, as 10 base pair DNA is insufficient to support dimerization. Furthermore, NMR results implicate the distal surface away from the DNA-binding site as a potential site of protein dimerization.

3.3 Introduction

Cellular fates and processes are the result of a complex series of signaling and binding events that govern gene expression. At the core of these tightly regulated events are transcription factors. These proteins are involved in the transcription of DNA into RNA and include a wide assortment of proteins with DNA-binding domains. The specific sequences of DNA recognized by the transcription factors, called promotor sequences, can be near the transcription start site or be thousands of base pairs from the gene being transcribed.
Additionally, these binding events can initiate or repress transcription to control gene expression (1,2). The Encyclopedia of DNA Elements (ENCODE) project details more than 150 such proteins, some of which are involved in hematopoiesis (3). Hematopoiesis, the name of a multi-step process involving the differentiation of a single progenitor cell, hematopoietic stem cell, into distinct blood cell lineages, is tightly regulated at the transcriptional level.

### 3.3.1 Role of PU.1 in biological systems

One of the ETS-family transcription factors is PU.1. This protein is considered to be one of the most essential hematopoietic regulators in ensuring the self-renewal of hematopoietic stem cells (4,5). However, like many other transcription factors, PU.1 has different roles at different stages of cell differentiation and is involved in both up- and down-regulation during expression. Elevated PU.1 activity is required to drive hematopoietic stem cells towards the myeloid lineages, while lower activity drives the progenitor cells towards the lymphoid lineages. Furthermore, up- or down-regulation of the transcription factor leads toward the macrophages and granulocytes (6) (continued PU.1 activity) or erythrocytes (7) (decreased activity) of the myeloid lineages or B- (high PU.1 activity) or T-lymphocytes (low PU.1 activity) of the lymphoid lineages (8). Unregulated or aberrant PU.1 activity is linked to a number of diseases including rheumatism (9), hematologic cancers (10-12) and Alzheimer’s disease (13). A fundamental understanding of the regulation of PU.1 is essential to understanding these diseases. Unlike other ETS DNA-binding family members which contain auto-inhibition, where helices near the DNA-binding domain reversibly reduce binding affinity (14), PU.1 lacks a known auto-inhibitory mechanism. Known inhibitory mechanisms for PU.1 include GATA1 inhibition, where GATA1 and PU.1 inhibit each other by both binding DNA, but their expressions only partially overlap (15).
3.3.2 Structure and dynamics of PU.1

ETS family proteins are characterized by a conserved 85 residue DNA-binding domain (16). These proteins share a winged helix-turn-helix binding motif (17). The PU.1 domain consists of a globular topology (33 x 34 x 38 Å$^3$) of three α helices and four-stranded antiparallel β-sheets (17). ETS family transcription factors share remarkable structural homology. Ets-1, another ETS member, and PU.1 represent extreme sequence divergence, with only 30% sequence homology (18), yet both have a high preference for the 5’-GGAA(A/T) sequence (19) and almost superimposable backbones when bound to high-affinity DNA (20). In a crystal structure, PU.1 binds DNA as a monomer, interacting with a 10 base pair region of DNA. Four of the amino acids which interact with the DNA are highly conserved among the ETS proteins, including: two arginine (Arg232 and Arg235) residues from the recognition helix (α3) lying in the major groove of the DNA, a lysine (Lys245) residue from the “wing” making contact with the DNA upstream of the consensus sequence, and a second lysine (Lys219) from the “turn” making a contact with the DNA downstream on the opposite strand (17). Arg232 and Arg235, of α3, make base contacts with the 5’-GGAA sequence. Lys245, of the “winged” β3, contacts the backbone of the minor groove, and Lys219, of the “turn”, forms a salt bridge with the phosphate backbone of the complementary strand downstream of the core sequence. These conserved residues are critical for DNA binding, as mutation studies with glycine at each site, eradicates binding (17). Additionally, water likely plays an important role in the protein-DNA recognition, as the crystal structure reveals 27 solvent molecules around the DNA, and solvent molecules involved in stability; Arg232 and Arg235 form direct and water-mediated contacts with the bases, as well as other residues including Thr226, Gln228 and Asn236 which are involved in water-mediated contacts to the DNA bases, but are not conserved for ETS proteins (17). PU.1
also shows remarkable dynamics in the backbone. The dynamics of the unbound PU.1 have been probed using hydrogen-deuterium exchange for amide protons and revealed very fast exchange and the possibility of local unfolding of the helix, α2 (21).

3.3.3 Evidence of multimeric complex formation in PU.1 and other ETS proteins

Diffusion studies by Jia, et al., found that in addition to dynamics in the backbone of unbound protein, while remaining a monomer at 0.3 mM concentrations, PU.1 oligomerized at high concentrations (2.5 mM) (21). Binding of 1:1 protein:DNA is universally observed in crystal structures of ETS family proteins, yet evidence for dimerization at a single DNA binding site by PU.1 has been observed by calorimetric titrations done by Poon, et al. (22). While other ETS family proteins have been found to dimerize at two sites, forming 2:2 complexes, dimerization at a single site is less common (23-26). Sequential dimerization of site-specifically bound PU.1 could act as a potential mechanism for inhibition, as dimerization would sequester any excess PU.1 from DNA.

3.4 Materials and methods

3.4.1 Protein and DNA NMR sample preparation

Proteins were expressed and purified by Shingo Esaki, from Dr. Gregory Poon’s group. Briefly, the recombinant construct representing the ETS domain of murine PU.1 (residues 167 to 272, PU.1ΔN167 or PU.1) was cloned with a thrombin-cleavable C-terminal 6xHis tag as described (27). Unlabeled protein was over-expressed in E. coli in LB medium. Uniformly $^{15}$N-labeled PU.1ΔN167 was expressed from 5 mL starter cultures in LB broth and grown at 37°C for 8 h. Cultures were inoculated into 250 mL of LB broth and grown at 37°C for 16 h before harvesting. The cell pellet was re-suspended in standard M9 medium containing $^{15}$NH$_4$Cl, MgSO$_4$, CaCl$_2$, trace metals, MEM vitamins (GE), and glucose. Protein expression was induced
with 0.5 mM IPTG overnight at 25°C. Both unlabeled and $^{15}$N labeled proteins were purified as described (28); cleared lysate was purified by immobilized metal affinity chromatography, cleaved with thrombin, dialyzed against 10 mM NaH$_2$PO$_4$/Na$_2$HPO$_4$ (pH 7.4) containing 0.5 M NaCl and purified on Sepharose SP (GE). Protein concentrations were determined by UV absorption at 280 nm using the following extinction coefficients: 22,460 M$^{-1}$cm$^{-1}$. The labeling efficiency of $^{15}$N-labeled constructs was over 98% as judged by mass spectrometry (5).

Synthetic DNA oligonucleotides were purchased from Integrated DNA Technologies. Single stranded DNA was dialyzed against 10 mM NaH$_2$PO$_4$/Na$_2$HPO$_4$ (pH 7.4) buffer with 150 mM NaCl for D$_2$O samples and 11 mM NaH$_2$PO$_4$/Na$_2$HPO$_4$ (pH 7.6), 167 mM NaCl for 10% D$_2$O samples. Concentrations of each strand were determined using extinction coefficients derived from the sum of the mononucleotides at an absorbance of 260 nm at 80°C on a Cary UV-Vis Spectrophotometer using the following extinction coefficients for A, C, G and T, respectively: 15500, 6980, 11200 and 9360 M$^{-1}$cm$^{-1}$. Duplexes were then annealed to form 16 base pair high-affinity (5’-GCAAGCGGAAGTGCAGC-3’), low-affinity (5’-GCAAGGGAAATGGAGC-3’), or nonspecific DNA duplexes (5’-GCAAGCGAGAGTGCAGC-3’) (ETS-specific core consensus in bold) with complementary strands. The 10 base pair strands are underlined for each duplex.

Purified PU.1ΔN167 (250 µM) and high- and low-affinity and nonspecific DNA were co-dialyzed against 10 mM NaH$_2$PO$_4$/Na$_2$HPO$_4$ (pH 7.4) containing 150 mM NaCl, lyophilized, and suspended to their previous volumes with 99.996% D$_2$O. The pH* (meter reading) of the reconstituted samples was 7.6 for all D$_2$O samples. Samples in H$_2$O/D$_2$O (90%/10%) were co-dialyzed in 11 mM NaH$_2$PO$_4$/Na$_2$HPO$_4$ and 167 mM NaCl, then samples were adjusted with 10% D$_2$O.
3.4.2 $^1$H NMR experiments

All NMR experiments were run on Bruker Avance 500 and 600 MHz spectrometers, equipped with a 5 mm TBI and QXI probe, respectively (Bruker). A 1D pre-saturation (zgpr) spectrum was measured for each titration prior to diffusion measurements for DNA and protein in 100% D$_2$O. Similarly, solvent suppression pulse programs, including Water-Gate (zggpW5) and a 1-1 Jump and Return (p11) was used for 10% D$_2$O samples to monitor protein folding and DNA base pairs, respectively. DNA base pairs were assigned using a 2D NOESY with 1-1 solvent suppression and excitation sculpting with 150 ms mixing time, 24 x 24 ppm sweep width. A 1D NOE using a 1-1 solvent suppression was used to confirm crosspeak assignments. All protons were referenced to an internal 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) standard.

3.4.3 Diffusion-ordered spectroscopy experiments

Diffusion experiment parameters ($\Delta$, $\delta$ and gradient strength) were optimized by running 1D diffusion experiments (stebpgp1sd) at 2 and 95% gradient strengths with 100 ms and 5 ms, $\Delta$ and $\delta$ diffusion times, respectively yielding ~10% signal retention. DOSY calibration, test samples and additional data is described in the Supplementary Information. Using these parameters, a pseudo-2D diffusion-ordered spectroscopy (DOSY) experiment using stimulated echo with bipolar gradient pulses (stebpgp1s) was acquired with 16k x 20 data points with a spoil gradient of 1.1 ms and 4.0 s relaxation delay from 2 to 95% gradient strength with a linear ramp. Data was then processed with Bruker Topspin T1/T2 software using manual peak picking. Care was taken to avoid NMR peaks that potentially overlap with free DNA at 1:1 (protein:DNA) and excess DNA titrations. The intensity $I$ of each picked peak was fitted to the following equation as a function of field gradient strength $G$:
\[ I(g) = I_0 \exp(-DQG^2), \]

where \( I_0 \) is the reference intensity, \( D \) the diffusion coefficient, and \( Q \) is a constant consisting of fixed parameters specific to the experimental configuration, listed in the Supplementary Information.

After initial DOSY measurements of free protein and DNA, titrations proceeded at various protein to DNA ratios. In the case of low-affinity and nonspecific DNA, precipitation was observed at ratios following DNA:protein = 1:3 immediately after titration that resolved overnight at room temperature.

### 3.4.4 \( ^1H-^{15}N \) HSQC of PU.1

The \( ^{15}N \) labeled protein constructs and DNA in 10% D2O were used for HSQC experiments, where DNA was titrated into protein at quarter ratios. \( ^1H-^{15}N \) correlated measurements were made using a phase sensitive, double incept transfer with a garp decoupling sequence and solvent suppression (hsqc3gpph19). Spectra were acquired with 1k x 144 data points and zero-filled to 4k x 4k.

### 3.4.5 Simulating \( ^1H \) methyl exchange

\( ^1H \) methyl resonances of PU.1 in the presence of various ratios of site-specific DNA were used to simulate a two state model in which the 2:1 (protein:DNA) complex is in intermediate exchange between unbound and 1:1-bound conformations using DNMR (Topspin 2.1). An envelope corresponding to no exchange was used to fit a superposition of the observed, normalized (to DSS) spectrum of experimental unbound and 1:1 bound at 7:3 population, based on a best fit of the data. The exchange rate constant, \( k_{ex} \), was then varied from to different exchange rates. Additional details can be found in the Supplementary Information.
3.5 Results

3.5.1 $^1$H NMR of PU.1

All 1D $^1$H NMR spectra of free, unbound proteins displayed characteristics of properly folded protein, including good chemical shift dispersion and far shifted methyl peaks, upfield of the proton standard DSS, Figure 3.1. Interestingly, upon titration of 16 base pair high-affinity DNA, the upfield methyl peaks (negative chemical shift), indicative of proper folding, disappear at the 1:0.5, protein:DNA, ratio, but reappear at molar equivalence and excess DNA, Figure 3.1, with slight chemical shift differences, corresponding to a different conformation, presumably the bound state. This disappearance can be due to partial structural rearrangement or chemical exchange, see Supplementary Information for more details.

3.5.2 Simulated exchange peaks

Chemical exchange was simulated using DNMR (Topspin) to recreate the disappearing methyl peaks and determine if these peaks disappear purely from a two state exchange model of free and bound states. The simulated coalescence envelopes, calculated from chemical shifts of the free and 1:1 bound methyl peak groups, were 28-81 Hz at 500 MHz (Supplementary Information). The most broadly coalesced envelope corresponding to intermediate exchange established a lower limit on the intensity which was significantly in excess of the observed signal for the experimental 2:1 methyl peaks, Figure 3.2. Relaxing the constraints of equal width for the three sets of peaks did not improve the fit to the 2:1 spectrum. Because the most broad, simulated coalescence point was still significantly more intense than the observed 1:0.5 ratio methyl peaks, the observation of the missing methyl peaks cannot purely be explained by exchange between free and 1:1 bound states. It is possible that at the 1:0.5 ratio (protein:DNA), the conformation of the protein is changed, and the methyl peaks are shifted and buried under different resonances.
3.5.3 $^1$H NMR of DNA substrates

Base pairing via the imino proton ($^1$H NMR) of DNA substrates was also investigated to yield insight into the binding of PU.1 and DNA. The free $^1$H spectra of the 16 and 10 base pair DNA was collected and each peak was assigned with NOESY spectra; data shown only for the 10 base pair high-affinity DNA, Figure 3.3. PU.1 at quarter molar equivalents was titrated into each DNA substrate and spectra were collected, Figures 3.4, 3.5 and 3.6.

The 16 base pair high-affinity DNA titration, showed expected free and bound resonances, corresponding to each titration until the 1:1 ratio, which showed only bound peaks (Figure 3.4); interestingly, the 1:1 and 2:1 (excess protein) titration curves showed no difference in the bound DNA. From the DNA perspective, PU.1 bound to high-affinity 16 base pair DNA does not change the DNA at the 1:1 or 2:1 ratio, suggesting that if PU.1 is a dimer at the 2:1 ratio, the dimerization interface is far from the DNA and the second protein does not make contacts with the bound DNA. Furthermore, at the 2:1 ratio (excess protein) a broadening of the bound DNA protons is observed, suggesting an overall larger complex. This observation is consistent with dimerization of the protein when bound to 16 base pair high-affinity DNA at excess protein concentrations. The 10 base pair high-affinity DNA titration reveals similar trends (Figure 3.5).

The titration of PU.1 into 16 base pair nonspecific DNA yielded different results, Figure 3.6. Unlike the high-affinity titrations, no distinct populations of free and bound DNA are observed, rather, all of the peaks appear to broaden out and shift in each titration.

3.5.4 Diffusion-ordered spectroscopy

DOSY spectra were acquired to characterize the PU.1/DNA complex hydrodynamic shapes. Diffusion measurements of PU.1 were measured with site-specific (high- and low-
affinity DNA) and nonspecific DNA (5). Protein concentrations were kept constant at ~250 μM concentration (sub 300 μM PU.1 is a monomer (21)). At 250 μM, PU.1 has a measured diffusion coefficient of $(9.1 \pm 0.1) \times 10^{-11} \text{m}^2/\text{s}$ in D$_2$O at 25°C, which is in good agreement with predicted PU.1 values, Figure 3.7 and Table 3.1 (5),(29). After measuring free protein, DNA substrates were subsequently titrated into each sample at various ratios.

Titration of PU.1 16 base pair high- and low-affinity DNA revealed similar trends, where the diffusion coefficient reached a minimum of $(5.9 \pm 0.1) \times 10^{-11} \text{m}^2/\text{s}$ at a ratio of 0.5 (PU.1:DNA = 2:1). Further addition of site-specific DNA past this point increased the diffusion coefficient to a value of $(7.5 \pm 0.2) \times 10^{-11} \text{m}^2/\text{s}$ at 1:1 ratios and excess DNA. The diffusion of a dimer, modeled as a pair of rigid spheres would have a diffusion coefficient of 75% of the monomer (30), or a diffusion coefficient of $6.8 \times 10^{-11} \text{m}^2/\text{s}$ (Supplementary Information). This value is consistent with the possibility of a dimer bound to 16 base pair DNA, which would have a slower diffusion than just a dimer of proteins. The diffusion values, Figure 3.7, are consistent with a model of sequential dimerization which starts at free protein, 2:1 (protein:DNA) complex, 1:1 (protein:DNA) complex, and is not consistent with a 2:2 complex, as no free protein at the 1:0.5 (protein:DNA) ratio is observed.

Nonspecific 16 base pair DNA was also titrated into PU.1 and revealed a contrasting trend, Figure 3.7 and Table 3.1 (5). The 16 base pair nonspecific DNA titration revealed the formation of only a single complex at the DNA:PU.1 0.5 ratio with a diffusion coefficient of $(6.6 \pm 0.2) \times 10^{-11} \text{m}^2/\text{s}$. This diffusion value is intermediate between the 16 base pair high- and low-affinity DNA 1:1 and 2:1 complexes. This value, at the 0.5 ratio, is believed to be the formation of a dimer, but with a different hydrodynamic shape than the site-specific dimer complexes. This
intermediate diffusion value is not likely to be a mixture of monomeric and dimeric complexes, as this would be concentration dependent at continued DNA additions.

Because the crystal structure of bound PU.1 to site-specific DNA revealed contacts with only 10 base pairs, 10 base pair DNA substrates containing core sequences of the 16 base pair counterparts were also titrated into PU.1 and DOSY spectra were acquired, Figure 3.7 and Table 3.1. In contrast with the 16 base pair counterparts, all of the 10 base pair DNA substrates, site-specific or not, revealed similar DOSY trends of the formation of a single complex at 1:1 molar equivalence and excess DNA (5). Interestingly, the diffusion values of the 10 base pair DNA substrate complexes were also at intermediate values between 16 base pair site-specific dimeric and monomeric complexes. Because the titration was at completion at the 1:1 ratio, the complex is believed to be that of a monomeric, 1:1 complex; however, although the complex is smaller (10 base pairs compared to 16 base pairs), the diffusion coefficient is lower than a 16 base pair monomer. Further evidence of a 1:1 and not 2:1 complex is supported by the imino $^{1}$H PU.1 into DNA titrations, Figure 3.5. The DOSY data reveals the formation of only a single complex and the imino $^{1}$H titration reveals that free DNA is not present at the 1:1 equimolar ratios, indicative of a 1:1 and not 2:1 complex. However, because the diffusion value is lower than predicted, a larger, or conformationally different 1:1 complex is predicted for the 10 base pair complex with PU.1.

3.5.5 $^{1}$H-$^{15}$N HSQC

To further probe structural and dynamic characteristics of the PU.1 complexes, $^{1}$H-$^{15}$N HSQC spectra of uniformly labeled ($^{15}$N) PU.1 was acquired. The free labeled PU.1 protein revealed a structured protein with well-dispersed crosspeaks, Figure 3.8 and 3.9. The temperature was increased every 10 degrees to reveal snapshots of the protein backbone. At 55°C, most of the
amide crosspeaks are no longer visible, which is consistent with the thermal stability of PU.1 (5). Thermal denaturation is reversible, as lowering the temperature back to 25°C brought back the HSQC crosspeaks. $^1$H-$^{15}$N HSQC crosspeaks assignments were from reported resonances by Jia et al. (21).

Addition of 16 base pair high-affinity DNA substrate at a 1:0.5 (protein:DNA) ratio resulted in the immediate disappearance of 80% of the NH resonances, Figure 3.9. Continued titration to 1:1 and excess DNA ratios restored the NH protons, with chemical shift perturbations, suggesting conformation differences of the protein. This loss of crosspeaks can be due to chemical exchange between conformations corresponding to free protein and bound 2:1 protein. No free protein is observed in the 1:0.5 ratio. The three distinct protein conformations of the HSQC mirrors the DOSY data, which suggests free protein, 2:1 (protein:DNA) dimeric complex, and 1:1 monomeric complex. Furthermore, many of the missing residues in the 1:0.5 ratio titration correspond to proteins found on the distal surface from the DNA, such as W192, W193, D197, etc. (Figure 3.10) (5). A dimerization interface far from the DNA is also consistent with the imino $^1$H spectra of the 16 base pair high-affinity titration, Figure 3.4, which shows no difference between the 2:1 or 1:1 (protein:DNA) titrations on the DNA.

The $^1$H-$^{15}$N HSQC titration of the 16 base pair nonspecific DNA revealed a marked loss of NH crosspeaks at the 1:0.5 (protein:DNA) titration ratio, and further addition of DNA substrate did not restore the crosspeaks, Figure 3.9. This is consistent with the DOSY data which suggests that a single complex forms at the 1:0.5 (protein:DNA) ratio and is unchanged even at excess DNA. Furthermore, no evidence of free protein is visible in the 1:0.5 ratio, consistent with a 2:1 dimeric protein complex. Although the majority of the NH crosspeaks have disappeared or are involved in exchange as with the 16 base pair site-specific DNA dimeric
complex, the pattern and chemical shifts are not identical; this further supports that PU.1 forms
different complexes with nonspecific DNA and site-specific DNA, as observed by the radically
different diffusion values (5).

The 10 base pair high-affinity DNA duplex was also titrated into uniformly $^{15}$N-labeled
PU.1 for HSQC titrations, Figure 3.9 (5). At the 1:0.5 (protein:DNA) ratio, free protein is still
visible, however, a loss of NH crosspeaks is observed at equimolar equivalence and excess
dNA. The visible free protein at the 1:0.5 ratio further supports the formation of a 1:1
monomeric complex over a 2:1 complex. Additionally, the loss of crosspeaks at 1:1 ratio could
be consistent with a dynamic protein backbone and a more extended hydrodynamic radius, which
could explain the lower than predicted diffusion data for the 1:1 DOSY data, Table 3.1.

3.6 Discussion

3.6.1 Evidence for dimerization of PU.1 in the presence of site-specific DNA

The DOSY data, together with the imino $^1$H and $^1$H-$^{15}$N HSQC titrations, supports the
formation of a dimeric complex between PU.1 and site-specific DNA (16 base pair) at excess
protein concentrations. The imino $^1$H spectra of the DNA titrated with PU.1 confirms a 1:1
monomeric complex at equimolar ratios, Figure 3.4. The HSQC and DOSY data reveals the
formation of two distinct complexes with different hydrodynamic radii and dynamics, Figures
3.7 and 3.9. Diffusion values are also consistent with the predicted complexes, Supplemental
Information. The formation of a 2:1 complex is also consistent with the data over a 2:2 complex,
as excess protein is not observed at 1:0.5 (protein:DNA) ratios in DOSY or HSQC data. The
disappearance of 80% of the NH amide crosspeaks in the HSQC spectra could be due to
chemical exchange between conformations; furthermore, the missing protein methyl peaks could
indicate a different structural conformation of the dimeric complex, as those residues are no
longer influenced by ring current from adjacent aromatic rings. Finally, the HSQC data and
imino $^1$H data of the 16 base pair high-affinity DNA titration give some insight into the structural
arrangement of the dimeric complex. Missing NH crosspeaks for residues on the distal surface
from the DNA suggests they are involved in dynamics and conformational exchange and could
be involved in the dimerization interface, Figure 3.10. This is supported by the $^1$H NMR spectra
indicating no change in the bound DNA in a monomeric or dimeric complex, save for line
broadening, Figure 3.4.

The sequential dimerization of PU.1 at site-specific DNA is also confirmed using a
number of different biochemical and spectroscopic techniques done by our collaborators (5).
Poon, et al. showed isothermal calorimetry titrations where site-specific DNA was titrated into
PU.1 and found a biphasic profile corresponding to the formation of a dimer and monomer
complex, which is not observed in Ets-1, a family homolog (5,22). Furthermore, fluorescence
anisotropy and gel shift assays also reveal sequential dimerization of PU.1 at site-specific DNA
>16 base pairs (5). Finally, a dimerization interface on the distal surface from the DNA binding
interface was confirmed with fluorescence anisotropy and gel mobility shifts of a mutant variant,
where residues $^{195}$DKDK$^{198}$ found on the predicted dimerization interface were mutated to
$^{195}$NINI$^{198}$ and only monomeric 1:1 binding was observed (5). Fluorescence intensity of 8-
-anilino-1-napthalenesulfonate (ANS) bound to PU.1 was also measured by Dr. Gregory Poon
and Shingo Esaki (5). ANS dye is an indicator of solvent-exposed hydrophobic basic residues,
producing a strong blue-shifted fluorescence upon binding. ANS bound to PU.1 (50 μM) results
in a high fluorescence intensity, which is reduced significantly at the 1:1 PU.1/DNA ratio (5);
this suggests that the ANS dye and DNA compete for the same hydrophobic pocket. At the 2:1
ratio (PU.1:DNA), the ANS dye still produces a binding associated intensity, albeit at a 3-fold
reduction compared to unbound 50 μM. This intensity is reduced compared to ANS bound to 25 μM concentration of PU.1. This confirms the 2:1 ratio represents a different binding scenario than simply a mixture of 1:1 complex and 25 μM of free PU.1. Furthermore, the fairly high fluorescence intensity at the 2:1 ratio (compared to the 1:1 fluorescence), indicates that there is still a solvent exposed hydrophobic region in the PU.1 protein which is not bound to DNA. This can be interpreted as a dimeric PU.1 protein-DNA complex which is less structured or more dynamic than the 1:1 complex, which is supported by an extended binding site exhibited by DNA foot-printing data (5). Alternatively, this solvent exposed binding pocket could be on the second protein of the dimer, not bound to DNA; this would further support the formation of a 2:1 (not 2:2) complex as well as shed some insight into the structure and dynamics of the second protein of the dimer.

3.6.2 Complex formation of PU.1 bound to nonspecific DNA is unique from the site-specific complex

Unlike the site-specific DNA, nonspecific DNA (16 base pairs) was shown to bind PU.1 in a different manner. DOSY data, Figure 3.7, revealed the formation of a single complex at 1:0.5 (protein:DNA) ratios, suggesting the formation of a single dimeric complex, which did not change upon further addition of DNA. The absence of any free protein in the \(^1\)H-\(^{15}\)N HSQC at this ratio further suggests the dimeric nature of this complex. Interestingly, this dimeric complex is unlike the one formed by the site-specific DNA and PU.1, and has a fundamentally different hydrodynamic radius. The dimeric nature of this complex is further evidenced by fluorescence anisotropy data which suggests the formation of only a dimeric species (5).
3.6.3 An extended DNA interface is necessary for sequential dimerization

Although the crystal structure of bound PU.1 revealed contacts along 10 bases of site-specific DNA (17), DOSY and HSQC data suggest 10 base pairs is too short of a DNA substrate to form a sequential dimer, as with the site-specific 16 base pair DNA. The diffusion data, Figure 3.7, reveals only a single complex forms at 1:1 ratios. This complex is not likely a 2:1 dimeric protein complex, as HSQC data shows free protein visible at 2:1 ratios, and the imino $^1$H data show no free DNA at the 1:1 ratio. The HSQC data at 1:1 of the complex has missing peaks suggesting some exchange. The idea of a larger hydrodynamic radius is consistent with the unique DOSY diffusion value, which is too slow for 1:1 with 10 base pair duplex. This finding is consistent with hydroxyl radical footprinting done by Dr. Gregory Poon, which, at single nucleotide resolution, revealed a footprint of a 2:1 complex spanning almost 20 base pairs of DNA at site-specific DNA (5). The dimer, even though the dimerization interface was far from the DNA resulted in an extended footprint. This finding explains why 10 base pairs is not sufficient for PU.1 dimerization, as flanking nucleotides downstream of the core sequence are involved in contacts with the protein (5). Fluorescence anisotropy data of the 10 base pair high-affinity duplex binds PU.1 at 1:1 ratios (5). Although not likely, as 16 base pair site-specific DNA binds PU.1 as a 2:1 complex, the possibility that PU.1 binds 10 base pair high-affinity DNA as a 2:2 complex cannot be entirely ruled out. Fluorescence anisotropy binding studies on the mutant $^{195}$NINI$^{198}$ should be done with the 10 base pair high affinity DNA; binding should confirm a 1:1 complex, whereas no binding could indicate a 2:2 complex.

3.7 Conclusion

The sequential dimerization of PU.1 bound to site-specific DNA at excess protein concentrations has been shown using NMR techniques. This dimerization is unique to PU.1 and
possibly other ETS family proteins, and is not universal for all ETS proteins. Additionally, although a crystal structure of PU.1 bound to site-specific DNA was shown to form a monomeric protein complex making contacts with 10 base pairs, NMR evidence suggests that additional flanking nucleotides may be necessary to induce sequential dimerization. Furthermore, NMR HSQC spectra suggests a dimerization interface on the protein, which is far from the DNA binding site, as 80% of NH protons disappear at the 2:1 (protein:DNA) ratio, including resonances for residues opposite of the DNA binding residues. This observation is consistent with imino $^1$H spectra of the complexed DNA at 1:1 and 2:1 (protein:DNA) ratios, which does not change from monomer to dimer complexes. This finding is confirmed by mutation studies done by Dr. Gregory Poon and Shingo Esaki, where residues at the distal surface away from the DNA were mutated and dimerization was abolished (5).

The ability of PU.1 to sequentially dimerize at a single site-specific region of DNA could shed insight in a mechanism for auto-inhibition, which is currently lacking in PU.1. Many ETS family proteins, including Ets-1 and ERG, share a similar mechanism of auto-inhibition, where inhibitory helices block DNA binding domains, which can be disrupted by binding partners (5). Because PU.1 lacks this inhibitory helix, this protein can potentially be locked in a permissive state. PU.1 forms a reversible, negatively cooperative 2:1 complex as a possible inhibition mechanism (5). Furthermore, even if the 2:1 complex retains the same functional activity as a 1:1 complex, a dimer complex sequesters free circulating protein, reducing the number of active PU.1 molecules in the system.
Table 3.1 Translational diffusion coefficients of PU.1, complexes and DNA substrates.

All diffusion coefficient values (x 10^{-11} m^2/s) as determined by DOSY experiments. DNA substrates include: 16 and 10 base pair high-affinity (5'-GCAAGCGGAAGTGAGC-3'), 16 and 10 base pair low-affinity (5'-GCAAAAGGAATGGAGC-3'), and nonspecific DNA (5'-GCAAGCGAAGTGAGC-3'). The underlined region for each sequence represents the core 10 base pairs for the 10 base pair duplex (5).

<table>
<thead>
<tr>
<th>PU.1 to DNA Substrate</th>
<th>16 base pairs</th>
<th>10 base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high affinity</td>
<td>low affinity</td>
</tr>
<tr>
<td>1:0</td>
<td>9.11±0.17</td>
<td>9.00±0.15</td>
</tr>
<tr>
<td>1:0.17</td>
<td>8.15±0.25</td>
<td>7.80±0.25</td>
</tr>
<tr>
<td>1:0.25</td>
<td>8.30±0.24</td>
<td>8.03±0.45</td>
</tr>
<tr>
<td>1:0.33</td>
<td>7.27±0.15</td>
<td>6.68±0.11</td>
</tr>
<tr>
<td>1:0.5</td>
<td>6.21±0.18</td>
<td>5.84±0.16</td>
</tr>
<tr>
<td>1:0.67</td>
<td>5.90±0.11</td>
<td>5.92±0.11</td>
</tr>
<tr>
<td>1:0.75</td>
<td>7.11±0.13</td>
<td>7.54±0.32</td>
</tr>
<tr>
<td>1:0.83</td>
<td>6.72±0.06</td>
<td>6.69±0.19</td>
</tr>
<tr>
<td>1:1</td>
<td>6.53±0.14</td>
<td>6.56±0.13</td>
</tr>
<tr>
<td>1:1.17</td>
<td>6.64±0.14</td>
<td>6.43±0.19</td>
</tr>
<tr>
<td>1:1.25</td>
<td>6.43±0.13</td>
<td>6.68±0.18</td>
</tr>
<tr>
<td>1:1.33</td>
<td>6.75±0.06</td>
<td>6.51±0.17</td>
</tr>
<tr>
<td>1:1.5</td>
<td>7.53±0.23</td>
<td>7.79±0.08</td>
</tr>
<tr>
<td>1:1.75</td>
<td>6.30±0.14</td>
<td>6.62±0.16</td>
</tr>
<tr>
<td>1:2</td>
<td>7.84±0.15</td>
<td>6.81±0.10</td>
</tr>
</tbody>
</table>
Figure 3.1 $^1$H NMR spectra of PU.1 (250 μM) titrated with 16 base pair high-affinity DNA (A) and methyl region (B).

Protein and substrate were co-dialyzed in 10 mM sodium phosphate, 150 mM sodium chloride with DSS internal standard reference and brought up in 100% D$_2$O, pH* 7.2. Top to bottom in each panel represents protein to DNA: 1:1, 1:0.5, 1:0.25 and 1:0 (unbound protein). Excess DNA titrations do not impact the resonances.
Figure 3.2 Methyl peaks were simulated to probe for chemical exchange to explain 2:1 methyl peak broadening in the 16 base pair site-specific titration.

Peaks were simulated assuming two state exchange between free and 1:1 bound complex. Experimental data is shown in color: free protein (blue trace), 1:1 complex (green trace), 2:1 protein:DNA complex (red trace). Simulated envelope corresponding to no chemical exchange between states is shown as a black trace. Simulated coalescence points are shown in grey scale for each \(k_{\text{coalescence}}\): 28 Hz (dark grey trace), 65 Hz (medium grey trace), 81 Hz (light grey trace).
Figure 3.3 NOESY spectra for 10 base pair high-affinity DNA imino protons. Resonances too broad to yield observable NOESY intensities were confirmed through a temperature series. Resonances which suffered from overlap were assigned by 1D NOE experiments.
Figure 3.4 Imino $^1$H spectra of the 16 base pair high-affinity DNA substrate (A) titrated with PU.1 (bottom to top) (B) at 293 K.

Titration (bottom to top) (B) are for DNA:PU.1: 0:1 (free DNA), 0.25:1, 0.5:1, 0.75:1, 1:1, 1.25:1, >2:1.
Figure 3.5 Imino $^1$H spectra of the 10 base pair high-affinity DNA substrate (A) titrated with PU.1 (bottom to top) (B) at 293 K.

Titrations (bottom to top) (B) are for DNA:PU.1: 0:1 (free DNA), 0.25:1, 0.5:1, 0.75:1, 1:1, 1.25:1. >2:1.
Figure 3.6 Imino $^1$H spectra of the 16 base pair nonspecific DNA substrate (A) titrated with PU.1 (bottom to top) (B) at 293 K.

Titrations (bottom to top) (B) are for DNA:PU.1: 0:1 (free DNA), 0.25:1, 0.5:1, 0.75:1, 1:1, 1.25:1, >2:1.
Figure 3.7 Diffusion coefficients of PU.1 titrated with DNA substrates in 100% D$_2$O at 25°C determined by DOSY (5).
Figure 3.8 Thermal denaturation of PU.1 monitored by $^1H-^{15}N$ HSQC. Thermal denaturation is reversible, as protein crosspeaks were restored after lowering temperature to 25°C.
Figure 3.9 HSQC titrations of PU.1 with DNA substrates (5).
Figure 3.10 Dimerization interface from HSQC titrations (A,B) and complexed PU.1 (C,D).

$^{1}$H-$^{15}$N HSQC spectra for PU.1 (green) and PU.1-DNA complex at 2:1 (red) and 1:1 (blue) with 16 base pair high-affinity DNA (A,B). Peaks that are observable in all titrations represent a stable core which is not directly involved in DNA binding or dimerization, labeled in cyan (A-D). Peaks that are observable only in free PU.1 and 1:1 complexed PU.1 (albeit shifted) which are not at the DNA-binding interface are believed to participate in dimerization, labeled in orange (A-D). The protein-DNA structure is 1PUE (PDB ID) (17).
Figure 3.11 Gradients calibrations were benchmarked from HOD $^1$H resonance peaks using a 100% D$_2$O sample at 298 K. Uncalibrated GCC diffusion curve (A) compared to new GCC diffusion curve (B).
Figure 3.12 DOSY parameters were optimized with sucrose (A) and DNA (B).
Figure 3.13 1D slices of the DOSY at varied gradient strengths (A) for unbound PU.1. Diffusion plots for select peaks (color coded) (B) for: water (blue) and two PU.1 resonances (red, green).
Figure 3.14 The structure of PU.1 from the crystal structure 1PUE (17).
3.8 References


3.9 Supplementary information

3.9.1 Pulse field gradient calibrations for DOSY

Diffusion by NMR can be achieved using pulse field gradients. In simple terms, magnetic field gradients can be used to spatial label a molecule in the NMR tube. The diffusion measurements can be achieved by varying delays between dephasing and refocusing pulse field gradients, running a series of gradient strengths, and then measuring resonance intensities as a function of gradient strength. For proper calculations, gradient strengths must be calibrated. All diffusion-ordered spectroscopy (DOSY) measurements were recorded on the 500 MHz NMR instrument (Bruker) using Topspin 2.1. Parameter delays and pulse lengths were optimized for all samples by varying the delay and gradient length, Δ and δ, respectively with simple 1D steppgp1s experiments using 2% and 95% gradient strength until signal retention at 95% gradient strength was around 10%. Gradient calibrations were performed using calibration setups in the Bruker Topspin 2.1 manual. Briefly, the instrument specific Gradient Calibration Constant (GCC) was found under the instrument configuration:

/opt/topspin/config/inst/gradient_calib

and used to calculate a new GCC based on the diffusion of $^1$H in a 100% D$_2$O standard, Figure 3.11. The diffusion of proton of the D$_2$O sample was measured at 25°C after adequate temperature calibration, as diffusion is dependent on temperature. The measured diffusion value was determined to be $2.161 \times 10^{-9}$ m$^2$/s and compared to a literature value of $1.872 \times 10^{-9}$ m$^2$/s (Bruker manual). To measure the diffusion of the HOD resonance in a 100% D$_2$O sample at
25°C, optimal $\Delta$ and $\delta$ parameters were found to be 10 ms and 2.5 ms, respectively. The values was used to calculate a new GCC:

$$\text{GCC}_{\text{new}} = (D_{\text{literature}}/D_{\text{measured}})^{1/2} \times \text{GCC}_{\text{old}}.$$  

The new GCC, 5.75 G/mm, was used to replace the old GCC, 5.35 G/mm in the instrument configuration. DOSY measurements of test molecules, including 30 mM sucrose and 9 base pair DNA were measured, Figure 3.12. For 30 mM sucrose measurements at 298 K, optimal $\Delta$ and $\delta$ parameters were found to be 50 ms and 2 ms, respectively. For 9 base pair DNA (0.5 mM in 10 mM sodium phosphate, 100 mM NaCl), the optimal $\Delta$ and $\delta$ parameters were found to be 100 ms and 2.5 ms, respectively. Acquisition of 16 data points was sufficient to yield a complete diffusion curve for each.

For the results of the present study, gradient parameters were optimized for PU.1 and DNA; however, water diffusion coefficients were still recorded as internal standards. Optimal $\Delta$ and $\delta$ parameters were found to be 100 ms and 5 ms, respectively. Peak intensities were measured as a function of gradient strength using the peak picking tool and T1/T2 software of Bruker Topspin 2.1. Care was taken to avoid peaks which overlap with free DNA, especially at equimolar and excess DNA, Figure 3.13.

### 3.9.2 Estimating hydrodynamic radii from diffusion measurements

The diffusion coefficient can be calculated from DOSY measurements, where each peak intensity is measured as a function of gradient strength. Data is fit with the following equation for a DOSY experiment with bipolar gradients (stebpgp1s):

$$I = I_0 \times \exp(-D (2\pi\gamma G\delta)^2*(\Delta - \delta/3)*10000),$$

where $I$ is the measured intensity, $I_0$ is the initial intensity, $D$ is calculated diffusion coefficient, $\gamma$ is the gyromagnetic ratio for the observed nucleic $^1\text{H}$ (4258 Hz/G), $G$ is the gradient strength.
(G/cm), \( \delta \) is the gradient length (ms) and \( \Delta \) is the delay between gradients (ms). The diffusion coefficient is proportional to the temperature and inversely proportional to the frictional coefficient:

\[
D = \frac{(k_b T)}{f},
\]

where \( k_b \) is the Boltzmann constant, \( 1.382 \times 10^{-23} \text{ m}^2 \text{kg} \text{s}^{-2} \text{K}^{-1} \), \( T \) is the temperature 298 K, and \( f \) is the frictional coefficient. The frictional coefficient can be estimated for a sphere:

\[
f = 6 \pi \eta r,
\]

where \( \eta \) is the viscosity (1.2514 cp for D\(_2\)O), and \( r \) is the radius of the sphere.

For free protein, which is roughly globular, a measured diffusion of \( 9.1 \times 10^{-11} \text{ m}^2/\text{s} \) is observed corresponding to a radius of 19.1 Å assuming a perfect sphere. This is consistent with the published topography for PU.1 (33 x 34 x 38 Å\(^3\)), Figure 3.14 (17). This measured diffusion coefficient is also consistent with a calculated diffusion coefficient based on the hydrodynamic radius for the protein based off the crystal structure of the 1:1 complex, which yields a value of \( 8.8 \times 10^{-11} \text{ m}^2/\text{s} \) (5,29). A dimer, modeled as two rigid spheres would have a diffusion coefficient of roughly 75% of that of a monomer:

\[
\frac{D_{\text{dimer}}}{D_{\text{monomer}}} = \frac{f_{\text{monomer}}}{f_{\text{dimer}}} \approx \frac{f_{\text{sphere}}}{f_{\text{prolate ellipsoid}}},
\]

\[
f_{\text{prolate ellipsoid}} = 6 \pi \eta b[(a/b)^2-1]^{1/2}/\ln[(a/b)+((a/b)^2-1)^{1/2}].
\]

Assuming that this dimer consists of two identical spheres, \( a = 2b \):

\[
f_{\text{prolate ellipsoid}} \approx 6 \pi \eta b[1.73/\ln(3.73)],
\]

\[
D_{\text{dimer}}/D_{\text{monomer}} \approx 75%.
\]

Based off the measured diffusion of the monomeric free PU.1, a theoretical dimer would have a diffusion of \( 6.8 \times 10^{-11} \text{ m}^2/\text{s} \). The measured diffusion for the 2:1 (PU.1:16 base pair high-affinity DNA) was \( 5.9 \times 10^{-11} \text{ m}^2/\text{s} \), which is reasonable for a 2:1 complex of dimeric protein with
additional 16 base pair DNA. Additionally, the measured diffusion coefficient for the 1:1 complex of PU.1 with 16 base pair site-specific DNA was $7.5 \times 10^{-11} \text{ m}^2/\text{s}$. This is consistent with a predicted diffusion value of $7.3 \times 10^{-11} \text{ m}^2/\text{s}$ for the 1:1 complex based off the crystal structure (5).

3.9.3 Simulating methyl peak exchanges

The far upfield shifting, characteristic of folded proteins, is due to ring current, which is an induced delocalization of $\pi$ electrons of an aromatic ring. These delocalized electrons are free to circulate, therefore responding more to a magnetic field. Placement of a proton in a folded protein can be such that the proton is highly shielded or highly deshielded due to the aromatic ring. Unfolded protein, or partial unfolding where the highly shielded (negative chemical shift) proton has shifted from shielding effects of the aromatic ring can result in the methyl proton shifting in a different region of the spectrum. Alternatively, chemical exchange can broaden signals. Coalescence is chemical shift and field strength dependent:

$$k_c = (\pi \Delta \nu) / (2)^{1/2},$$

where $\Delta \nu$ is the difference between two peaks in Hz. Assuming a two site chemical exchange at 600 MHz, coalescence was calculated for each set of methyl peaks. The first set of methyl peaks, where the chemical shifts of the free and bound states are -0.214 and -0.155 have a calculated coalescence of 65 Hz. The second set of peaks, free and bound states are -0.394 and -0.321 have a calculated coalescence of 81 Hz. The final set of peaks, free and bound states at -0.714 and -0.689 have a calculated coalescence point at 28 Hz. Envelopes which fit the chemical shift and peak shapes of each set of methyl peaks was built in DNMR (Topspin 2.1) along with DSS (used as an internal standard for experimental and simulated peaks). Assuming only a two state chemical exchange for each individual set, various exchange rates, including coalescence points,
were used in an attempt to recreate the broad methyl peaks observed at 2:1 ratio, from 0 Hz (no exchange) to 10,000 Hz (very fast exchange).
4 NMR METHODOLOGY APPLIED TO UNIQUE SYSTEMS: ARGinine PK$_A$ AND BASE PAIR OPENING OF INOSINE V. GUANOSINE

4.1 Preface

The two distinct projects described in this chapter were both aided by NMR to shed a unique perspective into their respective systems. The first mini-project follows the impact of methylation on the charge state of the side-chain of arginine (1). This project was initially conceptualized by Dr. Markus Germann and Dr. George Y. Zheng. The methylated arginine variants for the $^1$H study were synthesized and purified by his graduate students Diem Tran and Brandon Canup. The second mini-project follows the effect of an inosine base in place of a guanosine base on the base pair dynamics in a DNA hairpin. The structure of the guanosine hairpin was solved by Dr. Christopher Johnson and Dr. Subrata Mishra. NMR experiments and data analysis for the inosine containing hairpin were done in collaboration with Qiushi (Shauna) Li.

4.2 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is a powerful tool that provides a unique perspective into systems on a molecular level. The global and local structure and dynamics of nucleic acids and proteins can be probed by NMR; purity of molecules to subtle changes in torsion angles to application of translational motion for gauging complex size are all achievable via various pulse programs. Global and local structure and subtle dynamics are investigated in further detail in the previous chapter, Chapter 2. Similarly, Chapter 3 contains detail on the application of diffusion measurements for protein-DNA complexes. This chapter contains additional specific NMR methodologies to probe characteristics of two distinct systems: a small amino acid and a DNA hairpin.
4.2.1 Using NMR to measure the pK\textsubscript{a} of the side-chain of methylated arginine

Protein interactions are often governed by electrostatics. Charged residues are particularly common on protein surfaces and in catalytic active sites. However, ionizable groups have also been found in the hydrophobic interior of proteins. Arginine is one of the most frequently buried charged residues (2-4). The side-chain of arginine contains a guanidino head group, which retains a positive charge, even at elevated pH. Additionally, this residue contains five favorably positioned hydrogen bond donors, all of which can be methylated by a family of proteins, protein arginine methyltransferases (PRMT) (5). PRMTs catalyzes the formation of three distinct modified arginine variants: \(\omega \) NG monomethylarginine (MMA), \(\omega \) NG,NG asymmetric dimethylarginine (aDMA) and \(\omega \) NG,N’G symmetric dimethylarginine (sDMA), Figure 4.1 (5). The methylation of arginine is biologically relevant with as much as 2\% of all arginine residues methylated in rat liver nuclei (6). Deregulation of PRMTs has been linked to cancer (7), and so the impact of methylation on the characteristic traits of arginine is key in elucidating the interactions between methylated arginine and its ligands.

The process of methylation can modulate the characteristics of an amino acid side chain. This process introduces a bulky moiety which can sterically hinder the binding of a ligand, as well as impart a hydrophobicity to the residue. The methylation and subsequent removal of potential hydrogen bond donors disrupts the network of hydrogen bonds, potentially inhibiting binding (5). However, the effect of methylation on the pK\textsubscript{a} of arginine was unknown. The charge of an amino acid is governed by pK\textsubscript{a} and depends on pH environment. This characteristic can be modulated by environment; the pK\textsubscript{a} of internal residues, in less polar microenvironments can shift in the direction favoring the neutral species (8). Internal lysine residues have observable
pK_a shifts from 10.4 to 5 (9). Arginine, however, retains its charge even at pH 10; arginine is essentially always charged (10).

The pK_a of arginine has been difficult to determine with certainty. Although a commonly accepted pK_a of 12.48 has been reported in journals and texts since its publication in 1930, studies of arginine pK_a range from 11.5-15.0 (11-14). Traditionally, potentiometric methods are suitable for pK_a determination; however, readings at very high and very low pH levels are unreliable (15). The use of nuclear magnetic resonance (NMR) for pK_a determination is a method for residue specific pK_a determination at extreme ranges (16,17). During protonation/deprotonation of the amino acid, the electron distribution around the nuclei is affected, allowing for the use of NMR to observe the changing chemical shifts of side-chain protons. The changing chemical shifts with the respective pH values can be fit with a modified Henderson-Hasselbalch equation, adapted from a pK_a study in 2007 (17), to characterize pK_a values of methylated arginine variants to identify the contribution of methylation on interactions:

$$
\delta_{\text{obs}} = \chi_{\text{HA}} \delta_{\text{HA}} + \chi_{\text{A-}} \delta_{\text{A-}},
$$

$$
\delta_{\text{obs}} = (\delta_{\text{HA}} + \delta_{\text{A-}} 10^{\text{pH-pK_a}}) / (1 + 10^{\text{pH-pK_a}}),
$$

where the $\delta_{\text{obs}}$ is the observed chemical shift, $\delta_{\text{HA}}$ and $\delta_{\text{A-}}$ are the chemical shifts of the protonated and unprotonated species respectively, and $\chi$ is the mole fraction (1). The full derivation can be found in the Supplementary Information.

### 4.2.2 Using NMR to gauge local base pair dynamics of an inosine for guanosine substitution in a DNA hairpin

Inosine, one of the most common non-canonical nucleotides in DNA, is a nucleobase which is introduced into DNA or RNA through both endogenous and exogenous processes. In cells, inosine triphosphate is hydrolyzed to monophosphate to avoid incorporation into DNA, as
it can lead to miscoding (18). Regardless, incorporations in DNA occur, as adenosine is deaminated to inosine. In contrast, this nucleobase is a normal modification created by deaminases in RNA to introduce transcriptome diversity (18). Alternatively, this base can be used as a modification to investigate enzyme selectivity.

HIV integrase, a retroviral integrase, is an integral protein in the life cycle of the virus, catalyzing the integration of linear viral DNA (from the conversion of single-stranded HIV RNA via reverse transcriptase) into the host genome (19). Based off a crystal structure of full-length prototype foamy virus integrase (PFV-IN), a helix is inserted between two DNA strands, disrupting the 3’ terminal DNA base pairs (20). The DNA substrates for this enzyme contain a conserved CA dinucleotide motif opposite GT nucleotides. Studies in our lab confirm that it is not the local DNA structure of the CA motif but contacts with the exocyclic amino group of the conserved guanosine which is crucial for integrase binding, as replacing guanosine with inosine abolishes the 3’ processing (Qiushi Li, M.S. thesis).

Chemically, inosine and guanosine are identical, with the exception of the amino group, Figure 4.2. Structurally, the impact of the substitution is minimal, as both DNA hairpins are quite similar (Qiushi Li, M.S. thesis). Inosine also preferentially base pairs with cytosine, albeit with fewer hydrogen bonds. Thermally, this translates to a difference of ~ 10°C (T_m values of 333.8 K for G4-hairpin compared to 322.7 K for I4-hairpin) (Qiushi Li, M.S. thesis). Base pairing is crucial for nucleic acid structure and interactions. The kinetics of base pair opening can provide additional detail on the dynamics of nucleic acids. NMR can be used to measure the base pair opening rates (21). The exchange occurs when the base pair is in the open state. The lifetime of a base pair can be determined by measuring the T_1 longitudinal relaxation of each imino proton per base pair with the addition of catalyst (ammonia) to determine the exchange rate, k_{ex} or \(1/\tau_{ex}\):
\[
\frac{1}{T_1} = \frac{1}{\tau_{\text{ex}}} + \frac{1}{T_1^0},
\]
where \(T_1^0\) is the longitudinal relaxation time dependent on mechanisms other than exchange, ie, dipolar interactions (22). The longitudinal relaxation time, \(T_1\), is measured as a function of ammonia catalyst, and the base pair lifetime can be determined at infinite catalyst with a plot of \(1/\text{[catalyst]}\) vs \(\tau_{\text{ex}}\):

\[
\tau_{\text{ex}} = \tau_{\text{open}} + \frac{1}{(K_d\alpha\text{[catalyst]}k_i)},
\]
where \(K_d\) is the equilibrium constant for the fraction open, \(\text{[catalyst]}\) is the concentration of ammonia catalyst, \(\alpha\) is a correction factor for the accessibility difference of imino protons between the mononucleotide and open state of an oligonucleotide (value ~ 1), and \(k_i\) is the imino proton transfer rate or collision rate constant for ammonia depending on imino protons and ammonia, which has a value of \(2 \times 10^8 \text{ s}^{-1}\text{M}^{-1}\) (22).

### 4.3 Materials and methods

#### 4.3.1 Arginine pK\(_a\) sample preparation

The arginine variants for the \(^1\text{H}\) NMR titrations were synthesized by Diem Tram in Dr. George Y. Zheng’s lab. Briefly, solid phase synthesis using rink amide resins, with acetic anhydride used for capping was used to prepare arginine variants. Following the TFA cleavage, the crude products were purified by reverse phase HPLC and the products were confirmed through ESI-MS. The protected control arginine variant with amidation at the C-terminus and acetylation at the N-terminus for the \(^{13}\text{C}\) NMR experiment was purchased from Genscript. NMR samples were prepared in 90%/10% H\(_2\)O/D\(_2\)O at varying calculated pH and constant ionic strength 1.0 M adjusted with KOH/KCl; additional details and sample calculations can be found in the Supplementary Information (23). The pH was confirmed with a pH meter suitable for high pH measurements (24). Arginine concentrations for the \(^1\text{H}\) NMR titrations were 0.5 mM and 50
mM for the $^{13}$C natural abundance titrations. All NMR spectra were obtained on Bruker Avance 600 and 500 MHz NMR instruments equipped with 5 mm QXI and 5 mm TBI probes, respectively for the $^1$H titrations and a 10 mm BBO probe for the $^{13}$C natural abundance titrations. Spectra were recorded at 25 °C using a 1D presaturation $^1$H NMR pulse sequence. $^1$H chemical shifts were referenced to internal DSS at low pH, then the $^1$H$_{\alpha}$ chemical shift at high pH, since DSS showed inconsistencies at extreme pH. $^{13}$C natural abundance spectra were recorded using inverse gated $^1$H decoupling using a 45 degree flip angle. $^{13}$C chemical shifts were referenced to a CDCl$_3$ capillary. At high pH $>$11, degradation was observed at the C-terminal end of the protected arginine variants; spectra were obtained immediately after sample preparation and data from the degraded products were not used (1). Characterization of the degradation products can be found in the Supplementary Information.

Because of incomplete titration curves at pH $>$14, chemical shifts of protonated and unprotonated species of the arginine variants and imidazole control were calculated from Spartan 10 using a Hartree-Fock 6-31G* basis set. Electrostatic potential maps, hydrophobicity and volume were also calculated using Spartan 10 and Hartree-Fock 6-31G*. All modified variants were built by truncating arginine at the $\alpha/\beta$ carbons, to focus on the ionizable side chains. Each variant and tautomer was made and individually minimized prior to calculation. The lowest energy tautomer, where applicable, was reported (1).

**4.3.2 Inosine for guanosine base substitution for base pair opening rates**

The DNA hairpins were purchased from Integrated DNA Technologies, desalted using Hi-Trap desalting columns on FPLC and purity was gauged by gel electrophoresis using UV-shadowing (8 M urea, 15% polyacrylamide denaturing gels).
NMR spectra of the imino protons of the DNA hairpins was measured for temperatures ranging from 283 to 308 K using a 1-1 jump and return NMR pulse sequence (25). Imino protons were assigned from a 2D NOESY experiment (150 ms mixing time) and 1D NOE experiments. The $T_1$ inversion recovery experiments were performed at two temperatures: 278 and 293 K. These temperatures were selected on the basis of good chemical shift separation of the imino protons. Hairpin concentrations were 1 mM in 50 mM NaCl, 10 mM sodium phosphate, 0.1 mM EDTA at pH 7.8 with 10% D$_2$O. The pH was measured and adjusted, if necessary, with each catalyst titration. To prepare the ammonia catalyst, 6 M NH$_4$OH was prepared and the ammonia concentration was calculated from the equation (22):

$$[\text{NH}_3] = [\text{NH}_4\text{OH}] \times 10^{-\text{pKa}} / (10^{-\text{pH}} + 10^{-\text{pKa}}).$$

The control hairpin (with a guanosine at position 4), was titrated with catalyst concentrations (mM): 2, 2.5, 3.5, 5, 10, 20, 40 and 100. The inosine hairpin (with inosine at position 4), was titrated with catalyst concentrations (mM): 5, 10, 15, 30 and 60.

4.4 Results

4.4.1 Methylated arginine $pK_a$ study results

To investigate the effect of methylation, $^1$H NMR spectra of four arginine variants (Figure 4.1) were observed at different pH values. The calculated pH of each sample was prepared using concentrated KOH and then confirmed using a glass pH electrode appropriate for high pH. Ligand concentration and volume were kept constant to minimize systematic errors (23). In addition, the ionic strength was kept at 1.0 M for each NMR sample with KOH/KCl. This electrolyte pair was selected on the basis of having less self-association, compared to NaOH, and an overall higher pK$_w$ with the use of KOH (23, 24). The titration was conveniently monitored from the chemical shift of the delta protons of arginine near 3.25 ppm (Figures 4.3
and 4.4) (1). These protons are in close proximity to the charged guanidinium group and experience no overlap in that region of the $^1$H NMR spectra. In addition, the chemical shifts of the methyl peaks were also monitored for MMA, sDMA, and aDMA, Figures 4.5 and 4.6. The numerical pK$_a$ values were estimated using a Henderson-Hasselbalch equation modified for NMR titrations and adapted from a 2007 pKa study, Table 4.1 and Supplementary Information (26).

The $^1$H NMR titration curves show that all arginine variants behave similarly and that the titrations do not reach completion even at the highest pH, a clear indication of a pK$_a > 12.5$, Figure 4.5. Two control compounds with previously published high pK$_a$ values were also measured to demonstrate the validity of the method and reagents; imidazole, with a pK$_a$ of 14.5 (27) and trifluoroethanol, with a pK$_a$ near 12.5 (28) were prepared as previously described. A complete titration curve that did not require the estimation of the chemical shift of the unprotonated species was obtained for trifluoroethanol resulting in a pK$_a$ of 12.5 in agreement with the published data, Table 4.1 and Figure 4.7. Because the titration could not be completed due to pH limitations in aqueous solutions and degradation of the protected arginine (Supplementary Information), the chemical shifts of the unprotonated species of imidazole and the arginine variants could not be directly obtained from the curves and were therefore computationally assessed for a range of chemical shifts (Figures 4.6 and 4.7). The chemical shifts of protonated and unprotonated species were obtained from Spartan 10 using a Hartree-Fock 6-31G* basis set. For imidazole, an average chemical shift difference of 0.6 ppm was found while the difference was 0.5 ppm for the arginine variants (black lines). A previous publication reported a difference between protonated and unprotonated species of 0.2 ppm (29). We therefore calculated pK$_a$ values for 0.2-0.6 ppm to assess the impact of uncertainty in the
chemical shift difference. Note that even if the unprotonated species had unrealistically low chemical shifts difference (~0.1 ppm) that the pKₐ would still be > 13.5. For imidazole the pKₐ was estimated at 14.6 ± 0.3, in reasonable agreement with literature value of 14.5 for the second deprotonation of neutral imidazole to its anion form; note, imidazole has two pKₐ values, 7.0 and 14.5. The unmethylated arginine has a pKₐ of 14.2, as compared to the 14 pKₐ of arginine in a tripeptide from published NMR titrations (29). MMA, sDMA and aDMA had pKₐ values from delta protons of 14.3, 14.7 and 14.3, respectively (Table 4.1). For the methylated arginine derivatives, these values were further supported from the chemical shift changes of the methyl protons (Figure 4.6). Because of the unconventionally elevated pKₐ values, natural abundance ¹³C chemical shifts as a function of pH for the protected arginine control were also measured (Figure 4.8). The pKₐ estimated using a Δδ of 4 ppm for ¹³C₅ and 1 ppm for Δδ ¹³Cγ carbons using previously published ranges gave pKₐ values of 13.8 and 14.1 respectively, in agreement with the previously published arginine pKₐ in a tripeptide (14).

Since methylation does not markedly change the pKₐ of the guanidinium head group of arginine, altered molecular interaction with methylated arginines are not likely due to differences in charge. Methylation affects the hydrogen bonding properties, and thus protein-protein, peptide lipid and protein-ligand interactions, as shown by previous studies (30). Methylation also alters steric factors, hydrophobicity and charge distribution of the guanidinium head group (Figure 4.9). To further evaluate some of these properties, Spartan was used to calculate the alteration of the electrostatic potentials, volume and hydrophobicity in terms of logP. Tautomers for each variant were constructed and individually minimized. The introduction of a bulky moiety affects the overall volume of the head group, introducing steric hindrance; in the model constructs, this is represented by an increase of 17%, 33% and 34% for MMA, aDMA and sDMA with respect
to the control. Both the unmethylated and monomethylated models have similar charge
distribution. In comparison, both dimethylated constructs show electrostatic potential maps with
different and more diffuse charge localization. The charge distribution in each structure is
affected by methylation, which may result in altered interactions with proteins or nucleic acids.
The hydrophobicity of the molecules, described by the term logP, increases significantly from a
negative value of -0.34, as expected for a charged molecule in the control, to positive values of
0.18, 0.56 and 0.70 for the MMA, aDMA and sDMA model constructs (1).

4.4.2 Base pair opening results for a DNA hairpin containing an inosine substitution

The imino protons NMR spectra of the inosine and guanosine hairpins was similar, with
the exception of the inosine chemical shift, as expected, Figure 4.11. The imino proton
associated with inosine is a sharp resonance, indicative of a relatively stable base pair. Other
imino protons between the two hairpins are relatively similar, suggesting that the inosine in place
of a guanosine does not grossly affect the hairpin structure (Qiushi Li, M.S. thesis).

The T1 relaxation rates were measured for all possible imino protons for the guanosine-
containing (Figure 4.12) and inosine-containing (Figure 4.13) DNA hairpins. Many of the imino
protons relaxed too quickly to be measured after the first few additions of ammonia catalyst.
After a series of catalyst additions, the base pair opening lifetime, \( \tau_{\text{open}} \), was determined from a
plot at infinite catalyst addition, where \( 1/[\text{catalyst}] = 0 \). A comparison of the plots reveals that the
inosine base has a significantly shorter lifetime compared to a guanosine in the same hairpin
sequence. At 278 K, this is represented by 7 ms compared to 150 ms for the inosine and
guanosine bases respectively, Figure 4.14. To a lesser extent, G14 in each hairpin, a neighboring
base pair to the I4 or G4 base pair is also affected, exemplified by opening rates of 68 ms
compared to 100 ms for the inosine-containing and guanosine-containing hairpins, respectively. This trend is echoed at 293 K, Table 4.2.

4.5 Discussion and conclusion

4.5.1 The effect of methylation on the pK\textsubscript{a} of arginine

An understanding of the charge state of arginine and methylated arginine derivatives is crucial in understanding the mechanisms and functions of arginine and its derivatives in biological systems. NMR titrations of arginine reveal that the pK\textsubscript{a} values of free arginine and methylated arginine derivatives in water near 14 (1). This implies that in a biological environment the arginine side chain is fully charged regardless of the methylation state. Arginine’s characteristically high pK\textsubscript{a} explains its ability to remain charged even in highly nonpolar microenvironments at high pH. As a 2011 study has suggested, arginine is essentially always charged in biological systems (10). Its high propensity for charge and possibly due to its unique size and charge distribution allows arginine to remain an important residue with many functions (31). The placement of the methyl groups on the nitrogen groups does not have an appreciable impact on the pK\textsubscript{a}, with the possible exception of sDMA; however, the pK\textsubscript{a} differences for sDMA and aDMA are less than 0.5 pH units. This finding is supported by a report where guanidine and methylated guanidine derivatives were all found to have similar pK\textsubscript{a} values by potentiometric titrations of around 13.6-13.8 (32) and by a similar NMR study to probe the pK\textsubscript{a} of methylated lysine residues on Ca\textsuperscript{2+}-bound and apo calmodulin where variation in pK\textsubscript{a} between dimethylated and monomethylated lysine species were reported to be <0.8 pH units (33).

Although an extensively quoted value of 12.48 has been found in literature and textbooks since 1930, published arginine pK\textsubscript{a} results have ranged from 11.5-14 (11,12,14,28,29,34-42).
The experimental determination of arginine’s third pK\textsubscript{a} has been notoriously difficult, with few published studies citing measured pK\textsubscript{a} values using alternative methods until very recently. A 2007 study attempted to determine pK\textsubscript{a} of lysine and arginine on apo calmodulin using NMR titrations; they reported only slight chemical shift perturbations for arginine at pH 12.5, indicating that the pK\textsubscript{a} has not yet been reached (26). Also, several groups have recently published simulated or predicted arginine or guanidine variants pK\textsubscript{a} values with mixed results ranging from 12.5-16.1 (43-47).

Most recently, the pK\textsubscript{a} of free arginine using potentiometry and arginine using \textsuperscript{1}H, \textsuperscript{13}C and \textsuperscript{15}N NMR titrations was published at \geq 13.8 (29). Although this value is somewhat lower than the one reported here, it was stated that the reported value of 13.8 could be an underestimation. Also, the endpoint of our variants could not be reached due to degradation of the backbone mimic, we therefore estimated the chemical shift of the unprotonated species and we report an average pK\textsubscript{a} from a range of Δδ 0.2-0.6 ppm (1).

\subsection{4.5.2 The effect of an inosine substitution in place of a guanosine on local base pair dynamics}

Structurally the two DNA hairpins are quite similar by NMR (Qiushi Li, M.S. thesis), however the dynamics are quite different. Increased dynamics would be expected to aid in enzyme recognition; however the substitution abolishes 3’ processing of the HIV integrase protein, suggesting a crucial contact between the enzyme and exocyclic amino group (Qiushi Li, M.S. thesis). The base pair opening rate of an inosine-containing DNA hairpin is drastically different than a guanosine-containing hairpin. Although there is a decrease in the number of hydrogen bonds and T\textsubscript{m} for DNA with inosine substituted for guanosine (Qiushi Li, M.S. thesis), the significant difference in the base pair opening of the bases is still surprising. This finding is
supported by previous work where the substitution of inosine for guanosine has a significant effect on the base pair opening rate, although the effect on the base pair opening rates is highly sensitive to duplex conformation and structure (48). Substantial differences in stability, judged by $T_m$, are largely attributed to the difference in the number of hydrogen bonds; however, the 2-amino group, or lack-there-of, is also believed to play a role in the ordering of water molecules in the minor groove (48).

### 4.6 References


40. Greenstein, J. P. (1933) Studies of the peptides of trivalent amino acids: III. The apparent dissociation constants, free energy changes, and heats of ionization of peptides involving arginine, histidine, lysine, tyrosine, and aspartic and glutamic acids, and the behavior of lysine peptides toward nitrous acid. *Journal of Biological Chemistry* 101, 603-621.


4.7 Supplementary information

4.7.1 The Henderson-Hasselbalch equation modified for NMR pK_a titrations and sample calculations

The pK_a can be determined by monitoring chemical shift of a resonance upon protonation/deprotonation and fit to an equation based on the Henderson-Hasselbalch equation. Note, this equation is derived based on an equation from a publication by Andrè, *et al.* (17):

\[
\delta_{\text{obs}} = \chi_{\text{HA}} \delta_{\text{HA}} + \chi_{\text{A}} - \delta_{\text{A}},
\]

\[
\delta_{\text{obs}} = (1 - \chi_{\text{HA}}) \delta_{\text{A}} - \chi_{\text{HA}} \delta_{\text{HA}},
\]

\[
\delta_{\text{obs}} = \delta_{\text{A}} - \chi_{\text{HA}} \delta_{\text{A}} + \chi_{\text{HA}} \delta_{\text{HA}},
\]

\[
\delta_{\text{obs}} - \delta_{\text{A}} = \chi_{\text{HA}} (- \delta_{\text{A}} + \delta_{\text{HA}}),
\]

\[
(\delta_{\text{obs}} - \delta_{\text{A}}) / (\delta_{\text{HA}} - \delta_{\text{A}}) = \chi_{\text{HA}},
\]

\[
(\delta_{\text{obs}} - \delta_{\text{HA}}) / (\delta_{\text{A}} - \delta_{\text{HA}}) = \chi_{\text{A}},
\]

\[
\text{pH} - \text{pK}_a = \log [\text{A}^-] / [\text{HA}],
\]

\[
10^{\text{pH}-\text{pK}_a} = [\text{A}^-] / [\text{HA}],
\]

\[
10^{\text{pH}-\text{pK}_a} = [ (\delta_{\text{obs}} - \delta_{\text{HA}}) / (\delta_{\text{A}} - \delta_{\text{HA}})] [ -1 (\delta_{\text{HA}} - \delta_{\text{A}}) / (\delta_{\text{obs}} - \delta_{\text{A}}) ],
\]
10^{pH-pK_a} = (\delta_{\text{obs}} - \delta_{HA}) / (-\delta_{\text{obs}} + \delta_{A-}),

-\delta_{\text{obs}} (10^{pH-pK_a}) + \delta_{A-} (10^{pH-pK_a}) = \delta_{\text{obs}} - \delta_{HA},

-\delta_{\text{obs}} (10^{pH-pK_a}) - \delta_{\text{obs}} = (-\delta_{HA}) - \delta_{A-} (10^{pH-pK_a}),

-\delta_{\text{obs}} (10^{pH-pK_a} - 1) = (-\delta_{HA}) - \delta_{A-} (10^{pH-pK_a}),

\delta_{\text{obs}} = \delta_{HA} - \delta_{A-} (10^{pH-pK_a}) / (10^{pH-pK_a} - 1),

\delta_{\text{obs}} = (\delta_{HA} + \delta_{A-} 10^{pH-pK_a}) / (1 + 10^{pH-pK_a}).

Each NMR sample was kept at a consistent arginine concentration, ionic strength (1 M) and volume (500 μL). Sample calculation is shown below for two samples made for pH 13.8 based off a pK_w of 14.14 for KOH:

\begin{align*}
\text{pH} + \text{pOH} &= 14.14, \\
\text{pOH} &= 0.34 = -\log [\text{OH}^-], \\
[\text{OH}^-] &= 0.457 \text{ M}.
\end{align*}

To make a 0.457 M [OH\textsuperscript-] sample, 182.8 μL of 1.25 M KOH stock was used. To make this sample ionic strength 1 M, 135.8 μL of 2 M KCl was added. To this, arginine from a stock was added for a final arginine concentration of 0.5 mM (for \textsuperscript{1}H NMR) and 50 mM (for \textsuperscript{13}C NMR), along with 10% D\textsuperscript{2}O and additional H\textsubscript{2}O for a final volume of 500 μL. The pH was then confirmed by a micro pH meter suitable for high pH.

4.7.2 Degradation products observed for arginine with backbone mimics at N- and C-termini

At high pH, pH >12, degradation was evident from the \textsuperscript{1}H NMR spectra. The degradation products were carefully avoided and chemical shifts were not used in the pK_a calculations. As the pH increased, the arginine degraded in a shorter time, thus limiting the NMR experimental setup and run to less than 20 minutes at pH >14. After 24+ hours, the arginine variants were
completely degraded. Of note, degradation of unprotected arginine, was not observed, indicating that the flanking groups of the CA, MMA, sDMA, and aDMA were removed at high pH. These flanking groups, intended to mimic the environment of a peptide chain, consist of an additional amine at the C-terminal end and an acetyl group at the N-terminal end. To potentially identify the degraded product, a small aliquot of acetic acid (0.5 mM) was added to a previously prepared and degraded arginine sample at pH 11.62, Figure 4.10. The $^1$H NMR resonance around 2 ppm, characteristic of acetyl methyl peaks at the N-terminal end showed two peaks suggesting degradation; the two distinct singlet peaks are observed for two acetyl methyl peaks in different chemical environments. Integration of these peaks showed a degradation of around 25% and around 75% native amino acid. This integration was consistent with integrations for other resonances showing degradation. Assuming the degradation was at the N-terminal end, the degraded acetyl methyl peak should be identical to a singlet methyl peak corresponding to an acetate ion at the same pH. After the addition of acetic acid, a different singlet peak was observed, suggesting the absence of acetate in the original sample and further suggesting the degradation was at the other end of the protection group. Mass Spectrometry of the degraded product yielded inconclusive results. The visibility of a degraded product is an interesting benefit of using NMR for titration and pK$_a$ determination; other general methods, such as using a pH meter in standard potentiometric titrations could have possibly masked the presence of the degraded product at such high pH and interfered with measurement of the native product.

4.7.3 Base pair opening lifetimes determination using inversion recovery NMR

A T$_1$ inversion recovery experiment was used to determine the base pair opening lifetimes of base pairs in a DNA hairpin containing a guanosine or inosine at position 4. Briefly, a T$_1$ inversion recovery experiment consists of an inversion ($180^\circ$ pulse), delay, then excitation
(90° pulse), where delays are varied. The variation of the T₁ inversion recovery used was the Selective Inversion Recovery (Bruker pulse sequence selzgIR) (49). In this experiment, the selective pulses were I-BURP (Inversion) and E-BURP (Excitation) Band-selective Uniform-Response Pure-phase shaped pulses, selectively exciting only the imino protons. Example power levels for the I-BURP and E-BURP shaped pulses (1000 μs) on a 600 MHz NMR, based off a ¹H 90° pulse of 9.45 μs at 1 db power are 18.97 and 18.02 db, respectively. A series of delays was used (ms): 0.01, 0.1, 5, 10, 25, 50, 75, 150, 250, 500, 750, 1000, 1500, 3000. Intensities of each peak were measured and plotted as a function of delay (s). Normalized intensities were fit to the following equation to determine the T₁ longitudinal relaxation for each base pair, where τ is the delay:

\[ \frac{I(\tau)}{I(0)} = 1 - 2 \cdot \exp\left(-\tau/T_1\right). \]

The T₁ measured at 0 catalyst represents the T₁⁰, or the longitudinal relaxation depending on other factors, such as dipolar relaxation. After determining the T₁⁰ for each base pair at each temperature, catalyst (ammonia) was added in small quantities. T₁ measurements were taken at each titration for each imino proton (representing a single base pair). The exchange time, τₑₓ, was determined for each base pair and plotted as function of 1/[catalyst] to determine the base pair opening lifetime, τₒₚₑₜₐ at infinite catalyst.
Table 4.1 Extrapolated pK\textsubscript{a} values for the modified arginine variants and control compounds imidazole and trifluoroethanol (TFE). Values are based on the \textsuperscript{1}H NMR chemical shift data as a function of pH (1).

<table>
<thead>
<tr>
<th>Titrant</th>
<th>pK\textsubscript{a}</th>
<th>H\textsubscript{6}</th>
<th>H\textsubscript{met}</th>
<th>H\textsubscript{2,4/5}</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arginine</td>
<td>14.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MMA</td>
<td>14.3</td>
<td>14.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sDMA</td>
<td>14.7</td>
<td>14.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>aDMA</td>
<td>14.3</td>
<td>14.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Imidazole</td>
<td>-</td>
<td>-</td>
<td>14.6, 14.8\textsuperscript{a}</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TFE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>12.5</td>
<td>-</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Error of ±0.3 imidazole and ±0.4 for each arginine variant was estimated from a model fit.

\textsuperscript{a}Chemical shifts were monitored for H\textsubscript{2} and H\textsubscript{4}/H\textsubscript{5} of imidazole, yielding pK\textsubscript{a} values of 14.6 and 14.8, respectively.
Table 4.2 Base pair opening lifetimes for inosine and guanosine substitutions in a DNA hairpin. Rates were calculated in collaboration with Qiushi (Shauna) Li.

<table>
<thead>
<tr>
<th>Imino $^1$H (substrate)</th>
<th>Temperature (K)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>278</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>G4 (guanosine hairpin)</td>
<td>155 ms</td>
<td>19 ms</td>
<td></td>
</tr>
<tr>
<td>I4 (inosine hairpin)</td>
<td>7 ms</td>
<td>3 ms</td>
<td></td>
</tr>
<tr>
<td>G14 (guanosine hairpin)</td>
<td>100 ms</td>
<td>31 ms</td>
<td></td>
</tr>
<tr>
<td>G14 (inosine hairpin)</td>
<td>68 ms</td>
<td>14 ms</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.1 Arginine and methylated arginine variants. N- and C- termini contain protection groups to mimic a peptide environment.
Figure 4.2 Base pairing with guanosine and cytosine (A) found in the guanosine-containing DNA hairpin (B), in contrast to the inosine and cytosine base pair (C) found in the inosine-containing DNA hairpin (D).
Figure 4.3 $^1$H NMR spectrum of sDMA at pH 13.14. The delta protons (3.2 ppm) were monitored during titrations. Methyl peak (3 ppm) from the guanidino methylation. The methyl peak (2 ppm) is from the N-terminus of the residue.
Figure 4.4 600 MHz $^1$H NMR of the delta protons of 0.5 mM MMA at 25°C from pH 13.6-11.1 (top to bottom).
Figure 4.5 NMR pH titration of aDMA (black), sDMA (red), MMA (green) and control arginine (blue) (1).
Figure 4.6 Titration curves for (A-B) aDMA, (C-D) sDMA and (E-F) MMA. Experimental data are represented by data points, closed for delta proton chemical shifts and open for methyl proton chemical shifts.

Solid black lines are for estimated unprotonated species (0.4 ppm difference from the protonated species), dotted lines are for estimated unprotonated species (0.2 ppm difference from the protonated species) and dashed lines are for unprotonated species (0.6 ppm difference from the protonated species) (1).
Figure 4.7 Titration curves for (A) control compounds and (B) the control arginine.

(A) Monitoring the chemical shift and pH for $H_2$ of imidazole (top) and trifluoroethanol (bottom) and (B) $H_δ$ of arginine with a simulated curve for a hypothetical $pK_a$ of 12.48 (red trace). Experimental data are represented by circles for imidazole (A) and by diamonds for trifluoroethanol and (B) blue triangles for arginine. The lines are for simulated curves, assuming a difference in chemical shift between the protonated and unprotonated species of 0.2 (dotted), 0.4 (solid) and 0.6 (dashed) ppm for the arginine variants and 0.4, 0.5 and 0.6 ppm for imidazole (1).
Figure 4.8 Titration curves for the control arginine for natural abundance (A) $^{13}\text{C}_\zeta$ and (B) $^{13}\text{C}_\gamma$ chemical shifts for a 50 mM control sample.

Arginine contains an amide group at the C-terminus and acetyl at the N-terminus to mimic a peptide environment. Because degradation at the backbone was observed, data points above pH 14.04 were not recorded. The $\Delta \delta$ used were from previously reported values of 4 and 1 ppm, corresponding to $pK_a$ values of 13.8 and 14.1, respectively. Data were collected using a 10-mm BBO probe at 150.9 MHz (1).
Figure 4.9 Electrostatic potential maps of guanidinium head groups calculated using Spartan 10. The constructs shown above depict representative shortened, model structures of the four arginine variants: (A) unmethylated, (B) MMA, (C) aDMA and (D) sDMA.

The color scale for the electric potentials are shown in kJ/mol, where red represents the lowest electrostatic potential (electron rich) and blue represents the highest electrostatic potential (electron poor) regions. Tautomers for each variant were constructed and individually minimized; the lowest energy tautomer is shown (1).
Figure 4.10 $^1$H NMR spectra of a degraded sDMA at pH 11.62.

The top spectrum shows the degraded arginine variant (75% native, 25% degraded, as determined by calculating native and degraded methyl peaks around 2.03 and 2.05 ppm). The bottom spectrum shows the same sample after the addition of an acetate ion. The addition of a new peak around 1.9 suggests the degradation at the flanking group does not have the addition of an acetate ion.
Figure 4.11 Imino $^1$H NMR of guanosine-containing hairpin at 293 (A) and 278 K (C) and inosine-containing hairpin at 293 (B) and 278 K (D). Spectra were collected in collaboration with Qiushi (Shauna) Li.
Figure 4.12 Resonance intensity as a function of delay from $T_1$ inversion recovery experiments for guanosine-containing hairpin. Data for 0 mM catalyst at 278 (A) and 293 K (B) and 2 mM catalyst at 278 (C) and 293 K (D). Data was calculated in collaboration with Qiushi (Shauna) Li.
Figure 4.13 Resonance intensity as a function of delay from $T_1$ inversion recovery experiments for inosine-containing hairpin. Data for 0 mM catalyst at 278 (A) and 293 K (B) and 5 mM catalyst at 278 (C) and 293 K (D). Data was calculated in collaboration with Qiushi (Shauna) Li.
Figure 4.14 Base pair opening lifetimes for inosine-containing (red) and guanosine-containing (green) hairpins for residue 4 (A) and 14 (B) at 278 K. Data was calculated in collaboration with Qiushi (Shauna) Li.
5 CONCLUSIONS

5.1 Summaries and Future Outlook

In the previous chapters, NMR was used to study a number of biological systems. Chapter 2 focused on the structure and dynamics of a specific type of DNA damage. Through endogenous and exogenous processes, mistakes or errors are introduced into DNA, jeopardizing the genome. In Chapter 2, a subtle lesion, a single embedded ribonucleotide, which represents the most common type of DNA damage, was found to subtly alter the structure in a sequence dependent manner primarily in the sugar-phosphate backbone (1). Even though each DNA sequence studied contained a single riboguanosine at the core, the degree of structural distortion was unique based on the flanking nucleotides. Interestingly, although two of the sequences consisted of a riboguanosine flanked by a purine and pyrimidine (AxC and GxC, x=rG), the localization and degree of perturbation was different for the two duplexes. This possibility makes predictions of structural distortion and dynamics difficult and prompts the need for further studies on the localized structural impact of this lesion in additional sequence contexts.

Furthermore, the orientation of the hydroxyl groups in different sequence contexts was dependent on sequence, as the final hydroxyl group orientation was in the major groove for the rG_AxC and CxG duplexes and in the minor groove of the rG_GxC duplex. Because the primary ribonuclease of the repair pathway (RER), RNase H2, binds the minor groove of the DNA, making a number of contacts with the hydroxyl group, the efficiency of RNase H2 should be monitored to probe for a sequence effect on recognition and repair.

Chapter 3 detailed the binding of an ETS family transcription factor, PU.1, with different DNA substrates. When bound to site-specific DNA (consisting of 16 base pairs), the protein reversibly dimerizes at excess protein concentrations (μM range) (2). The structure of the
dimeric protein-DNA complex could not be determined by NMR, due to chemical exchange and the subsequent disappearance of 80% of the backbone amide crosspeaks ($^{1}$H-$^{15}$N HSQC); however, through the use of NMR, a potential protein-protein dimerization interface was revealed. Although only ~20% of the backbone amid crosspeaks remained, the identity of those crosspeaks comprised the protein core. Crosspeaks corresponding to residues at the distal surface to the DNA binding site which were involved in exchange (and disappeared) for the dimeric protein-DNA complex were implicated in dimerization. This interface was confirmed by further studies by Dr. Gregory Poon and Shingo Esaki, where dimerization was abolished after mutation of key distal residues (2). Because this dimeric protein-DNA complex is unique to PU.1 and is not shared by some other ETS family proteins, such as Ets-1 (2), this could represent a potential mechanism for auto-inhibition. Protein functional activity assays of the 2:1 (protein:DNA) complex could shed important insight into this potential mechanism.

Chapter 4 detailed two distinct projects: methylated arginine pK$_a$ and the base pair opening rate of inosine. The first project, the effect of methylation on arginine, found that methylation has little to no impact on the pK$_a$ of arginine (3). The functional impact of this modification includes other differences, such as changes in charge localization, a difference in hydrophobicity and differences in the overall size, which could introduce steric hindrance (3). In the second project, the base pair opening rates were measured for two DNA hairpins containing guanosine or inosine at the core. Structurally, this substitution had a modest effect. However, the inosine for guanosine base substitution significantly altered the base pair dynamics (rates of 7 ms compared to 155 ms for inosine and guanosine, respectively at 278 K). Even through the use of modern systems, the timescale of base pair opening rates are still outside of the scope of computational modeling. However, because this difference is quite significant, further
computational studies, possibly using accelerated molecular dynamics could prove to be useful in structural dynamics predictions.

5.2 References