MOLECULAR DYNAMICS SIMULATIONS OF RNA: PROVIDING A COMPUTATIONAL PERSPECTIVE TO AUGMENT EXPERIMENTAL DATA WHILE ADDRESSING INHERENT LIMITATIONS IN THE METHOD

by

Niel Michael Henriksen

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medicinal Chemistry

The University of Utah

December 2013

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Niel Michael Henriksen**

has been approved by the following supervisory committee members:

Thomas E. Cheatham, III	, Chair	**4/24/2013** Date Approved
Grzegorz Bulaj	, Member	**4/24/2013** Date Approved
Darrell R. Davis	, Member	**4/24/2013** Date Approved
Valeria Molinero	, Member	**4/24/2013** Date Approved
Eric W. Schmidt	, Member	**4/24/2013** Date Approved

and by **Darrell R. Davis** , Chair/Dean of

the Department/College/School of **Medicinal Chemistry**

and by David B. Keida, Dean of The Graduate School.

ABSTRACT

Advances in computer hardware have enabled routine MD simulations of systems with tens of thousands of atoms for up to microseconds (soon milliseconds). The key limiting factor in whether these simulations can advance hypothesis testing in active research is the accuracy of the force fields. In many ways, force fields for RNA are less mature than those for proteins. Yet even the current generation of force fields offers benefits to researchers as we demonstrate with our re-refinement effort on two RNA hairpins. Additionally, our simulation study of the binding of 2-aminobenzimidazole inhibitors to hepatitis C RNA offers a computational perspective on which of the two rather different published structures (one NMR, the other X-ray) is a more reasonable structure for future CADD efforts as well as which free energy methods are suited to these highly charged complexes. Finally, further effort on force field improvement is critical. We demonstrate an effective method to determine quantitative conformational population analysis of small RNAs using enhanced sampling methods. These efforts are allowing us to uncover force field pathologies and quickly test new modifications. In summary, this research serves to strengthen communication between experimental and theoretical methods in order produce mutual benefit.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Themes and Background

In a broad sense, the goal of the research presented in this dissertation is to improve techniques for the theoretical study of RNA structure and dynamics as well as to demonstrate the utility of integrating theoretical study with experimental data in order to advance hypotheses testing. The research is presented from a computational perspective and employs molecular dynamics (MD) simulations as the primary theoretical exploration environment. Although no experimental data were collected by the author, the evaluation of each study in the context of available experimental data is evident. Thus, the aim of a deeper, more reliable partnership between theoretical and experimental study of RNA becomes the theme throughout this dissertation. Ultimately, the work presented here will serve as a stepping stone towards making simulation techniques a routine part of studying biochemical systems and developing disease treatments. Currently, most RNA directed drugs are antibiotics acting on the ribosome while other RNA targets remain largely unstudied *(1)*. We anticipate the research in this dissertation will aid in developing RNA into a commonplace drug target.

It is important to justify whether MD simulations are in fact the best theoretical approach to partner with experimental studies of RNA structure and dynamics. While a variety of approaches can be used, MD simulations offer two key advantages. First, the system of interest (in this case RNA) is represented in atomic detail. Other approaches, in an effort to reduce computational expense, might rely on coarse-grained models in which groups of atoms are represented by a single particle. As the coarseness increases, the departure from a physically realistic representation of the system is also increased. In the MD approach, predictions about each atom in system can be made because they are physically represented. Of course, the MD approach is itself an approximation. Rather than explicitly representing the electronic structure of atoms through wave functions (i.e., quantum mechanics), the system is treated in classical physics approach wherein each atom is represented by a particle and the forces between particles are governed by Newton's equations of motion. This approximation leads to the second key advantage of MD simulations. Quantum mechanical calculations are extremely expensive computationally and are typically employed only to perform "snapshot" energy analysis on small systems rather than to propagate system dynamics over time. In contrast, the approximations used in MD simulations allow for the simulation of thousands to hundreds of thousands of atoms on timescales of tens of nanoseconds up to milliseconds. On these timescales, dynamic properties such as hinge-flexing, folding, and reorganization processes can be studied. The range and size of such processes is expected to increase as computer power

increases in the future. In summary, MD simulations appear to be ideally situated (given current computational resources) to provide atomic detail of biochemical system dynamics.

1.2 A Brief History of RNA MD Simulations

MD simulations are useful in studying biochemical systems because they fill a critical void in experimental data. Just as it might be difficult to fully understand how an internal combustion engine works if one only studied an inoperable engine, biochemistry is difficult to fully understand without an understanding of the system dynamics. For instance, although X-ray crystallography provides exquisite atomic level detail of biochemical systems, the constraints of the uniform crystal lattice make it difficult to study the dynamics necessary for biochemistry (for example, an enzymatic reaction). Other experimental techniques, such as nuclear magnetic resonance, circular dichroism, small angle X-ray scattering and others provide partial or indirect information about system dynamics, but bridging that data with an atomistic understanding requires some level of interpretation. MD simulations have increasingly performed this role over the last ten to fifteen years and can be expected to be more important in the future.

The quality of MD force fields, which are the parameters that dictate the classical physics interactions of atom in a simulation, are generally thought to be more mature for proteins than RNA. For example, recent MD simulations have provided deep insight into protein folding *(2-8)*, protein dynamics *(9, 10)*,

ligand-receptor interactions *(11-14)*, and structure-function relationships *(15)*. Achieving reliability for RNA force fields appears to be more challenging. When comparing protein structure with RNA structure, a variety of potential explanations are found. First, RNA is generally extremely flexible and dynamic which is likely an outgrowth of the large number of flexible torsion angles in the polymeric backbone (Figure 1.1) which adopt a variety of conformational suites *(16)*. Second, because RNA is a poly-anion, its structure is inherently coupled with water and ions and thus simulation accuracy is likely highly dependent on both RNA and solvent parameters. Third, RNA structure tends to be more linear than proteins, which are often globular, resulting in greater solvent exposure. This potentially complicates parameterization of force fields because the parameters must be simultaneously accurate in the context of both an RNA environment and a solvent environment. These difficulties are being addressed over time and current force fields are more reliable than ever.

It is instructive to give a brief overview on the progress of RNA MD simulations over the last twenty years. Prior to the mid-1990s, MD simulations were not only very short (on the picosecond timescale) but also unreliable. Simulations were often used with experimental restraints for structure refinements. Stable, unrestrained simulations of RNA were difficult until more accurate treatment of long distance charge interactions were introduced in the mid-1990s *(17)*. At the time, the simulations were on the order of one nanosecond, making reliable comparisons with experiment difficult. Examples of studies performed at this time included predicting loop geometries *(18, 19)*,

studying RNA-protein complexes in bound and unbound states *(20-22)*, and identifying sequence specific dynamics *(23)*. As computer power advanced in the early 2000s, the depth of analysis did too with simulations typically on the order of one to twenty nanoseconds. Studies at this time included the dynamics of RNA-protein systems *(24-27)*, deeper study of structure flexibility and base pair dynamics *(28-33)*, identification of solvent structure features including ion pockets *(34-38)*, diffusion studies *(39)*, and RNA-ligand binding *(40)*. For the most part, recent studies have expanded these approaches in an increasing variety of different systems. However, progress has been made to push the boundaries of what is achievable in terms of system size and timescale. In addition, the use of advanced algorithms for sampling, free energy analysis, and combination of quantum mechanics with MD have become feasible and commonplace. Examples of such literature include studies of RNA enzymes and riboswitches *(41-46)*, simulations of a complete, solvated virus *(47)*, folding and conformational ensemble studies of RNA hairpins *(48-50)*, constant pH simulations *(51, 52)*, and insight into RNA-protein complex cooperativity *(53)*.

Throughout this time, defects in the MD force fields of RNA (presumably due to the approximate nature of the parameters) have been uncovered and subsequent corrections have been shown to improve simulation results with respect to experimental data. All the research presented in this dissertation uses the AMBER simulation software and force fields *(54)* for which a number of force field improvements have been made over the years *(55-61)*. Considerable

effort has also been put into evaluating and improving the CHARMM biomolecular software *(62)* RNA force field as well *(63-65)*.

Taken together, the research in this dissertation seeks to build on previous efforts in RNA simulation, both in terms of offering insight into RNA structure and dynamics as well as developing methods to improve the quality of RNA force fields. In Chapter 2, we describe how explicitly solvated MD simulations offer a consensus picture of domain 5 of two group II introns. In Chapter 3, we report computational studies on the binding of several small-molecule inhibitors to hepatitis C viral RNA and our investigations into the differences in the available experimental structures of the RNA-inhibitor complex. In Chapter 4, we describe enhanced sampling techniques on a small RNA tetranucleotide and introduce a protocol for quantitatively identifying violations of experimental data due to force field errors. Finally, in the conclusion we give an overview of how this research will guide our future efforts.

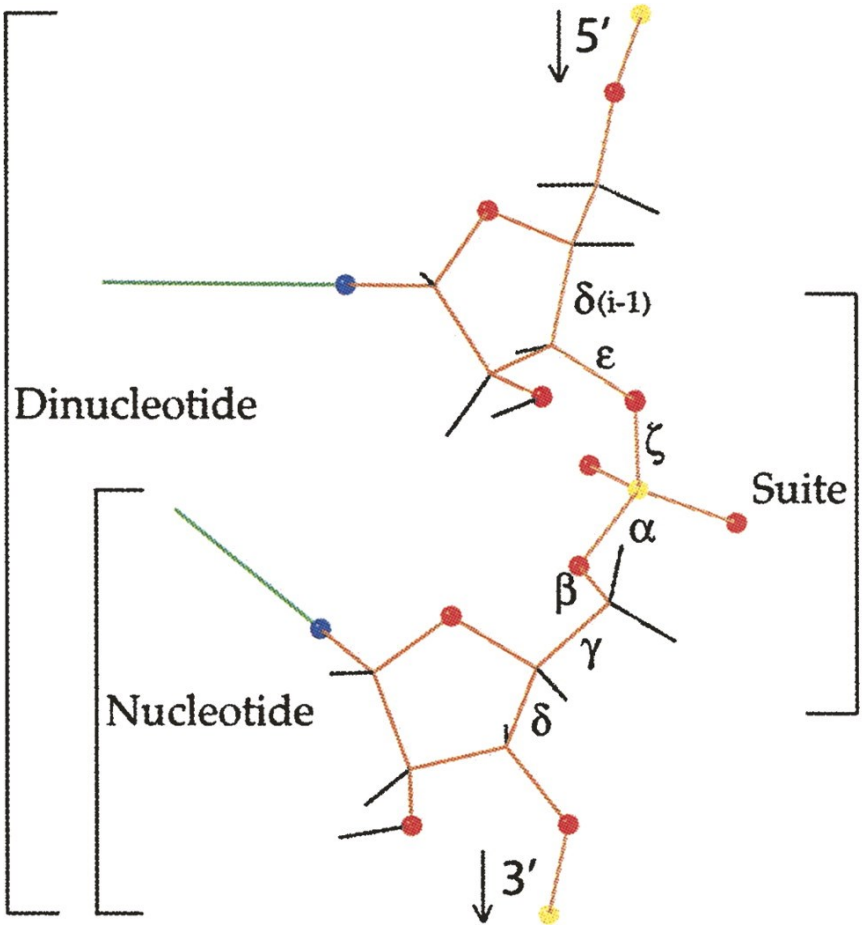

Figure 1.1 Complexity of the RNA dinucleotide structural unit. The backbone torsion angles are indicated in Greek letters. Figure adapted from Richardson et al. *(16)*.

1.3 References

1. Guan, L., and Disney, M. D. (2011) Recent advances in developing small molecules targeting rna, *ACS Chem. Biol. 7*, 73-86.

2. Bowman, G. R., and Pande, V. S. (2010) Protein folded states are kinetic hubs, *Proceedings of the National Academy of Sciences 107*, 10890-10895.

3. Bowman, G. R., Voelz, V. A., and Pande, V. S. (2010) Atomistic folding simulations of the five-helix bundle protein λ6−85, *J. Am. Chem. Soc. 133*, 664-667.

4. Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011) How fast-folding proteins fold, *Science 334*, 517-520.

5. Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., and Shaw, D. E. (2012) Structure and dynamics of an unfolded protein examined by molecular dynamics simulation, *J. Am. Chem. Soc. 134*, 3787-3791.

6. Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2013) Atomic-level description of ubiquitin folding, *Proceedings of the National Academy of Sciences*.

7. Voelz, V. A., Bowman, G. R., Beauchamp, K., and Pande, V. S. (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1−39), *J. Am. Chem. Soc. 132*, 1526-1528.

8. Voelz, V. A., Singh, V. R., Wedemeyer, W. J., Lapidus, L. J., and Pande, V. S. (2010) Unfolded-state dynamics and structure of protein L characterized by simulation and experiment, *J. Am. Chem. Soc. 132*, 4702-4709.

9. Gumbart, J. C., Teo, I., Roux, B., and Schulten, K. (2013) Reconciling the roles of kinetic and thermodynamic factors in membrane–protein insertion, *J. Am. Chem. Soc. 135*, 2291-2297.

10. Jensen, M. Ø., Jogini, V., Borhani, D. W., Leffler, A. E., Dror, R. O., and Shaw, D. E. (2012) Mechanism of voltage gating in potassium channels, *Science 336*, 229-233.

11. Buch, I., Giorgino, T., and De Fabritiis, G. (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations, *Proceedings of the National Academy of Sciences 108*, 10184-10189.

12. Dror, R. O., Arlow, D. H., Maragakis, P., Mildorf, T. J., Pan, A. C., Xu, H., Borhani, D. W., and Shaw, D. E. (2011) Activation mechanism of the ß2-adrenergic receptor, *Proceedings of the National Academy of Sciences 108*,

18684-18689.

13. Nury, H., Poitevin, F., Van Renterghem, C., Changeux, J.-P., Corringer, P.-J., Delarue, M., and Baaden, M. (2010) One-microsecond molecular dynamics simulation of channel gating in a nicotinic receptor homologue, *Proceedings of the National Academy of Sciences 107*, 6275-6280.

14. Vargiu, A. V., and Nikaido, H. (2012) Multidrug binding properties of the AcrB efflux pump characterized by molecular dynamics simulations, *Proceedings of the National Academy of Sciences 109*, 20637-20642.

15. Gasper, P. M., Fuglestad, B., Komives, E. A., Markwick, P. R. L., and McCammon, J. A. (2012) Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities, *Proceedings of the National Academy of Sciences 109*, 21216-21222.

16. Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., et al. (2008) RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution), *RNA 14*, 465-481.

17. Cheatham, T. E., III, Miller, J. L., Fox, T., Darden, T. A., and Kollman, P. A. (1995) Molecular dynamics simulations on solvated biomolecular systems: the particle mesh ewald method leads to stable trajectories of DNA, RNA, and proteins, *J. Am. Chem. Soc. 117*, 4193-4194.

18. Miller, J. L., and Kollman, P. A. (1997) Theoretical studies of an exceptionally stable RNA tetraloop: observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics, *J. Mol. Biol. 270*, 436-450.

19. Srinivasan, J., Miller, J., Kollman, P. A., and Case, D. A. (1998) Continuum solvent studies of the stability of RNA hairpin loops and helices, *Journal of Biomolecular Structure and Dynamics 16*, 671-682.

20. Hermann, T., and Westhof, E. (1999) Simulations of the dynamics at an RNA-protein interface, *Nat. Struct. Mol. Biol. 6*, 540-544.

21. Reyes, C. M., and Kollman, P. A. (2000) Structure and thermodynamics of RNA-protein binding: using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change, *J. Mol. Biol. 297*, 1145-1158.

22. Tang, Y., and Nilsson, L. (1999) Molecular Dynamics Simulations Of The Complex Between Human U1A protein and hairpin II of U1 small nuclear RNA and of free RNA in solution, *Biophys. J. 77*, 1284-1305.

23. Nagan, M. C., Kerimo, S. S., Musier-Forsyth, K., and Cramer, C. J. (1999) Wild-type RNA microhelixala and 3:70 variants: molecular dynamics analysis of local helical structure and tightly bound water, *J. Am. Chem. Soc. 121*, 7310-7317.

24. Blakaj, D. M., McConnell, K. J., Beveridge, D. L., and Baranger, A. M. (2001) Molecular dynamics and thermodynamics of protein−RNA interactions: mutation of a conserved aromatic residue modifies stacking interactions and structural adaptation in the U1A−stem loop 2 RNA complex, *J. Am. Chem. Soc. 123*, 2548-2551.

25. Cojocaru, V., Klement, R., and Jovin, T. M. (2005) Loss of G–A base pairs is insufficient for achieving a large opening of U4 snRNA K-turn motif, *Nucleic Acids Res. 33*, 3435-3446.

26. Pitici, F., Beveridge, D. L., and Baranger, A. M. (2002) Molecular dynamics simulation studies of induced fit and conformational capture in U1A–RNA binding: do molecular substates code for specificity?, *Biopolymers 65*, 424-435.

27. Réblová, K., Špačková, N. a., Koča, J., Leontis, N. B., and Šponer, J. (2004) Long-residency hydration, cation binding, and dynamics of loop E/Helix IV rRNA-L25 protein complex, *Biophys. J. 87*, 3397-3412.

28. Giudice, E., and Lavery, R. (2003) Nucleic acid base pair dynamics: the impact of sequence and structure using free-energy calculations, *J. Am. Chem. Soc. 125*, 4998-4999.

29. Noy, A., Pérez, A., Lankas, F., Javier Luque, F., and Orozco, M. (2004) Relative flexibility of DNA and RNA: a molecular dynamics study, *J. Mol. Biol. 343*, 627-638.

30. Pan, Y., and MacKerell Jr, A. D. (2003) Altered structural fluctuations in duplex RNA versus DNA: a conformational switch involving base pair opening, *Nucleic Acids Res. 31*, 7131-7140.

31. Sarzynska, J., Nilsson, L., and Kulinski, T. (2003) Effects of base substitutions in an RNA hairpin from molecular dynamics and free energy simulations, *Biophys. J. 85*, 3445-3459.

32. Schneider, C., Brandl, M., and Sühnel, J. (2001) Molecular dynamics simulation reveals conformational switching of water-mediated uracil-cytosine base-pairs in an RNA duplex, *J. Mol. Biol. 305*, 659-667.

33. Zacharias, M., and Engels, J. W. (2004) Influence of a fluorobenzene nucleobase analogue on the conformational flexibility of RNA studied by molecular dynamics simulations, *Nucleic Acids Res. 32*, 6304-6311.

34. Auffinger, P., Bielecki, L., and Westhof, E. (2004) Symmetric K+ and Mg2+ ion-binding sites in the 5 S rRNA loop E inferred from molecular dynamics simulations, *J. Mol. Biol. 335*, 555-571.

35. Auffinger, P., and Westhof, E. (2001) Water and ion binding around r(UpA)$_{12}$ and d(TpA)$_{12}$ oligomers - Comparison with RNA and DNA (CpG)$_{12}$ duplexes, *J. Mol. Biol. 305*, 1057-1072.

36. Auffinger, P., and Westhof, E. (2002) Melting of the solvent structure around a RNA duplex: a molecular dynamics simulation study, *Biophys. Chem. 95*, 203-210.

37. Golebiowski, J., Antonczak, S., Di-Giorgio, A., Condom, R., and Cabrol-Bass, D. (2004) Molecular dynamics simulation of hepatitis C virus IRES IIId domain: structural behavior, electrostatic and energetic analysis, *J Mol Model 10*, 60-68.

38. Réblová, K., Špačková, N. a., Štefl, R., Csaszar, K., Koča, J., Leontis, N. B., and Šponer, J. (2003) Non-watson-crick basepairing and hydration in RNA motifs: molecular dynamics of 5S rRNA loop E, *Biophys. J. 84*, 3564-3582.

39. Yeh, I.-C., and Hummer, G. (2004) Diffusion and electrophoretic mobility of single-stranded RNA from molecular dynamics simulations, *Biophys. J. 86*, 681-689.

40. Nifosì, R., Reyes, C. M., and Kollman, P. A. (2000) Molecular dynamics studies of the HIV-1 TAR and its complex with argininamide, *Nucleic Acids Res. 28*, 4944-4955.

41. Banáš, P., Sklenovský, P., Wedekind, J. E., Šponer, J., and Otyepka, M. (2012) Molecular mechanism of preQ1 riboswitch action: a molecular dynamics study, *The Journal of Physical Chemistry B 116*, 12721-12734.

42. Eichhorn, C. D., Feng, J., Suddala, K. C., Walter, N. G., Brooks, C. L., and Al-Hashimi, H. M. (2012) Unraveling the structural complexity in a single-stranded RNA tail: implications for efficient ligand binding in the prequeuosine riboswitch, *Nucleic Acids Res. 40*, 1345-1355.

43. Krasovska, M. V., Sefcikova, J., Špačková, N. a., Šponer, J., and Walter, N. G. (2005) Structural dynamics of precursor and product of the RNA enzyme from the hepatitis delta virus as revealed by molecular dynamics simulations, *J. Mol. Biol. 351*, 731-748.

44. Lee, T.-S., and York, D. M. (2010) Computational mutagenesis studies of hammerhead ribozyme catalysis, *J. Am. Chem. Soc. 132*, 13505-13518.

45. Priyakumar, U. D., and MacKerell Jr, A. D. (2010) Role of the adenine

ligand on the stabilization of the secondary and tertiary interactions in the adenine riboswitch, *J. Mol. Biol. 396*, 1422-1438.

46. Wong, K.-Y., Lee, T.-S., and York, D. M. (2010) Active participation of the Mg^{+} ion in the reaction coordinate of RNA self-cleavage catalyzed by the hammerhead ribozyme, *J. Chem. Theory Comput. 7*, 1-3.

47. Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A., and Schulten, K. (2006) Molecular dynamics simulations of the complete satellite tobacco mosaic virus, *Structure 14*, 437-449.

48. DePaul, A. J., Thompson, E. J., Patel, S. S., Haldeman, K., and Sorin, E. J. (2010) Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics, *Nucleic Acids Res. 38*, 4856-4867.

49. Garcia, A. E., and Paschek, D. (2007) Simulation of the pressure and temperature folding/unfolding equilibrium of a small RNA hairpin, *J. Am. Chem. Soc. 130*, 815-817.

50. Villa, A., Widjajakusuma, E., and Stock, G. (2007) Molecular dynamics simulation of the structure, dynamics, and thermostability of the RNA hairpins uCACGg and cUUCGg, *The Journal of Physical Chemistry B 112*, 134-142.

51. Goh, G. B., Knight, J. L., and Brooks, C. L. (2013) Toward accurate prediction of the protonation equilibrium of nucleic acids, *The Journal of Physical Chemistry Letters 4*, 760-766.

52. Goh, G. B., Knight, J. L., and Brooks, C. L. (2013) pH-dependent dynamics of complex RNA macromolecules, *J. Chem. Theory Comput. 9*, 935-943.

53. Kormos, B. L., Baranger, A. M., and Beveridge, D. L. (2006) Do collective atomic fluctuations account for cooperative effects? Molecular dynamics studies of the U1A−RNA complex, *J. Am. Chem. Soc. 128*, 8992-8993.

54. D.A. Case, T.A. Darden, T.E. Cheatham, I., C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, et al. (2012) AMBER 12, University of California, San Francisco.

55. Krepl, M., Zgarbova, M., Stadlbauer, P., Otyepka, M., Banas, P., Koca, J., Cheatham, T. E., 3rd, Jurecka, P., and Sponer, J. (2012) Reference simulations of noncanonical nucleic acids with different chi variants of the AMBER force field: quadruplex DNA, quadruplex RNA and Z-DNA, *J. Chem. Theory Comput. 8*, 2506-2520.

56. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER force field for

nucleic acids: improving the description of α/γ conformers, *Biophys. J. 92*, 3817-3829.

57. Yildirim, I., Stern, H. A., Kennedy, S. D., Tubbs, J. D., and Turner, D. H. (2010) Reparameterization of RNA x torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine, *J. Chem. Theory Comput. 6*, 1520-1531.

58. Yildirim, I., Kennedy, S. D., Stern, H. A., Hart, J. M., Kierzek, R., and Turner, D. H. (2012) Revision of AMBER torsional parameters for RNA improves free energy predictions for tetramer duplexes with GC and iGiC base pairs, *J. Chem. Theory Comput. 8*, 172-181.

59. Yildirim, I., Stern, H. A., Tubbs, J. D., Kennedy, S. D., and Turner, D. H. (2011) Benchmarking AMBER force fields for RNA: comparisons to NMR spectra for single-stranded r(GACC) are improved by revised x torsions, *J. Phys. Chem. B 115*, 9261-9270.

60. Tubbs, J. D., Condon, D. E., Kennedy, S. D., Hauser, M., Bevilacqua, P. C., and Turner, D. H. (2013) The nuclear magnetic resonance of CCCC RNA reveals a right-handed helix, and revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics, *Biochemistry 52*, 996-1010.

61. Cheatham, T. E., Cieplak, P., and Kollman, P. A. (1999) A modified version of the Cornell et al. Force field with improved sugar pucker phases and helical repeat, *Journal of Biomolecular Structure and Dynamics 16*, 845-862.

62. Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al. (2009) CHARMM: the biomolecular simulation program, *J. Comput. Chem. 30*, 1545-1614.

63. Denning, E. J., and MacKerell, A. D. (2012) Intrinsic contribution of the 2′-hydroxyl to RNA conformational heterogeneity, *J. Am. Chem. Soc. 134*, 2800-2806.

64. Denning, E. J., Priyakumar, U. D., Nilsson, L., and Mackerell, A. D. (2011) Impact of 2′-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA, *J. Comput. Chem. 32*, 1929-1943.

65. MacKerell, A. D., Banavali, N., and Foloppe, N. (2000) Development and current status of the CHARMM force field for nucleic acids, *Biopolymers 56*, 257-265.

CHAPTER 2

MOLECULAR DYNAMICS RE-REFINEMENT OF TWO DIFFERENT SMALL RNA LOOP STRUCTURES USING THE ORIGINAL NMR DATA SUGGEST A COMMON STRUCTURE

2.1 Chapter Notes

This chapter was adapted from the following published article:

> Niel M. Henriksen, Darrell R. Davis, Thomas E. Cheatham III. Molecular dynamics re-refinement of two different small RNA loop structures using the original NMR data suggest a common structure. *Journal of Biomolecular NMR*. **2012.** 53(4) 321-339.

N.M. Henriksen and T.E. Cheatham, III designed the research. N.M. Henriksen performed the research. N.M. Henriksen wrote the manuscript. N.M. Henriksen, D.R. Davis, and T.E. Cheatham, III revised the manuscript.

2.2 Introduction

Molecular dynamics (MD) simulations are often used with restraints derived from crystallography or nuclear magnetic resonance (NMR) experiments for the end-stage atomistic refinement of biological macromolecular structures *(1-5)*. Quite commonly, rather quick and standard MD refinement protocols are employed using codes such as X-PLOR *(6)*, XPLOR-NIH *(7, 8)* or CNS *(9, 10)*. A

refinement protocol might initiate a search for restraint-compatible structures via simulated annealing or distance geometry methods followed by very short (20-200 picosecond) gas-phase MD simulations with applied restraints to further relax and refine the structures. In these refinement protocols, often the force field—specifically the molecular mechanical parameters and force constants that define the covalent connectivity and atomic pair interactions—is rather simplified or crude and the MD simulations are performed *in vacuo* in the absence of solvent and mobile counter-ions. Despite the limitations of these simplified force fields, excellent results are generally obtained given a sufficiently robust set of experimentally derived data. This latter point is somewhat obvious noting that, if sufficient data from experiment has been collected to define the structure, the force field should not strongly influence or bias the results as the structure should be largely determined by the experimental data. However, when the experimental data are sparse, the structure is dynamic, or when solvent and mobile ions may be critically important elements of the structure, the arguably simple force fields and/or absences of experimental restraint data may lead to "loose" structure ensembles. These loose ensembles may suggest larger ranges of motion where data were absent and/or populate anomalous structures leading to an incorrect interpretation of the structure. A logical step to make up for missing data is to apply optimized biomolecular force fields in MD simulations with explicit solvent and modern simulation protocols. A relevant example involves the refinement of nucleic acid structures from NMR data, particularly in the

absence of residual dipolar coupling (RDC) information, where long-range restraint information is absent *(11)*. If the force field and simulation protocols are reasonably robust and ideally experimentally validated, they together with the experimentally derived restraint information should provide a better representation of the structure. Although NMR refinement using modern MD simulation protocols with experimentally derived restraints, optimized force fields for proteins and nucleic acids, and explicit solvent suggests this to be true *(12-18)*, such methods are not widely or routinely applied. For example, in recent years only a handful of NMR structures have been refined in explicit solvent using optimized force fields and modern simulation protocols, and these include the structure elucidation of a peptide *(19)*, an RNA hairpin *(20)*, a DNA naphthalimide adduct *(21)*, a designed metalloprotein *(22)*, and an RNA receptor/ligand complex *(23)*.

Although refinement of NMR structures using modern force fields and simulation protocols, including explicit solvent, appears to produce structures that provide excellent fits to the experimental data, these approaches are not without limitation. Complications relate to dynamics, the time scales of the dynamics, and whether these motions are even accessible on the timescale sampled during the molecular dynamics refinement. Moreover, RNA may populate multiple conformations under the given set of experimental conditions *(24-28)*. With traditional refinement, high restraint weights may lead to structural representations that are too tight, that hide dynamics, and that limit potential transformations between multiple conformations *(29)*.

Essentially, an average structure will be found that may not entirely satisfy all of the experimental data. Procedures such as time-averaged restraints *(30-32)*, selective enforcement of restraints over time *(33)*, or ensemble based refinement methods *(34, 35)* may help mitigate these issues. However, these methods will further depend on the reliability of the force field representation to correctly sample the accessible conformational space. Ultimately, given a reliable and validated force field, the molecular dynamics simulations without experimental restraints starting from the refined NMR structure should provide an accurate representation of the structure and dynamics over the simulation time scale. However, the force fields are not yet fully reliable, especially in the treatment of RNA *(36-40)*. Therefore, at present, there needs to be a careful balance between the relative weights of the force field compared to the experimental or structural restraints, with further care levied to understand the limitations of the force fields and implications of specific restraint choices.

In this work we further assess the reliability of more detailed MD structure refinement protocols through the re-refinement of two similar RNA molecules. As part of our larger force field assessment efforts, we have been investigating a variety of RNA structures in free, unrestrained MD simulation to better understand the reliability and flaws of the AMBER nucleic acid force fields *(40-46)* as compared to available experimental data. Ideal RNA structures for our investigation include those which display some noncanonical structure (i.e., nonhelical structure since the AMBER force fields appear to do a

reasonable job of modeling nucleic acid helices *(47-53)*) and importantly, structures where detailed NMR restraint information is available including NOE derived distance, J-coupling and RDC restraints. Our explorations led us to two published RNA structures that have nearly identical primary sequence, yet the published 3D structures differ significantly.

A comprehensive MD investigation of these two previously refined RNA structures from the PDB *(54)* was performed, specifically on structures with the PDB codes of 1R2P *(55)* and 2F88 *(56)*. These structures consist of a 34 residue segment derived from domain 5 (D5) of the group IIB intron ribozyme in yeast ai5γ *(55)* (ai5γ-D5) and *Pylaiella littoralis (56)* (PL-D5), respectively. Domain 5 serves an essential role in the core of the intron structure and contains the most important residues for catalysis *(57)*. The primary sequence of ai5γ-D5 and PL-D5 is mostly identical, except for three residues (Figure 2.1, noting the different boxed residues 8, 25, and 27). The structural elements of both D5s include a lower helix joined by a bulge region to an upper helix that is capped by a GAAA tetraloop (shown in Figure 2.1 in red, green, blue, and pink, respectively). Residues A2, G3, and C4 form what is known as the catalytic triad; this is a highly conserved region of interest noted for forming tertiary interactions and interactions with Mg^{2+}. According to the published structures, the conformation of the bulge region differs depending both on the sequence and experimental method (NMR vs. X-ray crystallography), while the lower helix, upper helix, and tetraloop features are all very similar. The bulge conformation as reported in the earlier crystal structure of ai5γ-D5 (PDB: 1KXK)

(58) shows G26 forming a wobble pair with U9, while A24 and C25 are unpaired and opened away from the helix. In contrast, the NMR structures of ai5γ-D5 and PL-D5 (referred to as ai5γ_NMR and PL_NMR from here on) both suggest that G26 is in a *syn* conformation that protrudes into the major groove. However, the positioning of the other bulge residues differs between the two NMR structures: while most of the ai5γ_NMR ensemble structures show residues 24 and 25 stacked into a narrow helix above U9, the PL_NMR structures show residues 24 and 25 in one of two conformations opposite, but not directly interacting with, U9. Both of these conformations of PL_NMR show a very wide bulge region accommodating A25 either stacked into the helix below A24 (as is shown in Figure 1.1, right) or packed against the minor groove out of the helix. The differences in these bulge structures are highlighted in the supporting information using molecular graphics (Figure 2.2) and annotated secondary structure representations (Figure 2.3). We note that the expected conformation of the bulge in the context of the full intron likely does not resemble any of the earlier X-ray or NMR structures. Despite sequence differences, the bulge conformation in the full intron will probably be similar to the more recent X-ray structures of the full *Oceanobacillus iheyensis* self-spliced group IIC intron where tertiary interactions stabilize a different bulge conformation (PDBs: 3BWP, 3EOG, 3EOH, and 3IGI) *(59, 60)*. Collectively, the various bulge conformations observed in the earlier NMR and crystal structures, and those we report upon re-refinement, suggest that the bulge structure clearly differs when in an isolated solution environment, is influenced by

sequence, crystal packing and tertiary packing, and is likely dynamic. Except for the bulge region, most of the other elements of the D5 structure are quite similar between ai5γ_NMR and PL_NMR. One remaining difference is the overall structural length of D5. The structures of ai5γ_NMR are more extended while those of PL_NMR are compact and similar in length to the X-ray structure of ai5γ-D5. The other difference of significance between the published studies relates to the determination of divalent ion binding as determined by NMR chemical shift perturbations upon addition of $MgCl_2$. It was not obvious to us whether real differences in ion binding or structure exist between the RNA constructs, or whether the published observations reflect subtle differences in NMR methods and/or experimental conditions, despite apparently strong similarities in the experiments and refinement protocols.

Together, these two structures and their previously accumulated and published data (chemical shifts, restraints, and chemical shift perturbations) provide an intriguing opportunity to validate the MD simulations, to assess simulation refinement protocols which use explicit solvent and modern force fields, and to ultimately determine whether the significant differences in these structures reported are real or artifacts from the refinement process. We show that re-refinement leads to two very similar structures that appear to better satisfy the NMR data. In addition, beyond the bulge region (which is strongly influenced by packing effects and tertiary interactions), the stem and loop regions around the bulge better match the ai5γ-D5 (PDB: 1KXK) crystal structure than the previously published NMR structures. Yet, the results also

suggest that a careful balance between relative weights of the force field compared to the experimental data is required, that the experimental data have to be carefully screened, that there are clear sampling limitations, and that there are still known and emerging force field limitations. Taken together, these observations preclude the use of automation to automatically refine structures. The previous and current success of these methods suggest that some published structures might be improved using explicit solvent MD refinements and that such techniques should be more routinely applied in future structure refinement projects.

2.3 Methods

2.3.1 Coordinates and restraint data

The coordinates and restraint data for the ai5γ_NMR and PL_NMR structures were retrieved from the RCSB Protein Data Bank website using the PDB codes 1R2P and 2F88, respectively. Of the ten conformers contained in each PDB file, only the first five were used in simulations. The distance restraint data were thoroughly checked for atom naming mismatches between the restraint file and the PDB files. Mismatched restraints were corrected based on visual inspection of the structures (common mismatches included H62 → H61 or H5'1 → H5'2 transpositions). Distance restraints were weighted at 20 kcal/mol-Å within 0.5 Å of the bounds of the flatwell restraint and at 20 kcal/mol-Å2 outside this range, unless otherwise noted. The applied dihedral restraint data from the two earlier refinements were consolidated and

reconfigured to produce a more consistent and liberal set of dihedral restraints and also to eliminate minor differences in the conventions used by the ai5γ_NMR and PL_NMR authors. Backbone torsions of the helical regions (torsions 1:γ – 8:γ, 10:ε – 14:γ, 19:ε – 23:γ, 27:ε-34:ε) were restrained to A-form values (± 15°) while those in the remaining noncanonical regions were left unrestrained. The x torsion angle was restrained to syn (70 ± 30°) for G26 and anti (-160 ± 15°) for all others. Sugar puckers were restrained (i.e., torsions δ, v1, and v2) in a similar manner as described in the original publications: for ai5γ-D5, A16 and C25 were restrained to C2'-endo, whereas G15, A17, and A18 were unrestrained, and the remaining were restrained to C3'-endo; for PL-D5, A16, A24, and A25 were restrained to C2'-endo, and the remaining were restrained to C3'-endo. Torsion restraints were weighted at 500 kcal/mol-rad within 1° of the bounds of the flatwell restraint and at 500 kcal/mol-rad2 outside this range, unless otherwise noted. Relative RDC restraint weighting, set with the "dwt" keyword in AMBER, was chosen to be the highest value that did not cause SHAKE errors during simulation (dwt=0.02 for ai5γ-D5, dwt=0.01 for PL-D5). Base pair planarity restraints were not applied. All restraints were converted to the AMBER formats using in-house scripts and scripts available in AmberTools. The restraint files used during the refinement are supplied in the Supporting Information.

2.3.2 Building, heating, and equilibrating solvated structures

All MD simulations were performed using the AMBER and AmberTools suites of software *(61, 62)*. The PDB structures were parameterized using the AMBER ff99bsc0 *(44)* force field, and were solvated in an icosahedral TIP3P *(63)* water box out to at least 10 Å in each direction from the solute followed by net-neutralization with Na+ ions and addition of ~200 mM NaCl using the Joung and Cheatham ion parameters *(64, 65)*. The Na+ cation was chosen for initial investigations due to the salt crystallization artifacts seen with the earlier K+ parameters *(66)*, despite the fact that K+ was used in the NMR buffer. The Na+ cation was used throughout this work, albeit with the improved parameters, in order to maintain consistency with our older data. In total, the solvated systems contained between 9000 – 12000 residues, corresponding to approximately 29000 – 37000 atoms. After building the coordinate and parameter/topology (prmtop) files, the positions of all the ions in the coordinate files were randomized using ptraj, ensuring that ions were at least 6 Å from an RNA atom and 4 Å from each other. The particle mesh Ewald method *(67)* was used to handle electrostatic interactions using a 9 Å cutoff with default parameters (including an ~1 Å grid spacing, cubic spline interpolation, and a direct space cutoff tolerance of 0.000001). Lennard-Jones interactions were also treated with a 9 Å cutoff and the pairlist built to 10 Å was automatically rebuilt if any atom moved more than 0.5 Å since the previous update. Each system was first relaxed with 1000 steps each of steepest descent and conjugate gradient minimization while RNA atom positions were restrained

with 25 kcal/mol-Å2 positional restraints. Continuing with the same positional restraints, the system was slowly heated from 100 K to 300 K over the course of 100 ps at constant volume. After heating, the system was repeatedly minimized (1000 steps each, steepest descent and conjugate gradient) and equilibrated at constant pressure (for 50 ps each round) using gradually weaker positional restraints (5.0, 4.0, 3.0, 2.0, 1.0, and 0.5 kcal/mol-Å2). For restrained simulations, distance and torsion restraints as previously described were enforced during each step of the equilibration process. A final equilibration step for the restrained simulations was included following the 0.5 kcal/mol-Å2 position restrained equilibration. This step consisted of a relaxation period of 2 ns at constant pressure without positional restraints and with distance and torsion restraints at 10% of normal strength.

2.3.3 Restrained production simulations

Production simulations were performed at constant pressure and temperature using the Berendsen algorithm *(68)* for scaling. The heat bath and pressure coupling time constants were set to a loose value of 5 ps. Chemical bonds to hydrogen atoms were constrained using the SHAKE algorithm *(69, 70)*, which permitted a time step of 2 fs for production simulations. Translational and rotational center-of-mass motion was removed every 500 steps. Coordinates of the system were recorded every picosecond during simulation. Distance/torsion angle (DA) restrained simulations were performed using the AMBER's PMEMD program. Restrained simulations using distance/torsion-

angle/RDC (DAR) restraints were performed using AMBER's sander program (PMEMD is generally faster than sander however PMEMD does not yet implement RDC restraints). DAR restrained simulations were not started from an independent equilibration, but rather were started from the final frame of the corresponding DA restrained simulation. To accomplish this, the RDC alignment tensor was first minimized to best fit the RNA structure. Then the DAR simulation was started using the tensor values obtained in the minimization. Every time a DAR simulation was restarted, the alignment tensor values were obtained from the final step of the previous output file and used as the starting values for the next calculation. A complete listing of the simulations is given in Table 2.1, noting that the five independent runs originated from the first five representative NMR structures from the 1R2P and 2F88 PDB files.

The structure refinement protocol presented here significantly extends the procedure used to generate the ai5γ_NMR *(55)* and PL_NMR *(56)* structures. Specifically, significantly longer MD simulations were performed including explicit solvent and mobile counterions, modern force fields, and proper treatment of the long range electrostatic interactions. Longer simulation, and in some case heating, provided significantly more sampling of potential RNA structure and helped identify where structures may have otherwise been trapped due to previously insurmountable barriers. In contrast, both earlier publications report using CNS to generate an extended structure, followed by the selection of 100 starting structures generated from different random initial velocities. The starting structures were relaxed using high-temperature,

torsion-angle dynamics, slow cooling using distance and angle restrained molecular dynamics, and minimization. Total molecular dynamics for each structure in the earlier refinement protocols did not exceed 250 ps. The PL_NMR structures were then further refined based on the RDC data using a more extensive protocol.

2.3.4 Analysis

All PDB and trajectory structures were visualized using UCSF Chimera *(71)*. Structure snapshots were also generated using Chimera. RMSD values were generated using AMBER's ptraj module and results were plotted using Grace *(72)* or Microsoft Excel. Distance and dihedral measurements were calculated and analyzed using ptraj and in-house scripts. Clustering was performed in ptraj *(73)* using the following settings: average-linkage algorithm, cluster count set to 5, rms similarity metric on base heavy atoms only, and sieve set to 5. To generate representative structures for the restrained simulations, the average structure of the dominate cluster for each trajectory was minimized with full restraints. Grid analysis *(74)* was performed using ptraj and visualized in Chimera. Occupancy analysis of water and Na+ was performed using the hbond command in ptraj.

2.4 Results

2.4.1 Initial results with unrestrained simulations

Prior to running the restrained simulations, we performed a set of ~100 ns unrestrained simulations using the ai5γ_NMR and PL_NMR starting structures in order to evaluate the AMBER ff99bsc0 force field; a summary of all of the simulations performed is provided in Table 2.1 and a figure highlighting the structural and sequence differences is shown in Figure 2.1. These simulations, named ai5γ_UR and PL_UR, respectively, were built and equilibrated using the same procedure as for the restrained simulations except without any of the steps related to the distance, torsion, or RDC restraints. The initial results from these simulations led us to begin a more thorough investigation using restrained simulations for two reasons. The first is related to the structural compactness of the ai5γ_NMR and PL_NMR structures. One of the more striking differences between the published ai5γ_NMR and PL_NMR structures is the overall structural length, as measured from the top of the tetraloop to the bottom of the lower helix. The PL_NMR structures display a compacted global conformation that is consistent with and almost as compacted as the X-ray structure of ai5γ-D5 *(56, 58)*, whereas the ai5γ_NMR structures are much more extended (Figure 2.2). In contrast, when we compared the average structures for the unrestrained ai5γ_UR and PL_UR simulations, both sets of structures adopted the more compact conformations. Visual inspection of the ai5γ_UR trajectories revealed that the end-to-end distance of the structures underwent an approximately 15 Å compaction within the first 10 ns of the simulations.

This rapid compaction on the MD simulation time scale suggests that the extended structure is not compatible with the force field which, as discussed in the introduction, is known to fairly reasonably model many RNA structures. The PL_UR structures, whose starting structures were already more compact, underwent no appreciable compaction during simulation. This likely leads to the rather significant difference in plateau RMSD values between the ai5γ_UR and PL_UR simulations (Figure 2.4). Note that although the RMSd values plateau and appear relatively small, at least in the case of the PL_UR structures with RMSd values in the ~3-6 Å range, structural disruption in the bulge and loop regions was evident. However, the similarity of the two sets of unrestrained simulation structures after MD simulation led us to wonder if perhaps the conformation of these two molecules was more similar than the conventionally refined structures suggest or if the minimalist gas-phase refinement protocol employed previously was insufficient to refine the structures.

The second reason these simulations encouraged us to perform a more detailed analysis was related to the localized loop and bulge structural features. We found that during both the ai5γ_UR and PL_UR simulations these regions experienced significant structural degradation. For instance, at various times in the five independent simulations the loop conformation would transition to a pathological, yet stable, geometry that often persisted for the rest of the trajectory. It was soon clear that accurate modeling of these molecules could not be achieved using the MD force field alone, and thus we

decided to perform restrained simulations using the original NMR data. While the RNA force field parameters have considerably improved *(40-45, 75, 76)*, our results suggest the geometries of the refined structures presented in the current study are primarily determined by the experimental restraints. The explicit solvent environment and updated force field play a secondary yet critical role, as the resulting structures appear to be improved compared to those obtained using conventional methods.

2.4.2 Restrained simulations produce conformational rearrangements in ai5γ-D5 and PL-D5

In addition to giving clues about structure compaction, our initial investigation of unrestrained simulations led us to hypothesize that the ai5γ-D5 and PL-D5 structures are more similar than the reported NMR structures suggest. To investigate this possibility we ran simulations with distance, angle, and residual dipolar coupling (RDC) restraints imposed (i.e., ai5γ_DAR and PL_DAR). The resulting trajectories were clustered and a representative structure from the most populated cluster for each trajectory was minimized. These minimized structures (five total, one for each of the five models of a given each simulation type), served together as representative structures for each type of simulation (summarized in Figure 2.5 and Table 2.2). A pairwise heavy atom RMSD measurement between the ai5γ_DAR and PL_DAR structures, which excluded the base atoms for the three differing residues, was much lower (3.33 Å) than the corresponding measurement between the ai5γ_NMR

and PL_NMR structures (6.03 Å) (RMSD data listed in Table 2.3). A significant portion of this decrease was likely due to the structure compaction of ai5γ_DAR.

In addition to a lower interstructure pairwise RMSD with PL_DAR, ai5γ_DAR also has a much lower intrastructure pairwise RMSD (2.10 Å) as compared to the original ai5γ_NMR structures (4.09 Å). Other than global compaction of the structure, the most significant differences between the ai5γ_NMR and ai5γ_DAR structures occur in the bulge region. In four out of the first five ai5γ_NMR structures, U9 is positioned below residues 24 and 25. For ai5γ_DAR, U9 is directly adjacent to A24 and appears to form Watson-Crick bonding in three of the five structures. The other two structures show U9 disengaged from A24 while still maintaining an adjacent stance. In all five ai5γ_DAR structures, C25 is pressed against the major groove side of the A8-U27 base pair, forming a hydrogen bond between A8 H62 and C25 N3 (Figure 2.6, top). This hydrogen bond seems unlikely due to the high angle between the base plane of A8 and C25, and is probably exaggerated by the force field which allows such bonding at any angle. The lower position of C25 in the ai5γ_DAR structures is likely caused by another interesting feature of the bulge region. The kink structure of backbone residues A24 – U27 is even more pronounced in the simulation structures than is observed in the NMR structures. The otherwise smooth curve of the backbone is disrupted between G26 and U27, forming a near right-turn in the helix when viewed from above (Figure 2.7, left). In the NMR structures, the kink maintains an upward direction

throughout the bulge region. In contrast, the ai5γ_DAR kink briefly travels downward as it cuts across the major groove and forces C25 into a very low position. The orientation of G26 also differs somewhat between the simulation and NMR structures. In most of the five ai5γ_NMR structures (models 1-5 from the PDB file), the base plane of G26 is close to perpendicular with the vertical axis of the lower helix. In contrast, most of the ai5γ_DAR structures have the base plane parallel to the vertical axis of the lower helix.

The differences between the PL_NMR structures and PL_DAR structures are not as drastic as those for ai5γ. In contrast to the results for ai5γ, the intrastructure pairwise RMSD is higher for PL_DAR (2.56 Å) than for the PL_NMR (1.30 Å). One of the most obvious differences between the PL_DAR and PL_NMR structures is that during equilibration and relaxation, residue A25 of model 4 shifted from a partially extruded position in the minor groove to a stacked position within the helix. Thus the PL_DAR ensemble has three of five structures with A25 stacked, whereas the PL_NMR structures have two of five (ignoring structures 6-10 in the published PDB structure file). All five representative structures for PL_DAR show U9 participating in a hydrogen bond with either A24, A25 or G26. In the two structures with A25 extruded, U9 interacts with A24. Of the three structures with A25 stacked in the helix, two show U9 interacting with A25 and one shows a hydrogen bond between U9 H3 and G26 N7. During simulations of the former case, U9 shifts back and forth between interactions with A24 and A25. In the latter, the U9 H3 - G26 N7 hydrogen bond is particularly stable throughout the trajectory, leaving A24 and

A25 stacked above U9. In all five simulations, these interactions close the "hole" described by Seetharaman et al. *(56)* in reference to the PL_NMR structures. Interestingly, for the three structures with A25 in the stacked position, G26 no longer "packs into the major groove against G8" (as in all five PL_NMR structures), but points away from the major groove with the base plane parallel with that of U9. For the two structures with A25 extruded into the minor groove, G26 remains oriented towards the major groove.

2.4.3 Troubleshooting problematic or unusual regions in the refined ai5γ-D5 and PL-D5 structures

On closer inspection of the refined models, some features of both the ai5γ_DAR and PL_DAR structures seemed problematic. For ai5γ_DAR, there were three distance restraints in the bulge region with consistently large upper bound violations during simulation (these restraints connected the following atoms: U7 H1' – A28 H2, G26 H8 – A28 H1', G26 H8 – U7 H3). Two of these involved the atom G26 H8. On closer inspection it seemed possible that these two restraints were responsible for the severe kink in the backbone noted by the authors of ai5γ_NMR structures as being a very unusual conformation. Given the high upper and lower bounds of these restraints, one of the corresponding authors (Samuel Butcher, personal communication) suggested to us that these restraints were derived from weak NOEs that may be mediated by spin diffusion. Even though the U7 H1' – A28 H2 restraint violation was probably a side effect of the other two restraints, we decided to investigate the effect

of removing all three problematic restraints. Identical simulations to ai5γ_DAR were then run with the three aforementioned restraints removed to generate the simulations that are designated as ai5γ_mDAR.

Additionally, the unrestrained simulations suggested that the positioning of the A25 residue in the PL_DAR structures might also be problematic. Although the unrestrained simulations are imperfect due to force field deficiencies, we never observed A25 extruded into the minor groove during the 100 ns of unrestrained simulations. As mentioned previously, we also found that one fully restrained simulation showed A25 move from the extruded position to the stacked position and we therefore considered whether this conformation might be preferred. To investigate this, we tested several different equilibration and relaxation conditions before choosing a procedure for the modified restrained simulations. Many of the conditions resulted in either one or two of the three extruded structures transitioning to a stacked structure. Critically, in none of these conditions did A25 transition from a stacked conformation to an extruded conformation. In one condition, all three of the extruded structures transitioned to stacked structures. This condition involved three alterations to the PL_DAR simulation procedure: 1) the removal of the Watson-Crick base pair restraints, but not the NOESY restraints, between residues G8 and C27 (which are not present for A8/U27 in ai5γ_DAR), 2) increasing the weight of distance restraints from 20 kcal/mol to 50 kcal/mol as well as increasing the G26 x dihedral restraint from 500 kcal/rad to 1000 kcal/rad, and 3) heating to 700 K with restraints at 8% strength to allow

structural relaxation, followed by a smooth increase of restraint weight to 100% strength prior to the production simulation. The rationale for removing the Watson-Crick base pair restraints was to allow structural transitions in the bulge region that may otherwise be hindered. The changes to the restraint weighting were made to ensure the enforcement of distance restraints and prevent G26 from flipping to the anti conformation (a frequent problem for previous attempts) while the heating period was intended to enhance conformational sampling. These simulations were named PL_mDAR.

2.4.4 Analysis of the bulge and loop regions in optimized, restrained simulations

The removal of the three problematic restraints from the ai5γ restraint list makes a significant difference in the bulge region of the ai5γ_mDAR structures. First, the intrastructure pairwise RMSD of the ai5γ_mDAR structures (1.04 Å) is significantly lower than ai5γ_DAR (2.10 Å) (Table 2.2). The sharp kink observed in the bulge of the ai5γ_DAR structures is replaced by a smooth, upward trending backbone in the ai5γ_mDAR structures (Figure 2.7, right). Rather than being drawn below the major groove face of the A8 – U27 base pair, C25 is positioned above A8 – U27 and its base plane is parallel with A8's in the ai5γ_mDAR structures (Figure 2.6, bottom). This positioning puts U9 directly in between the A24 and C25 bases and during simulation U9 alternates hydrogen bonding between A24 and C25. Occasionally, C25 N3 or O2 also form a hydrogen bond with the A8 H61, H62 atoms. In contrast to the ai5γ_DAR

structures, four of the five ai5γ_mDAR have the base plane of G26 perpendicular to the helical axis, thus pointing away from the major groove.

For the PL_mDAR structures, the removal of the A8-U27 Watson-Crick base pair restraints and subsequent heating yielded representative structures with a lower pairwise RMSD value (1.70 Å) than that of the PL_DAR structures (2.56 Å), although not as low as the original NMR structures (1.30 Å). During the five PL_mDAR simulations, three different bulge motifs dominate, all of which are more packed than the open bulge observed in the original PL_NMR structures. In Model 1, the U9H3 - G26N7 hydrogen bond forms (Figure 2.8A). As was the case for the same hydrogen bond seen before in model 2 of PL_DAR, this structure is quite stable during the simulation. Models 2 and 3 show U9 interacting with A25, yet U9 never quite reaches A24 (Figure 2.8B). This conformation is more dynamic, with U9 oscillating above and below A25. Finally, models 4 and 5 show U9 placed between A24 and A25 (Figure 2.8C). In these simulations, U9 alternates between interactions with A24, A25 and a shared interaction occurs between the two. Interestingly, the positioning of U9 is correlated with the positioning of G26. In models 1-3, where U9 is interacting with A25 or A26 (but not A24), G26 remains partially in the helical stack with its base plane perpendicular to the helical axis. However in models 4-5, with U9 interacting with A24 and A25, G26 is pushed out of the helix and the base plane is parallel to the helical axis. Contrary to the inferences from the ai5γ_NMR and PL_NMR structures, the ai5γ_mDAR and PL_mDAR structures are quite similar to each other. The average heavy atom pairwise RMSD (excluding

the base atoms of differing residues) is much lower (1.95 Å) than the first simulation structures (3.33 Å) and the published NMR structures (6.03 Å). In addition, the lower helix regions of the re-refined structures are very similar to each other and better match the earlier crystal structure (PDB: 1KXK). The average heavy atom pairwise RMSD to the crystal for the lower helix is 2.36 Å and 1.50 Å for the original NMR structures and 0.81 Å and 1.13 Å for the re-refined structures (for ai5γ and PL, respectively). The remaining structural differences in the bulge are likely related to the accommodation of the larger A25 in PL-D5 as opposed to C25 in ai5γ-D5.

The other noncanonical structure of interest, the GAAA tetraloop, is reasonably similar between the NMR structures and among each of the restrained simulation structures. These structures are also closer to the earlier crystal structure. Whereas the RMSD values for residues 10-23 are 1.64 Å and 1.79 Å when the original NMR structures are compared to the crystal, the re-refined structures yield values of 1.37 Å and 0.92 Å (for ai5γ and PL, respectively). However, some variation in the backbone torsions are observed, such as that of the γ torsion of residue A16, which we attribute to the bsc0 modifications of the AMBER ff99 parameters which disfavor γ in the trans configuration *(44)*. In both the ai5γ_NMR and PL_NMR structures, γ is in the trans configuration while in each of the ai5γ_mDAR and PL_mDAR structures this torsion is in the gauche+ position. In addition to the backbone differences, the base orientations in the tetraloop were compared between the NMR structures and the restrained simulation structures. Previous work by Correll

and Swinger identified aspects of the GNRA tetraloop that commonly vary between one of two conformations *(77)*. The first of these involves the planarity of the "NRA" portion of the tetraloop with respect to the underlying base pair of the upper helix (i.e., the planarity of residues 16 – 18 with respect to the base pair formed by residues 14 and 19 in the case of D5). When the NRA bases are planar with the underlying base pair, the conformation is referred to as the "standard orientation." If the NRA bases depart from planarity, typically tilting upward and away from the underlying base pair, the conformation is named the "altered orientation." In both cases, the three NRA bases remain stacked together. Visual inspection of both the NMR structures and representative structures from our restrained simulations reveal that nearly all of the tetraloops adopt the altered orientation. This contrasts with results from unrestrained simulations in which the standard orientation seems to be favored.

The second feature of interest identified by Correll and Swinger in GNRA tetraloops is the hydrogen bonding network of the first and fourth residues in the tetraloop (i.e., G15 and A18 for D5). In the first case (which we name the "outward orientation"), G15 N2 forms a bifurcated hydrogen bond between both A18 N7 and A18 O2P, while G15 N1 also hydrogen bonds with A18 O2P. In the second case (the "inward orientation"), G15 and A18 shift slightly relative to the backbone which allows for the same bifurcated hydrogen bond between G15 N2 – A18 N7 / A18 O2P as well as a hydrogen bond between G15 N3 and A18 N6. Interestingly, the ai5γ_mDAR structures seem to fit the outward

orientation, whereas the PL_mDAR structures fit the inward orientation (Figure 2.9). Both the ai5γ_NMR and PL_NMR structures also appear to adopt the inward orientation although the refinement did not seem to capture the fine detail of hydrogen bonds that stabilize the structure. For instance, in many of the submitted models for both NMR structures, the orientation of G15 N2 does not indicate a hydrogen bond is formed with A18 N7, although the atoms are close in space. The conformational differences found in the loop regions when comparing the ai5γ_mDAR and PL_mDAR structures are surprising given that the NMR conditions are similar and the upper helix and tetraloop sequence is identical. Slight differences in the restraint data likely lead to the observed differences, and as discussed below these conformational differences result in slightly different hydration and Na+ binding features. As both loop conformations are observed, and each is consistent with experimental data, it is not possible to directly ascertain which conformation is preferred and/or if the loops interconvert between the two conformations, although likely the two conformations are close in energy which suggests population of both.

2.4.5 Deviations from ideal geometry when using the AMBER force field

One problem which might deter researchers from using the AMBER force field is the slight deviations of covalent bond angles from ideal geometries. Deviations outside of the expected covalent bond angle range were observed for all of our AMBER refined structures as well as other recent structures *(78,*

79) when submitted to the ADIT-NMR (AutoDeposit Input Tool for NMR structures: http://deposit.bmrb.wisc.edu/bmrb-adit/) from the RCSB website (http://www.rcsb.org). The origin of the deviations results from the use of shared, transferable and rather generic atom types for the nucleobases (such as CT for all tetrahedral carbons and OS for O2', O3', O4', and O5') rather than specific types for each different atom to more accurately represent nucleoside geometry. Although the deviations (on the order of ~5^0 or less) are outside the range observed in experimental databases *(80, 81)*, these deviations are within the range of thermal fluctuation and likely have a small impact on the overall structural quality due to compensation by the many degrees of freedom in large biomolecules. Addressing these deviations will require an overhaul of the atom type naming system used by the AMBER force field and is the subject of ongoing research. In addition to the deviations observed for covalent bond angles, some deviations were also observed in base planarity (on the order of 0.1 Å rmsd or less), which likely reflect restraint strain on the relatively soft improper torsion parameters used in the AMBER force field to maintain planarity.

2.4.6 Comparison of representative ensembles with and without RDC restraints

A comparison between the simulations run with and without RDC restraints (DAR and DA, respectively), suggests that the RDCs primarily affect the structural compactness of these RNA structures, but not the local

conformations. Both the ai5γ_mDAR and PL_mDAR simulations resulted in an increase of overall structural length (measured by the distance between A16 P and C34 P). The average structure length for ai5γ_mDA and ai5γ_mDAR, taken from representative structures, was 47.0 and 54.5 Å, respectively. The increase was less dramatic for PL, for which the average structure length was 52.8 and 53.4 Å for PL_mDA and PL_mDAR, respectively.

These results are consistent with a recent study by Tolbert et al. *(78)* of a purely A-form RNA double helix, which suggested that a wide range of A-form structures are accessible when using only distance and torsion restraints. In contrast to what was observed by Tolbert et al., the addition of orientational restraints resulted in structural expansion, not contraction. However, the differences between the simulations (pure helical RNA vs. noncanonical RNA, implicit solvent vs. explicit solvent) preclude further conclusions.

2.4.7 Solvation and Na+ density during simulation

In addition to structural information, the publications describing both the ai5γ_NMR and PL_NMR solution structures also included data describing chemical shift perturbations observed upon titration of D5 RNA with Mg2+. The results for ai5γ_NMR, which only tracked 1D proton shift changes, differ somewhat from those for PL_NMR, which included more detailed 2D 1H – 13C and 1H – 15N chemical shift data. However, these differences could be attributed to either a real difference in structure, simply a difference in the experimental techniques, or perhaps a combination of these differences. A

detailed examination of Na+ position during the simulation provided a plausible explanation for the results observed in the Mg2+ binding experiments. Our simulations were performed with monovalent salt due to the lack of good parameters for divalent cations, the absence of polarization, and conformational sampling limitations. Although the replacement of Mg2+ with monovalent salt can destabilize RNA, for small RNAs high monovalent salt concentrations are generally a good substitute for physiological Mg2+, and typically provide similar structures *(82, 83)*. We studied Na+ and water binding during simulation using two techniques. First we generated density grids to map positions of highest density over the course of an entire set of simulations. Second we probed the occupancy of Na+ and water within a defined radius for atoms of interest.

One of the more significant differences between the results for ai5γ_NMR and PL_NMR involved the triad region (A2, G3, and C4). For ai5γ_NMR, Sigel et al. found that none of the protons in this region experienced a significant perturbation *(55)*. However, Seetharaman et al. reported that N7 of A2, G30, and G31 were significantly perturbed while N7 of G3 was not *(56)*. A comparison of the 3D grid structures for ai5γ_mDAR and PL_mDAR suggest that Na+ binding occurs in the tetraloop, bulge, and AGC triad region for both structures (Figure 2.10). These data support Seetharaman et al. *(56)* who suggest that probing for perturbations in the N7 atoms of ai5γ-D5 would also uncover these results. Furthermore, detailed analysis of Na+ and water density near each of the triad region base pairs suggests a possible explanation for why

N7 of G3 was not perturbed while N7 of A2, G30 and G31 were perturbed *(56)*. In the case of A2 N7 and G30 N7, a high density region of Na+ is positioned directly off the N7 atom (Figure 2.11). Occupancy analysis of Na+ within 2.8 Å of these two atoms shows that for both ai5γ_mDAR and PL_mDAR they rank among the highest in Na+ binding out of all N7 atoms (Figure 2.12). Although G31 N7 does not directly bind Na+, a large region of density exists nearby and the neighboring O6 atom does bind Na+ at a high occupancy and probably contributes to the chemical shift. In the case of residue G3, neither N7 nor O6 have a high Na+ occupancy, and grid analysis reveals two areas of high water density that form a barrier between the Na+ density and residue G3 (Figure 2.11).

Another region of interest is the bulge (residues 9, 24, 25, and 26; residues 8 and 27 can be included as well). Chemical shift perturbation analysis showed that H1' protons in residues 9 and 25 of ai5γ_NMR were greatly affected by Mg2+ titration, whereas only the aromatic carbon and nitrogen atoms in residue 24 of PL_NMR were affected. As was observed for the triad region, we suspect that carbon and nitrogen chemical shifts in residue 24 would have also been perturbed in ai5γ_NMR if they had been monitored. However, the shifts of C1' in residues 9 and 25 of PL_NMR were not affected by Mg2+ as was seen for the analogous H1' for ai5γ_NMR. The simulation data clearly account for this difference. Visual inspection of the Na+ density grid for ai5γ_mDAR reveals a large, high density region straddling the minor groove surface of residues 9, 25 and 27 (Figure 2.13, top). Na+ ions in this region are

likely stabilized by the O2 atom of U9, C25, and U27. The positioning of this high density region is therefore quite close to the H1' atoms of these same residues and occupancy analysis using a 5 Å cutoff reveals that these H1' atoms are among the nearest to Na+ during simulation (Figure 2.12). In contrast, no such high density region is observed for PL_mDAR (Figure 2.13, bottom), and the H1' atoms of U9, A25, and C27 are not near Na+ during the simulation (Figure 2.12). It is likely that Na+ binding in this region of PL-D5 is not supported for two reasons: 1) the occurrence of adenine at residue 25, rather than cytosine in ai5γ-D5, places a hydrogen atom in the binding region, rather than an oxygen atom, and therefore does not support metal binding and 2) replacement of the A8 – U27 base pair in ai5γ-D5 with the G8 – C27 base for PL-D5 produces a stronger base pair and does not allow the twisting conformation that permits ai5γ-D5's U27 O2 to interact with a Na+ atom.

The tetraloop region is the third area of high Na+ binding that is identified using grid analysis (Figure 2.14). As noted earlier, the ai5γ_mDAR and PL_mDAR loop structures adopt slightly different conformations (both adopt the altered orientation of the base planes, but ai5γ_mDAR displays the outward orientation of G15 and A18, whereas PL_mDAR adopts the inward orientation). A close inspection of the water and Na+ density in the tetraloop region reveals that these subtle differences in loop geometry lead to differences in ion binding and solvation (Figure 2.15). The solvation patterns from simulation closely match those observed in crystal structures by Correll and Swinger *(77)*. In the case of ai5γ_mDAR, which adopts the outward

orientation, a high density region of water between G15 N3 and A18 N6 likely mediates a hydrogen bonding network (Figure 2.15, top). For PL_mDAR, the inward orientation of G15 and A18 exclude this region of solvation, but a new interaction occurs wherein a water molecule mediates a hydrogen bonding network between G15 N1 and A18 O2P (Figure 2.15, bottom). The change in base orientation of the tetraloop also results in a shift in the uppermost region of Na+ density. For ai5γ_mDAR, the Na+ density is situated between N7 and O6 of G15 and lies directly below A17O2P, which appears to be coordinating the ion (Figure 2.14, top). For PL_mDAR, the Na+ is also located between N7 and O6 of G15, but A17 O2P is more distant and any possible coordination interaction is intervened by a region of water density (Figure 2.14, bottom). In this case, the highly ordered water density that forms around the Na+ density is also very similar to that seen in crystal structures, although the ion presence was not reported *(77)*.

The large simulated Na+ density near G15 explains the Mg2+ induced chemical shifts for G15 in both ai5γ_NMR *(55)* and PL_NMR *(56)* and its location in the major groove is consistent with previous NMR work with cobalt(III) hexamine *(84)*). It is less clear why A16 – G19 C1' shifts are so heavily affected in PL_NMR, whereas only the H1' of G15 is affected for ai5γ_NMR. For A16-A19, inspection of the 2D 1H – 13C spectra for H1' – C1' reveals that the majority of the chemical shift occurs in the carbon dimension (Supplementary Figure 2 in Seetharaman et al. *(56)*) and thus would not be revealed in the ai5γ_NMR results where only 1H shifts were measured. However, a significant shift does

occur in the proton dimension of the PL_NMR G19 H1' – C1' cross peak, while G19 H1' of ai5γ_NMR is unaffected by Mg2+. Another puzzle is why so many atoms of the tetraloop region have large chemical shift perturbations (according to the PL_NMR results) when the ion binding appears to be limited to the major groove according to the simulation Na+ density grid results. Inspection of the medium density solvation shell around the tetraloop region reveals that the base atoms are much more exposed to the bulk solvent than base atoms in the rest of the RNA molecule (data not shown). We propose that this lack of solvent shielding may explain the large Mg2+ induced chemical shift perturbations in the tetraloop.

2.4.8 Simulated annealing with the mDAR restraint sets

To compare the results of our explicit solvent refinement with traditional refinement methods, we performed simulated annealing on the completely extended ai5γ and PL RNA, both in vacuo and with Generalized-Born (GB) implicit solvent, using the mDAR restraint sets. One notable feature of in vacuo results is the presence of sharp kinks near the bulge for both the ai5γ and PL ensembles. Given that these kinks are not observed in the GB ensembles or the explicitly solvated ensembles, these results suggest that in regions with sparse distance restraints, such as the bulge, the lack of a solvation environment can lead to anomalous conformations. Other than the kinked bulge conformation in the in vacuo ensemble, the local conformation of both simulated annealing ensembles are very similar to the explicity solvated

ensemble. In contrast, comparison of the average structural length of the GB ensembles reveals that they are somewhat more extended than the explicit ensembles: for ai5γ_mDAR the average structural lengths are 56.9 and 54.5 Å for the GB and explicit, respectively; for PL_mDAR the average lengths are 58.2 and 53.4 Å. In contrast to what was observed for the explicit ensembles, the pairwise RMSD was higher for ai5γ_mDAR GB ensemble (2.54 Å) than that of the PL_mDAR GB ensemble (0.74 Å). The reasons for these differences likely lie with the representative structure selection method. Whereas the explicit solvation representative structures were chosen by their proximity to the centroid of the major cluster during a long simulation, the simulated annealing structures are simply the lowest energy structures which satisfy restraints from a few hundred simulated annealing cycles. It is possible that performing 10,000 or 20,000 simulated annealing cycles (a similar quantity to the frame count in the explicitly solvated simulations) would produce representative structures that are in better agreement with the explicit results. However this has yet to be investigated.

2.5 Discussion

The results presented in this study are immediately relevant to research in experimental structure determination and, more specifically, to refinement of RNA structure from NMR data. First, for structure refinement projects, we find that currently available MD tools with modern simulation protocols, force fields, inclusion of water and mobile counterions, and longer molecular

dynamics simulations, offer robust environments for probing structural features that may not be adequately modeled by older and more conventional structure refinement techniques. Our restrained simulations of ai5γ-D5 and PL-D5 produced a set of refined structures that differed significantly from the previously published NMR structures and offer new insights into the similarities and differences of these RNA molecules. For instance, the simulation refined ai5γ_mDAR structures are much more compact than the original NMR structures and more closely resemble both the NMR and simulation structures of PL-D5. Moreover, for regions outside the bulge region (which is strongly influenced by sequence, packing and tertiary interactions) the re-refined structures better match the ai5γ-D5 crystal structure (PDB: 1KXK). We also were able to identify and troubleshoot potentially incorrect regional conformations in the conventionally refined structures of both molecules. For instance, we uncovered three problematic long range distance restraints in the ai5γ-D5 bulge, which when removed generated a smoother backbone trajectory in the bulge region, lower RMSD values, and fewer restraint violations. We also found that in three of the five PL-D5 structures, residue 25 was apparently trapped in a partially extruded conformation that upon heat annealing converted to the stacked conformation having lower RMSD values and fewer restraint violations. Noticing the problematic regions required MD simulations orders of magnitude longer than were previously and typically applied in order to highlight trapped or metastable conformations. Finally, the pairwise RMSD values for the simulation ai5γ_mDAR structures are lower than those for PL_mDAR, which is

the reverse of what was found using conventional structure refinement. This is likely due to the greater flexibility in the bulge region of PL-D5 observed during the significantly longer and less-tightly restrained simulations used to re-refine the structure. The need to carefully evaluate the choice of NOE and restraint assignments (to not only check for misnaming, incorrect assignments, and/or potentially anomalous spin diffusion)—coupled with the potential for conformational trapping—suggest that automated NMR refinement of RNA structures remains a challenge. Our results suggest that the refinement of RNA structures requires a careful balance between the strength of the experimental restraints and the influence of the force field, and perhaps most importantly, the resulting structures require careful validation and assessment.

The publication of a self-spliced group IIc intron crystal structure *(59)* reveals that the D5 bulge adopts a different conformation in the context of the entire intron than in isolation. Therefore, while our refinement of the original NMR structures provides new insight into the isolated RNA hairpins, the functional relevance of these structures in the context of the intact intron remains unclear. However, the results clearly suggest that the bulge conformation is sensitive to the surroundings and sequence, and that the re-refinement leads to structures that are more consistent with the available crystal structures. Our simulations also suggest that these hairpin structures, which differ at only three positions, are much more similar than the older conventional refinement techniques indicated. These results also improve our understanding of how differences in primary sequence affect 3D structure, and

provide insight into conformational flexibility, solvation, and ion binding. The isolated D5 bulge apparently samples a range of conformations, while tertiary interactions in the context of the entire intron select and stabilize a flipped out conformation necessary to assemble a catalytically active intron *(59)*. MD simulations with explicit solvent and updated force fields can potentially uncover minor conformers that nevertheless have functional relevance. The flipped out conformation need not be the lowest energy structure observed by NMR, merely accessible with sufficient frequency to be captured by tertiary contacts in the intron. Accurate "ground state" structures determined from NMR data and explicit solvation MD provide a starting point from which to investigate the dynamical behavior of RNA. The methods employed in this paper should also aid researchers who use structural databases to further refine their models.

Remaining unanswered questions relate to potential disorder and/or dynamics in the bulge and loop regions. MD simulations without experimentally derived restraints are unable to maintain the expected structure; given this, movement away from the experimental structure does not represent true dynamics, but suggests deficiencies in the force field. Researchers may be interested in the minimal set of restraints required to maintain experimentally valid structures. We propose that the minimal restraint requirements for maintaining accurate structure using the AMBER force field would include as many distance restraints in noncanonical regions as possible and orientational restraints such as RDCs to maintain proper structural compaction. Although

structures consistent with the experimental data can be maintained via the application of restraints, such restraints will tend to inhibit conformational transitions and dynamics. Given this, it is unclear if the representative structures found in re-refinement completely represent the ensemble of frequently accessed conformations or simply represent the lowest energy structures without a proper depiction of the true disorder or dynamics sampled at room temperature. For instance, depending on the choice of experimental restraints, two GAAA tetraloop conformations, each with distinct solvation and ion binding properties were observed. Both are consistent with experiment. However, as exchange was not observed between “outward” and “inward” structures during the re-refinement, speculation on the relative populations or conformational dynamics is not possible. On the other hand, given that both are observed experimentally and are likely nearly isoenergetic, it is likely that both are populated and in dynamic equilibrium. To further resolve these questions through simulation will require improvements in the underlying nucleic acid force fields.

The AMBER force field is continually being developed and refined to produce improved simulation results. Generally these improvements are evaluated in the context of unrestrained simulations. However, evaluating force field performance is difficult given the huge diversity in RNA structure. Our unpublished work suggests that canonical A-form RNA is relatively stable in unrestrained simulations for long periods of time. In contrast, noncanonical regions frequently populate conformations which are not observed in

experimental structures suggesting a force field problem. One notable RNA motif for which this occurs is the UUCG tetraloop. Progress towards improving RNA torsional parameters is underway, including recent force field modifications that improve the glycosidic x in RNA *(40, 45, 46)*.

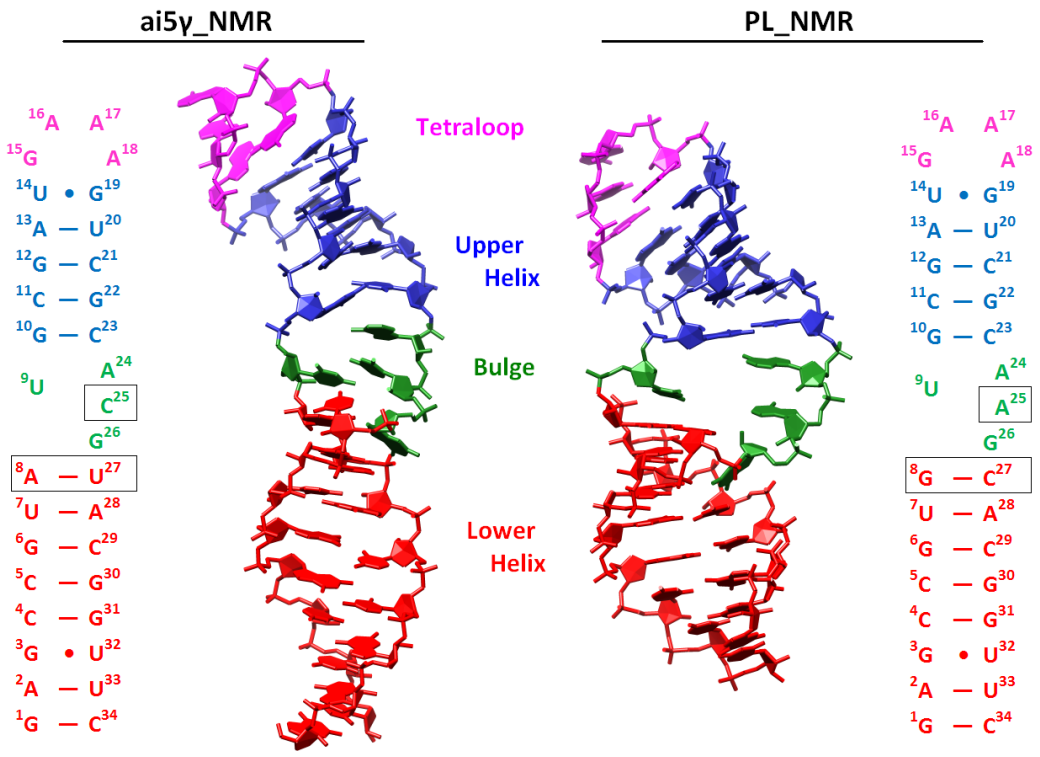

Figure 2.1. Comparison of the secondary and 3D structures for domain 5 of the Group II Intron from yeast ai5γ (ai5γ_NMR, PDB code 1R2P, *left*) and *Pylaiella littoralis* (PL_NMR, PDB code 2F88, *right*). The structural elements are indicated by color: tetraloop (pink), upper helix (blue), bulge (green), lower helix (red). Differences in sequence are boxed.

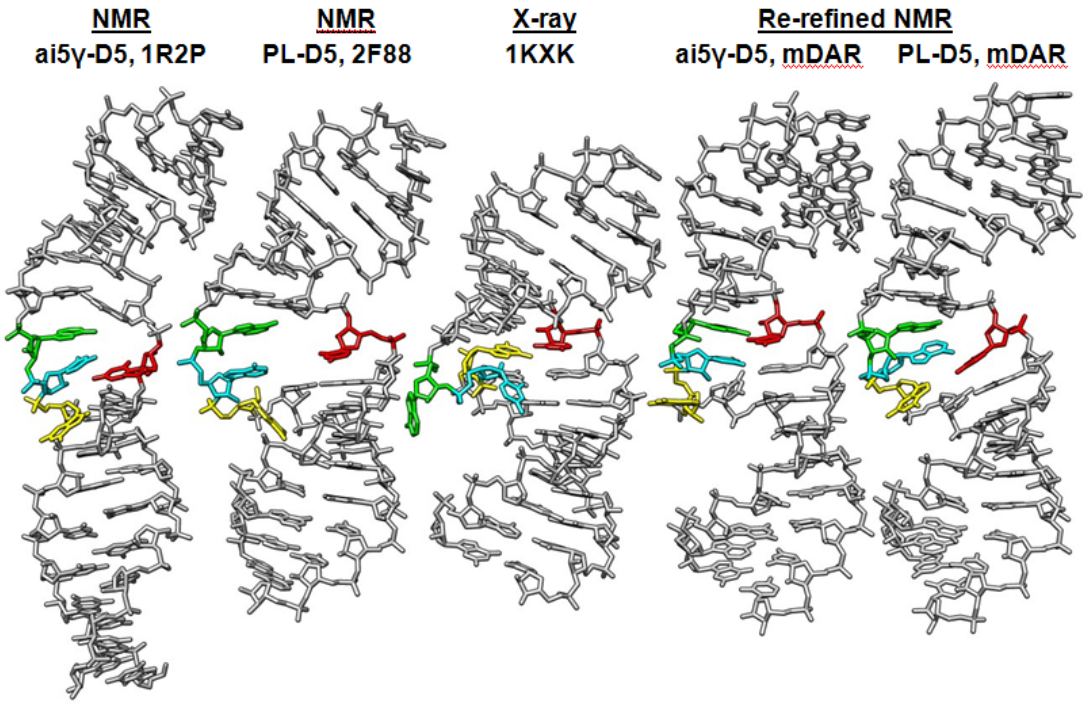

Figure 2.2. Molecular graphics representations of the heavy atoms of the previously and newly refined 34-residue portions of the domain 5 group II intron structures highlighting the differences in the bulge region. Shown are, from left to right, the earlier yeast ai5γ *(55)* (ai5γ-D5, PDB: 1R2P), the *Pylaiella littoralis (56)* (PL-D5, PDB: 2F88), a portion of the earlier crystal structure of ai5γ-D5 *(58)*, the first structure of the mDAR/DxAR (i.e., refined including distance and angle restraints, residual dipolar coupling restraints, and removal of a few bad NOE restraints due to spin diffusion as per the main text) re-refinement of ai5γ, and the first structure of the mDAR/DAR-heat (i.e. refined with all of the restraint information and an additional heating step for better sampling as per the main text) re-refinement of PL-D5. G26 is colored in yellow, C25 in cyan, A24 in green and U9 is in red.

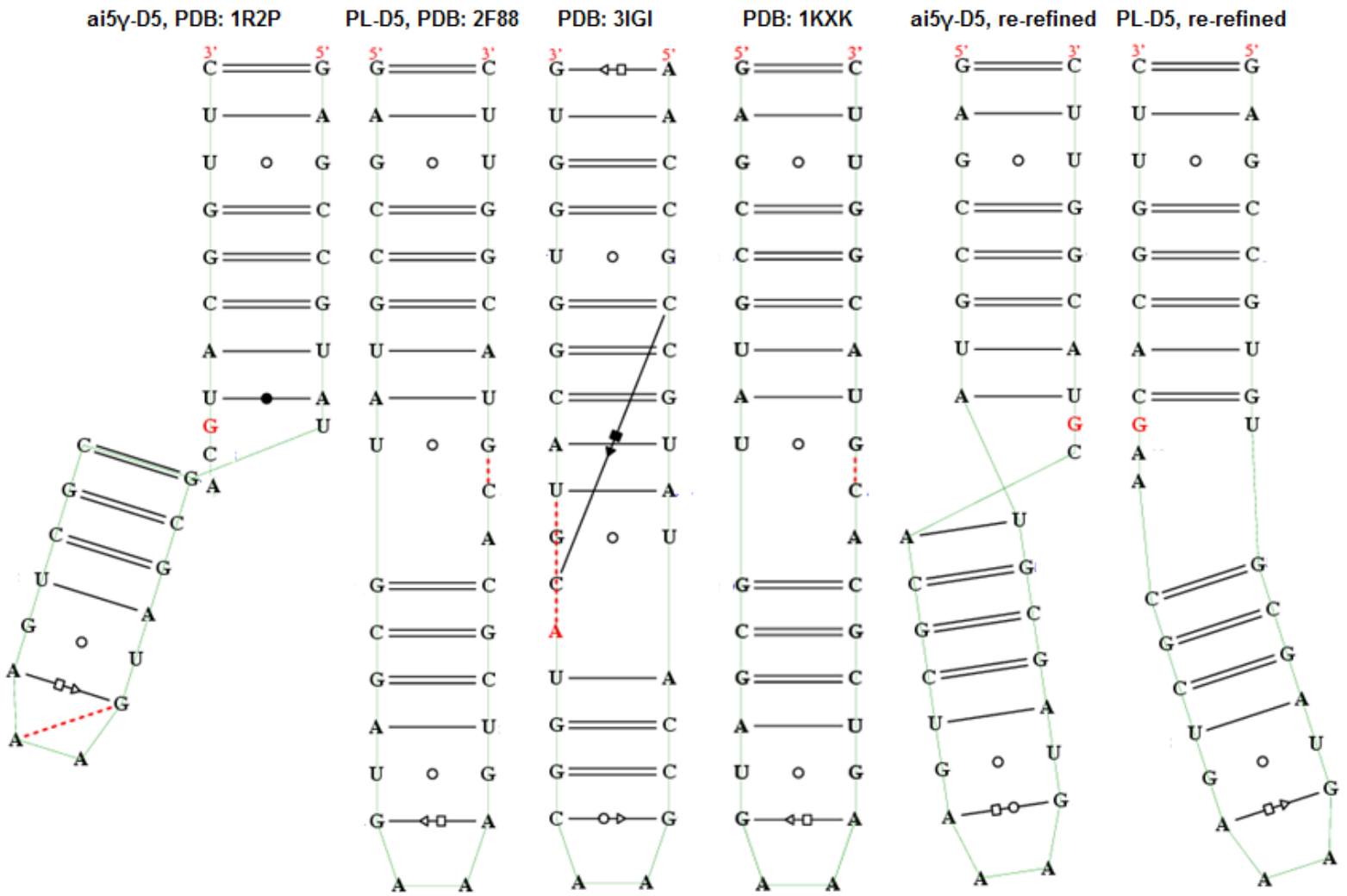

Figure 2.3 RNAML 2-D schematics of the annotated secondary structure of the various older and re-refined structures highlighting the significant differences in the bulge regions.

Table 2.1. Simulation details and nomenclature for refinements in this work.

Name	No. of Simulations	Simulation Notes	Time†
ai5γ_UR	5	Unrestrained simulation	111 ns
ai5γ_DA	5	Distance and angle restraints enforced	20 ns
ai5γ_DAR	5	Distance, angle, and RDC restraints enforced	8 ns*
ai5γ_mDA	5	Modified distance and angle restraints enforced	20 ns
ai5γ_mDAR	5	Modified distance, angle and RDC restraints enforced	8 ns*
PL_UR	5	Unrestrained simulation	92 ns
PL_DA	5	Modified distance and angle restraints enforced	20 ns
PL_DAR	5	Distance, angle, and RDC restraints enforced	8 ns*
PL_mDA	5	Modified distance and angle restraints enforced; additional heating step	23 ns
PL_mDAR	5	Modified distance, angle, and RDC restraints enforced	11 ns*

†The time listed is for the minimum trajectory length of the five models.
*DAR simulations started from the final frame of corresponding DA simulation.

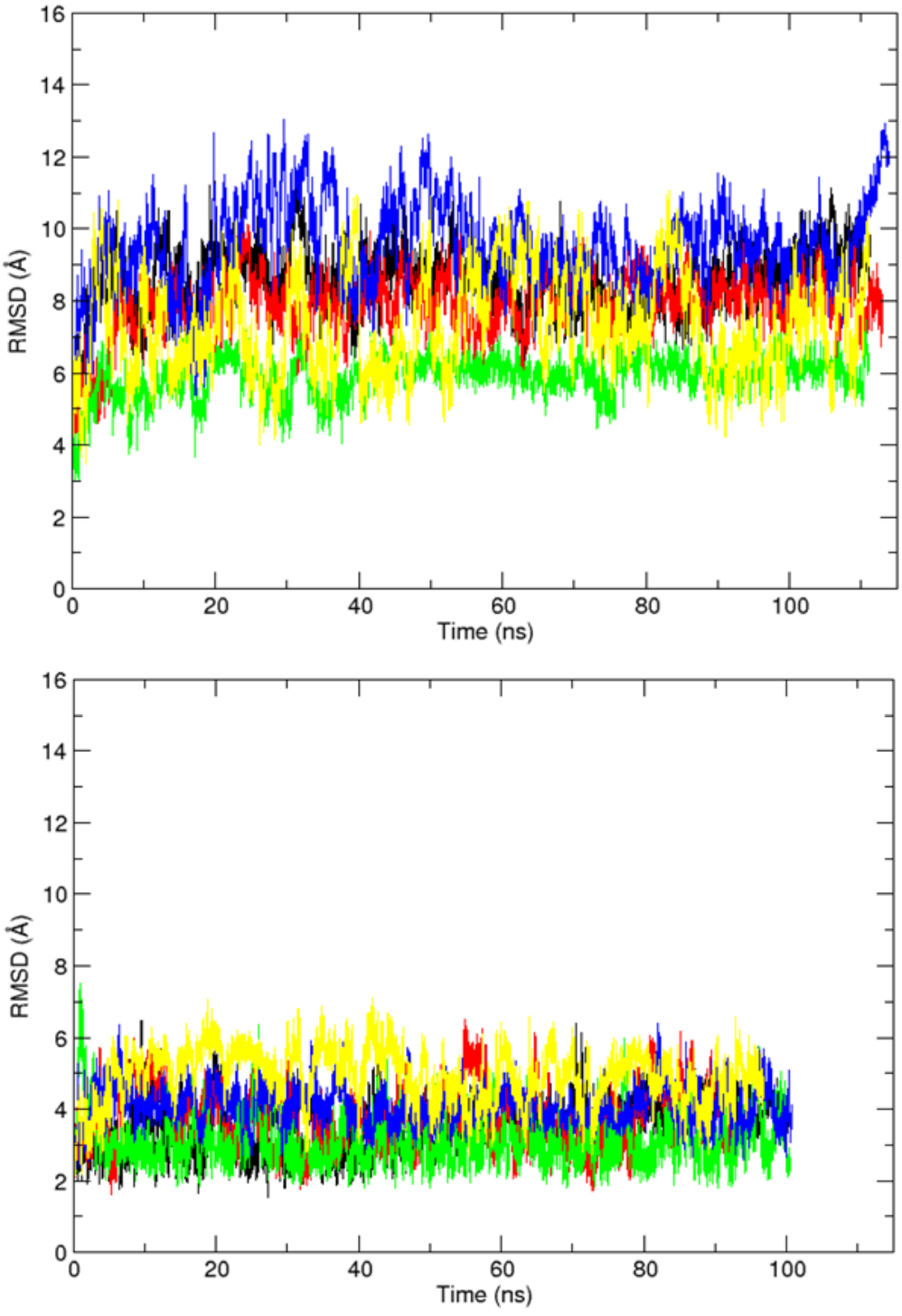

Figure 2.4. RMSD plot for the five unrestrained (UR) simulations of each PDB structure. PDB 1R2P or ai5γ_UR (*top*) and PDB 2F88 or PL_UR (*bottom*) simulations. RMSD values were calculated by fit to the initial structure. Although the RMSD values plateau, the RMSd values are relatively high indicating motion away from the starting structure.

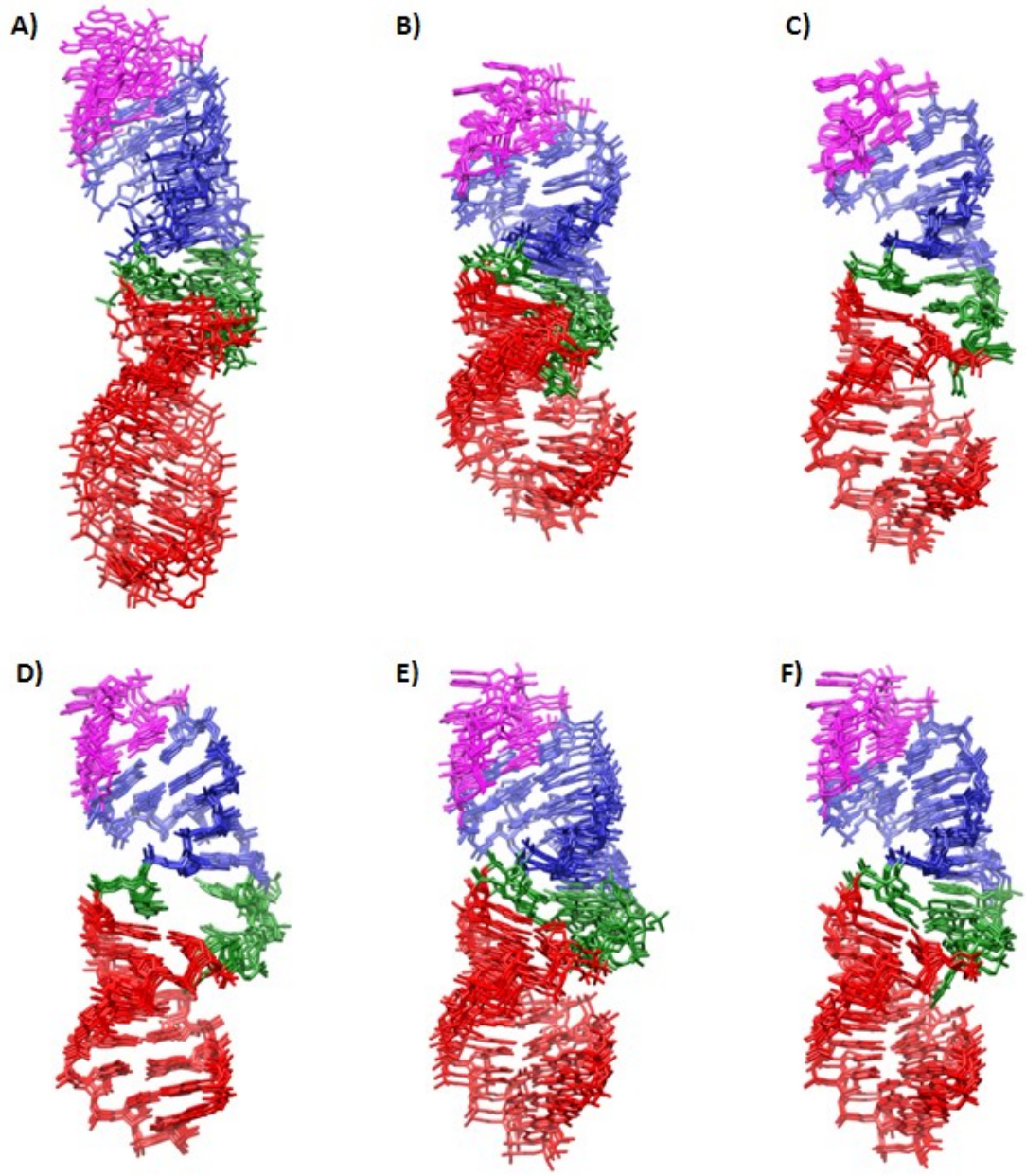

Figure 2.5. Representative structures for *A*) ai5γ_NMR, *B*) ai5γ_DAR, *C*) ai5γ_mDAR, *D*) PL_NMR, *E*) PL_DAR, and *F*) PL_mDAR. The "NMR" suffix denotes the original ensemble from earlier refinement (ai5γ_NMR is PDB: 1R2P, PL_NMR is PDB: 2F88). "DAR" refers to the representative structure from the dominant ensemble sampled during the five independent explicit water MD simulations with distance, torsion angle, and residual dipolar coupling restraints enforced. "mDAR" is identical to "DAR" except with small modifications to the restraint list or the equilibration protocol as discussed in the main text. See Table 2.1 for details.

Table 2.2. Structural statistics for ai5γ-D5 and PL-D5 representative structures.

	ai5γ-D5			PL-D5		
	NMR	DAR	mDAR	NMR	DAR	mDAR
No. of Structures	5	5	5	5	5	5
No. of Distance Restraints	595	595	592	549	549	543
No. of Dihedral Restraints	238	238	238	247	247	247
No. of RDC Restraints	24	24	24	37	37	37
Avg. RMSd of Distance (Å)	0.021	0.020	0.016	0.019	0.028	0.013
Avg. RMSd of Dihedral (°)	0.482	0.378	0.365	0.703	0.152	0.239
Avg. RMSd of RDC (Hz)	1.897	2.939	3.067	3.115	4.324	4.390
No. Distance Viol. >0.2 Å	4	5	0	2	13	0
No. Angle Viol >5.0 °	5	4	3	7	0	1
Overall Heavy Atom RMSD						
Avg. RMSd from mean	2.58±0.84	1.41±0.29	0.66±0.32	0.83±0.12	1.69±0.45	1.14±0.28
Avg. RMSd pairwise	4.09±1.20	2.10±0.87	1.04±0.46	1.30±0.24	2.56±1.04	1.70±0.76

Note: The corresponding structures are depicted in Figure 2.5. Definitions of NMR, DAR, and mDAR as per Table 2.1.

Table 2.3. Pairwise RMSD measurements.

Comparison	Pairwise RMSD
ai5γ_NMR vs. PL_NMR	6.03 ± 1.00
ai5γ_NMR vs. ai5γ_DAR	7.18 ± 1.52
ai5γ_NMR vs. ai5γ_mDAR	5.61 ± 1.22
ai5γ_NMR vs. PL_DAR	6.55 ± 1.69
ai5γ_NMR vs PL_mDAR	6.18 ± 1.34
PL_NMR vs. ai5γ_DAR	4.29 ± 0.52
PL_NMR vs. ai5γ_mDAR	2.71 ± 0.15
PL_NMR vs. PL_DAR	2.67 ± 0.63
PL_NMR vs. PL_mDAR	2.21 ± 0.40
ai5γ_DAR vs. PL_DAR	3.33 ± 0.61
ai5γ_DAR vs. PL_mDAR	3.23 ± 0.70
ai5γ_mDAR vs. PL_DAR	2.33 ± 0.82
ai5γ_mDAR vs. PL_mDAR	1.95 ± 0.40
ai5γ_DAR vs. ai5γ_mDAR	3.13 ± 0.66
PL_DAR vs. PL_mDAR	2.00 ± 0.88

Note: ensemble definitions as per Table 2.1.

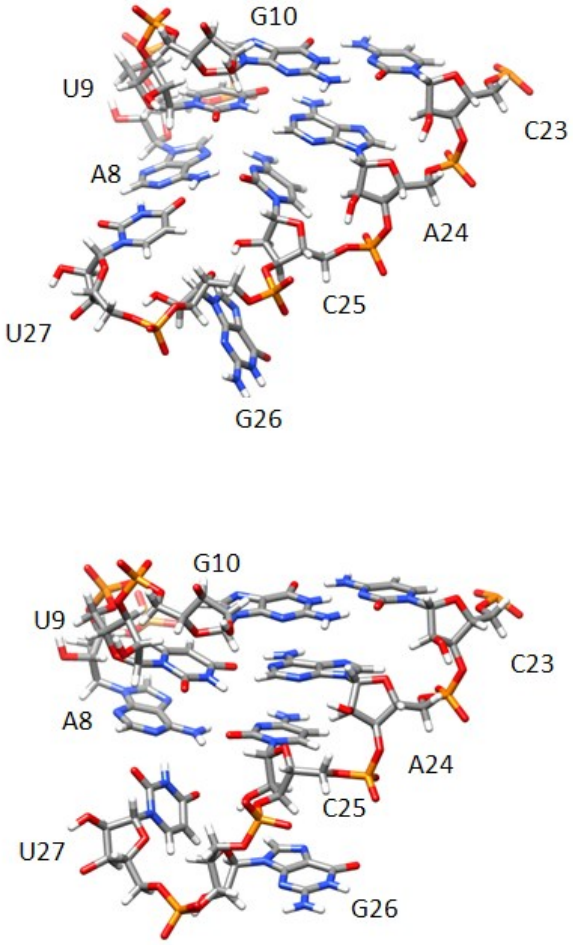

Figure 2.6. The bulge structures from the re-refined NMR structures: ai5γ_DAR (*top*) and ai5γ_mDAR (*bottom*).

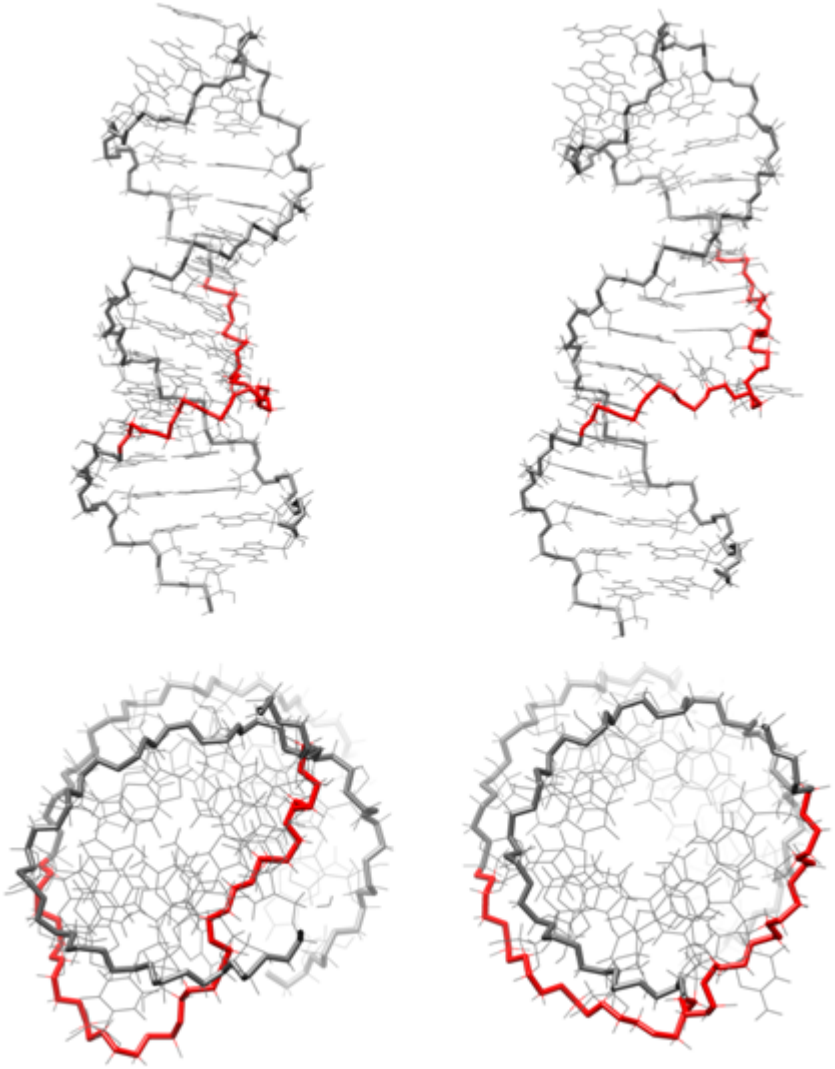

Figure 2.7. Side and top views of ai5γ_DAR (*left*) and ai5γ_mDAR (*right*). The top view has been truncated to focus on the bulge region. The backbone near the kink region has been highlighted red.

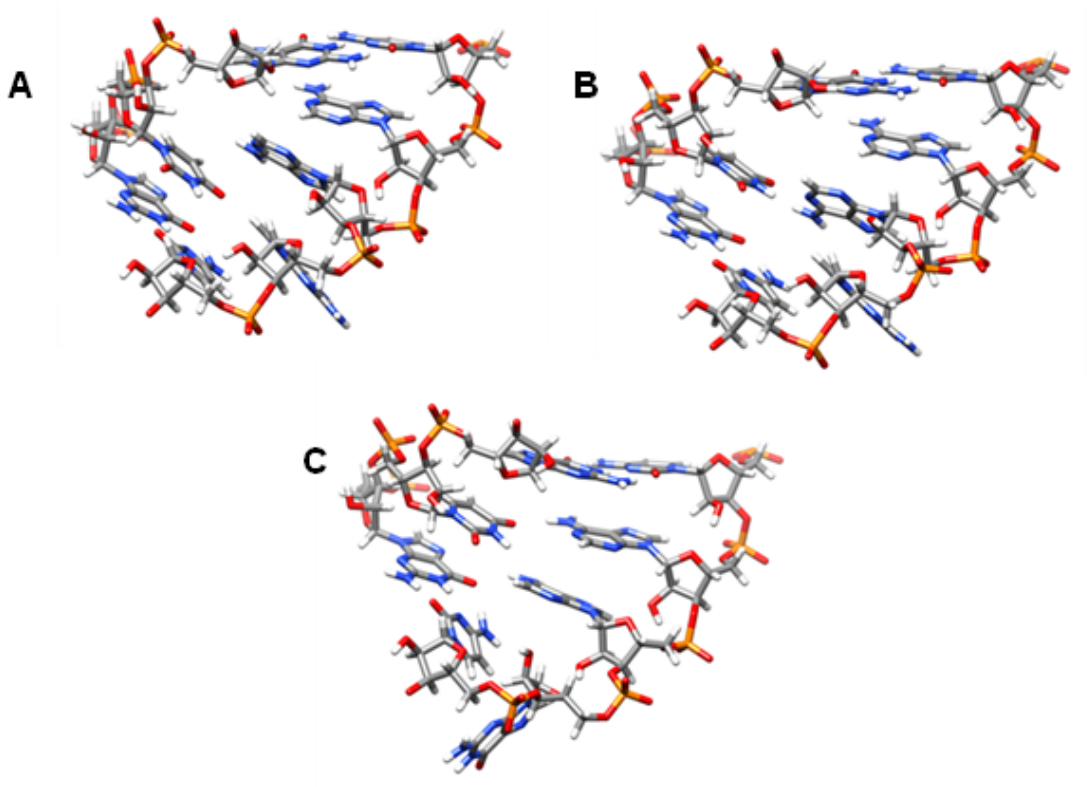

Figure 2.8. The three bulge conformations adopted by PL_mDAR.

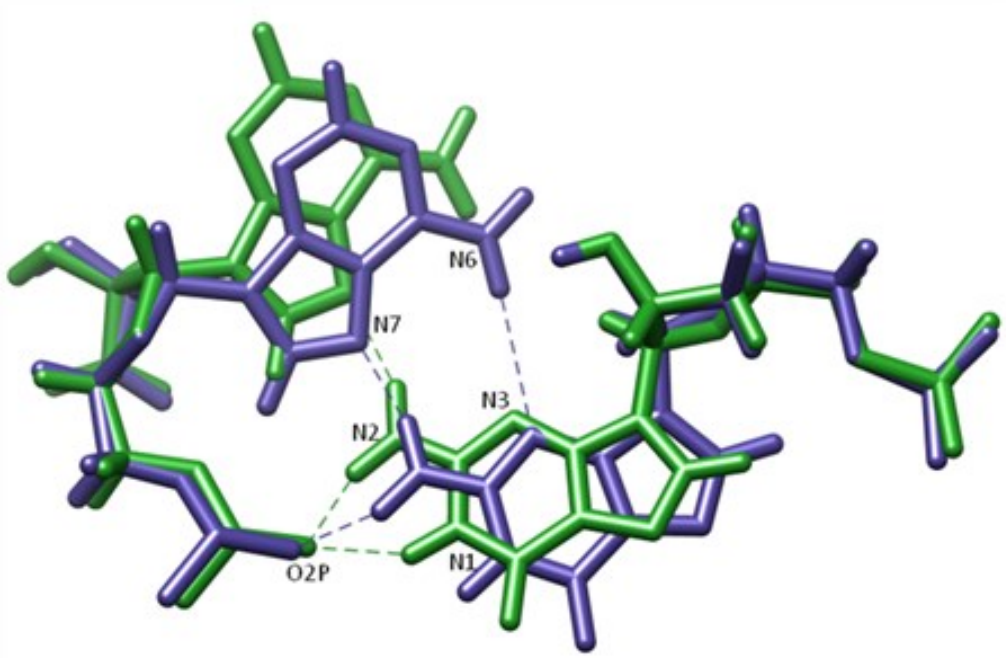

Figure 2.9. A comparison of the outward orientation (*green*) and inward orientation (*blue*) of the G15 – A18 basepair in the GAAA tetraloop.

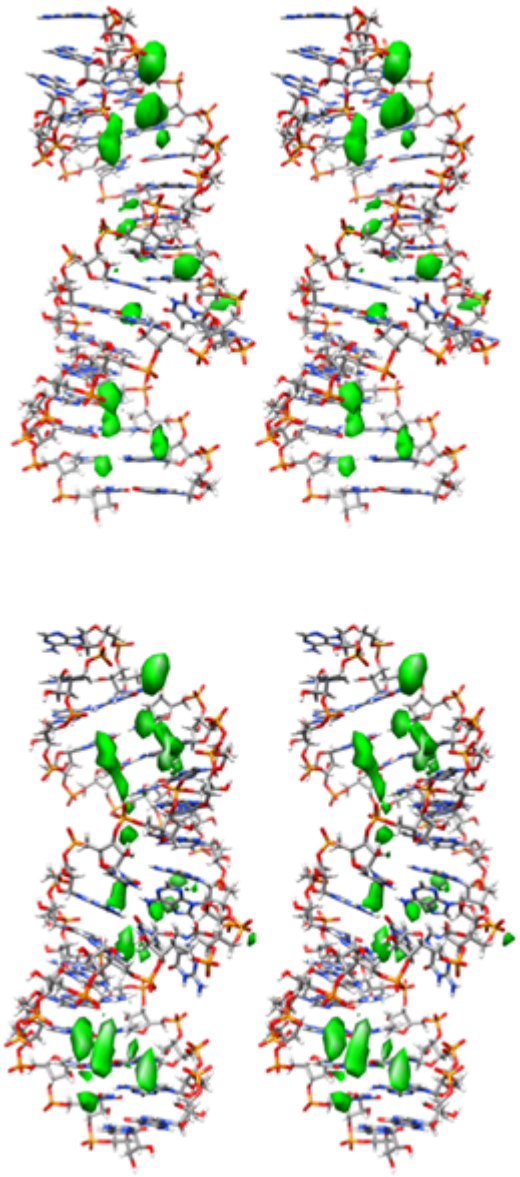

Figure 2.10. Stereo views of the Na^+ density grid maps for ai5γ_mDAR (*top*) and PL_mDAR (*bottom*). The isosurface grid density was chosen to show regions of ion localization which were higher than background levels.

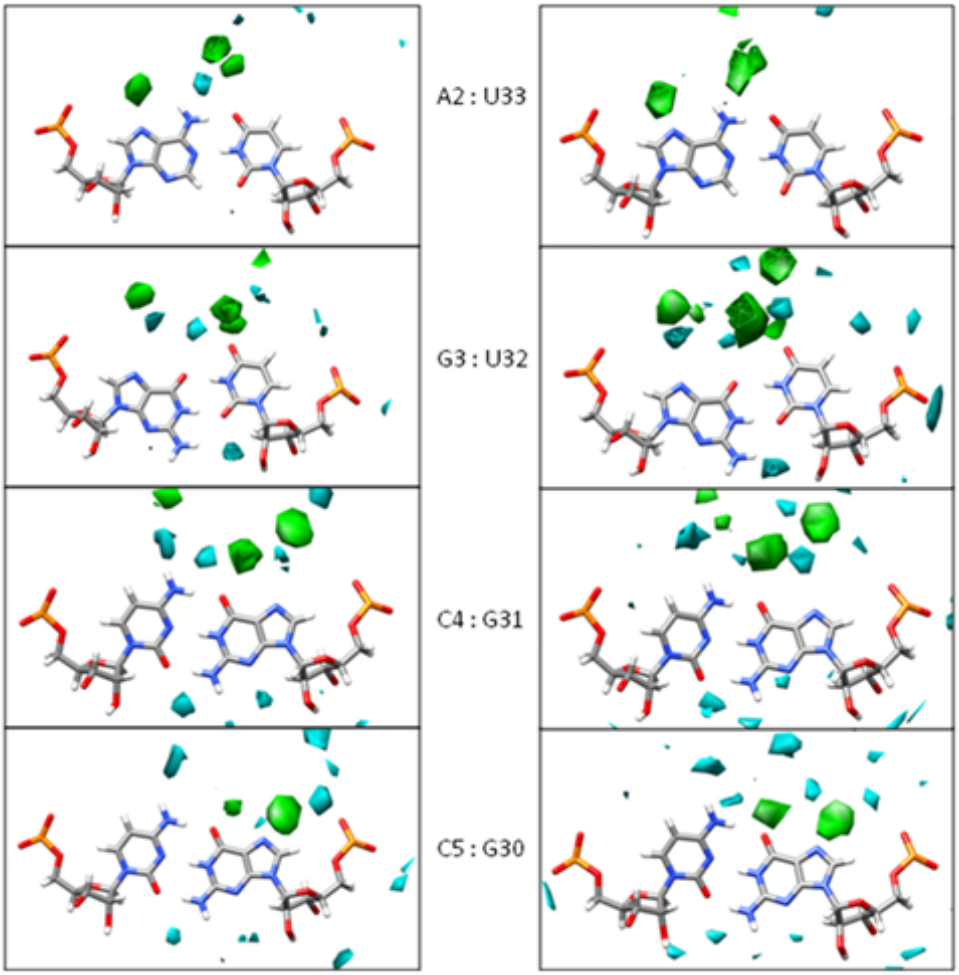

Figure 2.11. Comparison of water and Na^+ densities in the major groove of selected base pairs for ai5γ_mDAR (*left*) and PL_mDAR (*right*).

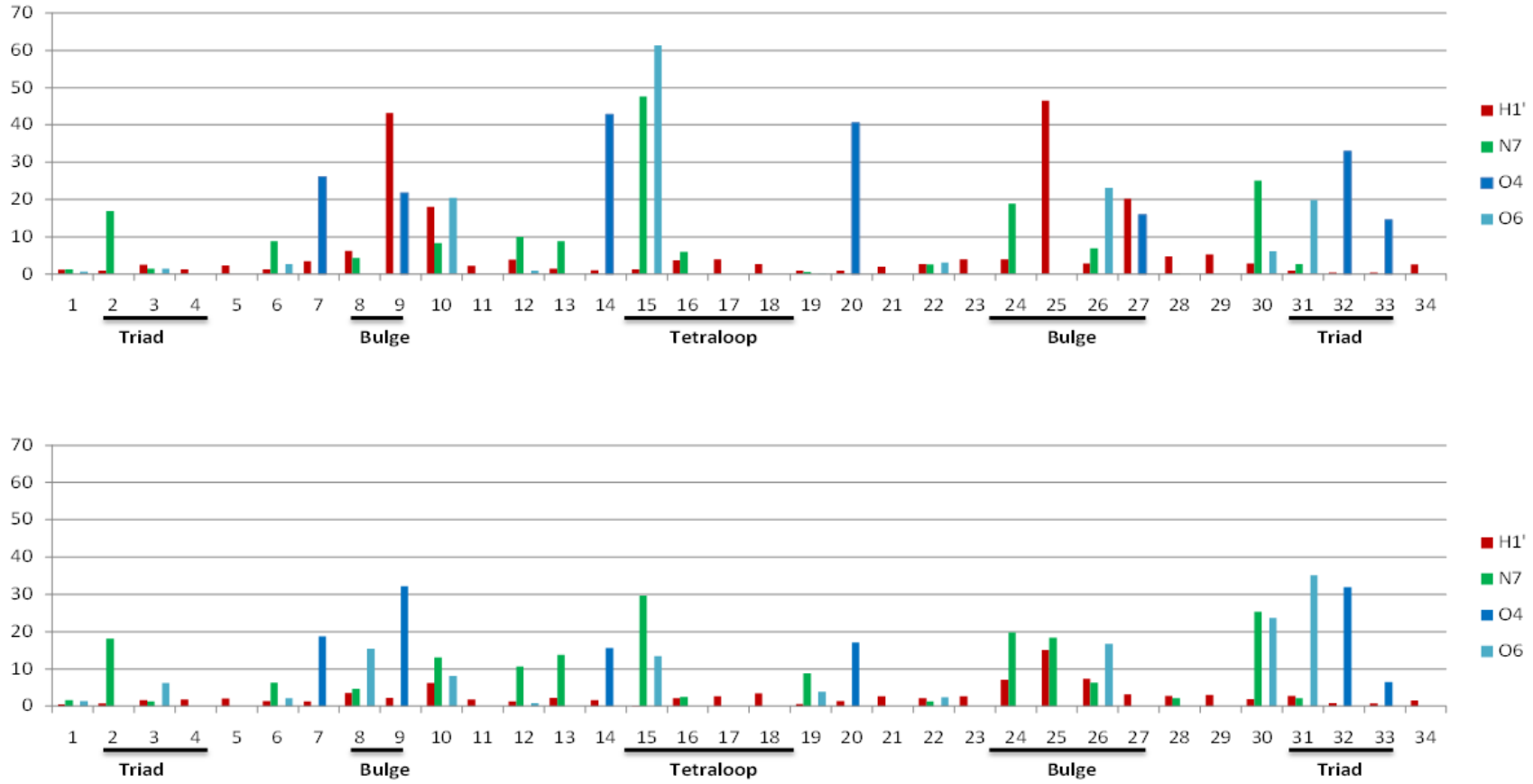

Figure 2.12. Percent occupation of Na^+ near selected atoms for all bases. Data is shown for ai5γ_mDAR (*top*) and PL_mDAR (*bottom*). The cutoff for H1' was 5 Å. The cutoff for N7, O4, and O6 was 2.8 Å.

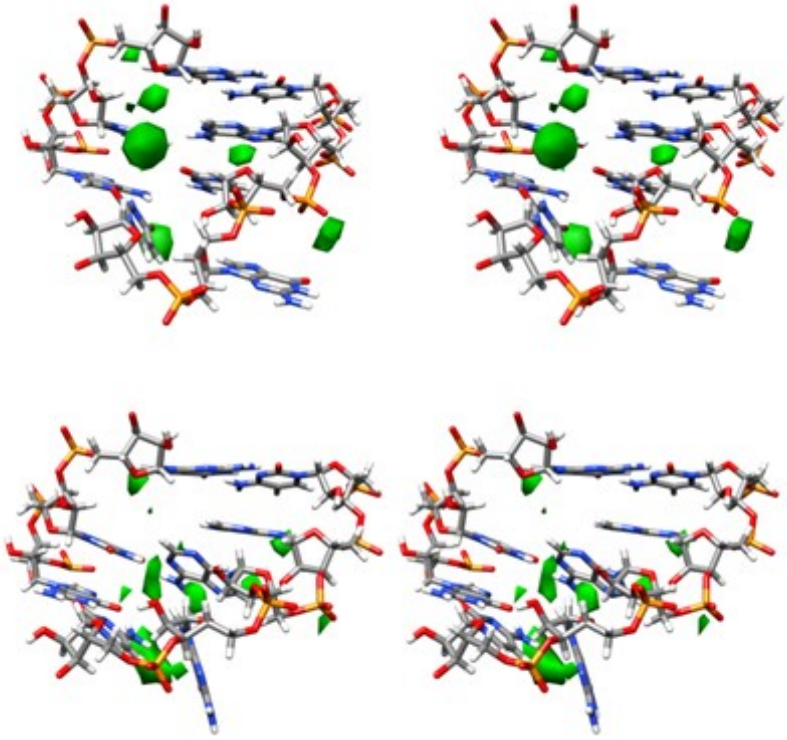

Figure 2.13. Stereo view of the Na^+ density grid in the bulge region of ai5γ_mDAR (top) and PL_mDAR (bottom).

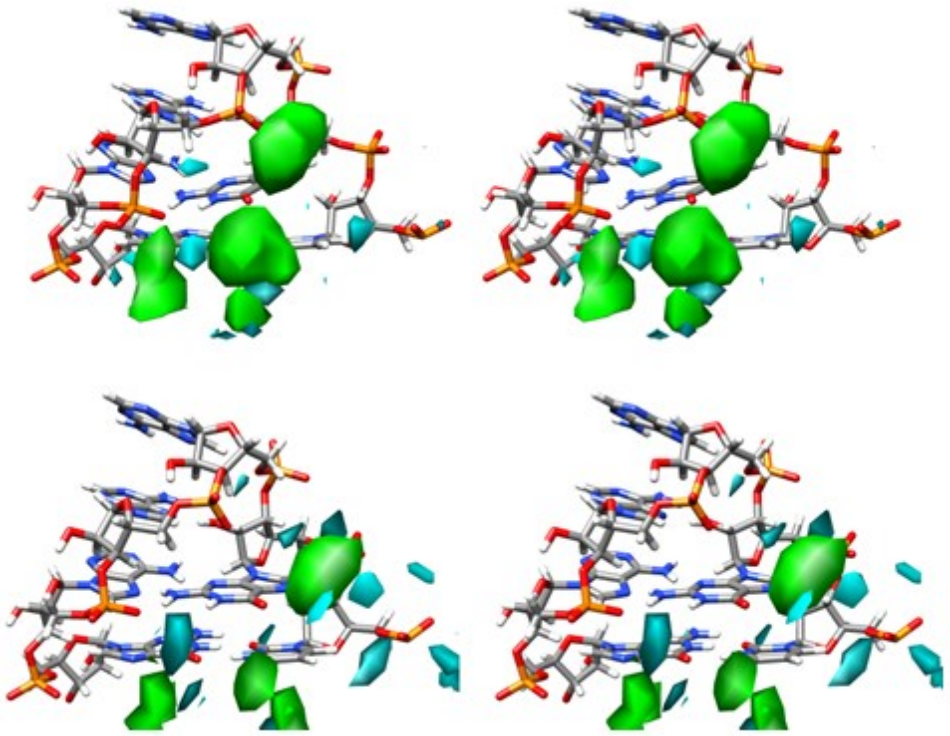

Figure 2.14. Stereo view of the Na^+ density grid in the loop region of ai5γ_mDAR (*top*) and PL_mDAR (*bottom*).

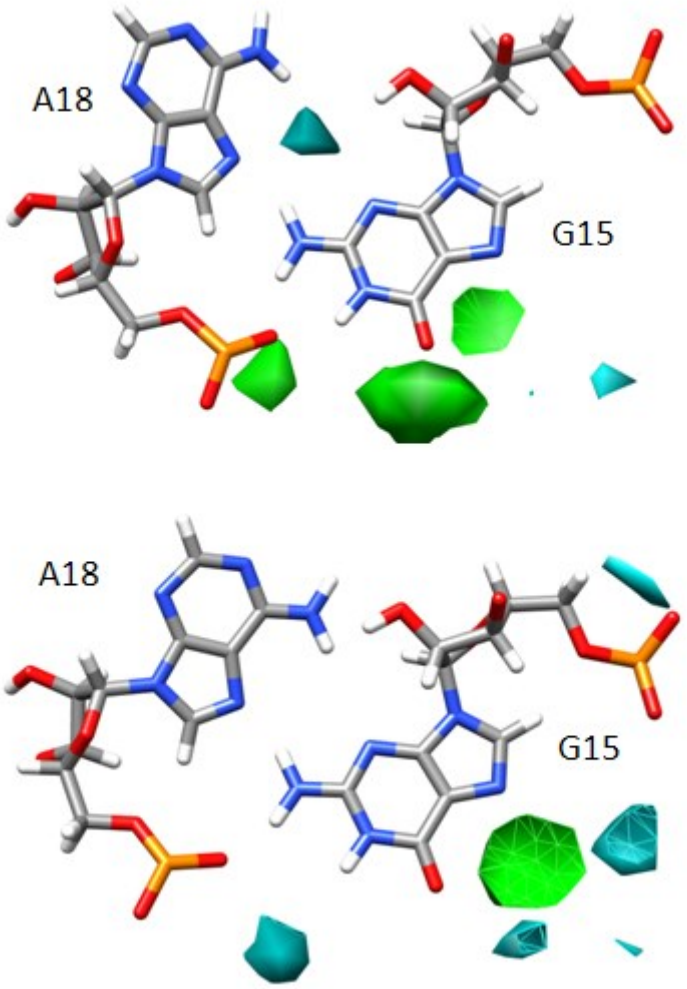

Figure 2.15. Water and Na^+ density grid in the region near the non-Watson-Crick G-A base pair of ai5γ_mDAR (*top*) and PL_mDAR (*bottom*).

2.6 References

1. Clore, G. M., Gronenborn, A. M., Brunger, A. T., and Karplus, M. (1985) Solution conformation of a heptadecapeptide comprising the DNA binding helix F of the cyclic AMP receptor protein of Escherichia coil. Combined use of 1H nuclear magnetic resonance and restrained molecular dynamics, *J. Mol. Biol. 186*, 435-455.

2. Brunger, A. T., Campbell, R. L., Clore, G. M., Gronenborn, A. M., Karplus, M., Petsko, G. A., and Teeter, M. M. (1987) Solution of a protein crystal structure with a model obtained from NMR interproton distance restraints, *Science 235*, 1049-1053.

3. Brunger, A. T., Kuriyan, J., and Karplus, M. (1987) Crystallographic R factor refinement by molecular dynamics, *Science 235*, 458-460.

4. Nilges, M. (1996) Structure calculation from NMR data, *Curr. Opin. Struct. Biol. 6*, 617-623.

5. Brunger, A. T., and Adams, P. D. (2002) Molecular dynamics applied to X-ray structure refinement, *Acc. Chem. Res. 35*, 404-412.

6. Brunger, A. T. (1992) X-PLOR, Version 3.1. A system for X-ray crystallography and NMR, In *Yale Univ. Press*, New Haven.

7. Schwieters, C. D., Kuszewski, J. J., and Clore, G. M. (2006) Using Xplor-NIH for NMR molecular structure determination, *Progr. NMR Spect. 48*, 47-62.

8. Schweiters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003) The Xplor-NIH molecular structure determination package, *J. Magn. Reson. 160*, 66-74.

9. Brunger, A. T., Adams, P. D., Clore, G. M., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J. J., Nilges, M., Pannu, N. S., Read, R. J., et al. (1998) Crystallography & NMR system (CNS), A new software suite for macromolecular structure determination, *Acta. Cryst. D 54*, 905-921.

10. Brunger, A. T. (2007) Version 1.2 of the crystallography and NMR system, *Nat. Protoc. 2*, 2728-2733.

11. Konerding, D. E., Cheatham, T. E., III, Kollman, P. A., and James, T. L. (1999) Restrained molecular dynamics of solvated duplex DNA using the particle mesh Ewald method, *J. Biomol. NMR 13*, 119-131.

12. Prompers, J. J., Folmer, R. H., Nilges, M., Folkers, P. J., Konings, R. N., and Hilbers, C. W. (1995) Refined solution structure of the Tyr41->His mutant of the M13 gene V protein. A comparison with the crystal structure, *Eur. J.*

Biochem. 232, 506-514.

13. Kordel, J., Pearlman, D. A., and Chazin, W. J. (1997) Protein solution structure calculations in solution: solvated molecular dynamics refinement of calbindin D9k, *J. Biomol. NMR 10*, 231-243.

14. Linge, J. P., and Nilges, M. (1999) Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation, *J. Biomol. NMR 13*, 51-59.

15. Gouda, H., Yamazaki, K., Hasegawa, J., and Hirono, S. (2001) Refinement of the NMR structures of alpha-conotoxin M using molecular dynamics simulation with explicit water and a full molecular force field, *Chem. Pharm. Bull (Toyko) 49*, 249-252.

16. Spronk, C. A., Linge, J. P., Hilbers, C. W., and Vuister, G. W. (2002) Improving the quality of protein structures derived by NMR spectroscopy, *J. Biomol. NMR 22*, 281-289.

17. Xia, B., Tsui, V., Case, D. A., Dyson, H. J., and Wright, P. E. (2002) Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water, *J. Biomol. NMR 22*, 317-331.

18. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M., and Nilges, M. (2003) Refinement of protein structures in explicit solvent, *Proteins 50*, 496-506.

19. Dolenc, J., Missimer, J., Steinmetz, M., and van Gunsteren, W. (2010) Methods of NMR structure refinement: molecular dynamics simulations improve the agreement with measured NMR data of a C-terminal peptide of GCN4-p1, *J. Biomol. NMR 47*, 221.

20. Nozinovic, S., Fürtig, B., Jonker, H. R. A., Richter, C., and Schwalbe, H. (2010) High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA, *Nucleic Acids Res. 38*, 683.

21. Rettig, M., Langel, W., Kamal, A., and Weisz, K. (2010) NMR structural studies on the covalent DNA binding of a pyrrolobenzodiazepine-naphthalimide conjugate, *Org. Biomol. Chem. 8*, 3179-3187.

22. Calhoun, J. R., Liu, W., Spiegel, K., Dal Peraro, M., Klein, M. L., Valentine, K. G., Wand, A. J., and DeGrado, W. F. (2008) Solution NMR structure of a designed metalloprotein and complementary molecular dynamics refinement, *Structure 16*, 210-215.

23. Paulsen, R. B., Seth, P. P., Swayze, E. E., Griffey, R. H., Skalicky, J. J.,

Cheatham, T. E., 3rd, and Davis, D. R. (2010) Inhibitor-induced structural change in the HCV IRES domain IIa RNA, *Proc. Natl. Acad. Sci. U. S. A. 107*, 7263-7268.

24. Al-Hashimi, H. M., and Walter, N. G. (2008) RNA dynamics: it is about time, *Curr. Opin. Struct. Biol. 18*, 321-329.

25. Hall, K. B. (2008) RNA in motion, *Curr. Opin. Chem. Biol. 12*, 612-618.

26. Baird, N. J., and Ferre-D'Amare, A. R. (2010) Idiosyncratically tuned switching behavior of riboswitch aptamer domains revealed by comparative small-angle X-ray scattering analysis, *RNA 16*, 598-609.

27. Solomatin, S. V., Greenfeld, M., Chu, S., and Herschlag, D. (2010) Multiple native states reveal persistent ruggedness of an RNA folding landscape, *Nature 463*, 681-684.

28. Stelzer, A. C., Frank, A. T., Kratz, J. D., Swanson, M. D., Gonzalez-Hernandez, M. J., Lee, J., Andricioaei, I., Markovitz, D. M., and Al-Hashimi, H. M. (2011) Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble, *Nat. Chem. Biol. 7*, 553-559.

29. James, T. L. (2001) NMR determination of oligonucleotide structure, *Curr. Protoc. Nucleic Acid Chem. Chapter 7*, Unit 7 2.

30. Torda, A. E., Scheek, R. M., and van Gunsteren, W. F. (1990) Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat, *J. Mol. Biol. 214*, 223-235.

31. Pearlman, D. A., and Kollman, P. A. (1991) Are time-averaged restraints necessary for nuclear magnetic resonance refinement? A model study for DNA, *J. Mol. Biol. 220*, 457-479.

32. Schmitz, U., Ulyanov, N. B., Kumar, A., and James, T. L. (1993) Molecular dynamics with weighted time-averaged restraints for a DNA octamer. Dynamic interpretation of nuclear magnetic resonance data, *J. Mol. Biol. 234*, 373-389.

33. Gorler, A., Ulyanov, N. B., and James, T. L. (2000) Determination of the populations and structures of multiple conformers in an ensemble from NMR data: multiple-copy refinement of nucleic acid structures using floating weights, *J. Biomol. NMR 16*, 147-164.

34. Bonvin, A. M., and Brunger, A. T. (1995) Conformational variability of solution nuclear magnetic resonance structures, *J. Mol. Biol. 250*, 80-93.

35. Schwieters, C. D., and Clore, G. M. (2007) A physical picture of atomic

motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data, *Biochemistry 46*, 1152-1166.

36. Hashem, Y., and Auffinger, P. (2009) A short guide for molecular dynamics simulations of RNA systems, *Methods 47*, 187-197.

37. McDowell, S. E., Spackova, N., Sponer, J., and Walter, N. G. (2007) Molecular dynamics simulations of RNA: an in silico single molecule approach, *Biopolymers 85*, 169-184.

38. Besseova, I., Otyepka, M., Reblova, K., and Sponer, J. (2009) Dependence of A-RNA simulations on the choice of the force field and salt strength, *Phys. Chem. Chem. Phys. 11*, 10701-10711.

39. Deng, N. J., and Cieplak, P. (2010) Free energy profile of RNA hairpins: a molecular dynamics simulation study, *Biophys. J. 98*, 627-636.

40. Banas, P., Hollas, D., Zagarbova, M., Jurecka, P., Orozco, M., Cheatham, T. E., III, Sponer, J., and Otyepka, M. (2010) Performance of molecular mechanics force fields for RNA simulations. Stability of UUCG and GNRA hairpins, *J. Chem. Theory Comput. 6*, 3836-3849.

41. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc. 117*, 5179-5197.

42. Cheatham, T. E., III, Cieplak, P., and Kollman, P. A. (1999) A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat, *J. Biomol. Struct. Dyn. 16*, 845-862.

43. Wang, J., Cieplak, P., and Kollman, P. A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?, *J. Comput. Chem. 21*, 1049-1074.

44. Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham, T. E., III, Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids. Improving the description of alpha/gamma conformers, *Biophys. J. 11*, 3817-3829.

45. Yildirim, I., Stern, H. A., Kennedy, S. D., Tubbs, J. D., and Turner, D. H. (2010) Reparameterization of RNA chi x torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine, *J. Chem. Theory Comput. 6*, 1520-1531.

46. Zgarbová, M., Otyepka, M., Sponer, J. i., Mládek, A. t., Banáš, P., Cheatham, T. E., and Jurečka, P. (2011) Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles, *J. Chem. Theory Comp. 7*, 2886-2902.

47. Csaszar, K., Spackova, N., Stefl, R., Sponer, J., and Leontis, N. B. (2001) Molecular dynamics of the frame-shifting pseudoknot from beet western yellows virus: the role of non-Watson-Crick base-pairing, ordered hydration, cation binding and base mutations on stability and unfolding, *J. Mol. Biol. 313*, 1073-1091.

48. Reblova, K., Spackova, N., Stefl, R., Csaszar, K., Koca, J., Leontis, N. B., and Sponer, J. (2003) Non-Watson-Crick basepairing and hydration in RNA motifs: molecular dynamics of 5S rRNA loop E, *Biophys. J. 84*, 3564-3582.

49. Beveridge, D. L., Barreiro, G., Byun, K. S., Case, D. A., Cheatham, T. E., III, Dixit, S. B., Giudice, E., Lankas, F., Lavery, R., Maddocks, J. H., et al. (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps, *Biophys. J. 87*, 3799-3813.

50. Dixit, S. B., Beveridge, D. L., Case, D. A., Cheatham, T. E., 3rd, Giudice, E., Lankas, F., Lavery, R., Maddocks, J. H., Osman, R., Sklenar, H., et al. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps, *Biophys. J. 89*, 3721-3740.

51. Perez, A., Luque, F. J., and Orozco, M. (2007) Dynamics of B-DNA on the microsecond time scale, *J. Am. Chem. Soc. 129*, 14739-14745.

52. Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T. C., Case, D. A., Cheatham, T., 3rd, Dixit, S., Jayaram, B., Lankas, F., Laughton, C., et al. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA, *Nucleic Acids Res. 38*, 299-313.

53. Ditzler, M. A., Otyepka, M., Sponer, J., and Walter, N. G. (2010) Molecular dynamics and quantum mechanics of RNA: conformational and chemical change we can believe in, *Acc. Chem. Res. 43*, 40-47.

54. Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P. E., and Berman, H. M. (2006) The RCSB PDB information portal for structural genomics, *Nucleic Acids Res. 34*, D302-305.

55. Sigel, R. K., Sashital, D. G., Abramovitz, D. L., Palmer, A. G., Butcher, S. E., and Pyle, A. M. (2004) Solution structure of domain 5 of a group II intron

ribozyme reveals a new RNA motif, *Nat. Struct. Mol. Biol. 11*, 187-192.

56. Seetharaman, M., Eldho, N. V., Padgett, R. A., and Dayie, K. T. (2006) Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA, *RNA 12*, 235-247.

57. Keating, K. S., Toor, N., Perlman, P. S., and Pyle, A. M. (2010) A structural analysis of the group II intron active site and implications for the spliceosome, *RNA 16*, 1-9.

58. Zhang, L., and Doudna, J. A. (2002) Structural insights into group II intron catalysis and branch-site selection, *Science 295*, 2084-2088.

59. Toor, N., Keating, K. S., Taylor, S. D., and Pyle, A. M. (2008) Crystal structure of a self-spliced group II intron, *Science 320*, 77-82.

60. Toor, N., Keating, K. S., Fedorova, O., Rajashankar, K., Wang, J., and Pyle, A. M. (2010) Tertiary architecture of the Oceanobacillus iheyensis group II intron, *RNA 16*, 57-69.

61. Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., Debolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structure and energetic properties of molecules, *Comp. Phys. Comm. 91*, 1-41.

62. Case, D. A., Cheatham, T. E., III, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs, *J. Comput. Chem. 26*, 1668-1688.

63. Jorgensen, W. L., Chandrasekhar, J., Madura, J., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys. 79*, 926-935.

64. Joung, I. S., and Cheatham, T. E., III. (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations, *J. Phys. Chem. B 112*, 9020-9041.

65. Joung, I. S., and Cheatham, T. E., 3rd. (2009) Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters, *J. Phys. Chem. B 113*, 13279-13290.

66. Auffinger, P., Cheatham, T. E., III, and Vaiana, A. C. (2007) Spontaneous formation of KCl aggregates in biomolecular simulations: a force field issue?, *J. Chem. Theory Comp. 3*, 1851-1859.

67. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and

Pedersen, L. G. (1995) A smooth particle mesh Ewald method, *J. Chem. Phys. 103*, 8577-8593.

68. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Comp. Phys. 81*, 3684-3690.

69. Miyamoto, S., and Kollman, P. A. (1992) Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models, *J. Comput. Chem. 13*, 952-962.

70. Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, *J. Comp. Phys. 23*, 327-341.

71. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera—A visualization system for exploratory research and analysis, *J. Comput. Chem. 25*, 1605-1612.

72. Grace. http://plasma-gate.weizmann.ac.il/Grace/.

73. Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E., III. (2007) Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms, *J. Chem. Theory Comput. 3*, 2312-2334.

74. Cheatham, T. E., III, and Kollman, P. A. (1997) Molecular dynamics simulations highlight the structural differences in DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes, *J. Am. Chem. Soc. 119*, 4805-4825.

75. Foloppe, N., and MacKerell, A. D. J. (2000) All-atom empirical force field for nucleic acids. 1) Parameter optimization based on small molecule and condensed phase macromolecular target data., *J. Comput. Chem. 21*, 86-104.

76. Denning, E. J., Priyakumar, U. D., Nilsson, L., and Mackerell, A. D., Jr. (2011) Impact of 2'-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA, *J. Comput. Chem. 32*, 1929-1943.

77. Correll, C. C., and Swinger, K. (2003) Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 A resolution, *RNA 9*, 355-363.

78. Tolbert, B. S., Miyazaki, Y., Barton, S., Kinde, B., Starck, P., Singh, R., Bax, A., Case, D. A., and Summers, M. F. (2010) Major groove width variations in RNA structures determined by NMR and impact of 13C residual chemical shift anisotropy and 1H-13C residual dipolar coupling on refinement, *J. Biomol. NMR 47*, 205-219.

79. Paulsen, R. B., Seth, P. P., Swayze, E. E., Griffey, R. H., Skalicky, J. J., Cheatham III, T. E., and Davis, D. R. (2010) Inhibitor-induced structural change in the HCV IRES domain IIa RNA, *Proc. Natl. Acad. Sci. 107*, 7263-7268.

80. Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K., and Berman, H. M. (1996) Geometric parameters in nucleic acids: nitrogenous bases, *J. Am. Chem. Soc. 118*, 509-518.

81. Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K., and Berman, H. M. (1996) Geometric parameters in nucleic acids: sugar and phosphate constituents, *J. Am. Chem. Soc. 118*, 519-529.

82. Draper, D. E. (2004) A guide to ions and RNA structure, *RNA 10*, 335-343.

83. Draper, D. E., Grilley, D., and Soto, A. M. (2005) Ions and RNA folding, *Annu. Rev. Biophys. Biomol. Struct. 34*, 221-243.

84. Rudisser, S., and Tinoco, I., Jr. (2000) Solution structure of Cobalt(III)hexammine complexed to the GAAA tetraloop, and metal-ion binding to GA mismatches, *J. Mol. Biol. 295*, 1211-1223.

CHAPTER 3

STRUCTURAL AND ENERGETIC ANALYSIS OF 2-AMINOBENZIMIDAZOLE INHIBITORS IN COMPLEX WITH THE HEPATITIS C VIRUS IRES RNA USING MOLECULAR DYNAMICS SIMULATIONS

3.1 Chapter Notes

This research was designed by Niel M. Henriksen and Thomas E. Cheatham, III. The simulations, free energy analysis, and docking of the RESP charged inhibitors were performed by N.M. Henriksen. Hamed S. Hayatshahi performed docking with the AM1-BCC charged inhibitors. The manuscript was written by N.M. Henriksen. The manuscript was revised by N.M. Henriksen, H.S. Hayatshahi, Darrell R. Davis, and T.E. Cheatham, III.

3.2 Introduction

RNA performs a vast array of functions in biological systems, including genetic encoding, regulation, and catalysis *(1-3)*, and yet very few drugs exist which target RNA *(4)*. This may be the result of many factors, including the relatively recent discovery of RNA's many biological roles and the difficulty in preventing RNA degradation during experiments, particularly by ribonucleases *(5, 6)*. Likewise, computational investigations of RNA-ligand binding are

comparatively rare (PubMed search of “protein binding simulations” as of December 2012 yields 6387 results, “rna binding simulations” yields 407 results) *(7, 8)*. In order to address this scarcity, this study reports the results of molecular dynamics (MD) simulations on a specific RNA-ligand system and aims to provide a more reliable foundation for future studies involving highly charged complexes, such as those described here.

The target of this research is the domain IIa RNA sequence from the hepatitis C virus internal ribosome entry site (HCV IRES) *(9)*. Experimental structures exist for both the unbound (or free) structure *(10, 11)* and also of the RNA in complex with 2-aminobenzimidazole inhibitors *(12, 13)*. This RNA-inhibitor complex is an attractive structure to study because it involves a relatively short RNA sequence bound to a drug-like molecule. This contrasts with typical structures that are often larger and more complex, such as RNA or ribo-protein molecules in complex with aminoglycosides *(14, 15)*. Moreover, distinct structural differences between the free and bound conformations are observed, and this is most notably characterized by the loss of a key bend in the RNA upon ligand binding that explains the inhibition mechanism *(16)*. Biologically, the structure is of interest due to both the high degree of sequence conservation in IRES elements and its importance in HCV genome translation and viral replication *(17)*. Rather than using the 5’ cap-dependent mechanism to initiate translation at the ribosome, as is typical in eukaryotes, the HCV IRES element is responsible for recruiting the 40S ribosomal subunits.

Thus, development of inhibitors of the IRES machinery could be useful tool in treating hepatitis C infections.

The 2-aminobenzimidazole inhibitors used in the experimental structures (J4 and J5) were discovered by Isis Pharmaceuticals, Inc. along with several others (Figure 3.1) using a high throughput mass-spectrometry assay and these inhibitors were found to reduce HCV RNA levels in a viral replication assay *(18)*. As part of their exploration of structure-activity relationships, a number of different derivatives were synthesized and binding constants estimated (Table 3.1). This provides a series of related inhibitors, studied by the same laboratory with equivalent and comparable experiments, which can be investigated in simulation to assess the biomolecular simulation protocols. Unfortunately, for the desired assessment, there are several drawbacks to the experimental data for these inhibitors, specifically: 1) the protonation state of the inhibitor upon binding is unknown, 2) several inhibitors were synthesized as racemic mixtures of enantiomers or diastereomers and the experimental binding data published do not distinguish the results based on the stereochemistry, and 3) the errors in the binding measurements were not reported. Not all of these issues preclude the assessment. For example, the protonation states can be estimated with reasonable accuracy using various pKa estimation software (see METHODS) and these calculations suggest that the inhibitors are all fully protonated in solution at physiological pH as depicted in Figure 3.1. With regard to stereochemistry, it is very easy to perform separate calculations on each of the enantiomers and diastereomers, and, when necessary, report the mean value for comparison

with experimental data. The lack of error analysis in the experimental binding constant data, however, does suggest the use of caution when making certain conclusions based on comparison of the experimental and computational results.

The experimental structures of the RNA-inhibitor complex, two derived from NMR data *(12)* and one from X-ray crystallography *(13)*, exhibit distinctly different binding modes. These differences cannot easily be attributed to the identity of the two different inhibitors used in the separate investigations as they only differ by a single -CH2- group (see J4 and J5 in Figure 3.1). These differences also cannot be attributed to the slight change in RNA sequence in the NMR versus crystal, as the sequence changes are only found at the ends of the molecule as shown and discussed in Figure 3.2A. Rather the differences must lie in the experimental method used to gather data, the structure refinement procedures, and/or potential crystal packing artifacts. The NMR structure can be described as an open conformation with stacking contacts formed below the inhibitor, a single phosphate-dimethylamino interaction, and a single base pair-dimethylamino interaction (Figures 3.2B, 3.3A). Lateral contacts between the benzimidazole ring and RNA are not observed. In contrast, the crystal structure is more compact, stacking interactions are formed both above and below the inhibitor, both dimethylamino groups interact with the RNA phosphate backbone, and two critical hydrogen bonds are formed between the benzimidazole ring and the residue G33 in the binding site (Figures 3.2C, 3.3B, 3.4).

In addition to the differences in the conformations of the NMR and crystal structures, there are potentially conflicting reports regarding the cation requirements necessary for the formation of the inhibitor bound complex. Magnesium is observed at core positions in both the unbound and bound crystal structures *(11, 13)*. It is also found that removal of magnesium from the FRET binding assay of the crystal structure yields a ~30 fold decrease in binding affinity *(13)*. Although these findings are consistent with the well-known relationships between RNA structure and cation binding *(19, 20)*, they should not be interpreted to suggest that specifically bound magnesium exclusively performs the stabilization role *(21)*. A variety of RNA tertiary structures are known to form in moderate levels of monovalent salt *(22)* and magnesium is known to compete with monovalent cations in stabilizing RNA *(23)*. In the case of the HCV IRES domain IIa RNA, although the addition of magnesium stabilizes the unbound solution structure *(10)*, no changes were observed in the NMR spectra of the bound complex upon addition of magnesium to a solution with a relatively high monovalent salt concentration *(12)*. Additionally, a fluorescence binding assay of the NMR sequence conducted in 0.15 M KCl and 0.15 M NaCl using the J4 inhibitor (Figure 3.1) yielded a dissociation constant in the equivalent range (2.4 μM) *(12)* as the value determined by FRET assay for the crystal sequence binding to the J5 inhibitor in 2 mM Mg^{2+} (EC_{50} = 3.4 μM) *(13)*. To address the discrepancies in the experimental data regarding inhibitor binding to the HCV IRES domain IIa, we have performed a variety of MD simulations using the reported NMR and crystal experimental conformations.

Stable MD simulations are observed for the crystal conformations, and we also observed the partial transition of an “unstable” NMR conformation to a more crystal-like conformation. Stable simulations are a necessary first step for future modeling studies of RNA-ligand interaction and computer-aided drug design, and stable MD simulations are not a guaranteed outcome using even the latest RNA force fields *(24, 25)*. Indeed, recent force field improvements are required to correct pathological simulation behavior of RNA *(26-30)* and further improvements are still likely necessary (unpublished data). In the case of the crystal structure simulations on the 200 ns time scale, however, the AMBER ff12SB force field (which includes the parmbsc0 *(26)* α/γ backbone and RNA XOL *(28)* modifications to the ff99 *(31-33)* force field) appears to be sufficiently accurate. Next, to test whether small changes in the inhibitor lead to discernible changes in the binding mode, we investigate the binding of six related inhibitors (yielding twelve stereochemically distinct compounds). We also examine whether magnesium ions are essential for stable simulations of the crystal conformation. Finally, in anticipation of further inhibitor development using computational models, we investigate methods for determining accurate binding energies, entropic penalties, and ligand docking procedures. We also investigate whether a novel drug scaffold is capable of forming stable interactions with known contacts in the receptor binding site.

3.3 Methods

3.3.1 Ligand parameterization

All inhibitors in this study were protonated at the dimethylamino and benzimidazole positions as indicated in Figure 3.1. The fully protonated state at pH 7.0 is consistent with pKa estimates by two different pKa prediction programs, SPARC *(34)* and Marvin Sketch *(35)*. Charge derivation was performed in a very careful manner due to the highly charged nature of the inhibitors: 1) A hand-built inhibitor model was geometry optimized at the QM HF/6-31G* level consistent with the AMBER ff10 and ff12SB force fields followed by restrained electrostatic potential (RESP) *(36)* charge fitting to determine initial atomic charges; 2) 50 ns of implicit solvent generalized-Born (GB) MD, using Hawking, Cramer, and Truhlar model *(37)*, was performed at 400K to sample relevant inhibitor conformations, and the resulting trajectory was clustered *(38)* using the "averagelinkage" algorithm into twenty clusters using AMBER's Ptraj program; 3) the representative structure from clusters whose occupancy was greater than 2% were then geometry optimized at the QM HF/6-31G* level; and 4) optimized structures whose energy was within ~0.5 kcal/mol of the minimum energy structure were used in a multiconformation, multi-orientation RESP fit using the RED program *(39)* to generate the final charges used in this study. Enantiomers were fit simultaneously to ensure identical charges. Bonds, angles, torsions, improper torsions and Lennard-Jones parameters were assigned from the general AMBER Force Field (GAFF) using the Antechamber and Parmchk programs *(40-42)*. Some torsions and improper torsion parameters

were modified because the default parameters did not maintain planarity at the C2 position of the 2-aminobenzimidazole ring. MOL2 files with GAFF atom types and charges, as well as "frcmod" files with the modified torsion parameters are provided in the Supplementary Information. All quantum mechanical calculations were performed using Gaussian09 *(43)* and all MD simulations were performed using AMBER12 *(44)*.

In addition to the more detailed approach to generate high quality atom charges discussed above, we were interested in the performance of more approximate charge parameterization methods. To test this, we performed molecular docking studies (discussed below) using both the RESP charges from the above procedure and AM1-BCC *(45, 46)* charges produced by AMBER's Antechamber program. In the latter case, charges were determined separately for each inhibitor conformation studied.

3.3.2 Initial RNA–inhibitor conformations

Experimentally determined atomic resolution structures exist for both the J4R and J4S inhibitors (as NMR structures) and also for the J5R inhibitor (as a crystal structure). To facilitate comparisons between the MD simulations, the crystal structure duplex was converted into a hairpin of identical sequence to the NMR structure. To accomplish this, the 3' dangling bases were removed, the C-G basepair at the base of the lower stem was converted to a G-C basepair, and a UUCG tetraloop was added to the opposite stem (Figure 3.2A). In order to generate a diverse set of binding conformations for the inhibitors in

Figure 3.1, the twenty representative conformations of each inhibitor identified by clustering during the charge derivation procedure were RMS fit to the benzimidazole core atoms of the experimental structure, either NMR or crystal. For the NMR structure, the first model from the PDB 2KU0 *(12)* ensemble was used as the reference structure. This procedure resulted in the generation of twenty "NMR-like" RNA-inhibitor conformations and twenty "crystal-like" RNA-inhibitor conformations for each of the twelve stereochemically distinct inhibitors. The selection of these conformations as initial structures for the various simulation sets in this work is described in the following section. Simulations which included magnesium used the exact coordinates from the crystal structure for the J5R inhibitor, magnesium, and RNA with the necessary sequence modifications included so as to match the NMR sequence. In the case of the novel ligands, a single inhibitor conformation was chosen and RMS fit to the benzimidazole core atoms in the crystal experimental structure.

3.3.3 Simulation sets

As described in Table 3.2, several sets of simulations were performed. For RNA-inhibitor studies, two strategies were employed: single long simulations and multiple short simulations. For the single long simulation sets (NMR1 and CRY1), a single initial structure was selected from the twenty initial conformations for each of the twelve inhibitors based on the minimum GB energy of the complex. For the multiple simulation set (CRY2), all twenty

initial conformations were used. In a few cases, bad initial conformations were replaced with good conformations due to severe atom overlap.

3.3.4 Building solvated models

The domain IIa RNA was parameterized using AMBER's ff12SB force field. The initial RNA-inhibitor conformations were first minimized for 2500 cycles using the steepest descent algorithm in implicit GB solvent and the resulting geometries were solvated. All simulations described in Table 3.2 were performed in TIP3P *(47)* water with net-neutralizing potassium ions and an additional ~200 mM KCl as parameterized by Joung and Cheatham *(48)*. The number of waters added was chosen to yield a periodic truncated octahedron with an approximately 12 Å minimum water shell between the solute and the box edge. In order to facilitate energetic comparison between inhibitors, the number of solvent atoms for systems in each simulation set were made to be identical using an in-house Perl script coupled to AMBER's LEaP program. In the case of the J1 inhibitor, which has +2 net charge rather than +3, direct energetic comparisons with the other inhibitors was not performed. Following solvation, the monovalent ion positions were randomized with AMBER's Ptraj program using the "randomizeions" command to remove bias from the initial ion placement. In the case of the MG simulation (Table 3.2), the crystallographic magnesium ions and water molecules were included using Allner, Nilsson, and Villa's magnesium parameters *(49)* in addition to 200 mM KCl.

3.3.5 Molecular dynamics simulations

All solvated simulations used a similar minimization, heating, and equilibration procedure: 1) The entire system was minimized for 1000 steps using the steepest descent algorithm followed by 1000 steps of conjugate gradient minimization while 25 kcal/mol-Å^2 positional restraints were enforced on the RNA and inhibitor benzimidazole core atoms, 2) the system was heated, with 25 kcal/mol-Å^2 positional restraints on the RNA and benzimidazole core atoms, from 10 to 150 K at constant volume with the Langevin thermostat over the course of 100 ps, 3) further heating and initial equilibration was performed from 150 to 298 K using constant pressure and the Langevin thermostat with 5 kcal/mol-Å^2 positional restraints on all solute atoms over the course of 100 ps, and 4) final equilibration at 298 K was performed for 2 ns using constant pressure and Langevin thermostat with 0.5 kcal/mol-Å^2 positional restraints on the RNA and benzimidazole core atoms. Production simulations were performed at 298 K at constant pressure using the weak-coupling algorithm for the thermostat and barostat *(50)*. The pressure relaxation times were 1 ps for the initial equilibration step, 5 ps for the final equilibration step, and 10 ps for production. For heating and both equilibration steps a collision frequency of 2 ps^{-1} was used for the Langevin thermostat. For production, the time constant of heat bath coupling was 10 ps using the weak-coupling algorithm. For all heating, equilibration, and production steps, the time step was 2 fs, the direct space sum used a cutoff of 8.0 Å, and SHAKE was applied to all bonds involving a hydrogen atom *(51)*. The default particle mesh Ewald *(52)* settings were used

to determine long-range charge interactions (which correspond to an ~1 Å grid spacing and a 10^{-6} direct space tolerance). Coordinates were recorded every picosecond during production simulations. All solvated simulations, with the exception of one performed with residual dipolar coupling (RDC) restraints, were performed using either the CPU or GPU version of the PMEMD program in AMBER12 *(44)*. A single simulation using the crystal conformation was performed using AMBER's Sander program with the NMR RDC restraints enforced (the use of RDC restraints is not yet implemented in the faster PMEMD program). A short minimization was performed on the equilibrated, solvated structure to best fit the RDC alignment tensor. The relative weighting for the alignment restraint was chosen to be 0.08 kcal/Hz^2, which represents an empirical determination of the largest value that does not produce simulation instability (e.g., integration errors).

3.3.6 Energy analysis

MM-GBSA and MM-PBSA are well-known postprocessing techniques for computing binding energies from simulation trajectories *(53-55)*. In this work, MM-GBSA and MM-PBSA analyses were performed with the MMPBSA.py program in AMBER12 using the single trajectory approach. For MM-GBSA, the Hawkins, Cramer, and Truhlar GB model *(37)* was used for implicit solvation with 200 mM salt concentration approximated using Debye-Huckel screening. For MM-PBSA, the following options were used: a level-set based dielectric model, a two-term nonpolar solvation free energy term based on a cavity and dispersion

calculation *(56, 57)*, 200 mM ionic strength, 1.4 Å solvent probe, 0.25 Å grid spacing, and the ratio between the longest grid dimension and the solute set to 6.0. Radii for inhibitors were chosen from a set of optimized radii to best match the atom types present *(56)*. For MM-GBSA, all frames from the trajectory were used, but for MM-PBSA only every 100th frame was used for computational efficiency reasons.

In addition to the MM-GBSA/MM-PBSA framework of energy analysis, we also calculated relative binding energies which included explicit solvent effects by subtracting the solvated inhibitor average potential energies (from simulations of the free ligands in explicit solvent, LIG) from the solvated RNA-inhibitor average potential energies (CRY1 and CRY2). These binding energies, which are formally an estimate of the binding enthalpy (the kinetic energy and pressure/volume contribution is assumed to be negligible), do not include additional entropy estimates. The inhibitor J1 was excluded from these calculations due to its +2 charge which complicates direct comparison due to differences in the number of counterions.

Besides binding energy calculations, we also performed two types of entropy analysis. In both cases, only the inhibitor entropy was considered and energy changes were computed by subtracting the free inhibitor entropy from the complexed inhibitor entropy (CRY1 minus LIG). The first method, quasi-harmonic analysis *(58)*, was computed using AMBER's Ptraj program. The second method, first-order configurational entropy analysis based on bond/angle/torsion probability distribution functions, was computed using the

ACCENT program developed by Gilson and co-workers *(59)*. Both methods ignore rotational and translational contributions to entropy. Due to the accumulative nature of these values, error bars are not given for the entropy estimates which were estimated over equivalent time windows.

Finally, inhibitor solvation energies were estimated by subtracting gas phase average potential energies from either GB implicit solvation energies or from potential energies of explicitly solvated trajectories. Although one could compute the gas phase energies from solvated trajectories by stripping the solvent, we performed 100 ns gas phase simulations for each of the twelve inhibitors represented in Figure 3.1 in order to ensure independence of these values.

3.3.7 Error analysis

Two approaches were used to estimate error depending on the simulation set. For the single long simulation sets, a previously described re-blocking procedure was used *(60)*. Briefly, a data set plotting the standard error of the mean (SEM) versus increasing block size is computed. Given sufficient sampling, the plot will plateau at a value that corresponds with the SEM. A trend line fit can be used to predict this value. However, due to assumptions about the correlation in the data, the result can be inaccurate. To be conservative, we use the maximum value observed in the plot. An example of this error analysis is given in Figure 3.5. For the multiple short simulation set (CRY2), we can consider the average value from each of the twenty separate

simulations as being independent, uncorrelated data points and compute the SEM in the traditional way by dividing the sample standard deviation by the square root of the number of data points (i.e., number of simulations). Error combining rules were as follows: when computing the difference of two mean values, the errors were added; when computing the average of two or more mean values, the errors were averaged.

3.3.8 Grid analysis

The regions of highest magnesium ion occupancy were determined using the "grid" command in AMBER's Ptraj program *(61)*. The simulation trajectory frames were centered, imaged, and RMS fit using the heavy atoms of residues 5, 11, 33, and 34 that form the binding region. Occupancy was determined by a three dimensional histogram approach using a 75 x 75 x 75 Å box with 0.5 x 0.5 x 0.5 Å resolution. The results are visualized on the average RNA-inhibitor structure from the simulation using UCSF Chimera package *(62)*. To choose the density surface contour level to be displayed, the contour level was increased until magnesium occupancy in the bulk solvent region was no longer observed, thus suggesting stable binding locations.

3.3.9 Docking

Docking was performed on the crystal receptor structure (modified to match the full NMR sequence, Figure 3.2A) using Dock 6.5 *(7)*. The top and bottom helical portions of the receptor were excluded from consideration as

they are known to not contain the binding site. This exclusion did not inappropriately limit docking poses to the known binding cavity since the entire backside region of the receptor was explored for docking. In order to include various ring pucker conformations in some of the inhibitors, for which Dock 6.5 is not able to search automatically, all of the inhibitor conformational clusters whose occupancy was greater than 2% (identified during the charge derivation procedure) were used as initial seed structures for docking. Two schemes were used to assign charges to the inhibitor for use during docking. The first scheme simply used the RESP charges that were derived for use in MD simulations. In the second scheme, which resembles a more typical docking procedure, charges were derived for each inhibitor conformation using the semi-empirical AM1-BCC charge model which can be accessed through the AMBER's antechamber program. The default grid-based method in Dock 6.5 was assigned as the primary scoring function.

3.4 Results

3.4.1 MD simulations

The NMR1 and CRY1 simulation sets (Table 3.2) were intended to evaluate the simulation stability of the two available experimental conformations in the context of twelve related inhibitors over a fairly long timescale (200+ ns). Visual inspection of the simulation trajectories reveals a stark contrast in the stability of the binding region. A quantitative measure of this difference is shown in Figure 3.6, where the binding region RMSD is plotted

versus simulation time using the initial conformation of the each production simulation as reference. For all twelve simulations in the CRY1 simulation set (shown in black) the RMSD value is low and very steady. The small fluctuations arise from inhibitor ring transitions and conformational searching by the dimethylamino groups. Throughout the CRY1 simulations all of the critical contacts depicted in Figure 3.4 are maintained in each simulation. In contrast to the CRY1 results, the NMR1 RMSD results (shown in red) reveal a high degree of fluctuation and departure from the initial structure. A variety of RNA-inhibitor poses are adopted during the twelve NMR1 simulations and do not point to a consensus alternative to the original NMR pose. Much of the structural instability is due to conformational transitions in the bulge residues 6-10. To visualize the structural difference between the NMR1 and CRY1 simulation sets, the final frame of each simulation was overlaid to produce an ensemble (Figure 3.7). Due to unfolding of the bulge residues, many of the NMR1 conformations do not retain the linear RNA orientation known to inhibit viral replication, but rather adopt the “L-shaped” conformation similar to the unbound structure (Figure 3.7A). Additionally, the inhibitor poses are smeared into a variety of orientations (Figure 3.7C). In contrast, the CRY1 simulations maintain the linear orientation and produce a rather tight ensemble of structures and inhibitor poses (Figure 3.7B,D).

One of the NMR1 simulations, that with the J5S inhibitor, is notable because it partially converts to the conformation of the crystal structure. All of the critical contacts in Figure 3.4 are observed, including the hydrogen bonds

between the inhibitor and G33 and the contacts between the dimethylamino groups and the phosphate backbone (Figure 3.8A). These distances are similar to those observed in the CRY1 simulation set (Figure 3.8B). The only interaction that is not formed is the base triple interaction between residue A6 and the Hoogsteen edge of A32, which forms a "roof" above the inhibitor. After 232 ns of simulation, the conformation of residues 7-9 continues to restrain the flexibility of A6 in such a way as to prevent the full formation of the base triple, although A6 and A32 are near enough to form direct contacts. Other simulations in the NMR1 set also partially adopt the crystal-like binding mode, usually by forming the hydrogen bonds between the inhibitor and G33, but none are as stable as the aforementioned simulation with J5S. Further research, likely involving enhanced sampling, is necessary to determine whether a complete transition is accessible on a reasonable timescale.

In addition to using long simulations, we also investigated if using many shorter simulations with diverse initial inhibitor conformations could produce comparable data. Due to the instability of the NMR conformation in simulation, we only discuss this approach for the crystal conformation which we term the CRY2 simulation set. Most comparisons are made in the energetic analysis portion of the results, but we also wanted to determine whether the RMSD space explored by the inhibitor was different between the two approaches. Figure 3.9 compares the mean RMSD value of the binding region, as well as the maximum and minimum value, between the CRY1 and CRY2 simulation sets using the same reference structure for each simulation of a given inhibitor

(values are listed in Table 3.3). For most inhibitors, the mean, minimum, and maximum RMSD values are similar. Only in the case of the weak binding J1 inhibitor does a large difference in maximum RMSD appear. These results do not guarantee that the exact same conformations are sampled by both approaches, nor do they indicate that the proportion of conformations sampled is similar. However, the results do indicate that long simulations do not explore RMSD space that is significantly farther away (higher RMSD values) than that sampled with an ensemble of diverse short simulations.

It is important to note that the crystal structure is published with six magnesium ions, five of them nearby the binding region *(13)*. The authors note three specific magnesium ions that seem particularly important structurally and also note that binding affinity dramatically decreases in the absence of magnesium. We chose not to include magnesium ions in this study, with the exception of the MG simulation, because the complex hydration and coordination geometries adopted by magnesium are difficult to model using a fixed charged model. To investigate whether simulations with magnesium differed from those without, a single 82 ns simulation was performed using the experimental coordinates for the J5R inhibitor, magnesium ions, crystallographic waters, and RNA (with the necessary sequence modifications to be consistent with the NMR structure). Also present was 200 mM KCl. No changes in the binding mode were observed. Figure 3.10 compares the regions of highest magnesium ion occupancy with the crystallographic locations of the magnesium ions. None of the highest occupancy locations observed in

simulation reproduce the exact coordination contacts observed in the crystal structure. The magnesium coordinated to N7 of A6 in the crystal structure moves to coordinate the phosphate groups of G30 and C31. The magnesium coordinated to both O4 of U14 and OP1 of U9 in the crystal structure loses the interaction to the phosphate and forms an interaction with O2 of C8 instead. Finally, both magnesiums located near G5 in the crystal structure move elsewhere. These results combined with the observation of stable simulations without magnesium ions suggest that magnesium ions do not appear to be critical in moderately high monovalent salt simulations (i.e., both had 200 mM KCl). This is consistent with *in vitro* binding assays in relatively high monovalent salt, and no magnesium, which observed expected binding behavior *(12, 18)*.

The highly stable solvated simulations of the RNA-inhibitor complex in the crystal conformation add support that the published structure is reasonably stable outside of the crystal lattice and therefore not strongly influenced by crystal packing forces. However, given the sampling limitations of MD simulations using current hardware, it is not possible to be sure that the crystal conformation represents the global minimum in the energy landscape. One way to investigate this is to determine whether the NMR distance restraints are satisfied by the crystal structure. To determine this, the r^6-averaged distances for all atoms identified as NOE pairs by NMR were computed for the 200+ ns simulation of J4R from the CRY1 simulation set (this inhibitor structure is identical to that of the published NMR structure). The crystal structure and

several of the NOE restraints appear to be incompatible. Most of the large violations occur between inhibitor and RNA atom pairs, with the largest being those between the dimethylamino arms of the inhibitor and RNA residues A6 and U12. It is likely that enforcing the NOE distance restraints during simulation of the crystal conformation would cause structural disruptions and form intractable "knots" due the large distance violations. However, a set of residual dipolar coupling (RDC) restraints were also obtained during collection of the NMR data *(12)*. It is not immediately obvious how well the crystal conformation fits the RDC data. To test this, we performed a 5 ns simulation with RDC restraints enforced. No significant changes in the global structure or inhibitor binding contacts were observed. This suggests that although significant NOE violations are observed for the crystal conformation binding mode, it is consistent with the RDC data. Further research will be required to determine whether the NOE violations are simply due to data misassignment or whether a more complicated situation, such as multiple solution binding modes or RNA conformational dynamics, is responsible for the discrepancy.

3.4.2 Energy analysis

A major goal of this study was to determine whether energetic binding analyses of the simulation trajectories could reproduce the experimental trends in binding energy. This is a challenging problem given the high ionization states of the ligands and the relatively tight range of K_D or $\Delta G_{binding}$ values estimated in experiment. Given the instability of the NMR1 simulation sets, we

only considered simulations performed with the crystal conformation. One of the quickest methods for obtaining binding energy data from MD simulations is to perform MM-GBSA/MM-PBSA analysis using the single trajectory method. The MM-GBSA and MM-PBSA results for both the CRY1 and CRY2 simulation sets are shown in Figures 3.11 and 3.12 (values are given in Table 3.4). Several observations can be made from the results. First, the magnitude of the binding energies is significantly larger than those observed experimentally. For MM-GBSA, the range of binding energies for the various inhibitors is around -45 to -65 kcal/mol. For MM-PBSA, the magnitude is slightly smaller with ranges around -30 to -55 kcal/mol. Obviously there is a significant difference between these values and the -5 to -8 kcal/mol range identified experimentally (Table 3.1). A variety of corrections can be considered to address this discrepancy. One explanation is that the free energy change upon RNA reorganization, which accompanies inhibitor binding, is not accounted for in the single trajectory approach. We have performed MD simulations and MM-GBSA analysis on the apo-RNA using the published NMR structure *(10)* as the initial conformation and can estimate this would add around +10 kcal/mol to the MM-GBSA binding energies. An additional energetic component, which we have not included in our MM-GBSA/MM-PBSA results, is the change in solute entropy upon binding. We have attempted to estimate the conformational entropy change for the inhibitors alone (estimates are between +0-25 kcal/mol, discussed below), but the size of the RNA presents a challenge to obtaining a full solute estimate. Additionally, the rotational and translational entropy loss upon ligand binding

would likely add in the range of +3-10 kcal/mol *(63-68)*. Together, these corrections are in the range of +13-45 kcal/mol and could, in theory, bring at least the MM-PBSA results closer to experimental ranges. However, the raw binding energies from MM-GBSA/MM-PGBSA not only have a large magnitude but also large spread in values. For example, if we exclude the results for the J1 inhibitor (due to uncertainty about its exact experimental binding energy), the range for MM-GBSA is ~10 kcal/mol and for MM-PBSA around ~20 kcal/mol. Given the similarity of the ligands and their binding mode with the RNA, it is unlikely that the aforementioned corrections for receptor reorganization and solute entropy could account for the large range in values. Finally, even if we ignore the incorrectly large range of binding energy, the relative trend in binding energy does not match what is observed experimentally (Table 3.1). The weakest binder, J1, can be distinguished from the other inhibitors, although less so with MM-PBSA than MM-GBSA. As for the other inhibitors, the trend does not correlate well. In particular, both MM-GBSA and MM-PBSA rank J2 as having an equal or lower binding energy than the J5 enantiomers, despite the fact that experimentally, J2 is among the weakest while J5 is among the strongest binders. It is not immediately clear why the MM-GBSA/MM-PBSA results do not match the experimental data, but it may be related to the failure of implicit solvent models to accurately model highly charged systems. In fact, it is known that the Hawkins, Cramer, and Truhlar GB model does not produce accurate values for salt bridges *(69)* and incorrectly models DNA helices *(70)*.

Given the poor performance of these implicit models, it is reasonable to seek an alternative calculation of the binding energy which includes the contribution of explicit solvent. We performed relative binding energy analysis on the explicitly solvated RNA-inhibitor systems using the total potential energy of the explicitly solvated trajectories directly. This was possible because all of the systems in the CRY1 and CRY2 simulation sets, with the exception of J1, had by design identical numbers of atoms (excluding the inhibitor atoms). Likewise, the LIG simulation set, which contained just the free inhibitors and solvent, also contained identical numbers of solvent atoms. Since the J1 inhibitor has a different charge and therefore different number of counterions, it was excluded from consideration. The relative energetic contribution to binding for the trivalent inhibitors (J2-J6) can be calculated by subtracting the average potential energy of the free inhibitor simulations (LIG set) from the average potential energy of the bound inhibitor simulations (CRY1 or CRY2 sets). Only the single long simulation strategy was used for the free inhibitor simulations (LIG) because the potential energies were tightly converged. Confirmation of this is demonstrated in the fact that the mean potential energy of all enantiomers are within expected error of one another (Table 3.5, fourth column). The results for the binding energy are shown in Figure 3.13. It should be noted that the absolute value of the binding energy is meaningless because it represents the force field energy difference of two systems with unequal atoms (i.e., ligand in solution and ligand bound to RNA). However, the relative energy differences can be compared because the relative difference between

the systems is identical, with the exception of the inhibitor atoms of interest. Several observations suggest that the relative binding energy values are more accurate than the MM-GBSA or MM-PBSA values. First, the range of binding energies is much smaller: ~6-7 kcal/mol. Second, the binding energies calculated from the CRY1 and CRY2 simulation sets are within error of one another, excepting J6SS which has a 0.9 kcal/mol difference in the error ranges. This was not the case for the implicit solvent approach, although it should be noted that the error values are larger for calculations involving explicit solvent. Third, the binding energies are internally consistent: the four J6 diastereomers, which contain two constrained rings of which one each is present in the J3 and J5 enantiomers, produce a binding energy trend which can be predicted from the J3 and J5 results. The binding energy of the S enantiomer is preferred for J5, whereas the R enantiomer is preferred for J3. Consistent with expectations, the J6SR inhibitor has the lowest binding energy. The lone exception to this internal consistency is J6SS of the CRY1 simulation set (the prediction is correct in the CRY2 results). Finally, if one averages the values of the enantiomers/diastereomers as an approximation of the experimental conditions (where stereochemistry was not considered), the trends match well with experiment (Figure 3.14, values are given in Table 3.6). The range of the binding energies calculated from simulation is somewhat larger than the range of binding energies from experiment (~4 kcal/mol vs. ~2 kcal/mol). However the correct trend is observed as well as agreement between the single long trajectory and multiple short trajectory approaches.

This last observation is critical, given the small number of data points. Two completely independent approaches produce very similar results, suggesting that the results accurately reflect the underlying force field terms used to model the system.

The difference between the implicit solvation results and the explicit solvation results can be highlighted by looking at cases where the inhibitor was improperly bound to the RNA. As noted in the methods section, the initial inhibitor poses were not chosen via a docking algorithm but rather by the crude RMS fit of the benzimidazole core atoms. While this ensured that critical hydrogen bonds were formed between the inhibitor and residue G33, it did nothing to prevent unfavorable clashes of the flexible inhibitor arms with the RNA. In nearly every initial inhibitor conformation, potential clashes were eliminated during the minimization step, and the resulting geometry resembled a reasonable binding pose. However, in a few cases, one of the inhibitor arms was improperly inserted through the back of the binding cavity. Despite the apparent strain in the geometry, some of these cases produced stable simulations without integration errors or energy instability. These cases were not included in the previously discussed energy analysis but they provide a useful test case in that an accurate representation of system energy should distinguish highly strained systems from those with an expected binding conformation. One such case of improper binding mode occurred with the J2 inhibitor. For the multiple short simulation set CRY2, the range of mean GB potential energies of J2 in normal binding poses was -7913 to -7891 kcal/mol.

The mean GB potential energy of the strained conformation was -7916 kcal/mol, suggesting that it was actually a lower energy conformation. The MM-GBSA binding energy computed using this strained conformation was -77 kcal/mol, about 15 kcal/mol stronger binding than the average MM-GBSA binding value for J2 from the CRY2 simulation set. Likely, some portion of implicit solvation model incorrectly modeled this interaction. In contrast, using explicit solvent potential energies accurately identifies the strained J2 conformation as a high energy outlier. The range of mean explicitly solvated potential energies for J2 from the CRY2 simulation set was -118790 to -118766 kcal/mol, whereas the mean explicitly solvated potential energy for the strained conformation was -118761 kcal/mol. This trend, in which explicitly solvated energies more reliably predicted strained conformations than implicitly solvated potential energies, was true for other cases of strained RNA-inhibitor conformations as well.

In order to understand whether solvation energies played a role in the errors of the MM-GBSA/MM-PBSA results, we have performed further analysis of the LIG simulation set (in which the inhibitor was simulated freely in explicit solution without RNA). After postprocessing the trajectories using MM-GBSA analysis, a comparison of the solvated potential energy versus the GB potential energy was made (Figure 3.15, top, values are given in Table 3.7). As indicated by the trend line fit, the relative potential energies are similar whether using explicit solvation energies or the GB solvation energies. An RMS fit of explicit solvation energy values onto the implicit solvation values reveals that

differences range between 0.06-1.60 kcal/mol between the two methods. First, this shows that the MM-GBSA values are primarily enthalpic since they are similar to the explicit results (which are purely enthalpic). Second, it suggests that the likely origin of errors in the implicit solvation model is found in calculating the energy of the RNA-inhibitor complex. Similarly, if the free energy of solvation is computed (neglecting solute entropic terms) by subtracting the gas phase inhibitor energies from the solvated energies (either explicit or implicit), very similar trends between the explicit and implicit solvation model are observed (Figure 3.15, bottom). The large difference in the free energy of solvation (~13 kcal/mol) between J2/J3 and the rest of the inhibitors is troubling. The distinguishing feature between the two sets of inhibitors is whether or not a ring is formed on the oxygen side of the benzimidazole ring. Further research is necessary to understand this difference.

As a final piece of energy analysis, we note that the experimental binding energy trends suggest that the use of ring constraints to reduce flexibility improves binding energy. Thus is it reasonable to expect that there are different entropic binding penalties for the inhibitors depending on the rigidity. Due to the difficulty in converging entropy estimates for large molecules, only the inhibitor was considered in these calculations. Two methods were used: quasi-harmonic analysis *(58)* and a configurational estimate based on bond, angle, and torsion probability distributions *(59)*. The entropic energy penalties upon binding are given for each inhibitor based on

calculations from the CRY1 and LIG simulation sets (Figure 3.16, values given in Table 3.8). Convergence plots for these values are shown Figures 3.17 and 3.18. The convergence plots demonstrate that the estimates are not reliably converged, even at greater than 700 ns for the free ligands, and thus must be interpreted very conservatively. One observation that is clear is that the quasi-harmonic estimates have a much larger magnitude (approximately 5 to 25 kcal/mol) than the conformational entropy estimates (0 to 5 kcal/mol). The large values produced by the quasi-harmonic approximation are likely due to an overestimation of the harmonic potential width for the free inhibitors because no attempt was made to differentiate between conformational energy minima and thus the estimated width is necessarily wider in order to cover a broader space. In contrast, the configurational estimate using probability distributions distinguishes both well width and conformational differences. In this case, the drawback is that individual degrees of freedom (bonds, angles and torsions) are assumed to be uncorrelated if only using the first-order approximation, which was done here. It is important to note that neither of these estimation techniques include the entropy loss due to changes in translational and rotational entropy upon ligand binding, which can be estimated to add an additional 3-10 kcal/mol *(63-68)*. Given the assumptions made in the estimates and the lack of clear convergence, it is unwise to treat these values as anything more than qualitative observations. Given that caveat, we do find the average entropic penalty for the fully constrained inhibitors (J6 diastereomers) to be less than that of the less constrained (J2, J3, J4, and J5).

3.4.3 Molecular docking

In order to test whether high throughput computational techniques could be used to predict accurate binding conformations between highly charged ligands and RNA, we performed docking analysis using the inhibitors shown in Figure 3.1 on the crystal RNA conformation with Dock 6.5. With the exception of J1, the weakest binding inhibitor considered, the docking pose with the best score for each inhibitor was consistent with the crystal structure binding mode and formed all of the critical binding contacts (Figure 3.19A,B). The J1 inhibitor binds incorrectly to a major groove pocket on the opposite side of the RNA as the correct ligand pocket. It is unclear why J1 was so poorly docked, but it may reflect its weak binding value. Not surprisingly, the trend in docking scores did not match the experimental binding energy trend (Table 3.9). However, the docking results suggest that fairly accurate binding poses can be obtained for highly charged systems at inexpensive computational cost. Given the somewhat elaborate procedure we used to obtain charges for the inhibitors, we were interested in whether similar docking results could be obtained using a more traditional approach in which the semi-empirical AM1-BCC charge model was used to assign atomic charged to the inhibitors. The results suggest that the more rigorous RESP approach can yield significant improvements in the results (Figure 3.19C and 3.20). This consideration is likely to be more important for multivalent ligands such as those studied in this work. Taken together, the results suggest that methods like Dock 6.5 using reliable scoring functions can be applied to generate reasonable binding modes which

may be further explored using more detailed simulations as presented in this work to get more reliable estimates of relative binding affinities.

3.4.4 Novel Ligands

Given the robustness of the crystal conformation across simulations of all the inhibitors in Figure 3.1, we were interested in how well ligands with a different scaffold could be accommodated by the binding site. Four novel ligands were designed which both exploit the known inhibitor interactions and, in certain respects, reduce the complexity of the ligand (Figure 3.21). These novel ligands have several advantages: less positive charge, fully aromatic rings which are less flexible, and no chiral centers. One of these ligands, N7, was stable in the receptor binding site during a 130 ns simulation (Figure 3.22). MM-GBSA analysis on that trajectory yielded a binding energy of -61 kcal/mol, which is in the same approximate range as the known inhibitors, although our caution regarding implicit solvent models applies. Visual inspection of the trajectories suggests that further optimization of the dimethylamino arm length and orientation would likely produce improved binding.

3.5 Discussion

The presented results are relevant for computational drug development targeting the HCV domain II site as well as research on RNA-ligand binding in general. First, we have demonstrated that the crystal structure conformation appears to be more suitable than the NMR conformation for future inhibitor

development studies and that it forms stable RNA – ligand binding complexes using the AMBER ff12SB force field and explicit solvent. This result highlights the need for accurate initial structures when performing MD simulations. Due to the rugged energy landscape of RNA, incorrect initial structures will likely not reach the global minimum on tractable timescales. Since this concern is coupled with known force field flaws for RNA (a subject of ongoing research in our lab), it necessitates caution when performing MD simulations. For example, prior to the publication of the crystal structure, it was unclear if the structural instability of the NMR structure was due to inherent flexibility of the complex, force field flaws, incorrect experimental structure, or some combination of the three. The stability of the crystal structure in simulation and the preservation of the critical RNA-inhibitor contacts identified in the crystal structure suggest that our computational model is accurate, yet use of caution cannot be overstated. For example, it is still unclear how to resolve conflicts between the NOE restraint data from the NMR studies and the crystal structure.

Even if an accurate simulation model is obtained, drug development efforts require accurate estimates of binding energy. Traditionally, MM-GBSA/MM-PBSA trajectory postprocessing techniques have been moderately successful at predicting the binding free energy of protein-ligand systems. But such studies with highly charged ligands and highly charged receptors (e.g., RNA) are rare. As has been noted elsewhere *(71)*, the binding free energy is largely determined by the difference between the desolvation energy and the bound complex energy. For a highly charged ligand-receptor interaction, both

of these values will be very large and thus errors in the method will dwarf the binding energy value. In this case, the error is likely not related to insufficient sampling (our estimated errors are reasonably small), but an error in the model used to describe the desolvation energy and the bound complex energy. As additional evidence, we have obtained very poor results from simulations of various RNA structures when using the Hawkins, Cramer, and Truhlar GB implicit solvent model (data unpublished). This could explain why the binding energies calculated from the explicit solvent systems compare more favorably with experiment than the binding energies obtained from implicit solvent approximations. Given the relative success of explicit solvent in comparison to implicit solvent for MD simulations of RNA, it is not surprising that energetic results utilizing the former solvation terms would produce better results. Along these same lines, the charge parameterization method appears to be crucial, at least for docking studies. This is not surprising given the large net charge of the inhibitors used in this study and the careful procedure reported here provides a useful framework for future charge parameterization on highly charged ligands. With care in ligand charge derivation, the docking results suggest that methods like Dock 6.5 with good scoring functions can accurately predict ligand binding modes.

Finally, the data suggest that the multiple short simulation approach offers efficiency benefits over the single long simulation approach for energetic analysis. The mean explicit solvent binding energy values computed using both approaches were within error of each other for ten out of the eleven inhibitors

considered. In the case that differed, it seemed likely that the multiple simulation approach was correct based on the argument for internal consistency of stereochemical binding. The aggregate simulation time for the multiple simulation approach used only 480 ns (CRY2 set), significantly less than the 2616 ns of aggregate time used for the single simulation approach (CRY1 set). The drawback to the multiple simulation approach is that it does not ensure binding stability throughout the simulation. The 2 ns trajectories used are not long enough to allow clear structural transitions. Thus, if we had just used 2 ns trajectories, even 240 of them, the difference in stability between the NMR conformation and the crystal conformation would not have been as clear as it was using the longer simulations.

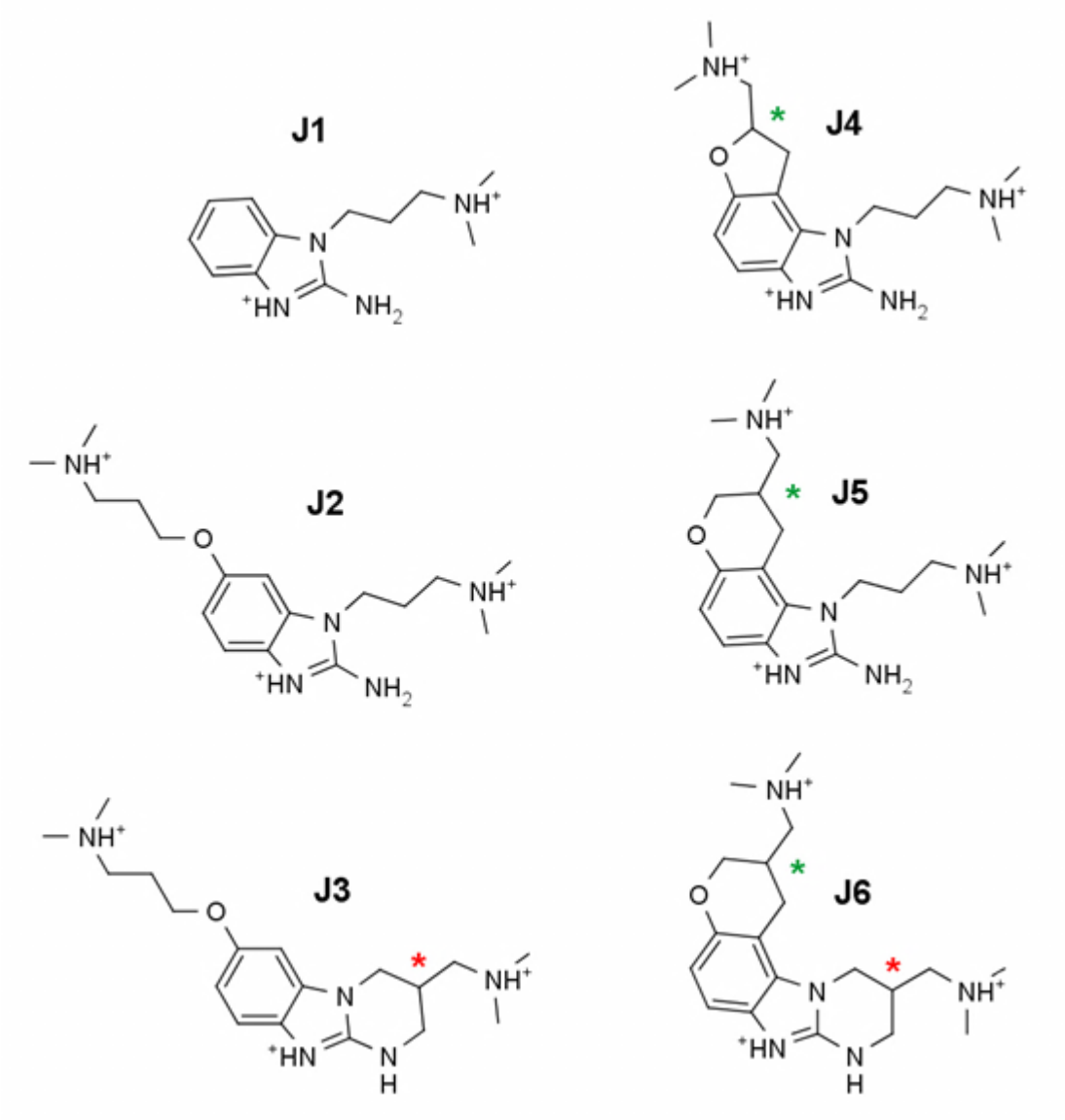

Figure 3.1. These previously identified inhibitors *(18)*, which are studied in this work, bind to the HCV IRES subdomain IIa. Twelve stereochemically distinct inhibitors can be derived from the six structures shown here. The protonation state used in simulation is as depicted. The location of a stereocenter is indicated by a colored star. Throughout this paper the identity of the stereoisomer is given by adding R or S to the name designator (i.e., J3R, J3S, J4R, J4S, J5R and J5S). The J6 diastereomers are identified in the following manner, with the color of the R/S corresponding with colored star location: J6RR, J6RS, J6SR, J6SS. The J4 inhibitor was used in the NMR study *(12)* whereas the J5 inhibitor was used in the crystallography study *(13)*.

Table 3.1. Experimental dissociation constants (μM) and the corresponding binding free energy (kcal/mol) determined previously by mass spectrometry *(18)*.

Inhibitor	K_D (μM)	$\Delta G_{binding}$ (kcal/mol)
J1	>100.00	-5.45 *
J2	17.00	-6.50
J3	3.50	-7.44
J4	1.70	-7.87
J5	0.86	-8.27
J6	0.72	-8.37

The free energy of binding was calculated according to the relation $\Delta G = RT \ln(K_D)$ at 298.15 K. *The exact dissociation constant is not known for J1, thus -5.45 represents the lower bound for the binding energy.

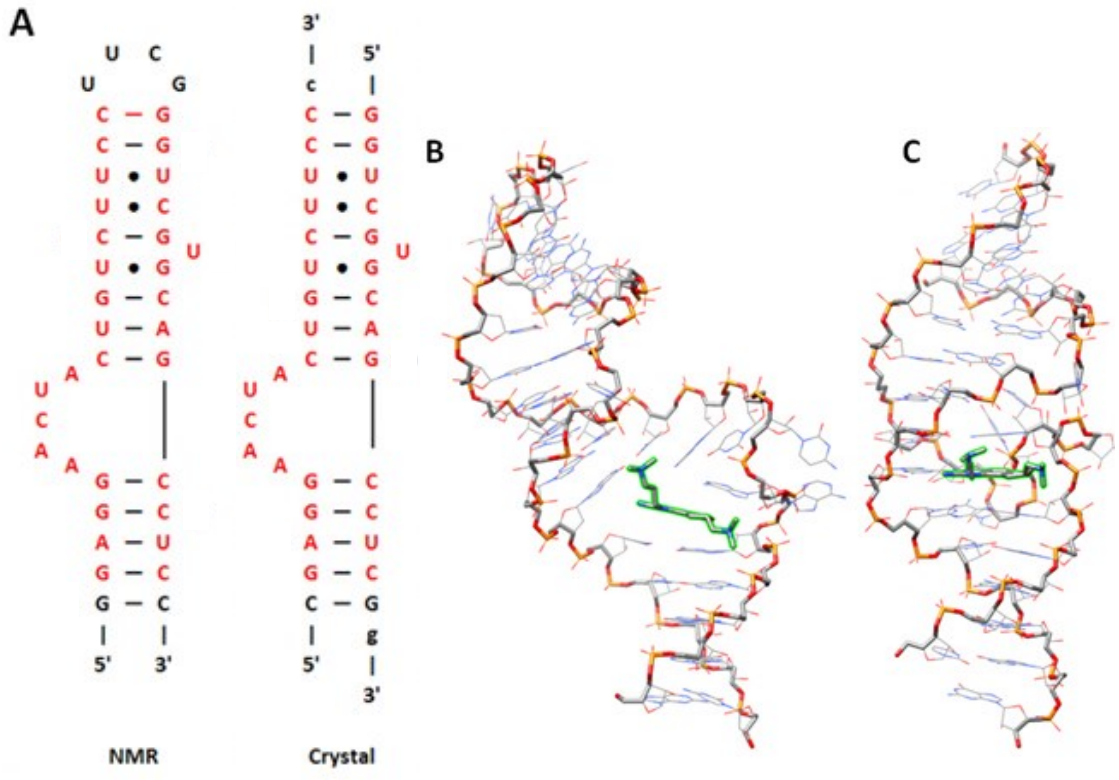

Figure 3.2. Despite similar sequences, the reported conformations of the inhibitor bound HCV IRES domain IIa differ between NMR and X-ray crystallography. (*A*) Secondary structure diagrams of the domain IIa constructs used in the NMR *(12)* and crystallography *(13)* studies. The hairpin sequence from the NMR study was used for all simulations in this study. The residues colored in red show the portion of the RNA which is identical in both published structures. Representative models depict the global structure of the NMR ensemble (*B*) and the crystal structure (*C*). The RNA backbone is emphasized with heavier width and the inhibitor is highlighted in green. The structural orientation was chosen to emphasize the global differences in the binding conformation.

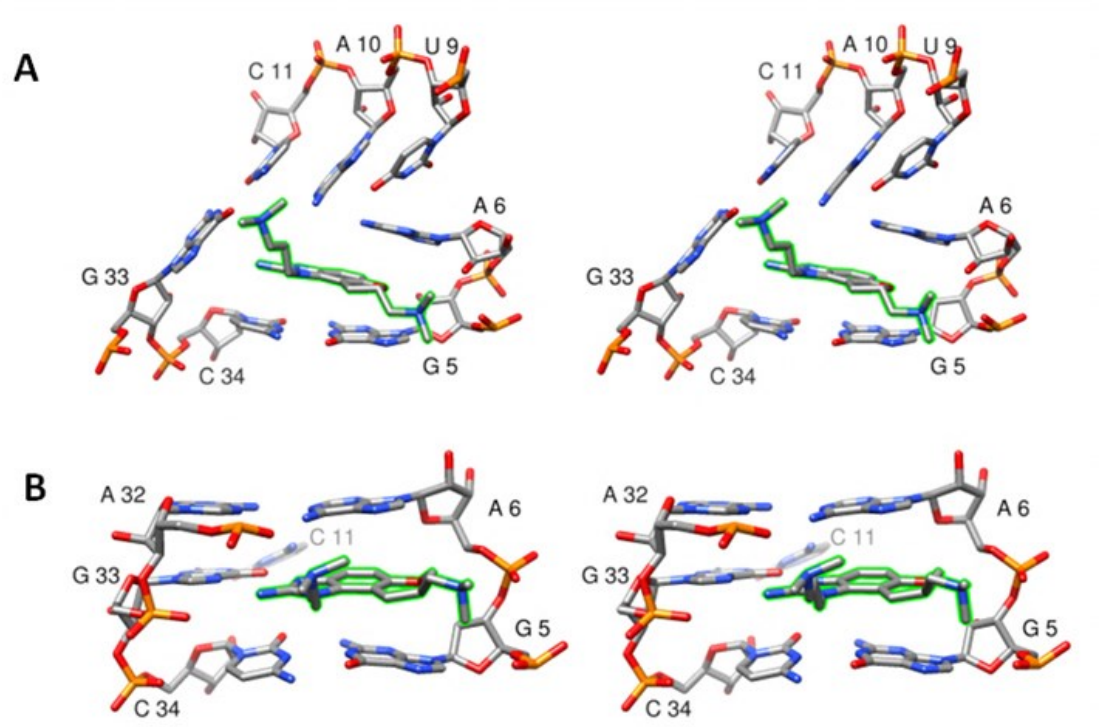

Figure 3.3. Stereo view (wall-eyed) of the inhibitor binding conformations observed in the experimental NMR ensemble (*A*) and the crystal structure (*B*). The inhibitor is highlighted in green.

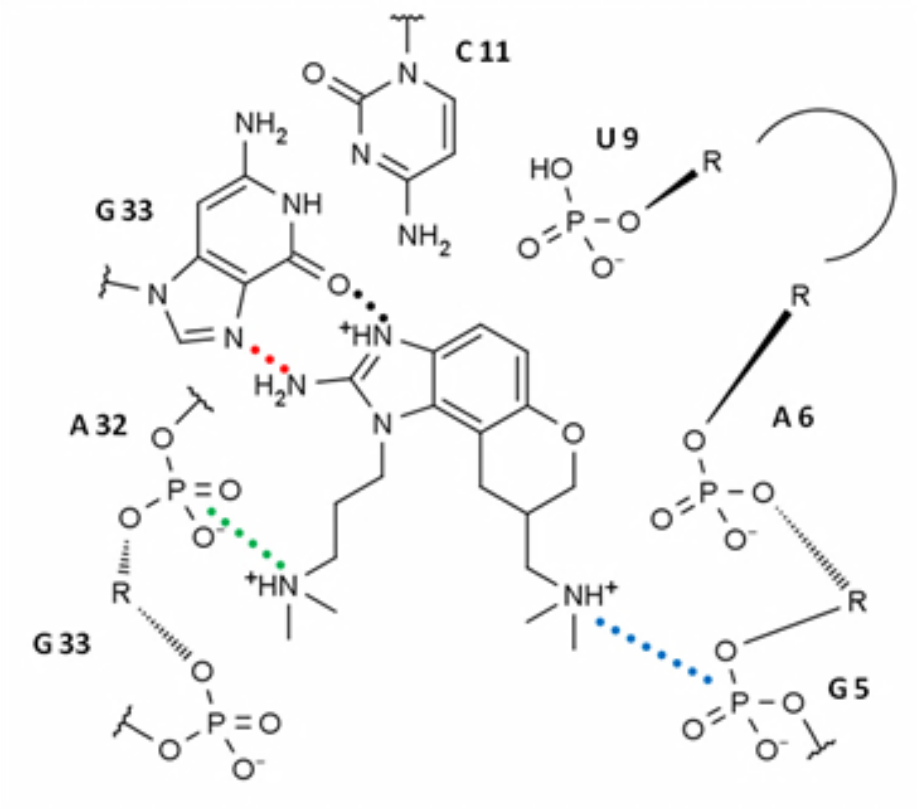

Figure 3.4. A schematic of the crystal structure binding site in the plane of the J5 inhibitor. Residue labels are numbered according to the NMR hairpin sequence (Figure 3.2A). Critical contacts are indicated in dotted, colored lines and the color corresponds to distances depicted in Figure 3.8.

Table 3.2. MD simulations performed in this work.

Simulation Set	System Contents	Ligands	Ligand Poses	Total Simulations	Simulation Length (ns)[5]
NMR1	RNA[1],Ligand[3],K^+,Cl^-,TIP3P	12	1	12	232
CRY1	RNA[2],Ligand[3],K^+,Cl^-,TIP3P	12	1	12	218 +
CRY2	RNA[2],Ligand[3],K^+,Cl^-,TIP3P	12	20	240	2
MG	RNA[2],J5R,Mg^{2+},K^+,Cl^-,TIP3P	1	1	1	82
RDC	RNA[2],J4R,K^+,Cl^-,TIP3P	1	1	1	5
NOV	RNA[2],Ligand[4],K^+,Cl^-,TIP3P	4	1	4	70
LIG	Ligand[3],K^+,Cl^-,TIP3P	12	1	12	594 +

[1]RNA receptor in the NMR conformation. [2]RNA receptor in the crystal conformation. [3]Indicates that the 12 ligands are those described in Figure 3.1. [4]Indicates that the 4 ligands are those described in Figure 3.11. [5]Simulation length represents the simulation time for each simulation (multiply Total Simulations by Simulation Length to get the aggregate time). The "+" indicates that the stated time is the minimum from among the simulation set.

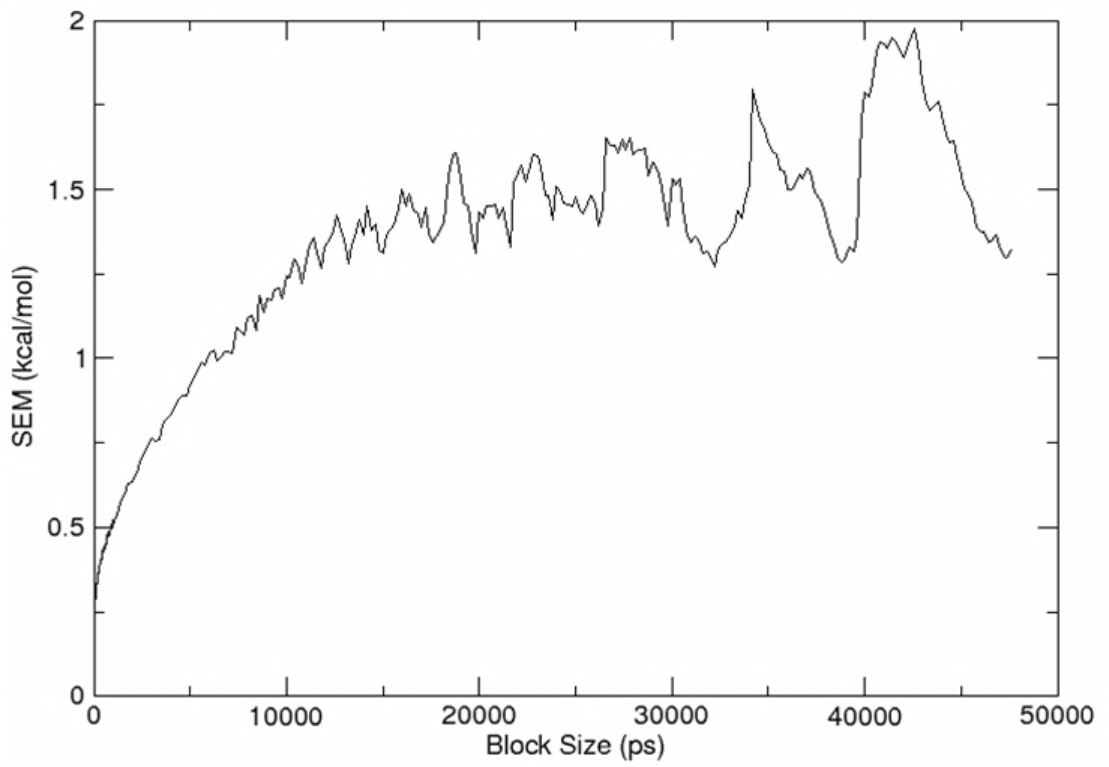

Figure 3.5. An example of error analysis using the re-blocking procedure. The convergence of the standard error of the mean (SEM) with increasing block size (ps) is depicted for the explicitly solvated potential energy (kcal/mol) taken from the J4R simulation of the CRY1 set. According to our protocol, where the maximum observed value is chosen as the error, this plot yields an error in the mean potential energy of 1.97 kcal/mol.

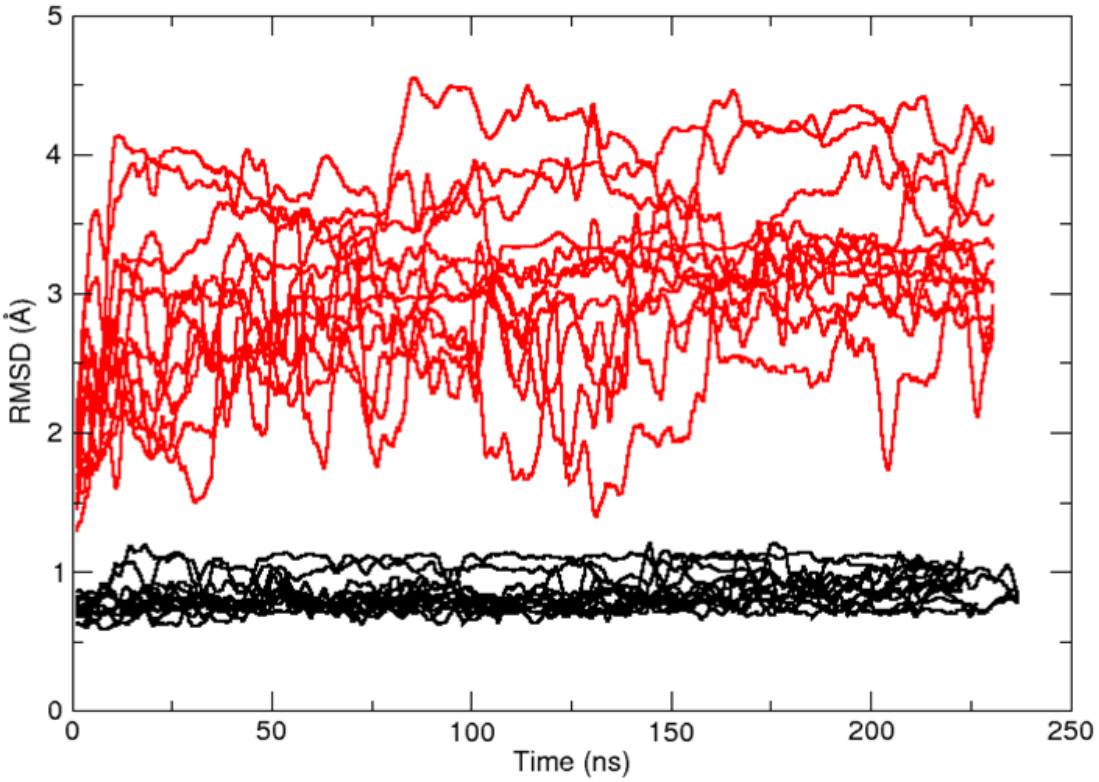

Figure 3.6. The binding region RMSD (Å) reveals that over the course of the simulation time (ns) the CRY1 (*black*) simulation set is much more stable than the NMR1 (*red*) set. Each simulation set is represented by twelve lines which correspond to twelve stereochemically distinct inhibitors (Figure 3.1). The atoms considered in the binding region are defined to be the heavy atoms in residues 5, 6, 32, 33, 34 and the inhibitor. The first frame of each production simulation was used as the RMSD reference structure for that simulation. For clarity, the RMSD values have been smoothed with a 2500 data point running average.

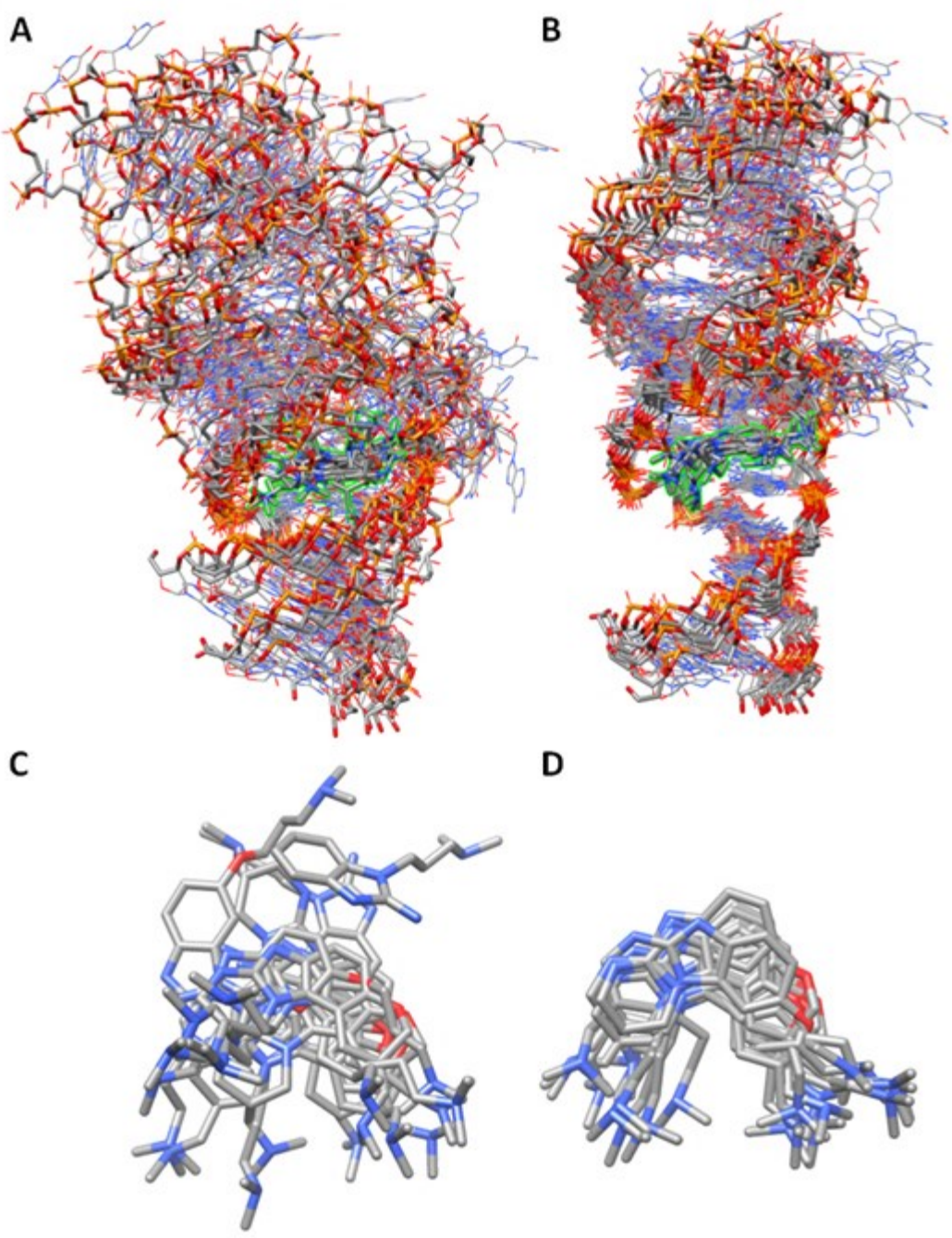

Figure 3.7. Structural ensembles made from the final frames of the NMR1 (*A*) and CRY1 (*B*) simulation sets as defined in Table 3.2. The ensembles were RMS fit using residues 5,6,33, and 34. The final inhibitor position is also shown for NMR1 (*C*) and CRY1 (*D*) using the same RMS fit as the full ensembles.

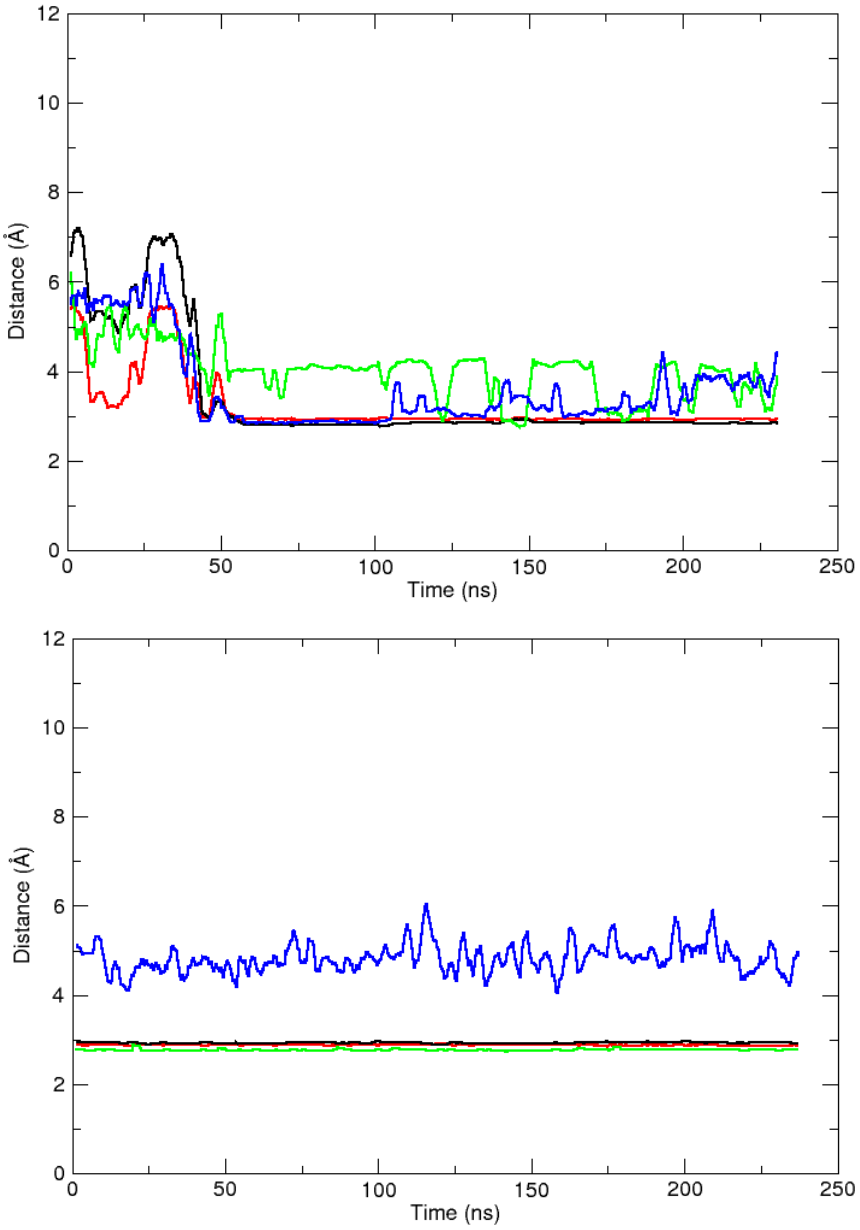

Figure 3.8. The NMR conformation can convert to a crystal-like conformation during simulation. Within 50 ns, the critical RNA-inhibitor distances (Å) of the J5R trajectory from the NMR1 simulation set (*A*) are similar to those observed in the J5R trajectory from CRY1 set (*B*). The line colors correspond to the distances indicated in Figure 3.4. Data are smoothed for clarity using a 2500 data point running average.

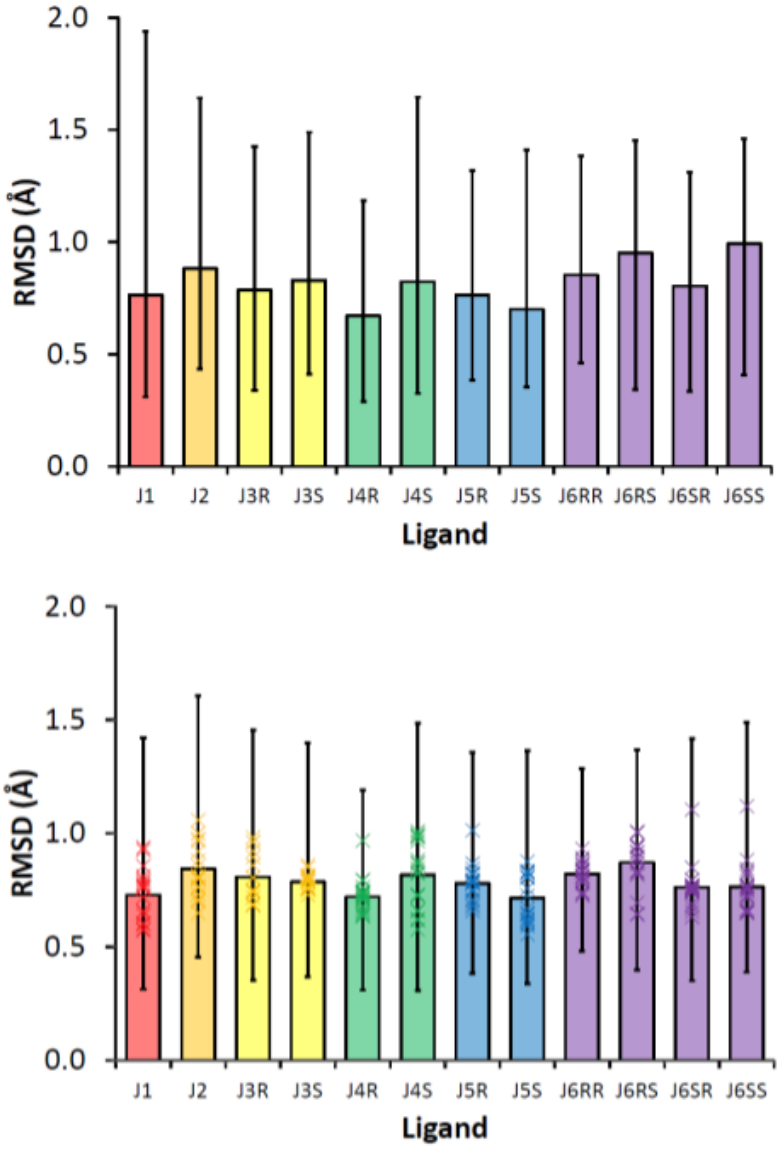

Figure 3.9. Comparison of the binding region RMSD (Å) space explored by the CRY1 (*top*) and CRY2 (*bottom*) simulations suggests that the conformations explored by each approach are not vastly different. The colored bars represent the mean value and the error bars show the minimum and maximum values. The mean value for each of the twenty individual CRY2 simulations is depicted by "x" data points (*bottom*), whereas the bar shows the overall mean value. A single reference structure, for each RNA-inhibitor complex, was used to compute the RMSD values. These values are given in Table 3.3.

Table 3.3. The binding region RMSD (Å) values for the CRY1 and CRY2 simulations sets.

Ligands	CRY1 RMSD			CRY2 RMSD		
	Avg	Min	Max	Avg	Min	Max
J1	0.8	0.3	1.9	0.7	0.3	1.4
J2	0.9	0.4	1.6	0.8	0.5	1.6
J3R	0.8	0.3	1.4	0.8	0.4	1.5
J3S	0.8	0.4	1.5	0.8	0.4	1.4
J4R	0.7	0.3	1.2	0.7	0.3	1.2
J4S	0.8	0.3	1.6	0.8	0.3	1.5
J5R	0.8	0.4	1.3	0.8	0.4	1.4
J5S	0.7	0.4	1.4	0.7	0.3	1.4
J6RR	0.9	0.5	1.4	0.8	0.5	1.3
J6RS	1.0	0.3	1.5	0.9	0.4	1.4
J6SR	0.8	0.3	1.3	0.8	0.4	1.4
J6SS	1.0	0.4	1.5	0.8	0.4	1.5

Note: these values correspond to those displayed in Figure 3.9.

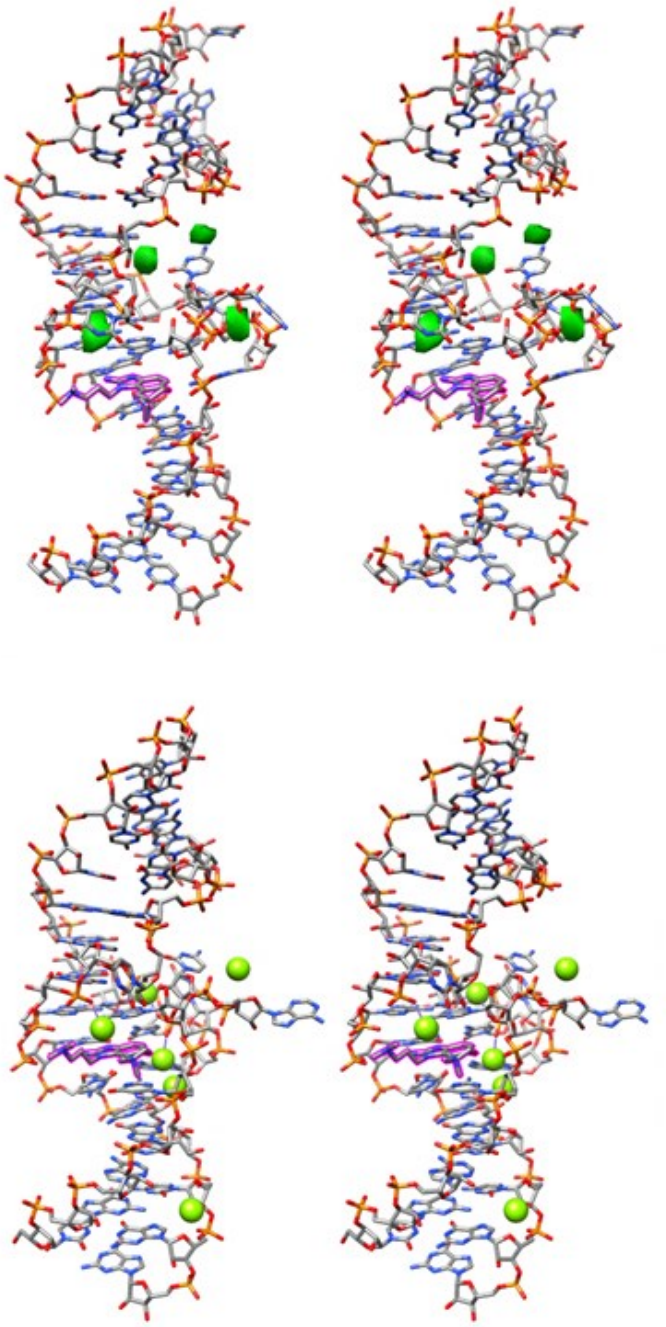

Figure 3.10. Stereo views (wall-eyed) of the Mg^{2+} binding locations in simulation (*top*) and in the experimental structure (*bottom*) reveal subtle differences in the binding location but very little difference in the inhibitor binding mode. In the simulation structure, the green surfaces indicate regions of high Mg^{2+} density as determined by grid analysis and are overlaid on the average structure of the RNA-inhibitor complex. Green spheres in the experimental structure indicate regions of electron density that correspond to Mg^{2+}. The inhibitor, J5R, is highlighted in pink.

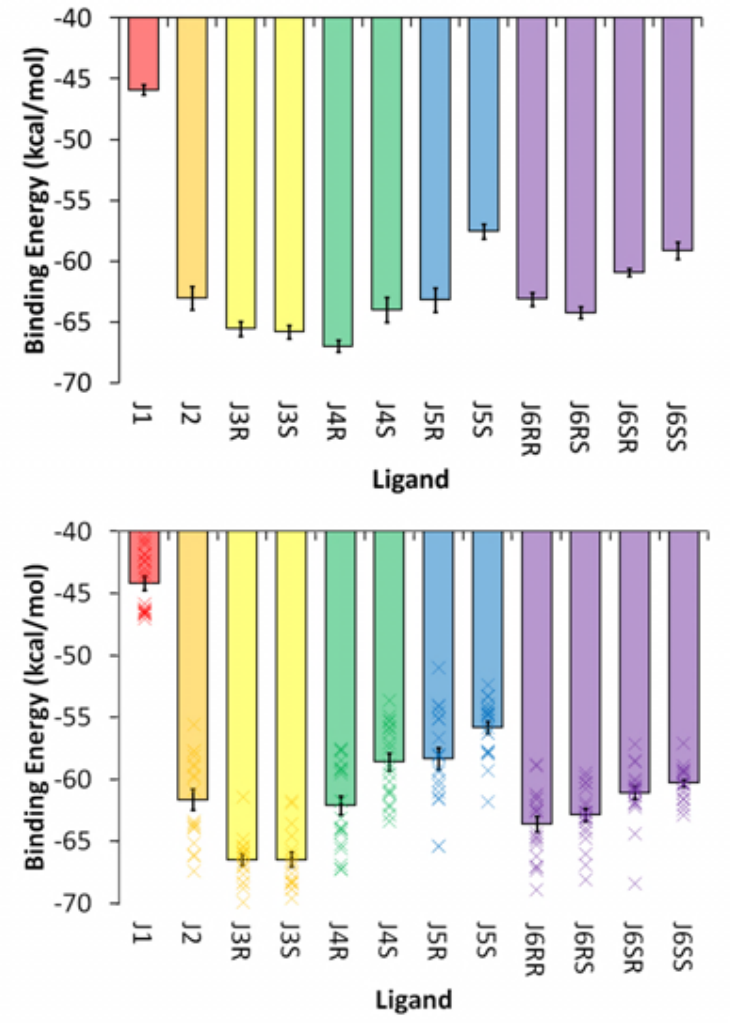

Figure 3.11. MM-GBSA binding energy results for the CRY1 (*top*) and CRY2 (*bottom*) simulation sets. The mean value for each of the twenty individual CRY2 simulations is depicted by "x" data points (*bottom*), whereas the bar shows the overall mean value. All data units are kcal/mol and the specific values for both figures are given in Table 3.4.

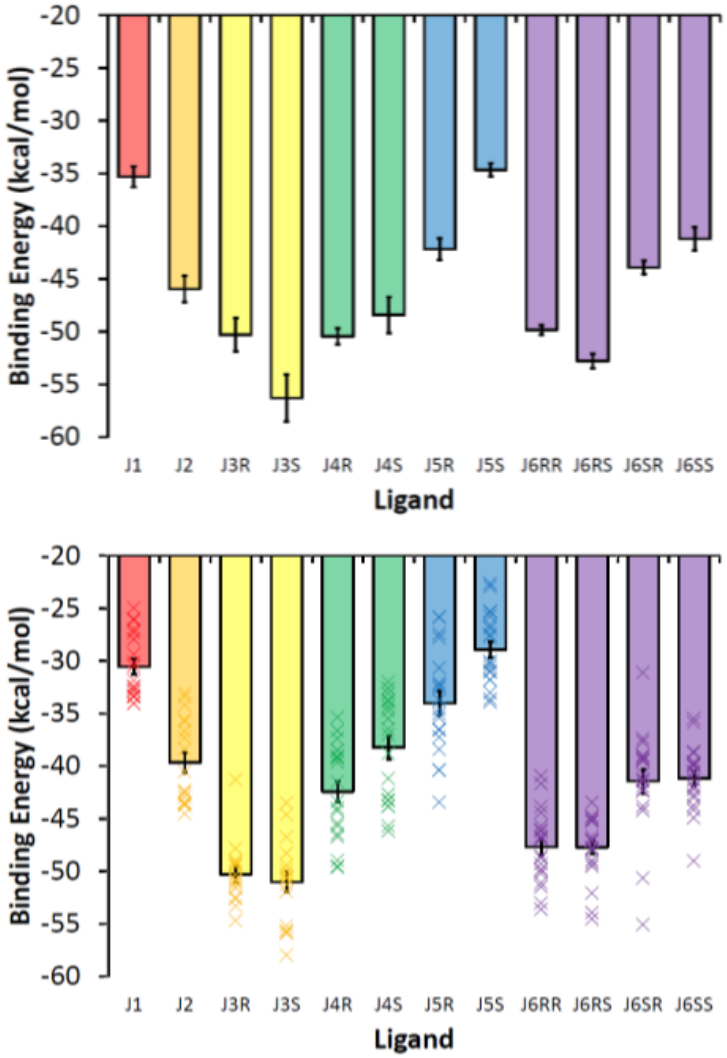

Figure 3.12. The MM-PBSA binding energy (kcal/mol) results for the CRY1 (*top*) and CRY2 (*bottom*) simulation sets. The average value for each of the twenty CRY2 simulations is depicted by "x" data points (*bottom*). Values are given in Table 3.4.

Table 3.4. The MM-GBSA and MM-PBSA binding energies (kcal/mol) for the CRY1 and CRY2 simulation sets.

Ligands	CRY1 MMGBSA		CRY2 MMGBSA		CRY1 MMPBSA		CRY2 MMPBSA	
J1	-45.9	(0.4)	-44.2	(0.5)	-35.3	(1.0)	-30.5	(0.7)
J2	-63.0	(1.0)	-61.6	(0.8)	-46.0	(1.3)	-39.7	(1.0)
J3R	-65.6	(0.6)	-66.5	(0.5)	-50.3	(1.6)	-50.3	(0.7)
J3S	-65.8	(0.6)	-66.5	(0.6)	-56.3	(2.2)	-51.0	(1.0)
J4R	-67.0	(0.5)	-62.1	(0.8)	-50.4	(0.8)	-42.4	(1.0)
J4S	-64.0	(1.0)	-58.6	(0.7)	-48.4	(1.7)	-38.2	(1.1)
J5R	-63.2	(1.0)	-58.3	(0.9)	-42.2	(1.0)	-34.0	(1.1)
J5S	-57.5	(0.6)	-55.8	(0.5)	-34.7	(0.6)	-28.9	(0.8)
J6RR	-63.1	(0.5)	-63.6	(0.6)	-49.8	(0.5)	-47.7	(0.8)
J6RS	-64.2	(0.5)	-62.9	(0.5)	-52.8	(0.7)	-47.7	(0.7)
J6SR	-60.9	(0.3)	-61.1	(0.5)	-43.9	(0.6)	-41.4	(1.1)
J6SS	-59.1	(0.7)	-60.3	(0.3)	-41.2	(1.1)	-41.2	(0.7)

Note: these data are depicted in Figures 3.11 and 3.12. The error is given in parentheses.

Table 3.5. The data values (kcal/mol) for the explicit solvent potential energy binding calculations from the CRY1, CRY2, and LIG simulation sets.

	CRY1	CRY2	LIG	CRY1-LIG	CRY2-LIG
Lig.	TIP3P EPtot	TIP3P EPtot	TIP3P EPtot	TIP3P EPtot	TIP3P EPtot
J1	-118851.9 (1.1)	-118850.1 (1.1)	-20523.7 (0.4)		
J2	-118774.4 (1.5)	-118774.4 (1.4)	-20756.2 (0.2)	-98018.2 (1.6)	-98018.2 (1.6)
J3R	-118750.2 (2.5)	-118747.0 (1.5)	-20727.0 (0.2)	-98023.1 (2.7)	-98020.0 (1.7)
J3S	-118744.6 (1.8)	-118745.4 (1.1)	-20726.8 (0.2)	-98017.9 (2.0)	-98018.7 (1.3)
J4R	-118772.3 (2.0)	-118772.4 (1.4)	-20750.4 (0.2)	-98021.8 (2.1)	-98022.0 (1.5)
J4S	-118768.9 (2.9)	-118770.9 (1.2)	-20750.0 (0.2)	-98018.9 (3.0)	-98020.9 (1.4)
J5R	-118761.7 (1.9)	-118760.5 (1.0)	-20739.9 (0.2)	-98021.8 (2.0)	-98020.6 (1.2)
J5S	-118762.0 (1.4)	-118762.6 (0.9)	-20739.5 (0.3)	-98022.5 (1.7)	-98023.2 (1.2)
J6RR	-118727.8 (2.0)	-118730.5 (1.1)	-20707.5 (0.3)	-98020.3 (2.3)	-98023.0 (1.4)
J6RS	-118734.2 (2.1)	-118736.9 (1.3)	-20717.4 (0.3)	-98016.8 (2.4)	-98019.5 (1.6)
J6SR	-118742.1 (1.6)	-118742.4 (1.2)	-20717.5 (0.1)	-98024.6 (1.7)	-98024.9 (1.3)
J6SS	-118732.8 (1.7)	-118728.1 (1.4)	-20707.8 (0.3)	-98025.0 (2.0)	-98020.4 (1.7)

Note: values in the fifth and sixth columns correspond to those displayed in Figure 3.13. The error is given in parentheses.

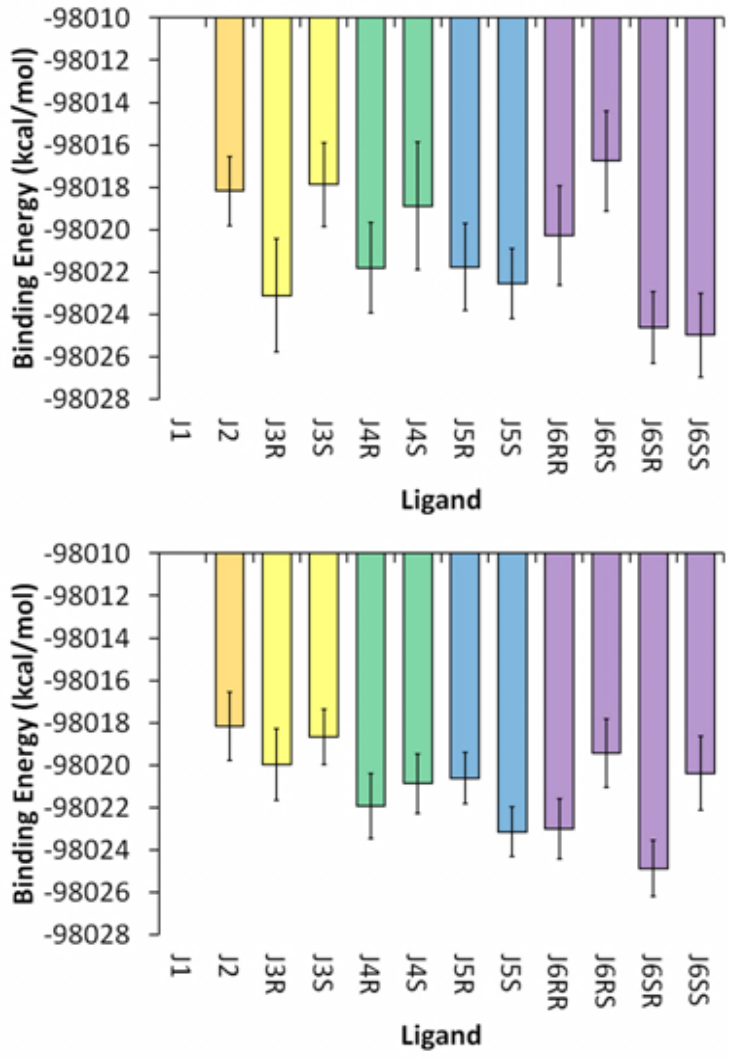

Figure 3.13. Relative binding energy using the solvated potential energy difference (see text) from the CRY1/LIG (*top*) and CRY2/LIG (*bottom*) simulation sets. Data for the J1 inhibitor are not shown due to the difference in charge with the rest of the inhibitors which results in systems that cannot be directly compared. All data units are kcal/mol and the specific values for both figures are given in Table 3.5.

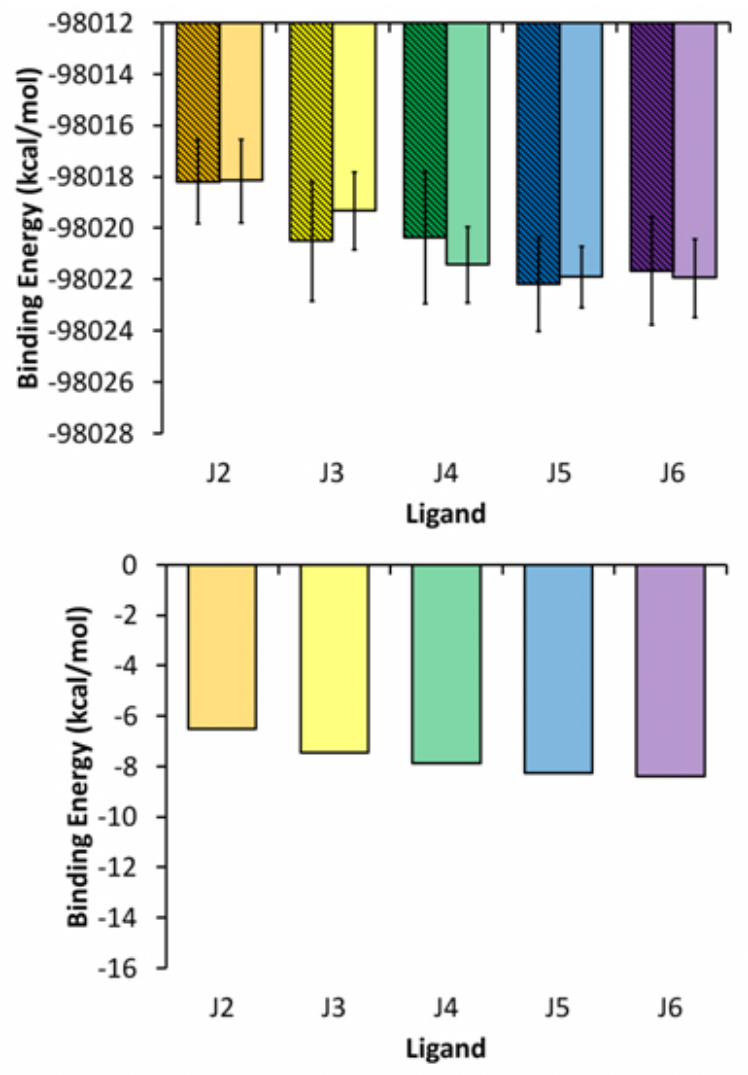

Figure 3.14. Stereochemically averaged relative binding energy using the solvated potential energy difference (see text) for the CRY1/LIG (*top, striped bars*) and CRY2-LIG (*top, solid bars*) simulation sets. Data for the J1 inhibitor are not shown due to the difference in charge with the rest of the inhibitors which results in systems that cannot be directly compared. (*Bottom*) Previously reported experimental binding energies *(18)*. All data units are kcal/mol and the specific values for both the top and bottom chart are given in Table 3.6.

Table 3.6. Stereochemically averaged simulation and experimental binding energies (kcal/mol).

	CRY1-LIG		CRY2-LIG		Experimental
Ligands	**TIP3P EPtot**		**TIP3P EPtot**		**Free Energy**
J2	-98018.2	(1.6)	-98018.2	(1.6)	-6.5
J3	-98020.5	(2.3)	-98019.3	(1.5)	-7.4
J4	-98020.4	(2.6)	-98021.4	(1.5)	-7.9
J5	-98022.2	(1.8)	-98021.9	(1.2)	-8.3
J6	-98021.7	(2.1)	-98021.9	(1.5)	-8.4

Note: the simulation binding energy values (from explicit solvent potential energies) were averaged based on stereochemistry from values listed in the fifth and sixth columns of Table 3.5. Errors are given in parentheses.

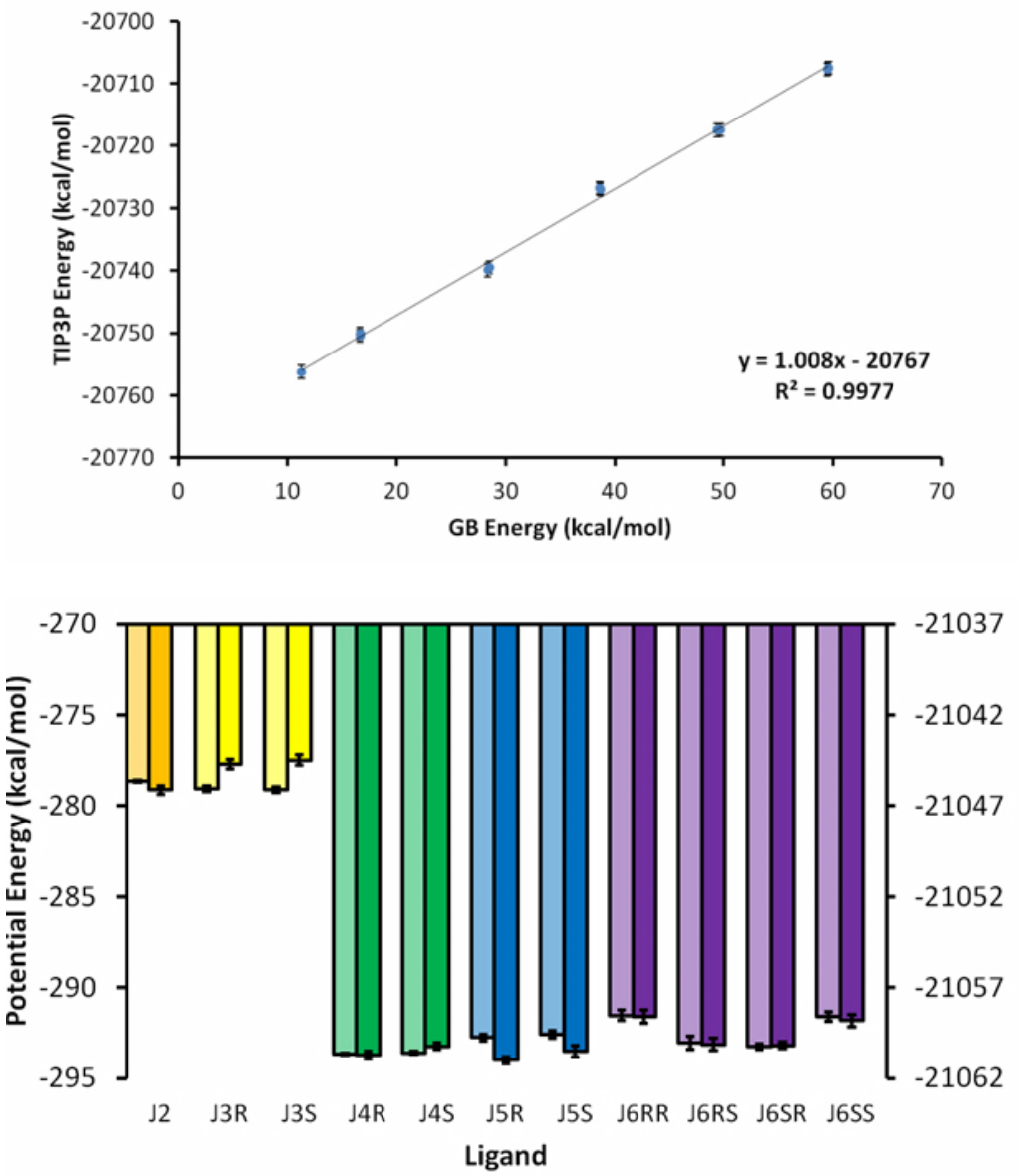

Figure 3.15. Inhibitor solvation energy (kcal/mol) analyses for the LIG simulation set reveal that implicit and explicit solvation models yield similar results. (*Top*) Plot of the average GB potential energy versus the average solvated potential energy. A linear regression trendline fit is shown. (*Bottom*) Solvation energies for GB solvent model (*left, light bars*) and TIP3P explicit solvent (*right, dark bars*). Solvation energy is determined by subtracting the gas phase potential energy from the solvated potential energy. Data values are listed in Table 3.7.

Table 3.7. The solvated potential energy (kcal/mol) and free energy of solvation (kcal/mol) for inhibitors in both implicit and explicit solvent.

Ligands	Ligand EPtot GB solvent		Ligand EPtot TIP3P		$\Delta G_{solvation}$ GB solvent		$\Delta G_{solvation}$ TIP3P	
J1	5.81	(0.02)	-20523.7	(0.4)	-159.5	(0.0)	-20689.0	(0.4)
J2	11.27	(0.02)	-20756.2	(0.2)	-278.6	(0.1)	-21046.1	(0.2)
J3R	38.62	(0.09)	-20727.0	(0.2)	-279.0	(0.1)	-21044.7	(0.2)
J3S	38.61	(0.06)	-20726.8	(0.2)	-279.1	(0.1)	-21044.5	(0.3)
J4R	16.61	(0.03)	-20750.4	(0.2)	-293.7	(0.1)	-21060.7	(0.2)
J4S	16.61	(0.02)	-20750.0	(0.2)	-293.6	(0.1)	-21060.2	(0.2)
J5R	28.33	(0.15)	-20739.9	(0.2)	-292.8	(0.2)	-21061.0	(0.2)
J5S	28.47	(0.19)	-20739.5	(0.3)	-292.6	(0.2)	-21060.5	(0.3)
J6RR	59.56	(0.22)	-20707.5	(0.3)	-291.5	(0.3)	-21058.6	(0.4)
J6RS	49.68	(0.30)	-20717.4	(0.3)	-293.0	(0.4)	-21060.1	(0.3)
J6SR	49.40	(0.10)	-20717.5	(0.1)	-293.3	(0.2)	-21060.2	(0.2)
J6SS	59.46	(0.24)	-20707.8	(0.3)	-291.6	(0.3)	-21058.8	(0.3)

Note: Data values for solvated potential energies and free energy of solvation as depicted in Figure 3.15. The free energy of solvation was determined by subtracting the solvated potential energy from the *in vacuo* potential energy (data not shown).

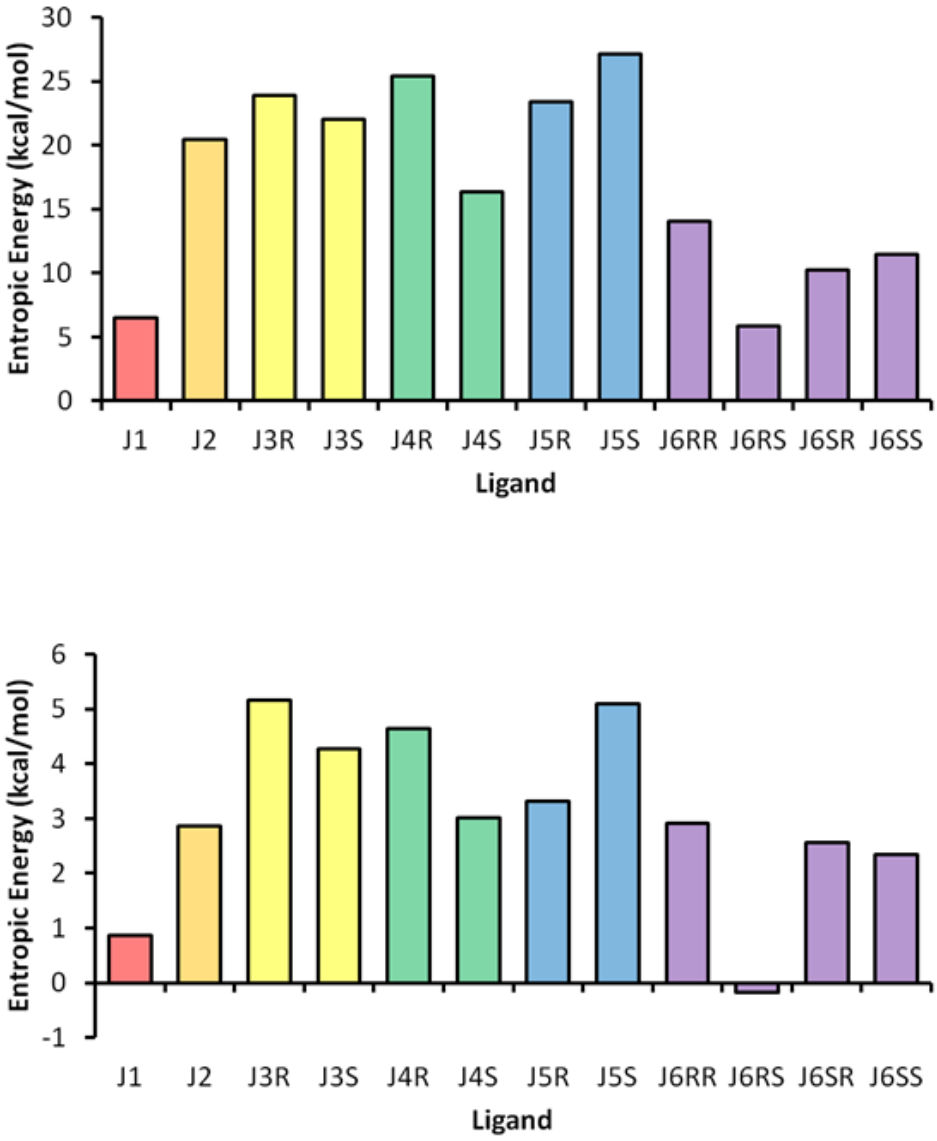

Figure 3.16. Similar trends, but differing magnitudes, are observed when comparing the ligand binding entropy penalty (kcal/mol) using the quasi-harmonic method (*top*) and the first-order configuration entropy (*bottom*). The penalty is calculated as –TΔS, where T is 298.15 K and ΔS is the absolute ligand entropy in free solution (from the LIG simulation set) subtracted from the ligand entropy in complex (from the CRY1 simulation set). Convergence of these values is depicted in Figures 3.17 and 3.18 and the data values are listed in Table 3.8.

Table 3.8. The binding entropy penalty (kcal/mol) calculated from the CRY1 and LIG simulation sets.

Ligands	Quasi-Harmonic	Configurational Entropy
J1	6.5	0.9
J2	20.4	2.9
J3R	23.9	5.2
J3S	22.0	4.3
J4R	25.4	4.6
J4S	16.4	3.0
J5R	23.4	3.3
J5S	27.2	5.1
J6RR	14.0	2.9
J6RS	5.8	-0.2
J6SR	10.3	2.6
J6SS	11.4	2.3

Note: the penalty is calculated as $-T\Delta S$, where T is 298.15 K and ΔS is the absolute ligand entropy in free solution (from the LIG simulation set) subtracted from the ligand entropy in complex (from the CRY1 simulation set). These data are depicted in Figure 3.16.

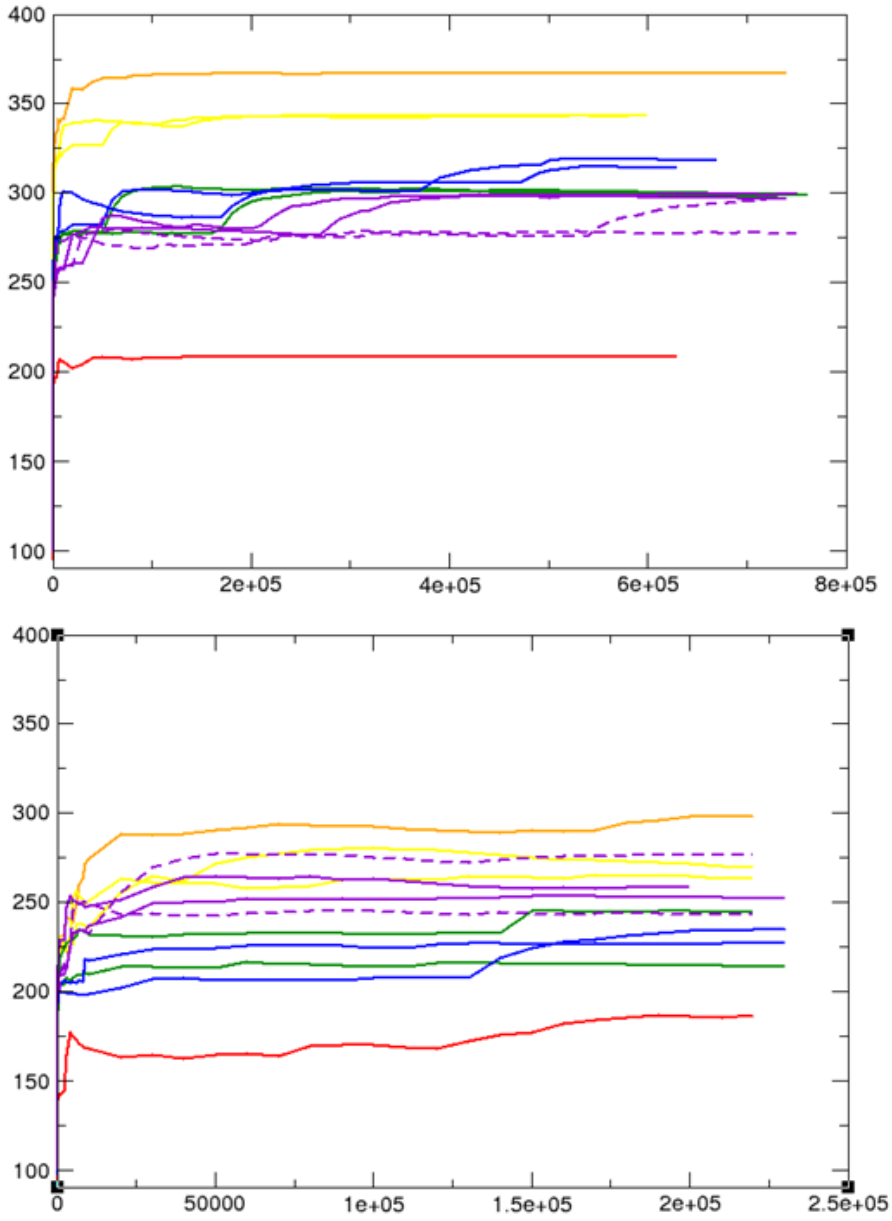

Figure 3.17. The quasi-harmonic entropy (cal/mol·K) for inhibitors in free solution (*top*) and in complex with RNA (*bottom*). The horizontal axis is time (ps). Colors correspond to the following inhibitors: J1 (*red*), J2 (*orange*), J3 (*yellow*), J4 (*green*), J5 (*blue*), J6RR and J6SS (*purple, solid*), J6RS and J6SR (*purple, dashed*). Data were taken from the LIG and CRY1 simulation sets.

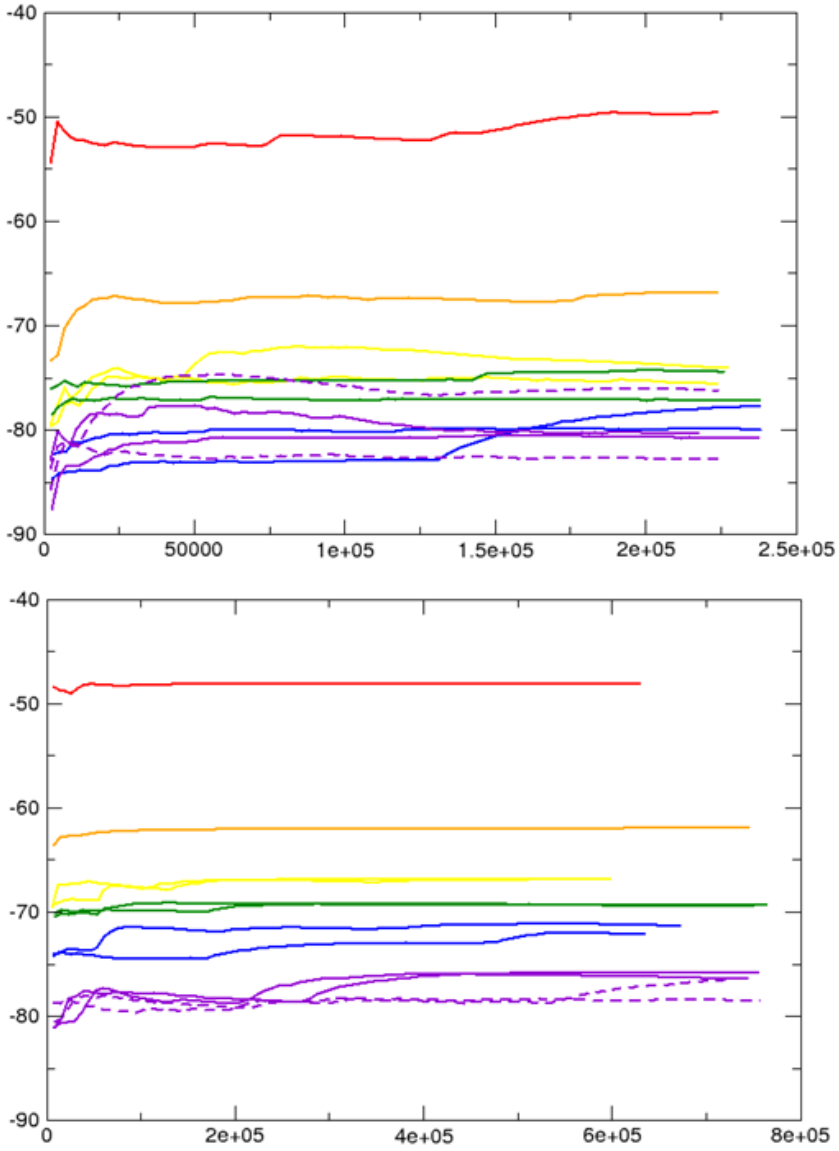

Figure 3.18. The first-order configurational entropy (cal/mol·K) for inhibitors in free solution (*top*) and in complex with RNA (*bottom*). The horizontal axis is time (ps). Colors correspond to the following inhibitors: J1 (*red*), J2 (*orange*), J3 (*yellow*), J4 (*green*), J5 (*blue*), J6RR and J6SS (*purple, solid*), J6RS and J6SR (*purple, dashed*). Data were taken from the LIG and CRY1 simulation sets.

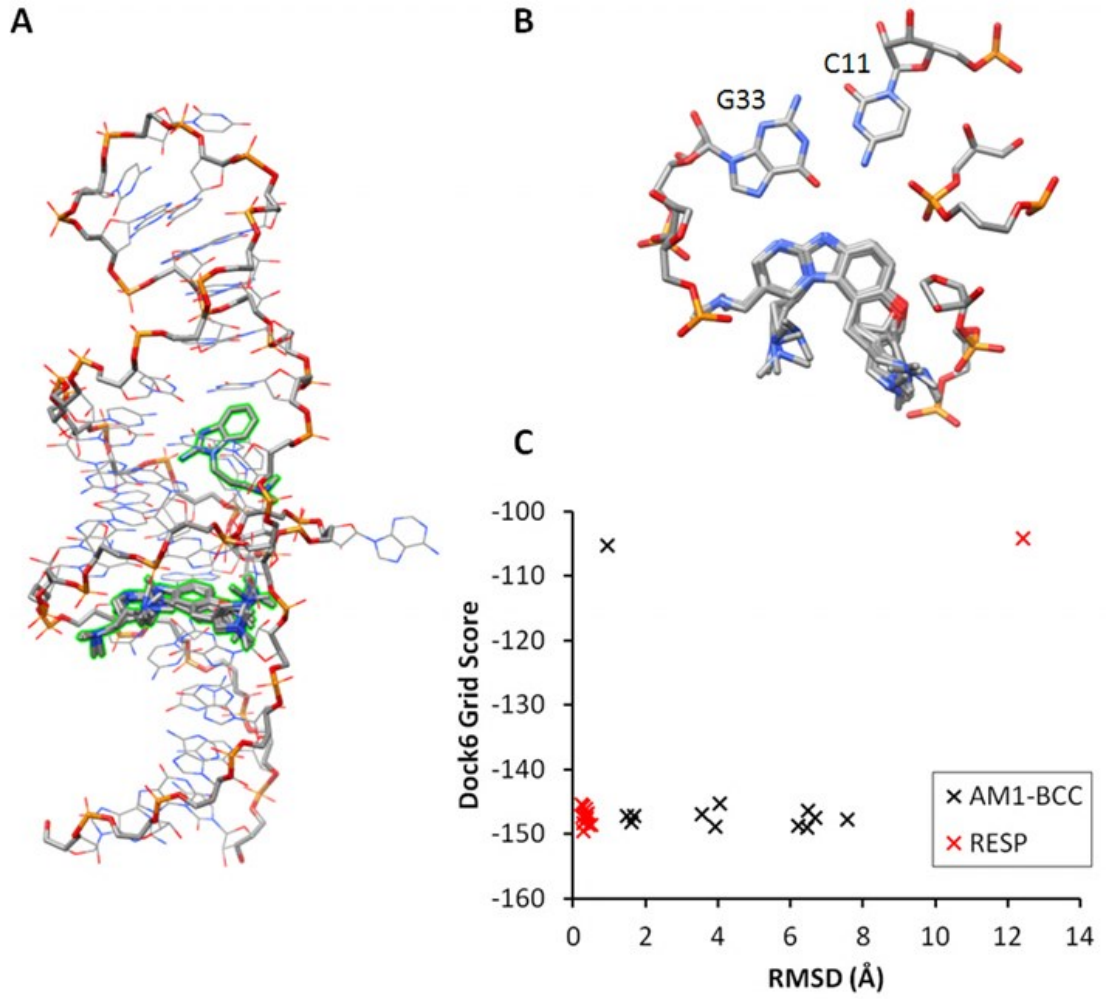

Figure 3.19. Docking results using the crystal structure receptor conformation. (*A*) Overlay of best scoring docking poses for each of the twelve stereochemically distinct inhibitors on the crystal structure receptor RNA. The only inhibitor that did not bind in the expected orientation was the weakest binder, J1. (*B*) Close-up view of the of the inhibitor docking poses in the binding site. (*C*) Comparison of the AM1-BCC (*black*) and RESP (*red*) charge methods: the best scoring pose for each of the twelve inhibitors is plotted with its corresponding RMSD value (experimental crystal structure used as reference, benzimidazole core atoms only). The outlier data point for the RESP values (*red*) is the weak binding J1 inhibitor.

Table 3.9. The best docking grid scores for the binding of the twelve stereochemically distinct inhibitors represented in Figure 3.1 to the RNA target in the crystal conformation.

Ligands	Docking Grid Score
J1	-104.0
J2	-147.5
J3R	-147.3
J3S	-145.1
J4R	-144.5
J4S	-142.7
J5R	-143.1
J5S	-143.7
J6RR	-140.2
J6RS	-142.3
J6SR	-141.9
J6SS	-140.6

Note: docking was performed with DOCK 6.5 and the data shown used the RESP charges (see section 3.2 Methods).

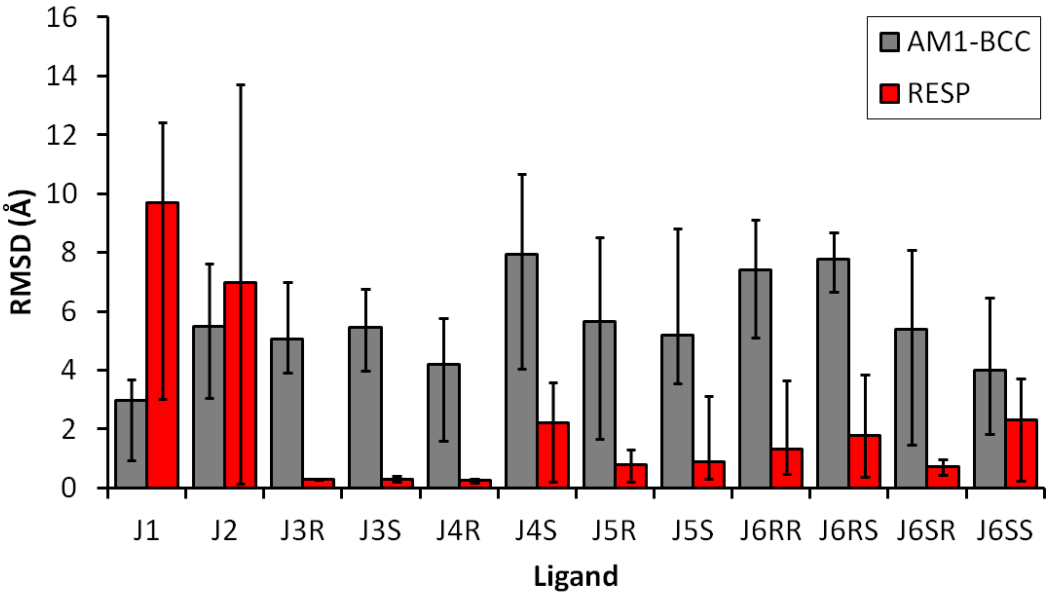

Figure 3.20. Comparison of the average core atom RMSD values (with respect to the crystal structure inhibitor core atoms) for the five best scoring docking poses of each inhibitor suggests that the RESP method yields improved results. Error bars indicate the minimum and maximum observed values.

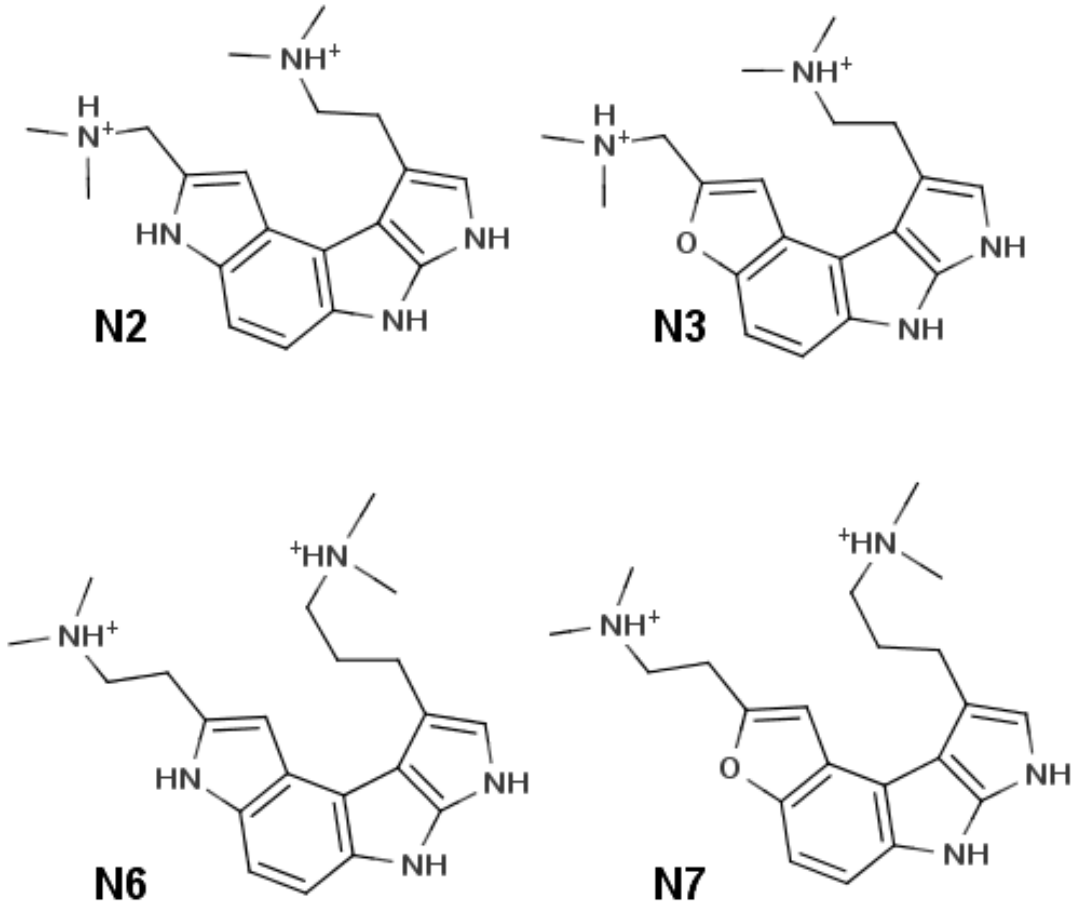

Figure 3.21. Novel ligands investigated in the NOV simulation set.

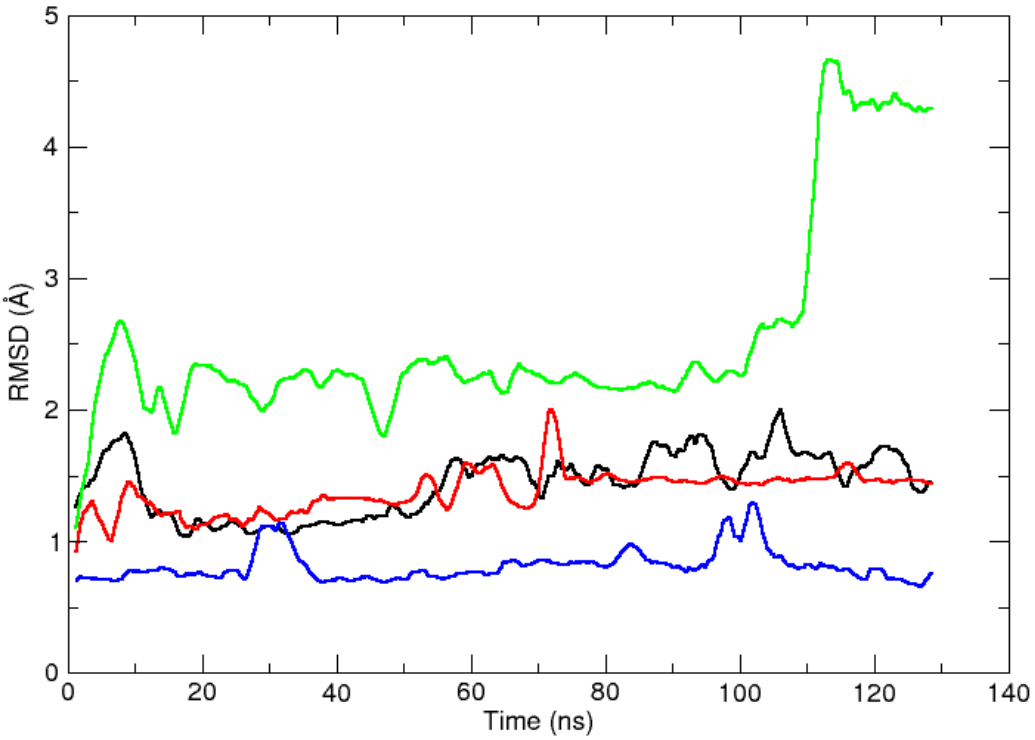

Figure 3.22. The binding region RMSD (Å) time series (ns) for the four trajectories in the NOV1 simulation set. The binding region is defined to include residues 5,6,32,33,34 and the inhibitor. The first frame of production simulation was used as the RMSD reference structure for each trajectory. The RMSD values have been smoothed with a 2500 data point running average. Colors refer to inhibitors listed in Figure 3.21 as follows: N2 (*black*), N3 (*red*), N6 (*green*), N7 (*blue*).

3.6 References

1. Serganov, A., and Patel, D. J. (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins, *Nat. Rev. Genet. 8*, 776-790.

2. Lilley, D. M. (2011) Catalysis by the nucleolytic ribozymes, *Biochem. Soc. Trans. 39*, 641-646.

3. Talini, G., Branciamore, S., and Gallori, E. (2011) Ribozymes: flexible molecular devices at work, *Biochimie 93*, 1998-2005.

4. Sucheck, S. J., and Wong, C.-H. (2000) RNA as a target for small molecules, *Curr. Opin. Chem. Biol. 4*, 678-686.

5. Claros, M. G., and Cánovas, F. M. (1999) RNA isolation from plant tissues: a practical experience for biological undergraduates, *Biochem. Educ. 27*, 110-113.

6. Rio, D. C., Ares, M., and Nilsen, T. W. (2011) *RNA: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

7. Lang, P. T., Brozell, S. R., Mukherjee, S., Pettersen, E. F., Meng, E. C., Thomas, V., Rizzo, R. C., Case, D. A., James, T. L., and Kuntz, I. D. (2009) DOCK 6: combining techniques to model RNA-small molecule complexes, *RNA 15*, 1219-1230.

8. Chen, L., Calin, G. A., and Zhang, S. (2012) Novel insights of structure-based modeling for RNA-targeted drug discovery, *J. Chem. Inf. Model. 52*, 2741-2753.

9. Davis, D. R., and Seth, P. P. (2011) Therapeutic targeting of HCV internal ribosomal entry site RNA, *Antivir. Chem. Chemother. 21*, 117-128.

10. Lukavsky, P. J., Kim, I., Otto, G. A., and Puglisi, J. D. (2003) Structure of HCV IRES domain II determined by NMR, *Nat. Struct. Mol. Biol. 10*, 1033-1038.

11. Dibrov, S. M., Johnston-Cox, H., Weng, Y.-H., and Hermann, T. (2007) Functional architecture of HCV IRES domain II stabilized by divalent metal ions in the crystal and in solution, *Angew. Chemie 119*, 230-233.

12. Paulsen, R. B., Seth, P. P., Swayze, E. E., Griffey, R. H., Skalicky, J. J., Cheatham III, T. E., and Davis, D. R. (2010) Inhibitor-induced structural change in the HCV IRES domain IIa RNA, *Proc. Natl. Acad. Sci. 107*, 7263-7268.

13. Dibrov, S. M., Ding, K., Brunn, N. D., Parker, M. A., Bergdahl, B. M.,

Wyles, D. L., and Hermann, T. (2012) Structure of a hepatitis C virus RNA domain in complex with a translation inhibitor reveals a binding mode reminiscent of riboswitches, *Proc. Natl. Acad. Sci. 109*, 5223-5228.

14. Magnet, S., and Blanchard, J. S. (2005) Molecular insights into aminoglycoside action and resistance, *Chem Rev 105*, 477-498.

15. Becker, B., and Cooper, M. A. (2012) Aminoglycoside antibiotics in the 21st century, *ACS Chem. Biol*.

16. Locker, N., Easton, L. E., and Lukavsky, P. J. (2007) HCV and CSFV IRES domain II mediate eIF2 release during 80S ribosome assembly, *EMBO J. 26*, 795-805.

17. Lukavsky, P. J. (2009) Structure and function of HCV IRES domains, *Virus Res. 139*, 166-171.

18. Seth, P. P., Miyaji, A., Jefferson, E. A., Sannes-Lowery, K. A., Osgood, S. A., Propp, S. S., Ranken, R., Massire, C., Sampath, R., Ecker, D. J., et al. (2005) SAR by MS: discovery of a new class of RNA-binding small molecules for the hepatitis C virus internal ribosome entry site IIA subdomain, *J. Med. Chem. 48*, 7099-7102.

19. Draper, D. E. (2008) RNA folding: thermodynamic and molecular descriptions of the roles of ions, *Biophys. J. 95*, 5489-5495.

20. Bowman, J. C., Lenz, T. K., Hud, N. V., and Williams, L. D. (2012) Cations in charge: magnesium ions in RNA folding and catalysis, *Curr. Opin. Struct. Biol. 22*, 262-272.

21. Tan, Z. J., and Chen, S. J. (2011) Importance of diffuse metal ion binding to RNA, *Metal Ions in Life Sciences 9*, 101-124.

22. Lambert, D., Leipply, D., Shiman, R., and Draper, D. E. (2009) The influence of monovalent cation size on the stability of RNA tertiary structures, *J. Mol. Biol. 390*, 791-804.

23. Shiman, R., and Draper, D. E. (2000) Stabilization of RNA tertiary structure by monovalent cations, *J. Mol. Biol. 302*, 79-91.

24. Yildirim, I., Stern, H. A., Tubbs, J. D., Kennedy, S. D., and Turner, D. H. (2011) Benchmarking AMBER force fields for RNA: comparisons to NMR spectra for single-stranded r(GACC) are improved by revised chi torsions, *J. Phys. Chem. B 115*, 9261-9270.

25. Tubbs, J. D., Condon, D. E., Kennedy, S. D., Hauser, M., Bevilacqua, P. C., and Turner, D. H. (2013) NMR of CCCC RNA reveals a right-handed helix and

revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics, *Biochem.*

26. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers, *Biophys. J. 92*, 3817-3829.

27. Banáš, P., Hollas, D., Zgarbová, M., Jurečka, P., Orozco, M., Cheatham III, T. E., Sponer, J., and Otyepka, M. (2010) Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins, *J. Chem. Theory Comp. 6*, 3836-3849.

28. Zgarbová, M., Otyepka, M., Sponer, J. i., Mládek, A. t., Banáš, P., Cheatham, T. E., and Jurečka, P. (2011) Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles, *J. Chem. Theory Comp. 7*, 2886-2902.

29. Yildirim, I., Stern, H. A., Kennedy, S. D., Tubbs, J. D., and Turner, D. H. (2010) Reparameterization of RNA x torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine, *J. Chem. Theory Comp. 6*, 1520-1531.

30. Yildirim, I., Kennedy, S. D., Stern, H. A., Hart, J. M., Kierzek, R., and Turner, D. H. (2012) Revision of AMBER torsional parameters for RNA improves free energy predictions for tetramer duplexes with GC and iGiC base pairs, *J. Chem. Theory Comp. 8*, 172-181.

31. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Amer. Chem. Soc. 117*, 5179-5197.

32. Cheatham, T. E., III, Cieplak, P., and Kollman, P. A. (1999) A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat, *J. Biomol. Struct. Dyn. 16*, 845-862.

33. Wang, J., Cieplak, P., and Kollman, P. A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?, *J. Comp. Chem. 21*, 1049-1074.

34. Hilal, S., El-Shabrawy, Y., Carreira, L., Karickhoff, S., Toubar, S., and Rizk, M. (1996) Estimation of the ionization pK_a of pharmaceutical substances using the computer program Sparc, *Talanta 43*, 607-619.

35. ChemAxon. (2012) http://www.chemaxon.com.

36. Bayly, C. I., Cieplak, P., Cornell, W. D., and Kollman, P. A. (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges- the RESP model, *J. Phys. Chem. 97*, 10269-10280.

37. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1995) Pairwise solute descreening of solute charges from a dielectric medium, *Chem. Phys. Lett. 246*, 122-129.

38. Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E., III. (2007) Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms, *J. Chem. Ther. Comp. 3*, 2312-2334.

39. Dupradeau, F. Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W., and Cieplak, P. (2010) The RED tools: advances in RESP and ESP charge derivation and force field library building, *Phys. Chem. Chem. Phys. 12*, 7821-7839.

40. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general AMBER force field, *J. Comp. Chem. 25*, 1157-1174.

41. Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations, *J. Mol. Graphics Mod. 25*, 247-260.

42. Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006) ANTECHAMBER, an accessory software package for molecular mechanical calculations, *J. Mol. Graphics Mod.*, [in press].

43. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., et al. (2009) Gaussian 09, Revision A1 ed., Gaussian, Inc. Wallingford, CT.

44. Case, D. A., Cheatham, T. E., III, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs, *J. Comp. Chem. 26*, 1668-1688.

45. Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000) Fast, efficient generation of high quality atomic charges. AM1-BCC model: I. Method, *J. Comp. Chem. 21*, 132-146.

46. Jakalian, A., Jack, D. B., and Bayly, C. I. (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation, *J. Comp. Chem. 23*, 1623-1641.

47. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparisons of simple potential functions for simulating

liquid water, *J. Chem. Phys. 79*, 926-935.

48. Joung, I. S., and Cheatham III, T. E. (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations, *J. Phys. Chem. B 112*, 9020-9041.

49. Allnér, O., Nilsson, L., and Villa, A. (2012) Magnesium ion-water coordination and exchange in biomolecular simulations, *J. Chem. Theory Comp. 8*, 1493-1502.

50. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Comp. Phys. 81*, 3684-3690.

51. Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, *J. Comp. Phys. 23*, 327-341.

52. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method, *J. Chem. Phys. 103*, 8577-8593.

53. Srinivasan, J., Cheatham, T. E., III, Cieplak, P., Kollman, P. A., and Case, D. A. (1998) Continuum solvent studies of the stability of DNA, RNA and phosphoramidate helices, *J. Amer. Chem. Soc. 120*, 9401-9409.

54. Hansson, T., Marelius, J., and Aqvist, J. (1998) Ligand binding affinity prediction by linear interaction energy methods, *J. Comp. Aided Mol. Des. 12*, 27-35.

55. Vorobjev, Y. N., and Hermans, J. (1999) ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model, *Biophys. Chem. 78*, 195-205.

56. Tan, C., Yang, L., and Luo, R. (2006) How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis, *J. Phys. Chem. B 110*, 18680-18687.

57. Tan, C., Tan, Y. H., and Luo, R. (2007) Implicit nonpolar solvent models, *J. Phys. Chem. B 111*, 12263-12274.

58. Schlitter, J. (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix, *Chem. Phys. Lett. 215*, 617-621.

59. Killian, B. J., Kravitz, J. Y., and Gilson, M. K. (2007) Extraction of configurational entropy from molecular simulations via an expansion approximation, *J. Chem. Phys. 127*, 024107.

60. Hess, B. (2002) Determining the shear viscosity of model liquids from molecular dynamics simulations, *J. Chem. Phys. 116*, 209.

61. Cheatham, T. E., III, and Kollman, P. A. (1997) Molecular dynamics simulations highlight the structural differences in DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes, *J. Amer. Chem. Soc. 119*, 4805-4825.

62. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *J. Comp. Chem. 25*, 1605-1612.

63. Gilson, M. K., Given, J. A., Bush, B. L., and McCammon, J. A. (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review, *Biophys. J. 72*, 1047-1069.

64. Hermans, J., and Wang, L. (1997) Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. Application to a complex of benzene and mutant T4 lysozyme, *J. Am. Chem. Soc. 119*, 2707-2714.

65. Lazaridis, T., Masunov, A., and Gandolfo, F. (2002) Contributions to the binding free energy of ligands to avidin and streptavidin, *Proteins: Struct., Funct., Bioinf. 47*, 194-208.

66. Luo, R., and Gilson, M. K. (2000) Synthetic adenine receptors: direct calculation of binding affinity and entropy, *J. Am. Chem. Soc. 122*, 2934-2937.

67. Page, M. I., and Jencks, W. P. (1971) Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect, *Proceedings of the National Academy of Sciences 68*, 1678-1683.

68. Yu, Y. B., Privalov, P. L., and Hodges, R. S. (2001) Contribution of translational and rotational motions to molecular association in aqueous solution, *Biophys. J. 81*, 1632-1642.

69. Geney, R., Layten, M., Gomperts, R., Hornak, V., and Simmerling, C. (2006) Investigation of salt bridge stability in a generalized Born solvent model, *J. Chem. Theory Comp. 2*, 115-127.

70. Gaillard, T., and Case, D. A. (2011) Evaluation of DNA force fields in implicit solvation, *J. Chem. Theory Comp. 7*, 3181-3198.

71. Wang, J., Dixon, R., and Kollman, P. A. (1999) Ranking ligand binding affinities with avidin: a molecular dynamics-based interaction energy study, *Proteins 34*, 69-81.

CHAPTER 4

RELIABLE OLIGONUCLEOTIDE CONFORMATIONAL ENSEMBLE GENERATION IN EXPLICIT SOLVENT FOR FORCE FIELD ASSESSMENT USING RESERVOIR REPLICA EXCHANGE MOLECULAR DYNAMICS SIMULATIONS

4.1 Chapter Notes

This chapter was adapted from the following published article:

> Niel M. Henriksen, Daniel R. Roe, and Thomas E. Cheatham, III. Reliable Oligonucleotide Conformational Ensemble Generation in Explicit Solvent for Force Field Assessment Using Reservoir Replica Exchange Molecular Dynamics Simulations. *The Journal of Physical Chemistry B*. **2013**. 117 (15), 4014-4027

N.M. Henriksen and T.E. Cheatham, III designed the research. N.M. Henriksen performed the simulations and analysis. D.R. Roe wrote many of the analysis programs. N.M. Henriksen wrote the manuscript. N.M. Henriksen, D.R. Roe, and T.E. Cheatham, III revised the manuscript.

4.2 Introduction

RNA takes on many essential roles in biology, from information encoding to catalysis to regulatory functions *(1)*. Molecular dynamics (MD) simulations provide a critical connection between structural theory and experimental data for RNA as well as other biomolecules *(2-5)*. Unfortunately, the force fields

which underlie the physical representation of RNA tend to be less reliable than those used for proteins and have required numerous refinements over the past several years *(6-11)*. This is likely due to a variety of factors which could include the highly charged and highly flexible RNA backbone. As we and others continue efforts at force field development, it has become apparent that a time-efficient and cost-effective method is necessary for producing biomolecular simulation data, from which quantitative results defining the ensemble of sampled conformations can be obtained and the underlying force field evaluated. Conventional MD simulations, typically performed at laboratory temperatures in order to match experiment, often require cost-prohibitive timescales and/or special purpose hardware to produce converged results *(12)*. Additionally, cost-saving approaches such as implicit solvation, which reduces the total degrees of freedom in the system, tend to result in major conformational distortions for many RNA systems *(13, 14)* (and as is demonstrated in the results of this work). To address the need for efficient simulation methods for generating well-converged conformational ensembles required for force field development and assessment, we turned to approaches from the widely studied field of enhanced sampling *(15-17)*.

Replica exchange MD (REMD) is a commonly used method for enhanced sampling which deploys an ensemble of independent system replicas, or MD simulations, that exchange various properties *(18, 19)*. By allowing a system replica at target conditions (such as the laboratory temperature) to exchange with system replicas at conditions which favor sampling (such as higher

temperatures), REMD enhances conformational sampling on a rugged landscape while maintaining a Boltzmann-weighted ensemble at each target condition *(17)*. Due to the computational cost of simulating multiple system replicas, temperature REMD is often used with implicit solvent which reduces the number of degrees of freedom in the system, requiring fewer replicas to span a given temperature regime while maintaining acceptable exchange ratios. Unfortunately, as noted above, RNA simulations in implicit solvent tend to produce poor results. There are a variety of small systems that have been studied using REMD with explicit solvent, including proteins *(20-28)*, DNA*(29, 30)*, and RNA *(31-34)*. In most cases, convergence of the REMD ensemble is discussed in qualitative terms, if at all, and typically only one REMD simulation is performed for each condition of interest. In this work we aim to provide a more detailed understanding of convergence in explicit solvent. Due to our interest in obtaining quantitative results, we chose to study two small systems in order to reduce computational cost: alanine dipeptide and a tetranucleotide RNA, rGACC (Figure 4.1). Alanine dipeptide is a frequently used test molecule due to its simplicity *(35-38)* and rGACC is optimal because it is small, has detailed NMR data published regarding its structure, and has been studied previously using conventional MD *(39)*.

Finally, in addition to traditional REMD, we also investigate two methods which were previously reported to offer the benefits of REMD at reduced resource cost: TIGER2 *(38)* and reservoir REMD (R-REMD) *(40)*. The TIGER2 method reduces the number of replicas required for the ensemble by

incorporating a velocity rescaling and thermal equilibration step both prior to and immediately following the exchange attempt. This added step shifts the potential energy distribution of high temperature replicas nearer to the distribution of the baseline temperature replica and allows exchanges to occur which would otherwise be improbable due to the large potential energy distribution spacing. The other method, R-REMD, enhances convergence by adding a high-temperature structure reservoir to the top of the REMD temperature ensemble. Thus, a replica at the highest target temperature can exchange with the pregenerated reservoir and the costly wait associated with traversing energy barriers is overcome by exchange. In effect, the reservoir drives convergence of the entire REMD ensemble. This approach has been reported previously with implicit solvent *(40, 41)*, and in explicit solvent with multiple reservoirs for the disordered Abeta(21-30) peptide *(42)*. Here, we apply the single reservoir approach and explore convergence for RNA. Quite surprisingly, even with a relatively small RNA system like rGACC, the results suggest that standard REMD methods may require over 2 μs per replica for convergence, and even with reasonable starting reservoirs, R-REMD still requires 20 – 100 nanoseconds per replica, depending on the required accuracy measures of the conformational ensemble populations for convergence.

4.3 Methods

4.3.1 System building

Two explicitly solvated systems were studied in this research: alanine dipeptide and the RNA tetranucleotide rGACC (Figure 4.1). Alanine dipeptide was built by capping an alanine residue with ACE (i.e., CH_3CO-) and NME (i.e., $-NHCH_3$), using AMBER's LEaP program and parameterized with the AMBER ff12SB force field *(43, 44)*. To solvate the system, 544 TIP3P water molecules *(45)* were added to a cubic periodic box around the solute. The system size was chosen to match and be consistent with the systems used in previous studies with the TIGER2 protocol *(38)*.

The initial conformation for the RNA tetranucleotide, rGACC, was taken from an A-form portion of a RNA crystal structure (PDB: 3G6E, residues 2623-2626). The RNA, parameterized with the ff12SB force field (note, for nucleic acids this is identical to ff10 which was released with AMBER11), was solvated with 2500 TIP3P water molecules in a truncated octahedron periodic box and the total system charge was neutralized with three sodium ions. Additional salt was not added to be consistent with previous computational studies by Turner and co-workers *(39)*. This RNA was previously studied with AMBER ff99 force field as well as a modified variant by Turner and co-workers *(9)*, which includes corrections to the x torsion parameters. We chose to use the ff12SB force field for the following reasons: it includes backbone torsion corrections *(6)*, higher accuracy x torsion corrections (in comparison to the Turner variant) *(8)*, and has been tested on larger, more representative RNA structures *(7, 46, 47)*.

4.3.2 System heating and equilibration

All simulations were performed with AMBER12 *(43)* using either the PMEMD or SANDER programs. PMEMD was used when possible due to its superior computational performance and parallel efficiency. However, PMEMD does not (yet) support R-REMD simulations and thus SANDER was used for these simulations. Additionally, TIGER2 simulations are not implemented in AMBER and thus an in-house script was used as a wrapper to implement the simulation cycle (details discussed later). Prior to production simulations, both the alanine dipeptide and RNA systems were energy minimized and equilibrated. Energy minimization was performed with 25 kcal/mol-$Å^2$ atomic positional restraints on the solute and consisted of 1000 steps with the steepest descent algorithm followed by 1000 steps with the conjugate gradient algorithm. Following minimization, the systems were heated from 10 K to 150 K at constant volume with 25 kcal/mol-$Å^2$ atomic positional restraints on the solute using 1 fs timestep. This heating step was accomplished with 100 ps of MD simulation for the alanine dipeptide system and 1 ns for the RNA system. During heating, the temperature was controlled using a Langevin thermostat with a collision frequency of 2.0 ps^{-1}. Further heating to the target temperature (300 K in most cases) was performed at constant pressure using a weak-coupling algorithm *(48)*, a pressure relaxation time of 1 ps, 5 kcal/mol-$Å^2$ atomic positional restraints on the solute, and the same thermostat, timestep, and duration settings as the previous heating step. After reaching the target temperature, the system was further equilibrated (1 ns for alanine dipeptide, 5 ns for the

RNA) with 0.5 kcal/mol-Å^2 positional restraints on the solute using weaker-coupling constant pressure relaxation time of 5.0 ps. The integration time step during this equilibration step was 2 fs and temperature regulation the same as previous steps. For all explicit solvent MD in this research, the nonbonded direct space cutoff was set to 8.0 Å and the default AMBER12 particle mesh Ewald *(49)* settings were used to control reciprocal space calculations. SHAKE constraints *(50)* with a tolerance of 0.00001 Å were used to eliminate short timescale bond vibrations between hydrogen atoms and heavy atoms.

4.3.3 Production MD

Table 4.1 provides a list of the production simulations discussed in this research. In the text below, we provide details on the settings for each type of simulation: Conventional MD, TIGER2, REMD, and R-REMD.

Conventional MD: Both constant pressure (NPT) and constant volume simulations (NVT) were performed in the conventional manner. The constant pressure simulations (ALA-NPT and RNA-NPT in Table 4.1) were intended to match typical simulation protocols. The time step was set to 2 fs and a weak coupling algorithm governed both the temperature and pressure regulation with a relaxation time of 10 ps. Constant volume simulations (ALA-NVT, ALA-398, RNA-398) used 2 fs time step and the Langevin thermostat with the collision frequency set to 10 ps^{-1} for alanine dipeptide, to be consistent with previous work *(38)*, and 2 ps^{-1} for the RNA.

TIGER2: A method for enhanced sampling, named TIGER2 *(38)*, has been previously described and offers the benefits of temperature REMD with reduced computational cost. The TIGER2 algorithm is not implemented in AMBER, but can be achieved fairly easily with a simple wrapper script. Our implementation, which closely followed the published description, uses repeated cycles of PMEMD calculations for the MD steps and a Perl script to perform the velocity rescaling and Metropolis selection steps. Four replicas were used at the following temperatures: 300, 377, 476, and 600 K. A typical cycle consists of four steps starting from initial systems all at 300K: 1) velocity rescaling to one of the four assigned temperatures followed by thermal equilibration, 2) dynamics sampling at the assigned temperature, 3) velocity rescaling of all replicas back to 300 K followed by thermal equilibration, and 4) exchange attempt and temperature reassignment. The last step consists of the following substeps: a candidate system from the three highest temperatures is randomly selected (using Perl's rand function) for an exchange attempt with the baseline temperature replica (300 K), the exchange probability is determined by the Metropolis criterion, the result assigns one of the two considered replicas to the baseline temperature, and the remaining systems are assigned to the higher temperatures according to their system potential energies (higher energies are given a higher temperature). For all TIGER2 simulations, the heating and production sampling time periods were 1 ps each. The cooling period was varied between three time periods, 1ps, 2ps, and 5ps, as this time period had a large affect on exchange acceptance (ALA-TIG-1, ALA-TIG-2, and

ALA-TIG-3 in Table 4.1, respectively). Similar to what was used in the reference publication, we used a Langevin thermostat collision frequency of 25 ps^{-1} during the heating and quenching portions of the cycle and 10 ps^{-1} during the sampling portion.

REMD: Traditional REMD calculations were performed using AMBER's PMEMD program. The temperature intervals between replicas were estimated using an online generator at http://folding.bmc.uu.se/remd *(51)* and are listed in Table 4.2. These distributions led to exchange rates between 0.15-0.30 for the ten replica simulations and 0.15-0.40 for the twenty-four replica simulations. All REMD simulations were performed at constant volume as AMBER does not support constant pressure for REMD. It has been noted that the hydrophobic effect will be increased at high temperatures for constant volume simulations *(52, 53)*. The extent of this phenomenon at high temperatures for the systems studied here and its influence on lower temperature conformational results remains unclear. Prior to production simulations, a 200 ps heating period was employed to equilibrate each replica to the assigned initial temperature. Temperature was controlled with the Langevin thermostat set to a collision frequency of 10 ps^{-1} for alanine dipeptide and 2 ps^{-1} for the RNA. A time step of either 1 or 2 fs was used and is noted in Table 4.1. The 2 fs timestep is typically the preferred value for computational performance reasons, but 1 fs was used for simulations with high temperatures (~600 K) out of caution. The exchange attempt interval was set to either 0.2 or 1.0 ps and is noted in Table 4.1. An exchange attempt interval of 1 ps is more commonly

used. However, the use of a shorter interval may improve convergence *(54, 55)*. The value we chose was arbitrary in some cases but dependent on the hardware being used in others. For instance, the GPU simulations are fast enough that we were concerned about a performance drop if the exchange attempt interval was too small. To be consistent with the GPU simulation (RNA-REMD-1), we used a 1 ps exchange attempt interval for the R-REMD simulations of the RNA system even though they were performed on CPUs. In the case of alanine dipeptide, a 0.2 ps exchange attempt interval was used for all REMD simulations except ALA-24REMD, which was intended to mimic a more traditional/conservative approach (1 ps exchange attempt interval and 1 fs timestep). In addition to explicit solvent REMD simulations, we also performed one implicit solvent REMD simulation (RNA-REMD-GB). Implicit solvation was implemented using the Hawkins, Cramer, and Truhlar generalized Born (GB) model *(56, 57)* with a surface area contribution to the solvation term computed by the LCPO model *(58)*. A salt concentration of 200 mM was approximated using Debye-Hückel screening. The Langevin thermostat was used to control temperature with a collision frequency of 2 ps^{-1}. An infinite cutoff was employed for the nonbonded cutoff and SHAKE constraints were used to eliminate high frequency bond vibrations between hydrogen atoms and heavy atoms. All analysis of the traditional REMD simulations with RNA discarded the first 50 ns as equilibration (the initial structure of each replica was identical so we use the term equilibrate here to mean that a variety of structures are being

sampled at each temperature level as determined by inspecting the RMSD plot (Figure 4.2)).

R-REMD: In contrast to traditional REMD, the highest temperature replica in R-REMD simulations exchanges with a pregenerated reservoir of structures. This method was previously introduced by Okur *et al. (40, 41)* and tested in implicit solvent. In the current work, reservoirs were generated at 398 K using conventional MD simulations and consisted of high precision coordinate/velocity frames saved every 10 ps. The potential energy of each frame in the reservoir was computed and used in the exchange calculation during the R-REMD simulation with a 1/N non-Boltzmann weighting assumed for each frame. Exchange between highest temperature replica and the reservoir was ~0.47 for the alanine dipeptide simulations and ~0.33 for RNA simulations. All other simulation settings were identical to those used in the traditional REMD simulations. Note, the "-S" and "-L" suffixes stand for small and large reservoirs, respectively, in Table 4.1.

4.3.4 Hardware

All simulations were performed using traditional CPU clusters at a variety of resource locations with the exception of the RNA-NPT, RNA-398, and RNA-REMD-1 simulations, which were performed on GPUs. The CPU clusters include NICS Kraken, SDSC Gordon, and the University of Utah CHPC clusters. The GPU simulations were performed using the CUDA enabled PMEMD code *(59)*

on GPU accelerated nodes (NVIDIA Tesla M2090 GPUs) at either NICS Keeneland or the University of Utah CHPC.

4.3.5 Conformational analysis

The REMD algorithm implemented in AMBER specifies that when an exchange attempt is successful, replicas will exchange thermostat temperatures (the alternative being the exchange of system coordinates). This results in each replica containing simulation data from a variety of temperatures. Often, the researcher is most interested in data for a specific temperature and thus the REMD trajectory ensemble must be sorted such that contiguous data are obtained for each temperature level. In this chapter, we refer to such a process as "sort by temperature." Alternatively, it is occasionally of interest to study the data directly for each replica without sorting by temperature (as we do for our ensemble RMSD profile analysis). We will refer to this as "sort by replica." In order to sort by temperature we used a development version of AMBER's Cpptraj. Conformational analysis, including RMSD profiles, clustering, principal component analysis, and torsion and distance calculations, were performed using a combination of AMBER's Ptraj and Cpptraj programs and in-house Perl scripts. All RMSD analysis was made using a common reference structure, which ensured that results could be compared. The common reference structure for both alanine dipeptide and rGACC was generated by GB energy minimization of initial build structure for each molecule. RMSD analysis was performed with mass-weighting using only

heavy atoms for alanine dipeptide and all atoms for the RNA. Clustering of the RNA simulations was performed with Ptraj using the "averagelinkage" agglomerative algorithm, a critical distance epsilon value of 2.3 Å, and a variable sieve value which ensured that the initial clustering pass contained ~5000 frames. The sieve, which uses an initial subset of randomly chosen frames to define the clustering divisions, was required because memory limitations do not allow complete clustering of a large number of frames. Thus five independent cluster analyses were performed for each simulation of interest and an average and standard deviation was reported. Animation of the PCA eigenvectors was performed using PCAsuite *(60)*. Molecular graphics images were generated using UCSF Chimera *(61)*.

4.4 Results and Discussion

4.4.1 Alanine dipeptide conformations and convergence

Due to its small size and simple structure, a solvated alanine dipeptide system is convenient for quick testing and demonstrating proof of principle. Thus we began by studying the feasibility of various explicitly solvated REMD simulations with this system. The primary purpose of these initial investigations was to demonstrate that a given simulation can reliably sample the conformational space of the solute of interest. One convenient method for observing the sampled space is to generate a histogram profile of the atomic RMSD values with respect to a common reference structure. We found that conventional MD (both constant volume and constant pressure), traditional

REMD, and R-REMD produce nearly indistinguishable RMSD profiles (Figure 4.3). We note here that RMSD profiles represent a necessary but not sufficient test of convergence. Thus they are useful as an initial test to compare whether ensembles from two independent simulations overlap. More detailed conformational analysis is required (and provided later) to confirm this in situations where multiple conformations occupy the same RMSD space.

In order to gauge the convergence of these simulations, a more in-depth analysis is required. By plotting the cumulative RMSD profile over simulation time (Figure 4.4), it is possible to observe the time convergence at any point along the profile, such as the frequency maxima (Figure 4.5). From this convergence plot we find that REMD calculations require significantly less time to reach convergence than conventional MD and R-REMD calculations are nearly converged from the beginning of the simulation (the latter observation was also made by Okur *et al. (40)*). A comparison of the ten replica REMD simulation (ALA-10REMD) with the twenty-four replica REMD simulation (ALA-24REMD) suggests that the larger ensemble takes longer to converge to the same precision than the smaller ensemble, although near quantitative results are obtained fairly quickly (Figure 4.6).

Unfortunately, cumulative-type plots of convergence don't always indicate that the ensemble has converged to the true value dictated by the force field if for some reason an individual replica becomes improperly trapped in a given conformation. A potentially more revealing test of convergence is found by plotting the RMSD profile of each replica trajectory prior to sorting

the data by temperature. In an ideal scenario, a converged REMD ensemble would consist of replicas which each traverse temperature space many times and therefore, over the course of a long simulation, would sample identical conformational space. Thus, the RMSD profile of each replica would be identical to one another. Nearly complete convergence of the replica RMSD profiles can be qualitatively demonstrated for the REMD simulations of alanine dipeptide (Figure 4.7) and we suggest such plots should be regularly included in publications. As we show later in the results for the RNA system, replicas which become conformationally trapped are easily spotted using this method.

To quantify the conformational distribution of the alanine dipeptide simulations, the phi/psi torsion space of alanine dipeptide was divided into six regions (Figure 4.8, left) and the percent occupancy was calculated (Table 4.3). The regions were identified by locating the minima (which correspond to regions of highest density) in a population based free energy plot of the phi/psi space (Figure 4.8, right). A comparison of the conformational populations between the conventional MD simulations (ALA-NVT, ALA-NPT) and the traditional REMD simulations (ALA-10REMD, ALA-24REMD) suggests that the two approaches yield very similar conformational results (Table 4.3). Also, the use of constant pressure or constant volume does not seem to significantly affect the results for conventional MD. Additionally, it appears that REMD with ten replicas and a maximum temperature of 398 K (ALA-10REMD) yields similar results to the twenty-four replica REMD with a maximum temperature of 600 K (ALA-24REMD). It is interesting to note that the REMD methods appear to

sample the rare F conformation more frequently than conventional MD. This may suggest that that the conventional MD simulations require more simulation time than was collected to adequately sample the F conformation and that REMD efficiently traverses the energy barriers to this conformation.

When performing R-REMD simulations, in which a predefined reservoir quickly drives convergence of the REMD ensemble, the size and conformational diversity of the reservoir is an important concern. To understand how the composition of the reservoir affects the results, we examined a variety of reservoir sizes and compositions and report two examples here. The first, used in the ALA-R-REMD-L simulation, was a larger reservoir and consisted of 4764 frames collected every 10 ps from a 47 ns simulation at 398 K. Comparison of the RMSD profile of this reservoir with the 397.7 K temperature replica from the ALA-10REMD simulation suggests that the reservoir is a reasonable approximation of a converged ensemble at 398 K, although its profile is not smooth due to the relatively small number of frames used (Figure 4.9). Using this reservoir, the ALA-R-REMD-L simulation generates a conformational ensemble at 300 K very similar to those observed using traditional REMD (Table 4.3).

We also studied a much smaller reservoir (ALA-R-REMD-S), containing 538 artificially selected frames from the larger reservoir, in which structures from each of the three RMSD profile maxima are present but not at the correct relative frequencies (Figure 4.9). The resulting conformational distribution at 300 K from the ALA-R-REMD-S simulation is perturbed slightly relative to the

other REMD and R-REMD simulations (Table 4.3), yet the small difference suggests that the R-REMD method is fairly robust at recovering the correct conformational frequencies even when the reservoir is sparse and non-Boltzmann weighted.

The R-REMD method offers significant savings in computational resources over traditional REMD. An even greater savings can potentially be found in the previously published TIGER2 method *(38)*. Briefly, the TIGER2 method involves a small number of replicas for which a subset are rapidly heated from the baseline temperature to a higher sampling temperature and after a brief period rapidly quenched and equilibrated back to the baseline temperature. Following this heating and quenching cycle, the replicas are exchanged with the baseline replica with a probability based on the Metropolis criterion. The heating step allows the system to rapidly overcome energy barriers at high temperatures and the quenching step allows for replica exchange to occur at a reasonable rate at the baseline temperature. The TIGER2 algorithm is not implemented in AMBER, yet its simplicity makes using a wrapper script feasible, which is the approach we took. After testing our implementation of TIGER2, it was immediately clear that the equilibration period following the quenching step had a large effect on the exchange probability. We tested three quench-equilibration periods, 1, 2, and 5 ps, which resulted in exchange success rates of 0.066, 0.277, and 0.475, respectively (corresponding to ALA-TIG-1, ALA-TIG-2, and ALA-TIG-3). The conformational distribution of the TIGER2 simulations is similar in the overall trend to the other methods tested, yet there is a

noticeable difference in the frequency of the B and F conformations (Table 4.3). It is not clear what causes this difference and further investigation will be necessary to understand the discrepancy. For this reason the TIGER2 method was not used to investigate the larger and more complex RNA system.

4.4.2 rGACC RNA conformational analysis

As part of an overall effort to improve RNA force fields, we wanted to focus computational efforts on a minimal RNA system for which conformational convergence could be quickly obtained and experimental data were available. The rGACC RNA has previously been studied in solution using NMR and in simulation using other force fields *(39)*. Although it is very small, it still populates (at least two) A-form-like conformations and is therefore an ideal test case. The difficulty with using conventional MD simulations is that conformational convergence at the temperature of interest, even with the latest accelerated hardware, is difficult to achieve. We performed a 5 μs simulation at 300 K (RNA-NPT) and the atomic RMSD versus time plot indicates that only a few transitions occur between the major conformations on that timescale (Figure 4.10, top).

Identification of four major conformations was performed by clustering and a representative structure of each is shown in Figure 4.11 in both molecular graphics and simplifying cartoon. Additional conformations were identified other than the four described here, but due to their population frequencies being 3% or less, they were not studied further in great detail.

Quantitative frequencies of the four conformations are given in Table 4.4. The most populated conformation in the RNA-NPT simulation, a non-A-form conformation which we term the "Intercalated structure," does not fit the NMR data and was also observed to dominate a microsecond timescale simulation using the modified force field by Yildirim *et al*. *(39)*. The next most populated conformations, termed the "NMR Minor" and "NMR Major," are consistent with the NMR data as their names suggest and are essentially A-form structures. However, Yildirim *et al*. concluded that the NMR Major structure should dominate in solution with a small fraction of the NMR Minor structure present as well *(39)*. In contrast to the experimental results, we observed a greater frequency of the NMR Minor structure compared to the NMR Major structure. The fourth structure, which was not described in the previous computational study of rGACC but was reported in a recent study of rCCCC *(62)*, is termed the "Inverted" structure. It is a non-A-form conformation and is also not consistent with the observed NMR data.

At 5 μs of conventional MD, the RNA-NPT simulation is still not converged enough for reliable force field comparison purposes. Even at 100 ns/day simulation speeds, attainable through use of GPU accelerated simulations, this simulation required 50 days to complete. Thus a more convenient, reliable, and cost effective method for studying the conformational landscape of complex structures is of interest. REMD simulations provide an attractive alternative. By coupling high temperature simulations which traverse energy barriers quickly with low temperatures simulations at experimentally relevant conditions, a

converged conformational ensemble is expected to be obtained quicker than by conventional MD alone *(63)*. REMD is typically performed with implicit solvent because the number of replicas increases with the size of the system. Unfortunately, as of present, RNA simulations generally do not behave well in implicit solvent. To demonstrate this, we performed a 500 ns per replica REMD simulation in GB solvent using six replicas spanning a temperature range from 277-463 K. At lower temperatures, the RNA nearly exclusively adopts the Inverted conformation whereas at the maximum temperature the RNA is largely unstructured (Figure 4.12). These results are consistent with our previous experience with RNA simulations in implicit solvent (unpublished) and suggest that explicit solvent is necessary for (even crudely accurate) simulations in the foreseeable future.

To evaluate the performance of explicitly solvated REMD simulations, we performed three simulations: one 2 μs per replica timescale simulation using GPU acceleration (RNA-REMD-1) and two shorter, 400-500 ns per replica simulations using traditional CPU hardware (RNA-REMD-2 and RNA-REMD-3). Inspection of the RMSD versus time plot for RNA-REMD-1 suggests that frequent conversion between the four major conformations occurs over the course of the simulation (Figure 4.10, bottom). Comparison of the RMSD profile at 277 K for the three traditional REMD simulations reveal the difficulty in obtaining converged results (Figure 4.13). For instance a large peak at an RMSD value of 4.0 Å is observed in the RNA-REMD-2 simulation but is much smaller in the other two simulations. The corresponding structure to this peak (Figure 4.14) is

not one of the previously mentioned conformations. Although it is present in low frequencies in the other REMD simulations it seems to dominate the RNA-REMD-2 simulation. A small number of replicas are the main contributors to the overpopulation of this conformation in the ensemble and those replicas remain trapped for significant time periods despite regularly traversing the complete temperature space (Figure 4.15). This problem may be exacerbated by a short exchange attempt interval which will tend to reduce the average duration a replica stays at a given temperature. If a conformational transition requires some minimum time to proceed and also requires a high temperature to make traversal of an energy barrier more likely, a short exchange attempt interval may cause the replica to stay conformationally trapped. Extending the exchange attempt interval and/or increasing the number of target temperatures intervals above 398 K may alleviate this problem.

As we mentioned in the results for alanine dipeptide, a convenient method to examine REMD ensemble convergence is to study the replica RMSD profiles for the ensemble data sorted by replica rather than sorted by temperature. In the ideal case, a converged REMD ensemble will generate an identical curve for each replica. We demonstrate the tendency towards a converged RMSD profile in Figure 4.16. At 200 ns, the replica RMSD profiles are dispersed and a consensus profile is difficult to distinguish. By the end of the 2 μs simulation, the RMSD profiles are much more consistent although a few outliers remain, suggesting that longer simulation is required for complete convergence. A comparison can also be made between the three traditional

REMD simulations (Figure 4.17). These results also suggest that the 2 μs simulation (RNA-REMD-1) is closer to convergence than the shorter simulations (RNA-REMD-2 and RNA-REMD-3), although a single consensus profile is still not achieved.

The conformational landscape of the rGACC structure is much more complicated than alanine dipeptide and cannot easily be represented by the equivalent of a phi/psi plot. Instead we used principal component analysis to identify the primary motional modes of the RNA in the RNA-REMD-1 simulation (Figure 4.18A,B). The division of clustering along these PCA axes can be visualized (Figure 4.18C) and a population based free energy plot reveals the relative minima (Figure 4.19). These plots show that the two conformations consistent with NMR data, NMR Major and NMR Minor, are nearly overlapping in the primary motional axes whereas the two non-A-form conformations are distant from the NMR structures, suggesting a force field problem which leads to incorrect population sampling.

The results of traditional REMD suggest that even on fairly long timescales by current standards and at significant resource cost, conformational convergence is not achieved. Thus, this is not a very feasible method for force field development which requires multiple such simulations run in a linear fashion. Given the success we found with R-REMD simulations of alanine dipeptide, we decided to investigate whether the method was also feasible for larger more complicated systems like rGACC. To generate the reservoir, a 1.4 μs simulation of the RNA at 398 K was performed saving frames

every 10 ps (RNA-398). A plot of the RMSD versus simulation time reveals that all four conformations are sampled many times during the simulation and a smooth RMSD profile is generated (Figure 4.20). This is an important observation because it shows that conventional MD at 398 K does not exhibit the conformational trapping behavior which was observed for the traditional REMD approach. As we demonstrate below, the R-REMD method also does not suffer from conformational trapping due to the frequent exchange with this high temperature reservoir.

Three independent R-REMD simulations were run using this reservoir: RNA-R-REMD-1 (205 ns/replica), RNA-R-REMD-2 (160 ns/replica), and RNA-R-REMD-3 (250 ns/replica). Analysis of the RMSD profile for these simulations shows that the conformational detail at 277 K emerges from the smooth, distinctly different reservoir profile at 398 K (Figure 4.21). Inspection of the replica RMSD profiles (sorted by replica, not temperature) for the three R-REMD simulations suggests improved convergence (Figure 4.22) as do the similar quantitative results obtained from cluster analysis (Table 4.4). Comparison of the quantitative clustering results for the R-REMD method at three temperatures, 277, 299, and 328 K, reveals an interesting trend. Increasing the temperature significantly reduces the frequency of the two NMR consistent structures but not the non-A-form structures. Only a small decrease is seen for the Intercalated structure and an increase is observed for the Inverted structure. This suggests that the entropic component of the free energy favors

the non-A-form structures more than the NMR structures and thus this preference increases with temperature.

In order to test how the composition of the reservoir affects the R-REMD results for the RNA system, we performed an additional simulation with a much smaller reservoir. This simulation, RNA-R-REMD-S, contained only 10,000 frames from the first 100 ns of the RNA-398 simulation. In this time period, three of the four major conformations were sampled including the Intercalated, NMR Minor, and NMR Major (Figure 4.20). The Inverted conformation was not sampled until ~150 ns and thus was not present in this reservoir. As expected, quantitative conformational analysis of the RNA-R-REMD-S simulation shows that the Inverted conformation is not present in any significant amount (Table 4.4). This underscores the importance of having representatives from each significantly populated structural class present in the R-REMD reservoir.

4.4.3 Comparison of rGACC RNA simulation data with NMR data

Given that the R-REMD simulations appear to be a reliable method for obtaining an accurate conformational data about rGACC in simulation, we decided to make a deeper comparison between the published NMR data and the simulation data (specifically for RNA-R-REMD-3). Two easy comparisons include the sugar pucker and base orientation preferences. Values at three different temperatures for these metrics are given in Table 4.5.

The sugar pucker values observed in simulation are consistent with the NMR measurements for the first three residues, while residue C4 significantly underpopulates the C3'-endo conformation compared to experiment at 277 K. At 328 K, both residues G1 and C4 are out of the experimental range, with G1 overpopulating C3'-endo and C4 still underpopulating C3'-endo. The deviation of the simulation sugar pucker from experiment at C4 is likely due to the overpopulation of the NMR Minor structure, which prefers C2'-endo at C4. Quantitative values for the other metric of interest, base orientation, were not obtained in the experimental publication; however the NOESY spectrum indicated that all four residues preferred the *anti* conformation at 275 K. Simulation values are consistent with this observation, with the possible exception of residue G1. It seems likely that a stronger NOE would have been observed if the *anti*/*syn* ratio was 64/36, as was observed in simulation. Taken together, these two metrics offer only a weak indication that the force field is incorrectly modeling the solution structure, primarily by highlighting the under-population of the C3'-endo sugar pucker state by residue C4.

In order to make a more detailed comparison with the NMR data, we investigated whether any of experimentally observed NOEs were violated. To study this, the r^6-averaged distances for all possible NOE pairs were calculated for the RNA-R-REMD-3 simulation at 277 K. Of the 21 experimentally supported atom distance restraints, only one pair was in violation: atoms G1 H8 and A2 H8. The experimental restraint lower and upper bounds are 2.0 – 5.0 Å, while the simulation r^6-average is 5.3 Å. (Note: the lower/upper bounds listed for this

restraint in the publication SI were written as 2.0 – 10.0 Å, but this was done to "search broader conformational space" during annealing. In fact, the NOESY data indicate that the upper bound should be less than 5.0 Å *(39)*.) A violation rate of 1/21 would ordinarily be fairly high; however, in this case it leaves us with just one data point. Thus it is important to inspect atom pairs for which the simulation results would predict an NOE signal but where none is observed experimentally. This approach was noted in the previous computational study when discussing the non-A-form Intercalated structure *(39)*. We identified fifteen atom pairs for which the r^6-average value was 5.0 Å or less, but for which no NOE signal was identified by NMR (Table 4.6). A visual depiction of these predicted atom pairs, overlaid on both the NMR Major and Intercalated conformations, as well as the single restraint violation, are shown in Figure 4.23.

These data strongly suggest that the force field overpopulates the non-A-form structures. We acknowledge that even if the non-A-form structures were truly present in solution, it is unlikely that NOE signals would be observable for all fifteen identified atom pairs due to the limitations of experimental studies. However, we do show that there exist many atom pairs for which an NOE signal should arise if the non-A-form conformation is really present. Taken together, our comparison to the published NMR data suggests that the ff12SB force field requires further improvements in order to adequately model this RNA structure.

4.5 Conclusion

In this work we have demonstrated a tractable method for reliable generation of conformational ensembles from explicitly solvated simulations. This represents a necessary step towards cost-effective force field development of RNA. Conventional MD, even with advances in accelerated hardware, is still not a feasible method for research that requires both quick turnaround and extensive sampling of a rugged conformational landscape. REMD, which has been one of the traditional solutions to this problem, is not nearly as cost effective when explicit solvent is required as is the case for most RNA simulations. In addition, we have shown that even REMD simulations of significant length (2 μs) may be slow to converge for certain systems. In other words, although traditional REMD has improved convergence with respect to conventional MD, that does not mean it is guaranteed to converge. Thus we have turned to the R-REMD method, in which a pregenerated, high temperature reservoir drives the convergence of a REMD ensemble. At high temperatures, the rugged conformational landscape is much more easily traversed than at lower temperatures (compare Figure 4.10, top and 4.20) and thus the reservoir is quickly and relatively cheaply generated. Use of the reservoir reduces the simulation time required for the resource intensive REMD step and thus leads to converged results which are cheaper than traditional REMD. It should also be mentioned that the aggregate simulation time for all replicas is far less than traditional REMD and comparable to that of the brute force conventional MD, and thus is a more efficient use of computational resources.

Care must be used when employing R-REMD method. For example, the reservoir must sufficiently sample the energy landscape in order to include all of the major conformations. We observed that a deliberately small reservoir completely eliminated one of the major rGACC conformations. Despite this, a perfectly converged reservoir does not seem to be necessary. We showed that a reservoir with skewed populations (albeit one with that included all major conformational classes) only slightly modified the conformational distribution at the baseline temperature. Finally, we suggest it is necessary to plot RMSD profiles for the data sorted by replica, rather than temperature, in order to determine how much simulation time is required for convergence. Significant discrepancies between these profiles can indicate that the ensemble has not yet converged and might contain trapped replicas.

After we determined that the R-REMD method was a reliable method for generating consistent data regarding the conformational landscape of rGACC, we were able to make specific comparisons with the published NMR data. These comparisons suggest that the ff12SB force field overpopulates non-A-form conformations for the rGACC tetramer. The source of this error is not immediately clear, although we suspect that refinements to the sugar pucker torsions or backbone torsions may improve the results. In addition to studying force field refinements, we can also easily study the effect of various water models, salt composition, and salt concentration on the results. The R-REMD method also holds promise for studying flexible portions of much larger molecules. For instance, a noncanonical region of a larger RNA could be studied

by generating a high temperature reservoir in which base pair restraints are used to prevent complete unfolding. The resulting R-REMD simulation would allow a focus on the conformational landscape of regions where the current force field is flawed (i.e., noncanonical regions), while limiting unfolding in helical regions where the force field behaves better. This approach may also be useful in studying ligand binding modes in which receptors undergo significant rearrangement.

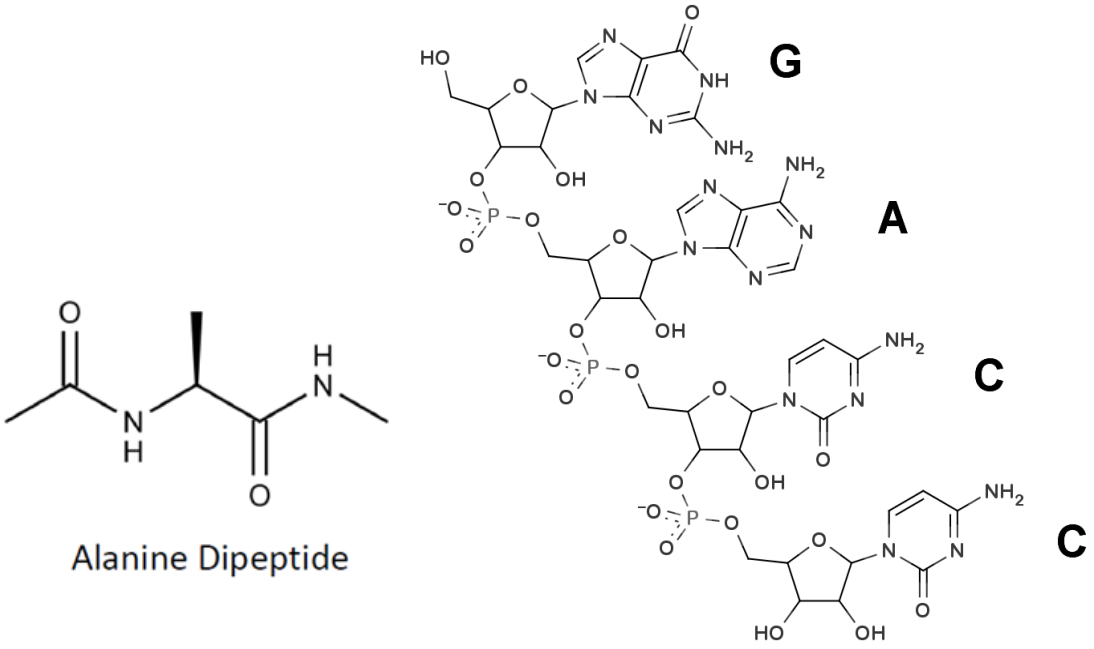

Figure 4.1. Molecules investigated in this research: alanine dipeptide (*left*), RNA tetranucleotide rGACC (*right*).

Table 4.1. Simulations performed in this work.

Simulation ID	Simulation Details	Temp.	Rep.	Length
ALA-NVT	NVT, dt=2fs, cwi=1ps	300	1	100
ALA-NPT	NPT, dt=2fs, cwi=1ps	300	1	314
ALA-398	NVT, dt=2fs, cwi=10ps	398	1	47
ALA-10REMD	REMD, NVT, dt=2fs, eai=0.2ps, cwi=1ps	300 - 398	10	138
ALA-24REMD	REMD, NVT, dt=1fs, eai=1ps, cwi=1ps	300 - 600	24	96
ALA-R-REMD-S	R-REMD, NVT, dt=2fs, eai=0.2ps, cwi=1ps, r=538	300 - 398	10	240
ALA-R-REMD-L	R-REMD, NVT, dt=2fs, eai=0.2ps, cwi=1ps, r=4764	300 - 398	10	84
ALA-TIG-1	TIGER2, NVT, dt=1fs, eai=3ps, cwi=1ps	300 - 600	4	75
ALA-TIG-2	TIGER2, NVT, dt=1fs, eai=4ps, cwi=1ps	300 - 600	4	100
ALA-TIG-3	TIGER2, NVT, dt=1fs, eai=7ps, cwi=1ps	300 - 600	4	105
RNA-NPT	NPT, dt=2fs, cwi=2ps	300	1	5000
RNA-398	NVT, dt=2fs, cwi=10ps	398	1	1405
RNA-REMD-GB	REMD, dt=2fs, eai=0.2ps, cwi=1ps	277 - 463	6	500
RNA-REMD-1	REMD, NVT, dt=2fs, eai=1ps, cwi=1ps	277 - 396	24	2010
RNA-REMD-2	REMD, NVT, dt=2fs, eai=0.2ps, cwi=1ps	277 - 396	24	500
RNA-REMD-3	REMD, NVT, dt=2fs, eai=0.2ps, cwi=1ps	277 - 396	24	400
RNA-R-REMD-S	R-REMD, NVT, dt=2fs, eai=1ps, cwi=1ps, r=10000	277 - 396	24	46
RNA-R-REMD-1	R-REMD, NVT, dt=2fs, eai=1ps, cwi=1ps, r=140510	277 - 396	24	205
RNA-R-REMD-2	R-REMD, NVT, dt=2fs, eai=1ps, cwi=1ps, r=140510	277 - 396	24	160
RNA-R-REMD-3	R-REMD, NVT, dt=2fs, eai=1ps, cwi=1ps, r=140510	277 - 396	24	250

Note: Simulation IDs beginning with "ALA-" contained one alanine dipeptide molecule solvated in TIP3P water. Those beginning with "RNA-" contained one rGACC molecule, three Na^+ counterions, and TIP3P water (with the exception of RNA-REMD-GB which used implicit solvent). Column titles and abbreviations are as follows: **Temp.**: the temperature range covered in the simulations, **Rep.**: the number of replicas used, **Length**: the per replica simulation time length in nanoseconds, **NVT**: constant volume/temperature, **NPT**: constant pressure/temperature, **REMD**: replica exchange molecular dynamics, **R-REMD**: reservoir replica exchange molecular dynamics, **dt**: the simulation time step, **cwi**: the trajectory coordinate writing interval, **eai**: the replica exchange attempt interval, **r**: number of frames in the reservoir.

Table 4.2. Temperatures used for REMD and R-REMD simulations.

ALA-24REMD											
300.00	309.83	319.91	330.22	340.79	351.63	362.74	374.11	385.76	397.70	409.94	422.47
435.33	448.51	462.02	475.87	490.06	504.60	519.52	534.83	550.53	566.64	583.14	600.07

ALA-10REMD, ALA-R-REMD-S, ALA-R-REMD-L									
300.00	309.83	319.91	330.22	340.79	351.63	362.74	374.11	385.76	397.70

All Explicit Solvent RNA REMD or R-REMD											
277.00	281.30	285.70	290.20	294.70	299.40	304.00	308.80	313.60	318.50	323.50	328.60
333.70	338.90	344.20	349.60	355.10	360.70	366.30	372.00	377.90	383.80	389.80	395.90

RNA-REMD-GB					
277.00	308.16	342.06	379.00	419.19	462.87

Note: All temperatures are given in Kelvin. The distributions were obtained from an online generator at http://folding.bmc.uu.se/remd/. Refer to Table 4.1 for more information about each simulation ID, noting that the number refers to the number of replicas and "S" and "L" for the R-REMD simulations refer to small and large reservoirs, respectively.

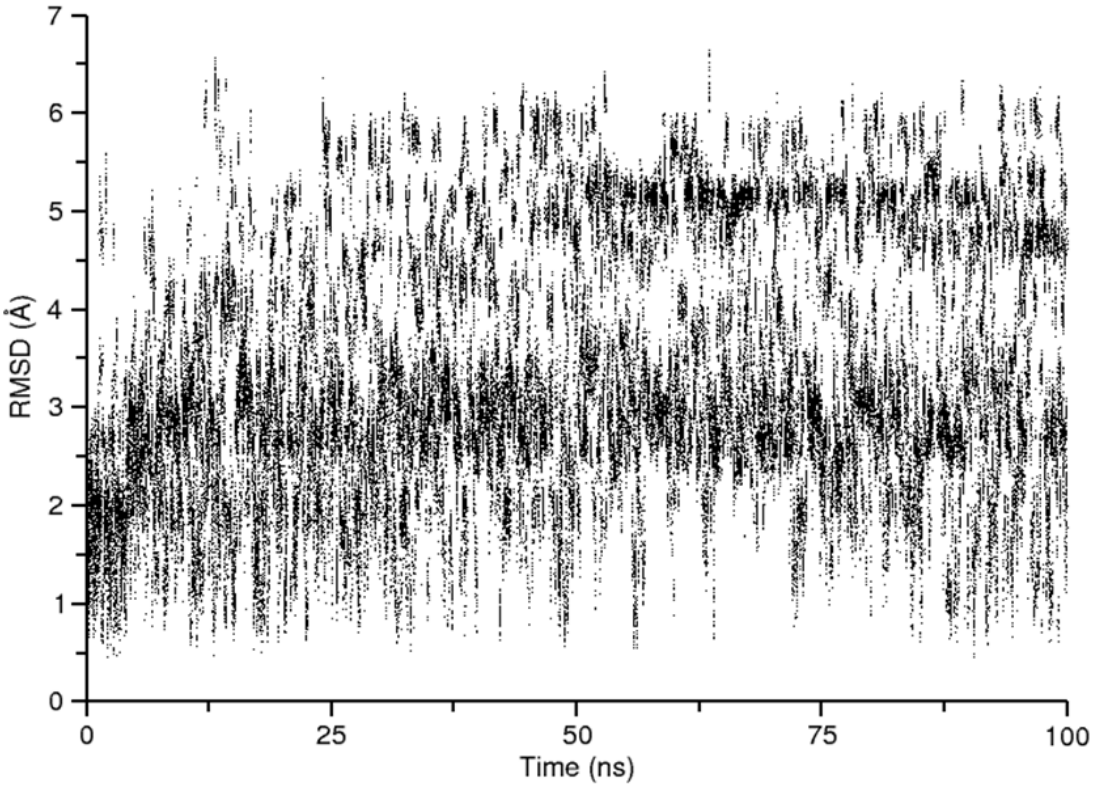

Figure 4.2. RMSD versus time at 277 K for the first 100 ns of the RNA-REMD-3 simulation. The plot demonstrates that about 50 ns are required for the baseline temperature to begin evenly distributed sampling of the RMSD space.

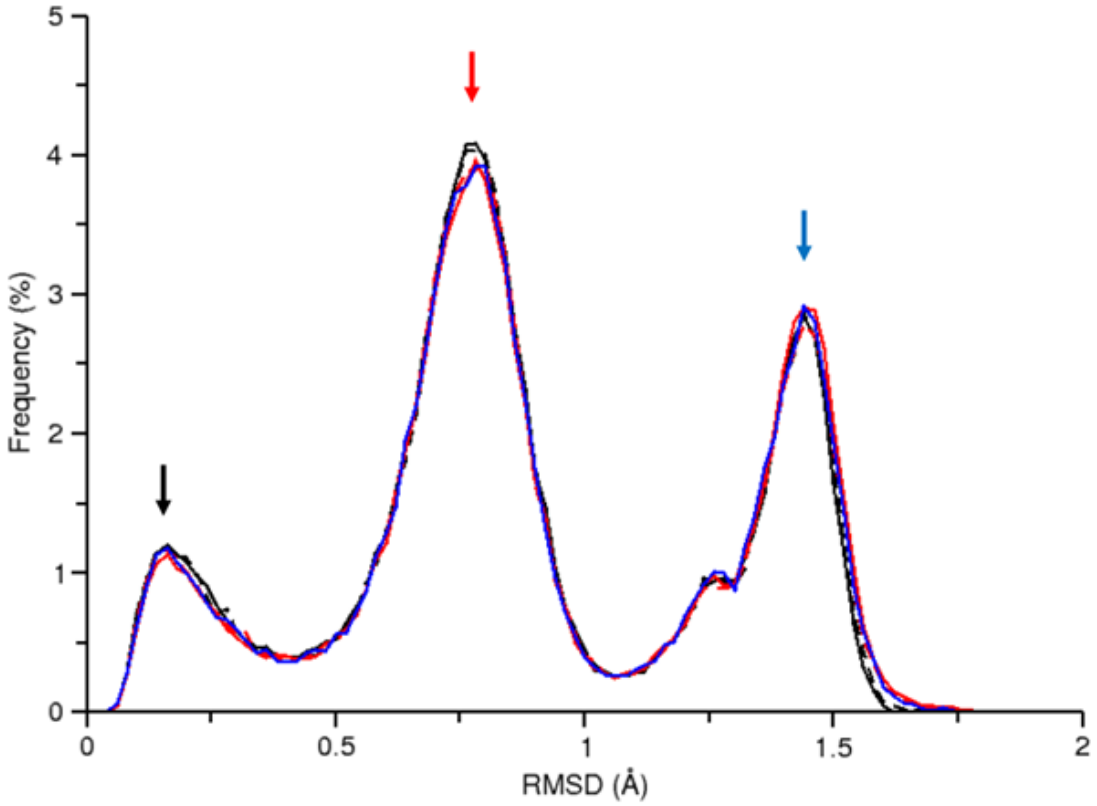

Figure 4.3. RMSD profile for the 300 K simulation data from the following alanine dipeptide simulations: ALA-NVT (*solid black*), ALA-NPT (*dashed black*), ALA-10REMD (*solid red*), ALA-24REMD (*dashed red*), ALA-R-REMD-L (*solid blue*). Tight overlap in the profiles suggests that these simulations explore very similar RMSD space. The colored arrows indicate RMSD profile points studied for the time convergence displayed in Figure 4.5. The points were chosen due to their variability as observed in Figure 4.4.

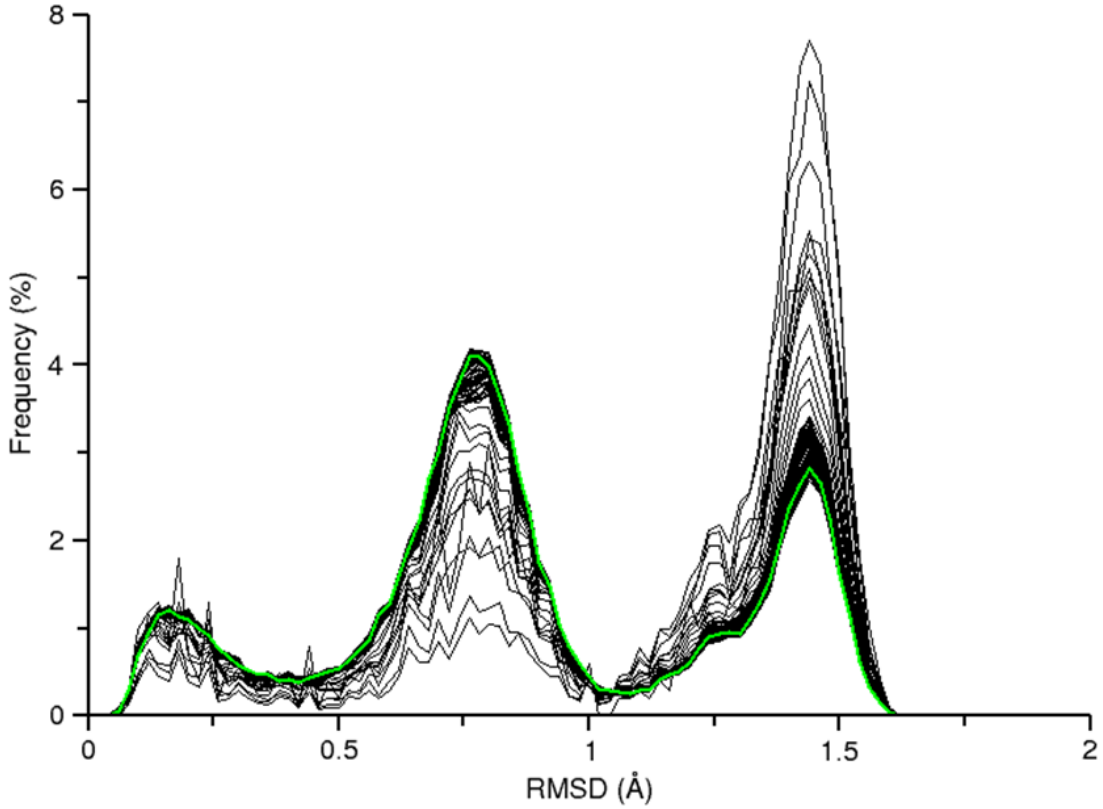

Figure 4.4. Convergence of the RMSD profile of ALA-NVT. The cumulative RMSD profiles, plotted each nanosecond of the 100 ns simulation, is given (*black lines*) along with the final profile (*green line*).

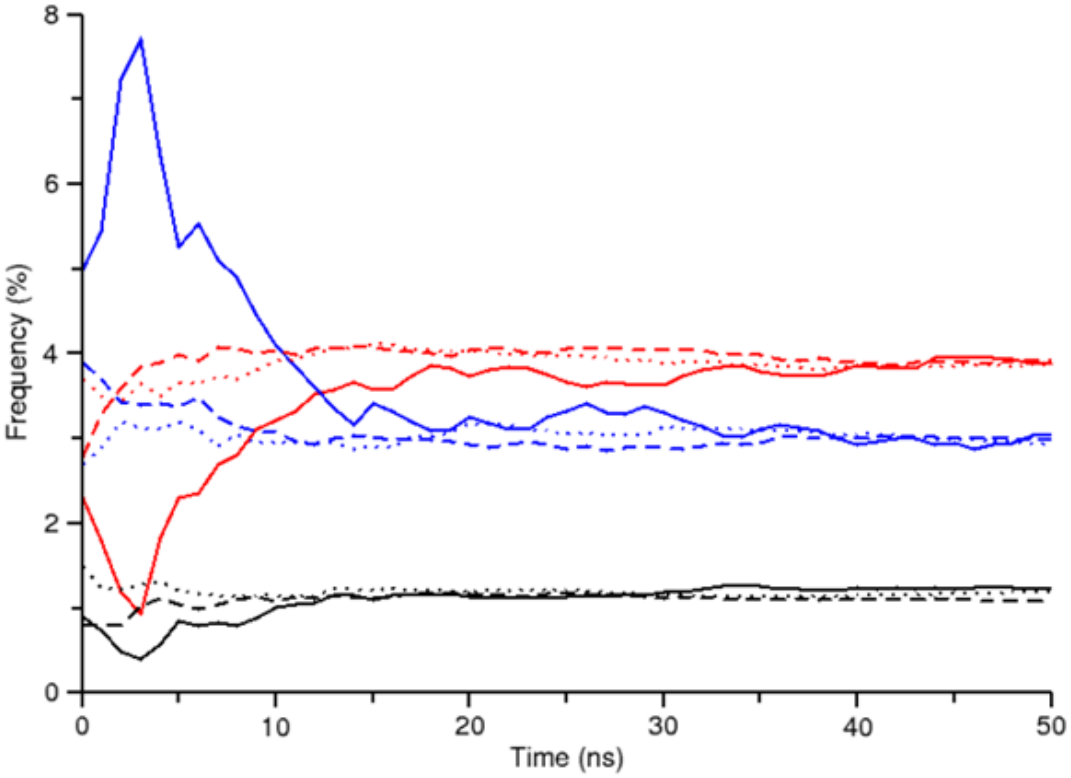

Figure 4.5. Convergence of alanine dipeptide RMSD profile maxima versus simulation time at 300 K for the following simulations: ALA-NVT (*solid*), ALA-10REMD (*dashed*), ALA-R-REMD-L (*dotted*). Colors correspond to the three RMSD profile maxima observed in Figure 4.3 at the following values: 0.16 (*black*), 0.78 (*red*), 1.44 Å (*blue*). Only the first 50 ns of each simulation are shown to emphasize the initial convergence period.

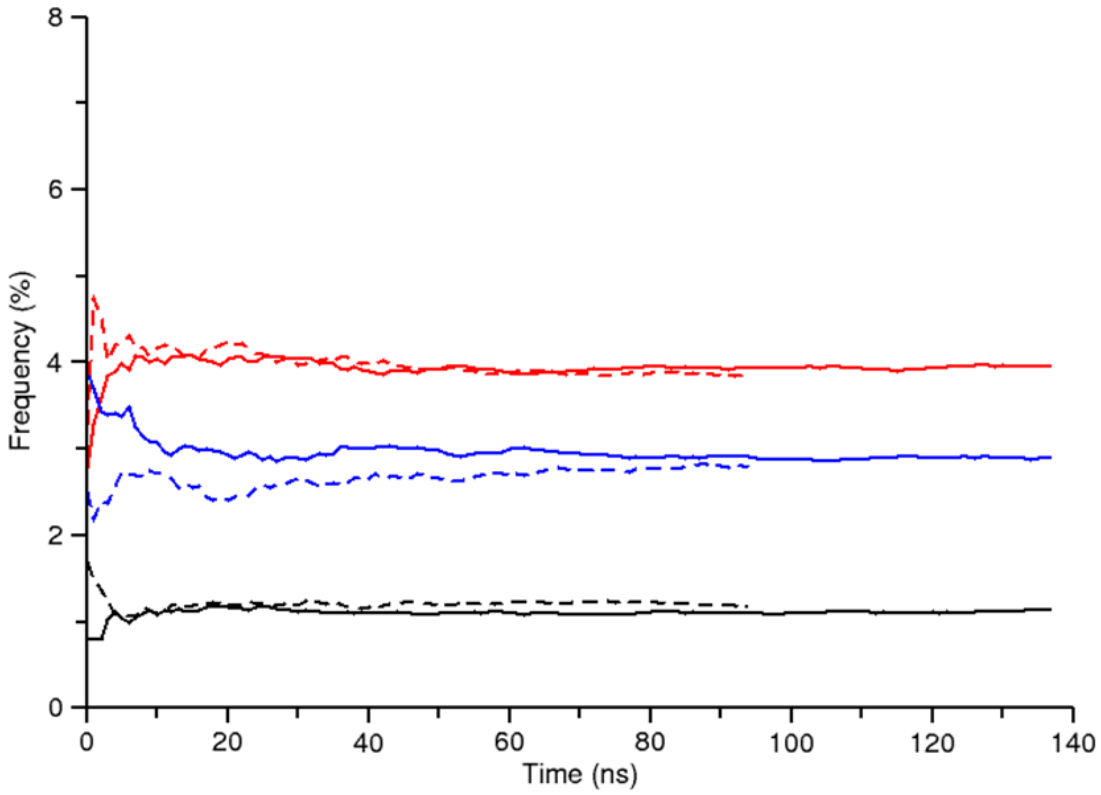

Figure 4.6. Convergence of alanine dipeptide RMSD profile maxima versus simulation time at 300K for the following simulations: ALA-10REMD (*solid*) and ALA-24REMD (*dashed*) Colors correspond to the three RMSD profile maxima observed in Figure 4.2 at the following values: 0.16 (*black*), 0.78 (*red*), 1.44 Å (*blue*). The vertical axis maximum value was set to 8% for comparison with Figure 4.5.

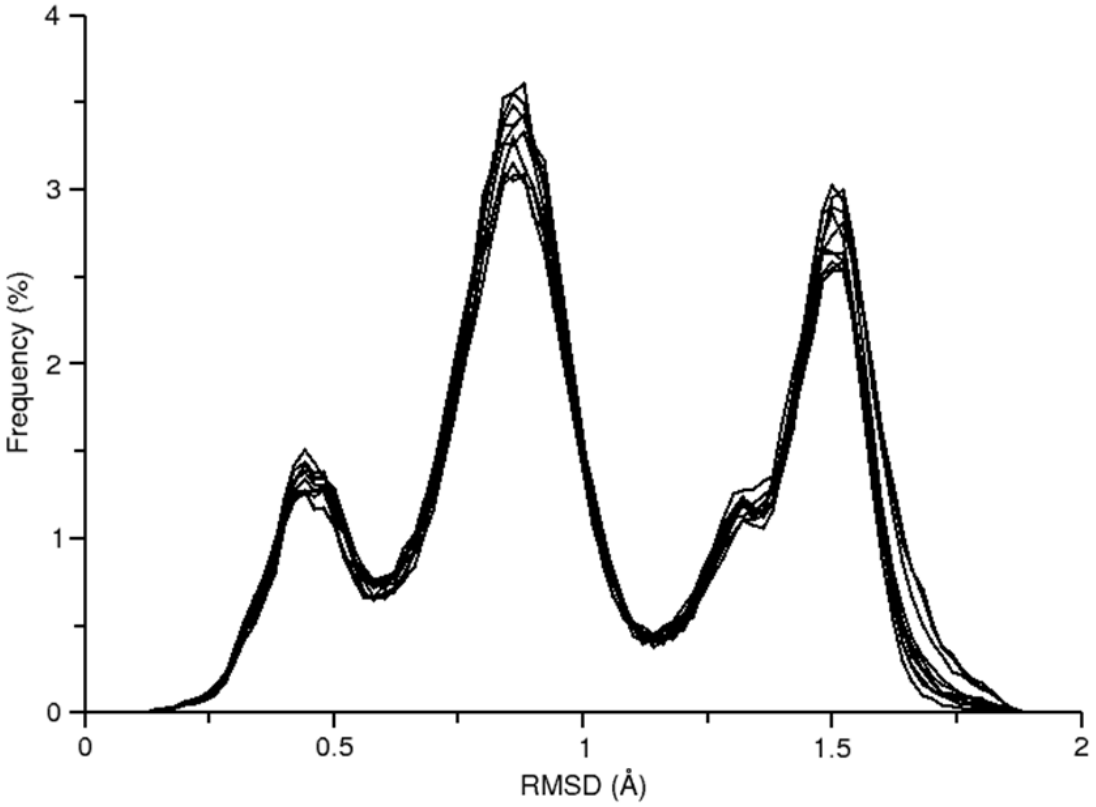

Figure 4.7. Overlay of the ten replica RMSD profiles for the ALA-10REMD simulation. Convergence of the simulation is indicated by the tight overlap of the RMSD profiles suggesting that all replicas have sampled similar RMSD space. Temperature sorting was not performed when generating this plot and thus each of the ten replicas contains simulation data from all available temperatures.

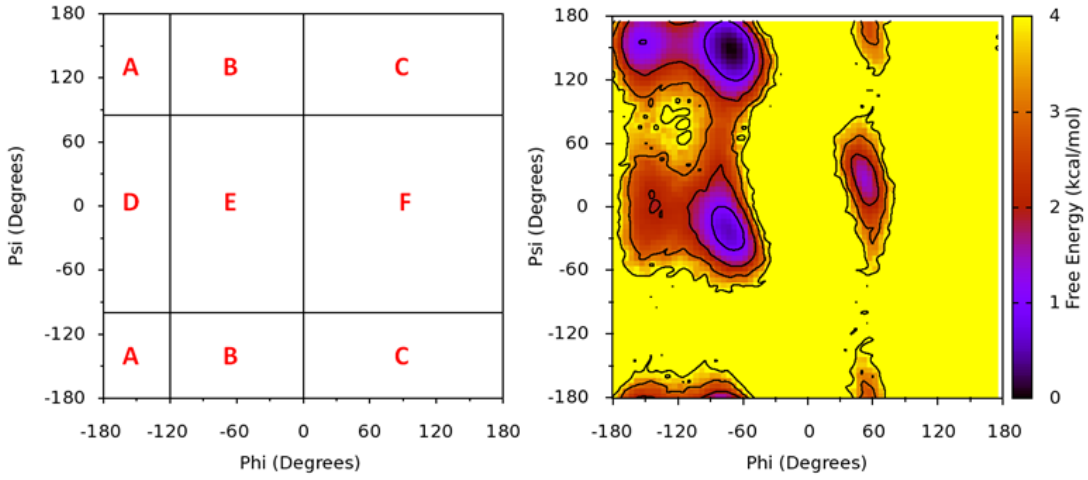

Figure 4.8. Identification of alanine dipeptide phi/psi conformational divisions. (*Left*) Six regions were identified and labeled A-F. (*Right*) Population based free energy plot based on phi/psi dihedral angles for the ALA-10REMD simulation at 300 K. Free energy estimates are calculated using the following equation: $G_i = -k_B T \ln(N_i/N_{tot})$.

Table 4.3. Alanine dipeptide conformational distribution at 300 K for simulations in this work.

	Alanine Dipeptide Conformational Frequency (%)					
Simulation	**A**	**B**	**C**	**D**	**E**	**F**
ALA-NVT	12.3 (0.2)	56.6 (1.9)	0.0 (0.0)	4.1 (0.1)	27.0 (1.7)	0.0 (0.0)
ALA-NPT	12.2 (0.0)	56.9 (1.2)	0.2 (0.1)	4.0 (0.1)	25.1 (0.6)	1.7 (0.3)
ALA-10REMD	11.3 (0.2)	54.2 (0.1)	0.6 (0.1)	3.8 (0.0)	26.0 (0.3)	4.1 (0.7)
ALA-24REMD	11.8 (0.5)	55.6 (0.1)	0.3 (0.1)	3.9 (0.0)	25.5 (0.3)	3.0 (0.7)
ALA-R-REMD-S	13.3 (0.2)	58.9 (0.1)	0.1 (0.0)	3.4 (0.0)	23.0 (0.1)	1.2 (0.2)
ALA-R-REMD-L	11.6 (0.0)	54.9 (1.0)	0.3 (0.0)	4.1 (0.2)	26.3 (1.2)	2.9 (0.5)
ALA-TIG-1	11.3 (0.7)	49.5 (2.9)	0.9 (0.3)	4.2 (0.2)	28.9 (1.5)	5.2 (2.1)
ALA-TIG-2	12.3 (0.8)	45.9 (0.1)	1.3 (0.2)	4.8 (0.4)	28.6 (1.9)	7.1 (2.7)
ALA-TIG-3	11.3 (0.2)	46.0 (2.0)	2.0 (0.5)	4.7 (0.1)	27.8 (0.2)	8.2 (2.0)

Note: Letters A-F indicate regions of the phi/psi space indicated in Figure 4.8, left. Given in parentheses is a crude approximation of error obtained by comparing the average frequency values from the first and second halves of a simulation according to the following equation: Error = abs(FirstHalf - SecondHalf)/2

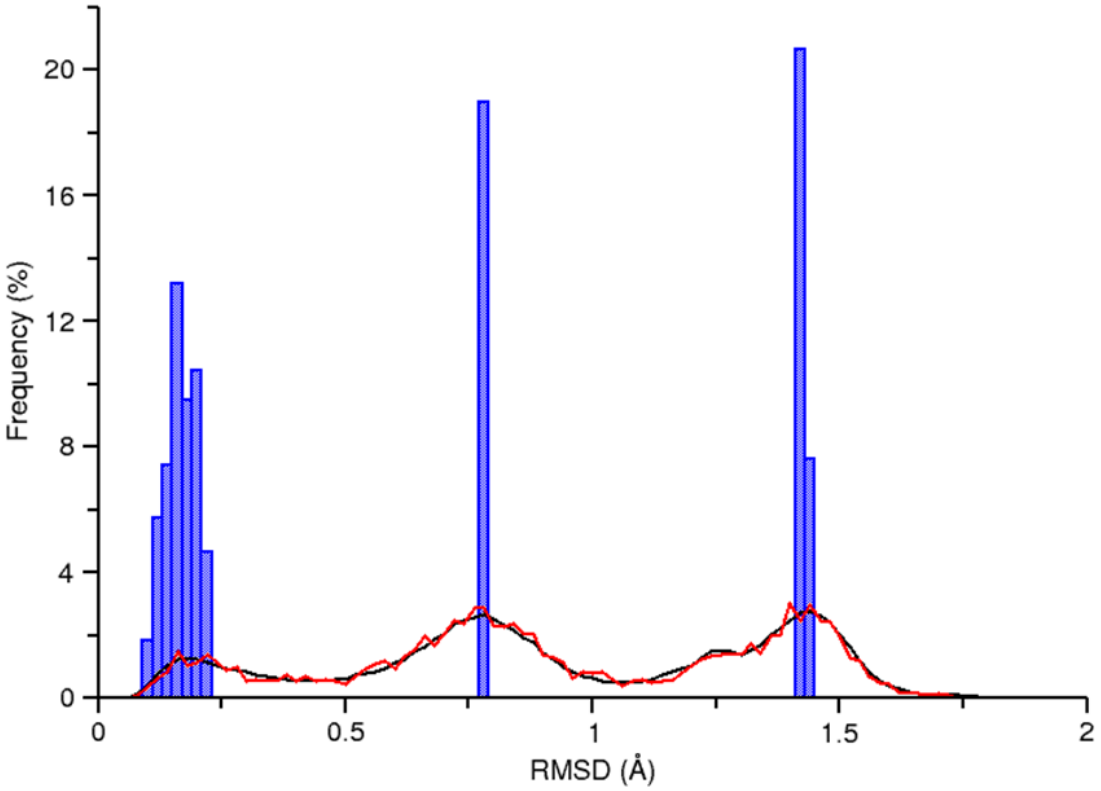

Figure 4.9. RMSD profiles for the following simulations: ALA-10REMD at the 397.7 K temperature level (*black*), reservoir of ALA-rREMD-L (*red*), reservoir of ALA-rREMD-S (*blue bars*). The ALA-rREMD-L reservoir (*red*) contains 4764 frames from a 47 ns simulation at 398 K, whereas the ALA-rREMD-S reservoir (*blue bars*) contains 538 artificially selected frames from the larger ALA-rREMD-L reservoir. Bars are used to depict the ALA-rREMD-S reservoir due to the sparseness of the data.

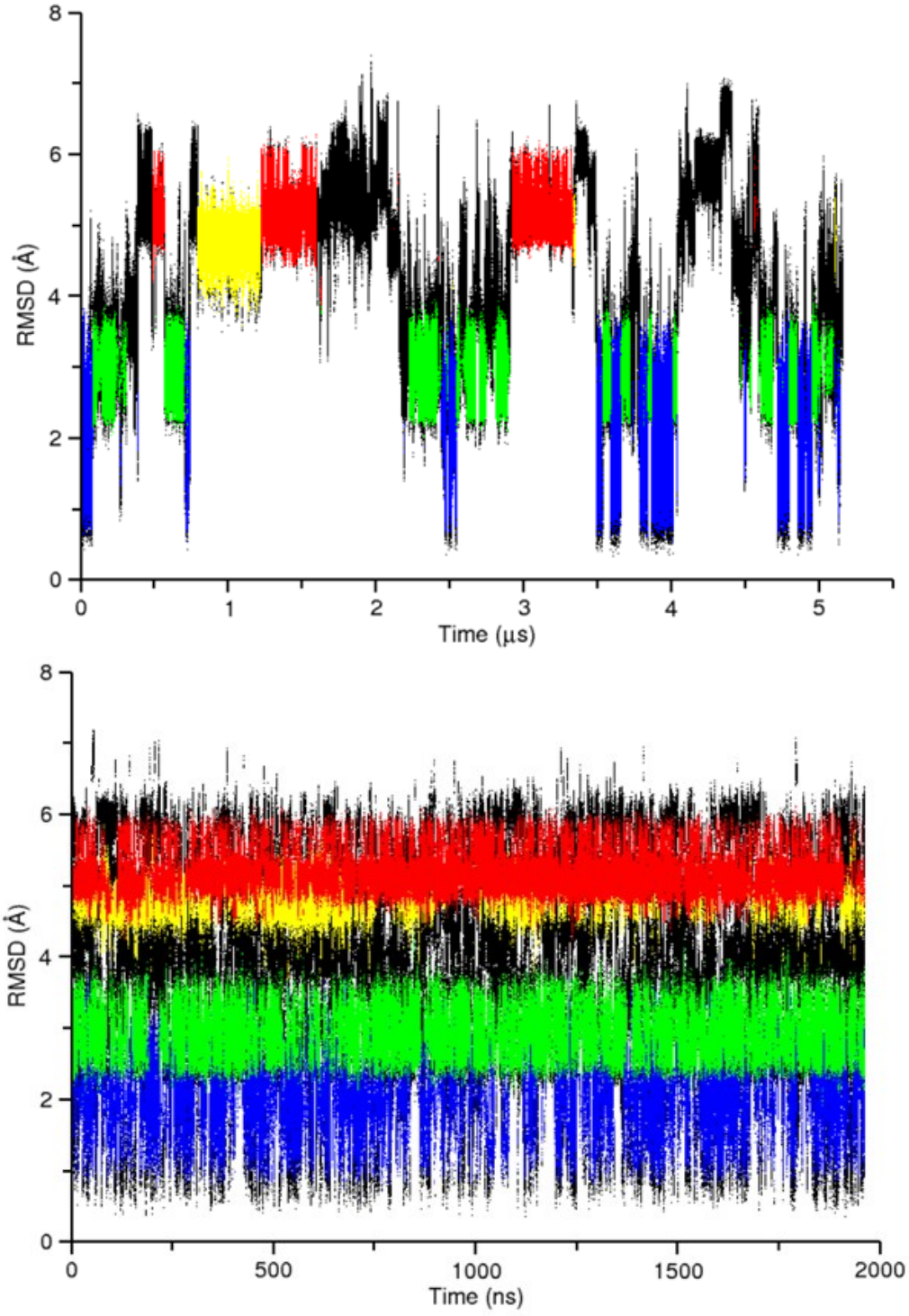

Figure 4.10. RMSD analysis of the RNA-NPT (*top*) and the RNA-REMD-1 (*bottom*) simulations. The data for RNA-REMD-1 are taken exclusively from the 277 K temperature level. The colored regions correspond to the four most populated conformational clusters depicted in Figure 4.11 as follows: Intercalated (*red*), NMR Minor (*green*), NMR Major (*blue*), Inverted (*yellow*), other conformations (*black*). (*Right*) RMSD profile for the data at left.

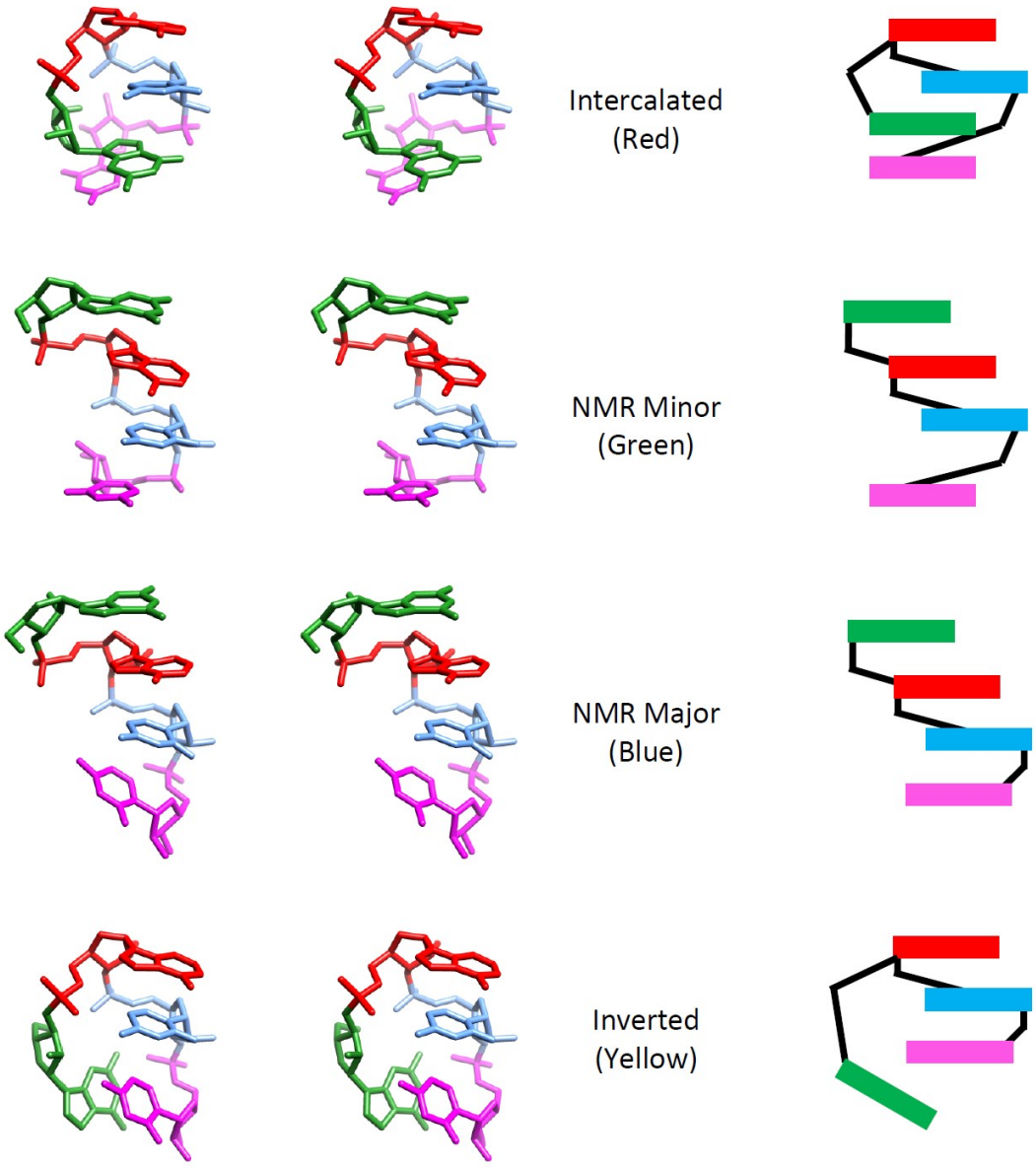

Figure 4.11. Stereoview and cartoon representation of the representative structures for the four most populated conformational clusters of the RNA-NPT simulation. Each structure is labeled with a reference name and color which is used in other plots to indicate to the conformation.

Table 4.4. Conformational frequency of the RNA simulations determined by cluster analysis.

	rGACC Conformational Frequency (%)			
Simulation ID	Intercalated	NMR Minor	NMR Major	Inverted
RNA-NPT[1]	16.0 (0.3)	12.9 (0.7)	9.2 (0.4)	8.4 (0.1)
RNA-398[2]	6.2 (0.4)	3.5 (0.3)	3.1 (0.1)	7.1 (0.5)
RNA-REMD-GB	-- --	-- --	-- --	92.9 (0.7)
RNA-REMD-1	24.5 (0.9)	15.9 (0.7)	11.8 (0.6)	7.6 (0.0)
RNA-REMD-2	24.2 (1.2)	10.5 (1.0)	8.8 (0.5)	9.9 (0.1)
RNA-REMD-3	18.8 (0.9)	16.3 (1.0)	13.1 (0.5)	7.3 (0.1)
RNA-rREMD-S	29.4 (0.1)	28.3 (1.1)	12.0 (0.2)	-- --
RNA-rREMD-1	18.7 (0.3)	15.5 (0.7)	13.1 (0.4)	11.3 (0.1)
RNA-rREMD-2	18.5 (0.1)	15.3 (1.0)	13.6 (0.1)	10.9 (0.0)
RNA-rREMD-3	18.7 (0.5)	14.6 (0.7)	14.0 (0.4)	10.2 (0.1)

Note: The four conformational categories are structurally depicted in Figure 4.6. Data shown are for the 277 K temperature level, except for the RNA-NPT and RNA-398 simulations, for which the temperature is indicated (below). An error estimate is given in parentheses which corresponds to the standard deviation of five independent clustering calculations. [1]Simulation performed at 300 K. [2]Simulation performed at 398 K. The "-- --" indicates that the conformation was not observed in the top fifteen most populated clusters.

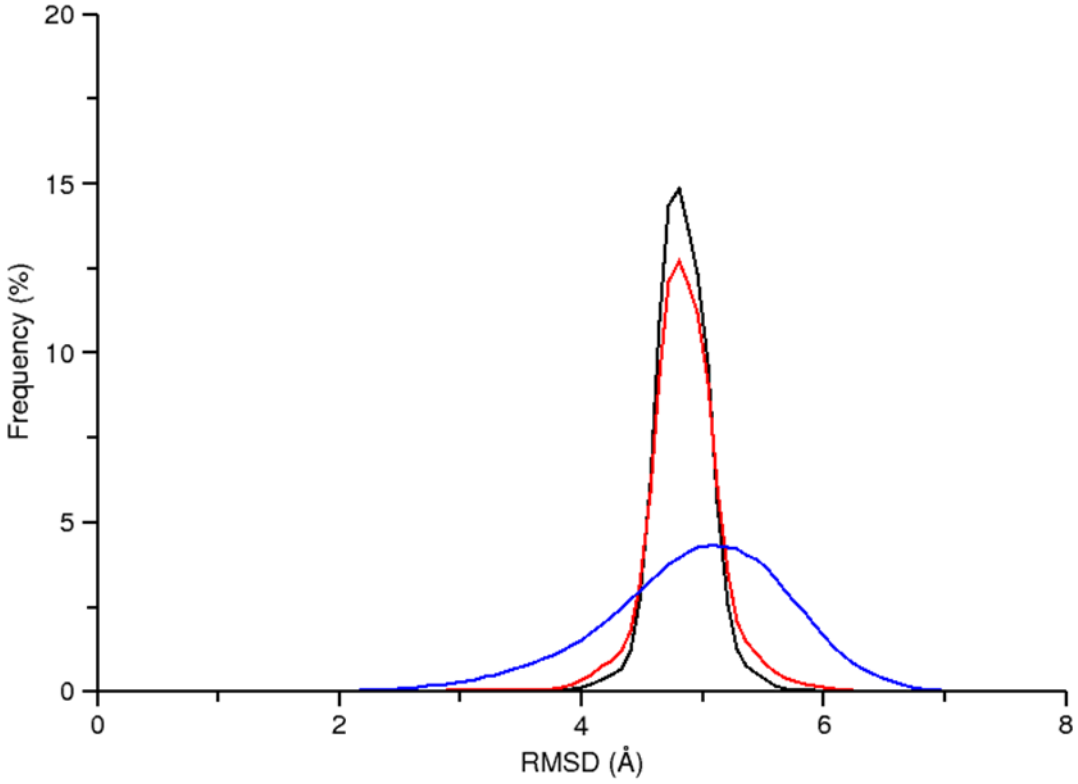

Figure 4.12. RMSD profile for RNA-REMD-GB at 277 K (*black*), 308 K (*red*), and 462 K (*blue*). The two lowest temperature profiles (*black and red*) exclusively adopt the Inverted conformation (refer to Figure 4.11). The highest temperature simulation (*blue*) is mostly unstructured.

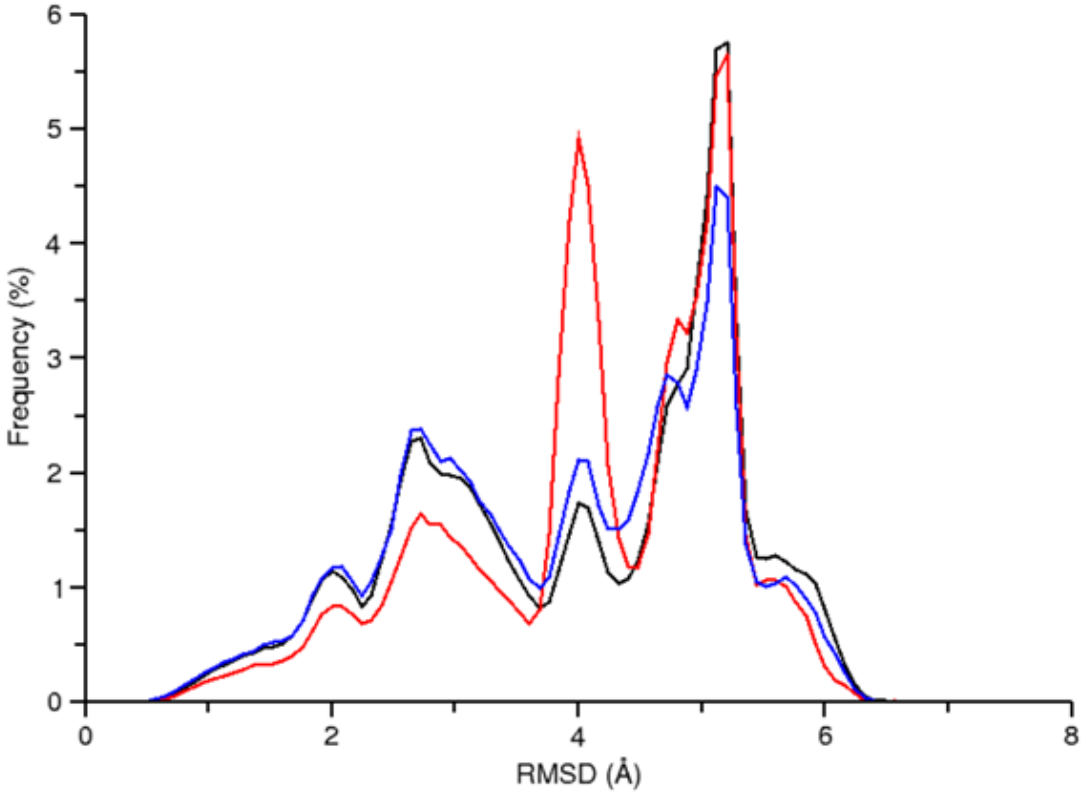

Figure 4.13. RMSD profile for data collected at 277 K from the following simulations: RNA-REMD-1 (*black*), RNA-REMD-2 (*red*), and RNA-REMD-3 (*blue*).

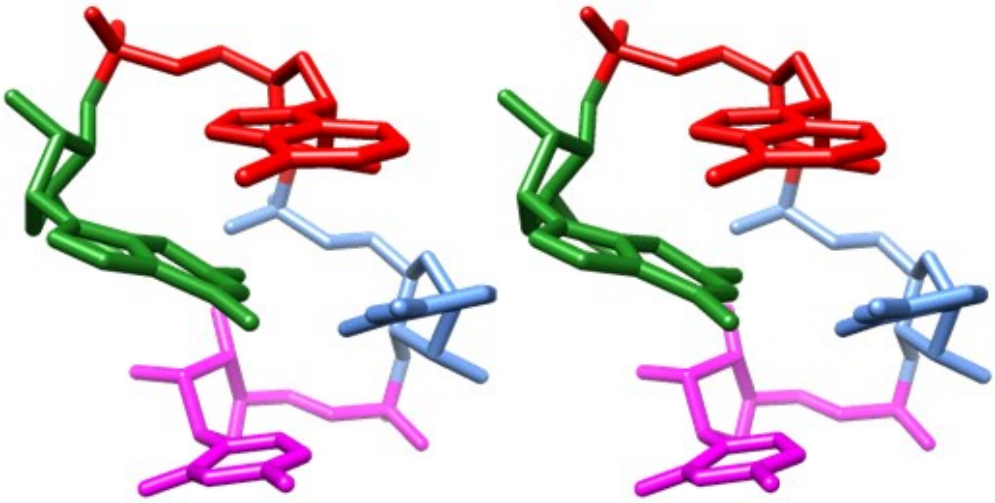

Figure 4.14. Molecular graphics stereoview of the additional RNA conformation which forms an RMSD profile peak at 4.0 Å in the RNA simulations from this work. The conformation appears to dominate some replicas in the traditional REMD simulations. It is also present in both the RNA-398 and the three R-REMD simulations but at low populations (<2%).

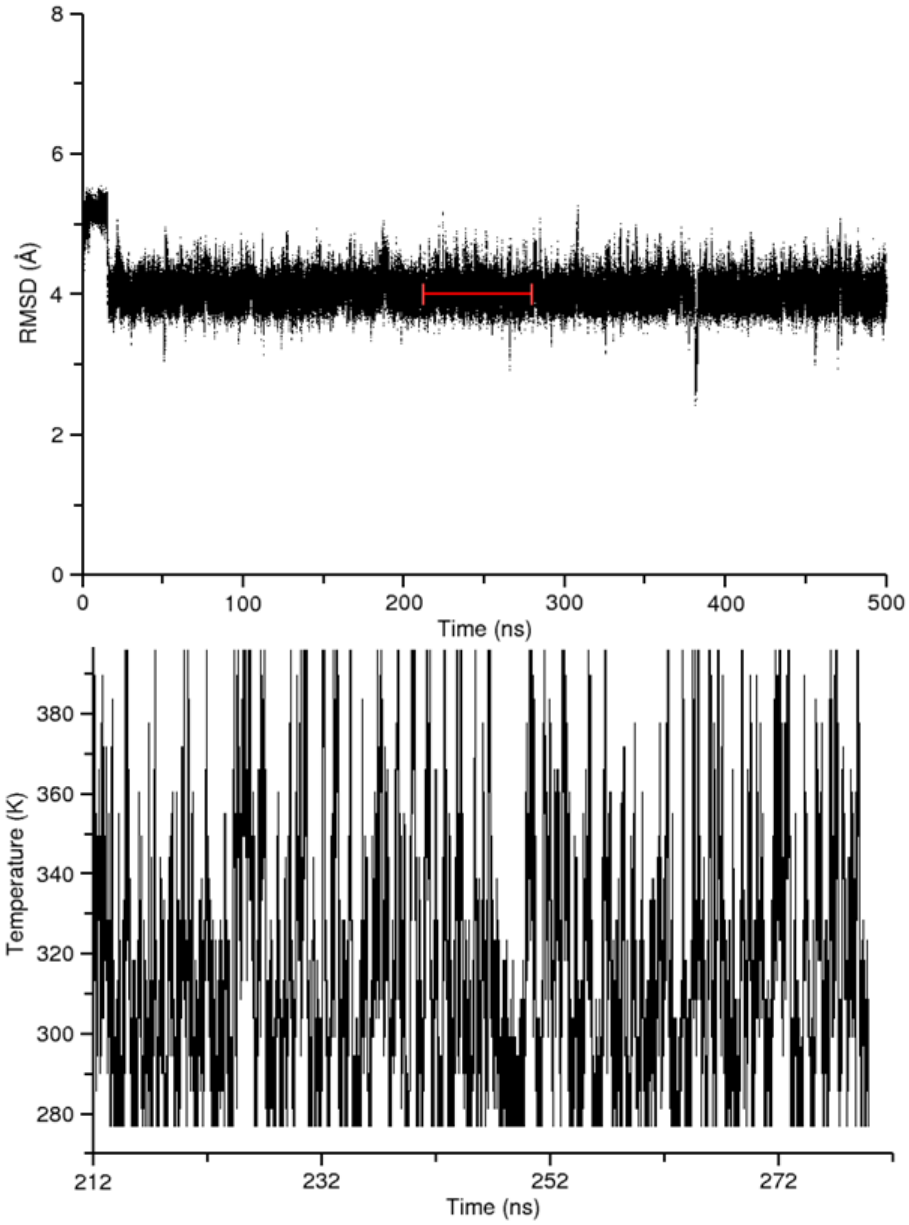

Figure 4.15. An example of a trapped replica in the RNA-REMD-2 simulation. *(Top)* RMSD versus time plot for replica 16. The primary conformation sampled, around 4.0 Å, is depicted in Figure 4.14. These data come directly from the replica trajectory (prior to temperature sorting) and thus include all the temperatures from the simulation. To demonstrate this, we analyzed the target temperature of the replica during time period indicated by the red line segment. *(Bottom)* The target temperature of replica 16 from RNA-REMD-2 during the simulation time period indicated in the top plot. The replica regularly moves between all target temperatures specified for the REMD ensemble.

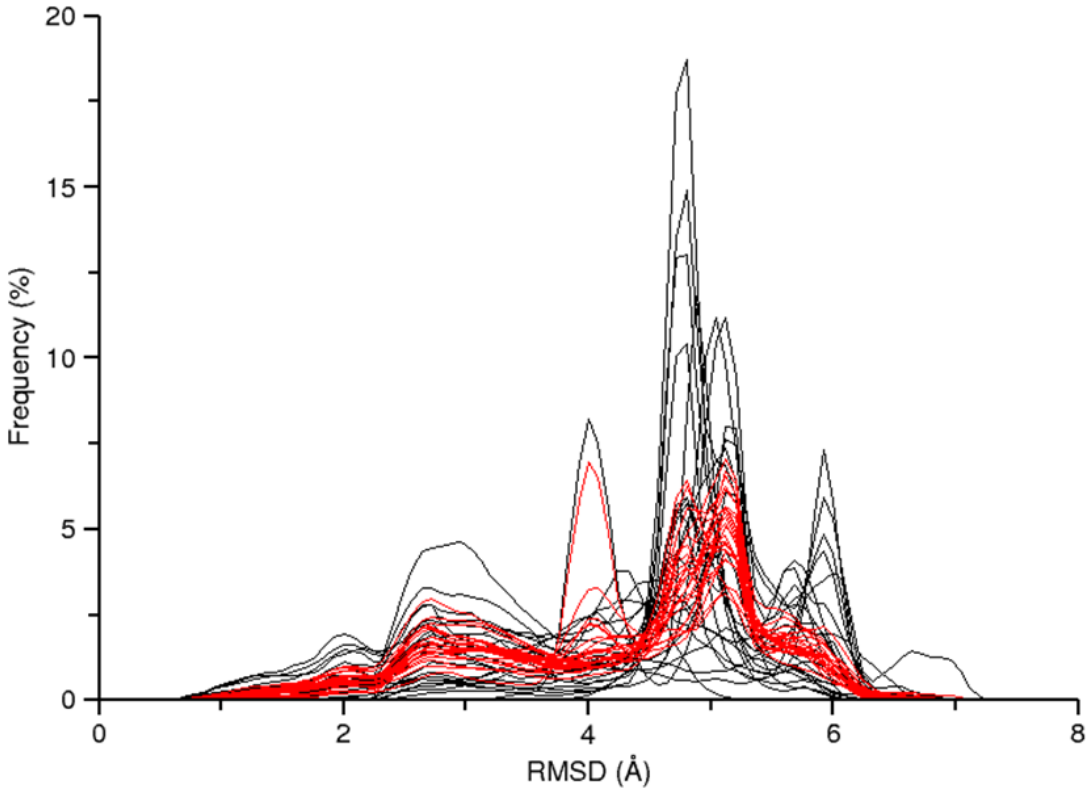

Figure 4.16. Overlay of the twenty-four replica RMSD profiles for the RNA-REMD-1 simulation after 200 ns (*black*) and for all the data (*red*). Temperature sorting was not performed when generating this plot and thus each of the replicas contains simulation data from all target temperatures.

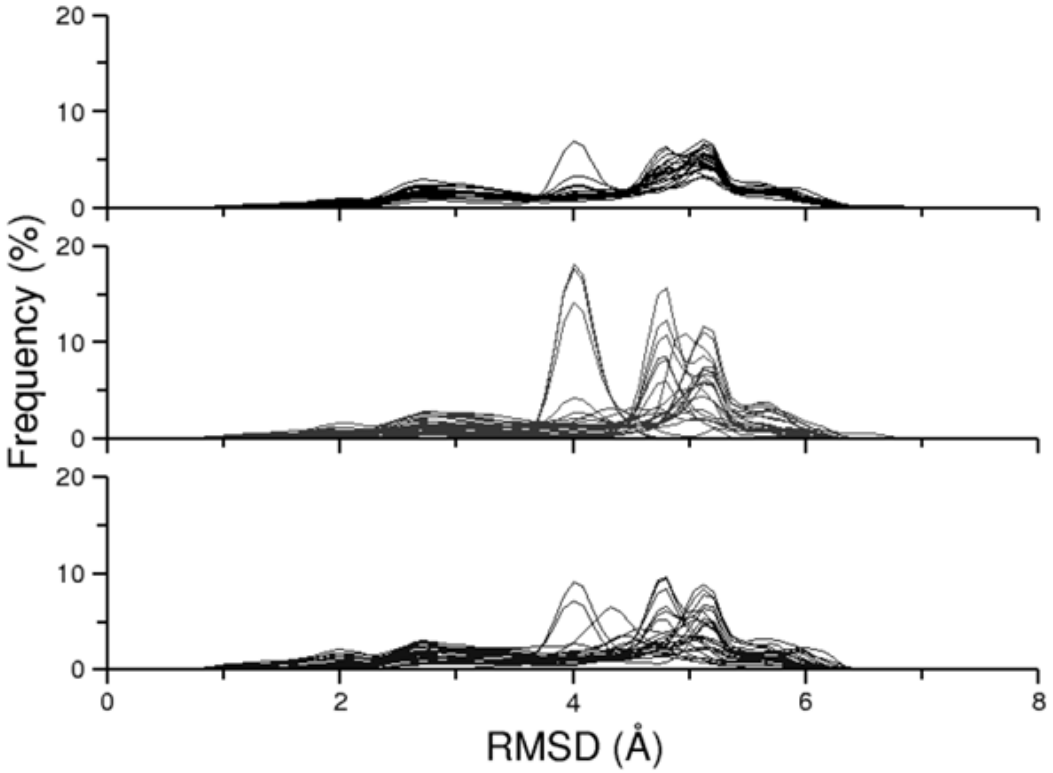

Figure 4.17. Overlay of the twenty-four replica RMSD profiles for the following simulations: RNA-REMD-1 (*top*), RNA-REMD-2 (*middle*), RNA-REMD-3 (*bottom*). A fully converged REMD ensemble should produce an identical curve for each replica and thus disorder in the plot indicates that the simulation has not yet fully converged.

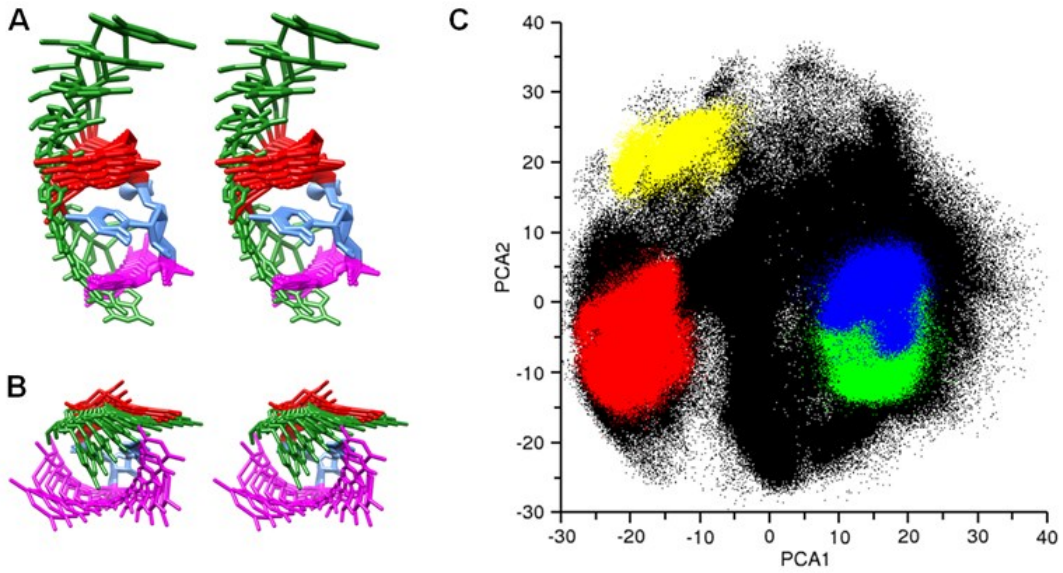

Figure 4.18. PCA analysis of the RNA-REMD-1 simulation at 277 K. Stereoview depiction of the motion described by the first (*A*) and second (*B*) eigenvectors determined by principal component analysis. (*C*) Distribution of the simulation data along the first two eigenvectors identified by PCA. Colored regions correspond to the four most populated conformational clusters depicted in Figure 4.11 as follows: Intercalated (*red*), NMR Minor (*green*), NMR Major (*blue*), Inverted (*yellow*), other conformations (*black*).

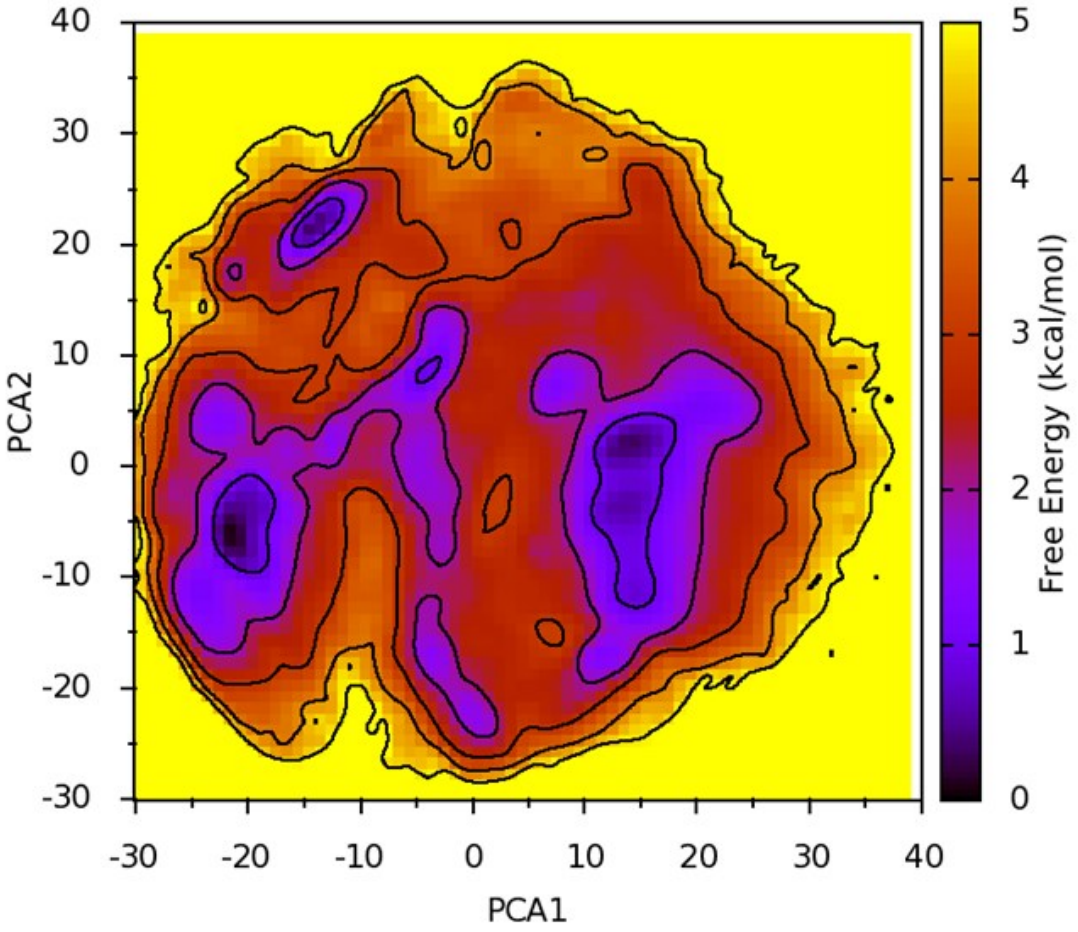

Figure 4.19. Conformational analysis of the RNA-REMD-1 at 277 K using a population based free energy plot. Free energy estimates are calculated using $G_i = -k_B T \ln(N_i/N_{tot})$. Compare with Figure 4.18C to observe clustering locations.

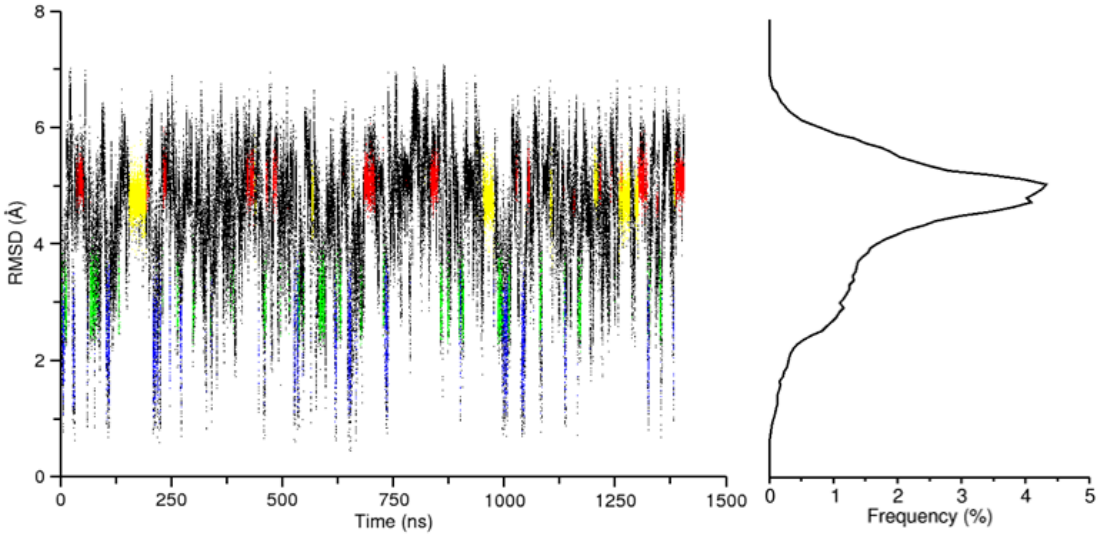

Figure 4.20. Conformational analysis of the RNA-398 simulation. This simulation was used to generate structures for the R-REMD simulations of the RNA. (*Left*) RMSD versus simulation time. Colored regions correspond to the four most populated conformational clusters depicted in Figure 4.11 as follows: Intercalated (*red*), NMR Minor (*green*), NMR Major (*blue*), Inverted (*yellow*), other conformations (*black*). (*Right*) RMSD profile for the data at left.

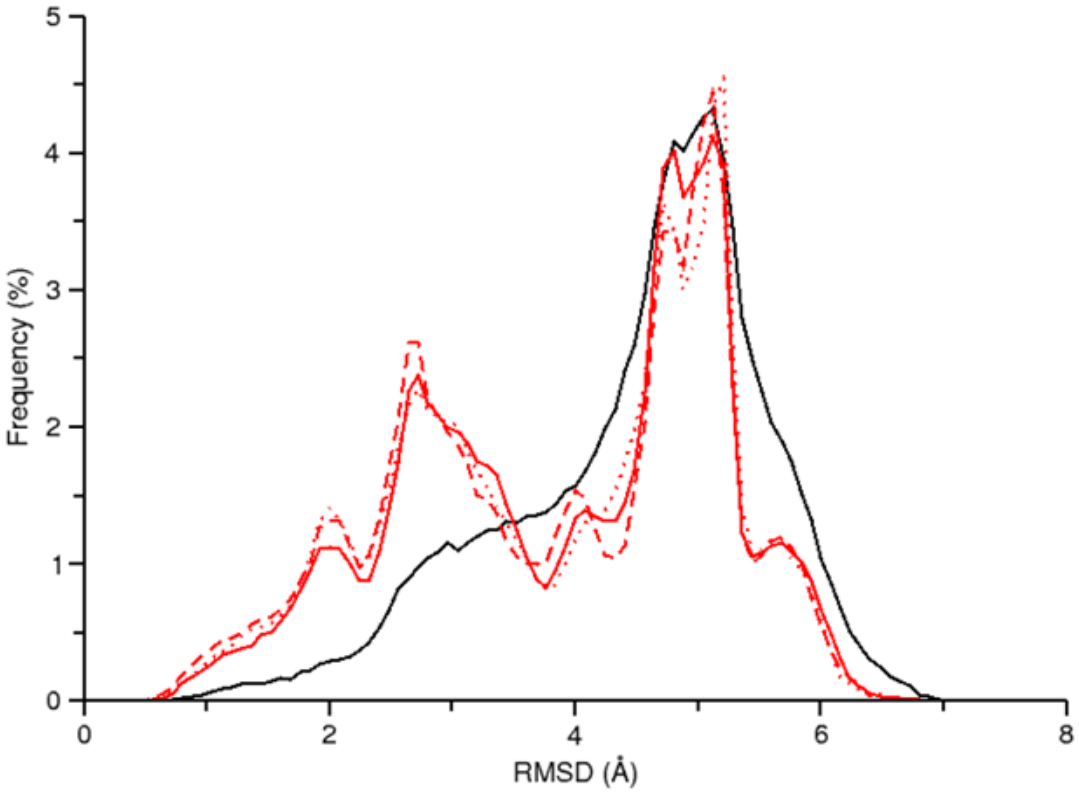

Figure 4.21. RMSD profile at the 277 K temperature level for RNA-R-REMD-1 (*solid red*), RNA-R-REMD-2 (*dashed red*), RNA-R-REMD-3 (*dotted red*) as well as the profile for the structural reservoir used in these simulations (which was generated by conventional MD at 398 K).

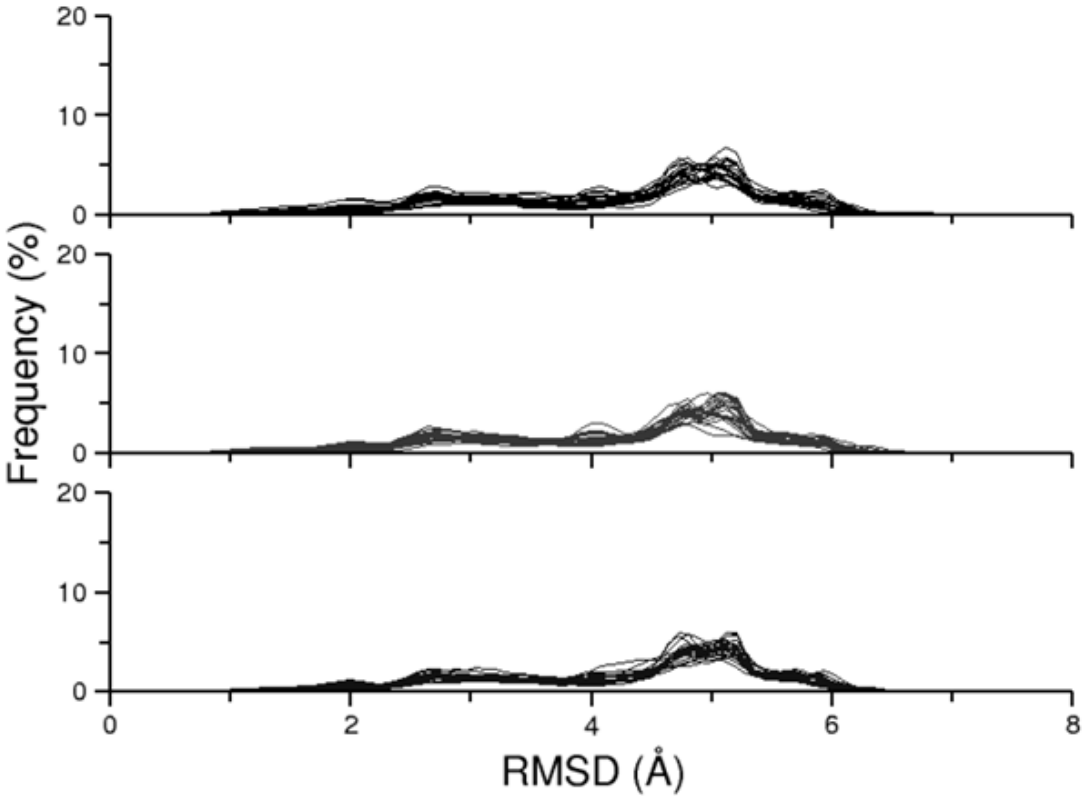

Figure 4.22. Overlay of the twenty-four replica RMSD profiles for the following simulations: RNA-R-REMD-1 (*top*), RNA-R-REMD-2 (*middle*), RNA-R-REMD-3 (*bottom*). The replica profiles for these R-REMD simulations seem to be closer to convergence than the traditional REMD simulations (see Figure 4.17).

Table 4.5. Sugar pucker and base orientation at three temperatures from the RNA-R-REMD-3 simulation.

	277 K		299 K		328 K	
Residue	**% C3'-endo**	**% Anti**	**% C3'-endo**	**% Anti**	**% C3'-endo**	**% Anti**
G1	91 (80-90)	<u>64</u>	89 (60-70)	57	<u>84</u> (50-60)	48
A2	81 (80-90)	96	79 (80-90)	94	76 (70-80)	92
C3	87 (80-90)	98	84 (80-90)	98	80 (70-80)	97
C4	<u>52</u> (70-80)	99	50 (70-80)	99	<u>49</u> (60-70)	99

Note: In parentheses are the experimentally determined ranges for the sugar pucker determined by NMR at the following temperatures: 278, 298, and 328 K *(39)*. The base orientation, determined by NMR, was anti for all four residues at 275 K. Values discussed in the main text are underlined.

Table 4.6. NOEs predicted by the RNA-R-REMD-3 simulation at 277 K but not observed experimentally by NMR.

Atom 1	Atom 2	r^6 Avg.	Int.	Min.	Maj.	Inv.
:3@H3'	:4@H3'	2.9	**2.6**	**2.8**	4.8	5.0
:3@H3'	:4@H2'	2.9	4.8	**2.9**	6.1	6.5
:1@H2'	:3@H5	3.2	3.6	6.6	5.8	**3.1**
:1@H5'	:4@H2'	3.4	**2.7**	12.6	14.1	11.2
:1@H8	:4@H3'	3.8	**2.9**	12.5	14.4	7.5
:1@H2'	:4@H5	3.8	6.3	13.6	8.4	**2.6**
:1@H8	:4@H2'	3.9	**3.2**	12.1	13.9	6.7
:1@H8	:3@H3'	4.5	**3.2**	12.2	11.7	8.5
:1@H3'	:3@H6	4.6	**4.0**	7.3	6.5	**4.3**
:1@H8	:4@H6	4.7	**3.5**	14.1	11.9	6.7
:1@H5'	:4@H3'	4.7	**4.5**	12.9	14.6	10.6
:1@H8	:4@H5	4.8	**4.0**	14.9	10.1	5.0
:1@H8	:3@H2'	4.9	**4.0**	12.6	12.3	8.7
:1@H8	:3@H6	4.9	**4.2**	9.6	8.9	8.5
:1@H5"	:4@H2'	5.0	**4.0**	11.0	12.7	11.0

Note: the first three columns indicate the atom pairs and the r^6-averaged distances obtained from the simulation. The last four columns list the corresponding distance observed in the most populous representative structures which are depicted in Figure 4.11. Conformational abbreviations: **Int**. (Intercalated). **Min**. (NMR Minor). **Maj**. (NMR Major). **Inv**. (Inverted).

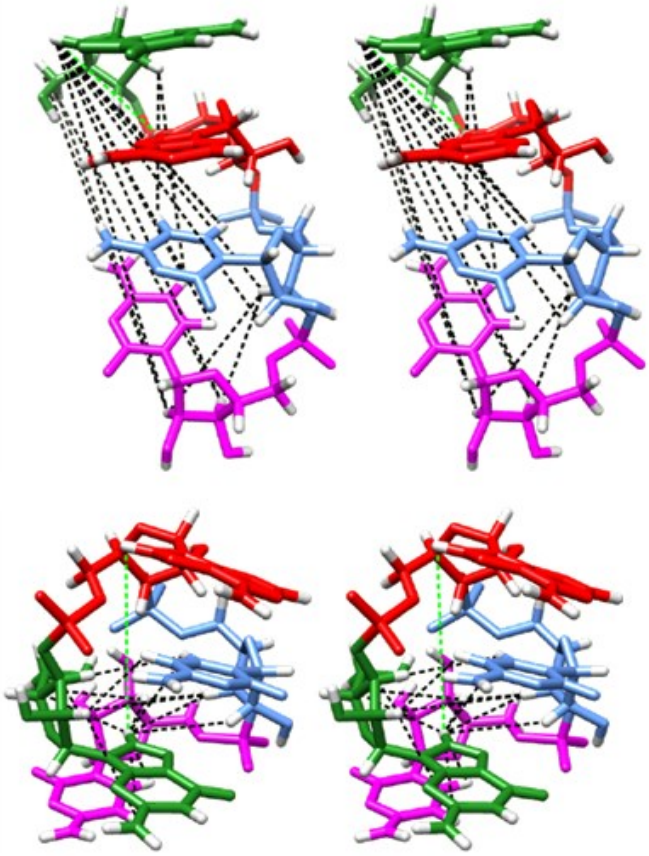

Figure 4.23. Stereo views of the simulation predicted NOEs (*black lines*) for RNA-R-REMD-3 at 277 K mapped onto the NMR Major conformation (*top*) and the Intercalated conformation (*bottom*). Only the NOEs for which the Intercalated conformation is the primary contributor are shown. The single NOE violation between G1 H8 and A2 H8 is also depicted (*green line*).

4.6 References

1. Meister, G. (2011) *RNA Biology: An Introduction*, Wiley.

2. MacKerell Jr, A. D., and Nilsson, L. (2008) Molecular dynamics simulations of nucleic acid-protein complexes, *Curr. Opin. Struct. Biol. 18*, 194-199.

3. Cheatham III, T. E., and Kollman, P. A. (2000) Molecular dynamics simulation of nucleic acids, *Annu. Rev. Phys. Chem. 51*, 435-471.

4. Cheatham, T. E., 3rd, and Young, M. A. (2000) Molecular dynamics simulation of nucleic acids: successes, limitations, and promise, *Biopolymers 56*, 232-256.

5. Whitford, P. C., Ahmed, A., Yu, Y., Hennelly, S. P., Tama, F., Spahn, C. M. T., Onuchic, J. N., and Sanbonmatsu, K. Y. (2011) Excited states of ribosome translocation revealed through integrative molecular modeling, *Proc. Natl. Acad. Sci. U. S. A. 108*, 18943-18948.

6. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers, *Biophys. J. 92*, 3817-3829.

7. Banáš, P., Hollas, D., Zgarbová, M., Jurečka, P., Orozco, M., Cheatham, T. E., Šponer, J. i., and Otyepka, M. (2010) Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins, *J. Chem. Theory Comput. 6*, 3836-3849.

8. Zgarbová, M., Otyepka, M., Šponer, J. i., Mládek, A. t., Banáš, P., Cheatham, T. E., and Jurečka, P. (2011) Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles, *J. Chem. Theory Comput. 7*, 2886-2902.

9. Yildirim, I., Stern, H. A., Kennedy, S. D., Tubbs, J. D., and Turner, D. H. (2010) Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine, *J. Chem. Theory Comput. 6*, 1520-1531.

10. Yildirim, I., Kennedy, S. D., Stern, H. A., Hart, J. M., Kierzek, R., and Turner, D. H. (2012) Revision of AMBER torsional parameters for RNA improves free energy predictions for tetramer duplexes with GC and iGiC base pairs, *J. Chem. Theory Comput. 8*, 172-181.

11. Denning, E. J., Priyakumar, U. D., Nilsson, L., and Mackerell, A. D., Jr. (2011) Impact of 2'-hydroxyl sampling on the conformational properties of RNA:

update of the CHARMM all-atom additive force field for RNA, *J. Comput. Chem. 32*, 1929-1943.

12. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009) Long-timescale molecular dynamics simulations of protein structure and function, *Curr. Opin. Struct. Biol. 19*, 120-127.

13. Dong, F., Wagoner, J. A., and Baker, N. A. (2008) Assessing the performance of implicit solvation models at a nucleic acid surface, *Phys. Chem. Chem. Phys. 10*, 4889-4902.

14. Gong, Z., and Xiao, Y. (2010) RNA stability under different combinations of AMBER force fields and solvation models, *J. Biomol. Struct. Dyn. 28*, 431-441.

15. Kelso, C., and Simmerling, C. (2006) Enhanced sampling methods for atomistic simulation of nucleic acids, In *Computational Studies of RNA and DNA* (Šponer, J., and Lankaš, F., Eds.), pp 147-167, Springer Netherlands.

16. Zuckerman, D. M. (2011) Equilibrium sampling in biomolecular simulations, *Annu. Rev. Biophys. Biomol. Struct. 40*, 41-62.

17. Sugita, Y., and Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding, *Chem. Phys. Lett. 314*, 141-151.

18. Mitsutake, A., Sugita, Y., and Okamoto, Y. (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers, *Biopolymers 60*, 96-123.

19. Nymeyer, H., Gnanakaran, S., and Garcia, A. E. (2004) Atomic simulations of protein folding, using the replica exchange algorithm, *Methods Enzymol. 383*, 119-149.

20. Kannan, S., and Zacharias, M. (2009) Folding simulations of Trp-cage mini protein in explicit solvent using biasing potential replica-exchange molecular dynamics simulations, *Proteins: Struct., Funct., Bioinf. 76*, 448-460.

21. Nguyen, P. H., Stock, G., Mittag, E., Hu, C.-K., and Li, M. S. (2005) Free energy landscape and folding mechanism of a β-hairpin in explicit water: a replica exchange molecular dynamics study, *Proteins: Struct., Funct., Bioinf. 61*, 795-808.

22. Paschek, D., Nymeyer, H., and García, A. E. (2007) Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water, *J. Struct. Biol. 157*, 524-533.

23. Periole, X., and Mark, A. E. (2007) Convergence and sampling efficiency

in replica exchange simulations of peptide folding in explicit solvent, *J. Chem. Phys. 126*, 014903.

24. Sanbonmatsu, K. Y., and García, A. E. (2002) Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics, *Proteins: Struct., Funct., Bioinf. 46*, 225-234.

25. Sgourakis, N. G., Merced-Serrano, M., Boutsidis, C., Drineas, P., Du, Z., Wang, C., and Garcia, A. E. (2011) Atomic-level characterization of the ensemble of the Aß(1-42) monomer in water using unbiased molecular dynamics simulations and spectral algorithms, *J. Mol. Biol. 405*, 570-583.

26. Kannan, S., and Zacharias, M. (2010) Application of biasing-potential replica-exchange simulations for loop modeling and refinement of proteins in explicit solvent, *Proteins: Struct., Funct., Bioinf. 78*, 2809-2819.

27. Jimenez-Cruz, C. A., Makhatadze, G. I., and Garcia, A. E. (2011) Protonation/deprotonation effects on the stability of the Trp-cage miniprotein, *Phys. Chem. Chem. Phys. 13*, 17056-17063.

28. Paschek, D., Day, R., and Garcia, A. E. (2011) Influence of water-protein hydrogen bonding on the stability of Trp-cage miniprotein. A comparison between the TIP3P and TIP4P-Ew water models, *Phys. Chem. Chem. Phys. 13*, 19840-19847.

29. Kannan, S., and Zacharias, M. (2007) Folding of a DNA hairpin loop structure in explicit solvent using replica-exchange molecular dynamics simulations, *Biophys. J. 93*, 3218-3228.

30. Kannan, S., and Zacharias, M. (2011) Role of the closing base pair for d(GCA) hairpin stability: free energy analysis and folding simulations, *Nucleic Acids Res. 39*, 8271-8280.

31. Villa, A., Widjajakusuma, E., and Stock, G. (2007) Molecular dynamics simulation of the structure, dynamics, and thermostability of the RNA hairpins uCACGg and cUUCGg, *J. Phys. Chem. B 112*, 134-142.

32. Zuo, G., Li, W., Zhang, J., Wang, J., and Wang, W. (2010) Folding of a small RNA hairpin based on simulation with replica exchange molecular dynamics, *J. Phys. Chem. B 114*, 5835-5839.

33. Garcia, A. E., and Paschek, D. (2007) Simulation of the pressure and temperature folding/unfolding equilibrium of a small RNA hairpin, *J. Am. Chem. Soc. 130*, 815-817.

34. Kirmizialtin, S., and Elber, R. (2010) Computational exploration of mobile ion distributions around RNA duplex, *J. Phys. Chem. B 114*, 8207-8220.

35. Ioannou, F., Archontis, G., and Leontidis, E. (2011) Specific interactions of sodium salts with alanine dipeptide and tetrapeptide in water: insights from molecular dynamics, *J. Phys. Chem. B 115*, 13389-13400.

36. Kwac, K., Lee, K. K., Han, J. B., Oh, K. I., and Cho, M. (2008) Classical and quantum mechanical/molecular mechanical molecular dynamics simulations of alanine dipeptide in water: comparisons with IR and vibrational circular dichroism spectra, *J. Chem. Phys. 128*, 105106.

37. Vymeťal, J. i., and Vondrášek, J. i. (2010) Metadynamics as a tool for mapping the conformational and free-energy space of peptides — The alanine dipeptide case study, *J. Phys. Chem. B 114*, 5632-5642.

38. Li, X., Latour, R. A., and Stuart, S. J. (2009) TIGER2: an improved algorithm for temperature intervals with global exchange of replicas, *J. Chem. Phys. 130*, 174106.

39. Yildirim, I., Stern, H. A., Tubbs, J. D., Kennedy, S. D., and Turner, D. H. (2011) Benchmarking AMBER force fields for RNA: comparisons to NMR spectra for single-stranded r(GACC) are improved by revised x torsions, *J. Phys. Chem. B 115*, 9261-9270.

40. Okur, A., Roe, D. R., Cui, G., Hornak, V., and Simmerling, C. (2007) Improving convergence of replica-exchange simulations through coupling to a high-temperature structure reservoir, *J. Chem. Theory Comput. 3*, 557-568.

41. Roitberg, A. E., Okur, A., and Simmerling, C. (2007) Coupling of replica exchange simulations to a non-Boltzmann structure reservoir, *J. Phys. Chem. B 111*, 2415-2418.

42. Ruscio, J. Z., Fawzi, N. L., and Head-Gordon, T. (2010) How hot? Systematic convergence of the replica exchange method using multiple reservoirs, *J. Comput. Chem. 31*, 620-627.

43. D.A. Case, T.A. Darden, T.E. Cheatham, I., C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, et al. (2012) AMBER 12, University of California, San Francisco.

44. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple AMBER force fields and development of improved protein backbone parameters, *Proteins 65*, 712-725.

45. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys. 79*, 926-935.

46. Banas, P., Sklenovsky, P., Wedekind, J. E., Sponer, J., and Otyepka, M.

(2012) Molecular mechanism of preQ1 riboswitch action: a molecular dynamics study, *J. Phys. Chem. B 116*, 12721-12734.

47. Krepl, M., Zgarbova, M., Stadlbauer, P., Otyepka, M., Banas, P., Koca, J., Cheatham, T. E., 3rd, Jurecka, P., and Sponer, J. (2012) Reference simulations of noncanonical nucleic acids with different chi variants of the AMBER force field: quadruplex DNA, quadruplex RNA and Z-DNA, *J. Chem. Theory Comput. 8*, 2506-2520.

48. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. (1984) Molecular dynamics with coupling to an external bath, *J. Chem. Phys. 81*, 3684.

49. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method, *J. Chem. Phys. 103*, 8577-8593.

50. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, *J. Comput. Phys. 23*, 327-341.

51. Patriksson, A., and van der Spoel, D. (2008) A temperature predictor for parallel tempering simulations, *Phys. Chem. Chem. Phys. 10*, 2073-2077.

52. Hummer, G., Garde, S., García, A. E., Paulaitis, M. E., and Pratt, L. R. (1998) The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins, *Proc. Natl. Acad. Sci. U. S. A. 95*, 1552-1555.

53. Nymeyer, H., Gnanakaran, S., and Garcia, A. E. (2004) Atomic simulations of protein folding, using the replica exchange algorithm, *Methods Enzymol. 383*, 119-149.

54. Sindhikara, D., Meng, Y., and Roitberg, A. E. (2008) Exchange frequency in replica exchange molecular dynamics, *J. Chem. Phys. 128*, 024103.

55. Sindhikara, D. J., Emerson, D. J., and Roitberg, A. E. (2010) Exchange often and properly in replica exchange molecular dynamics, *J. Chem. Theory Comput. 6*, 2804-2808.

56. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1995) Pairwise solute descreening of solute charges from a dielectric medium, *Chem. Phys. Lett. 246*, 122-129.

57. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1996) Parameterized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium, *J. Phys. Chem. 100*, 19824-

19839.

58. Weiser, J., Shenkin, P. S., and Still, W. C. (1999) Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO), *J. Comput. Chem. 20*, 217-230.

59. Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., and Walker, R. C. (2012) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born, *J. Chem. Theory Comput. 8*, 1542-1555.

60. Banas, P., Hollas, D., Zagarbova, M., Jurecka, P., Orozco, M., Cheatham, T. E., III, Sponer, J., and Otyepka, M. (2010) Performance of molecular mechanics force fields for RNA simulations. Stability of UUCG and GNRA hairpins, *J. Chem. Theory Comput. 6*, 3836-3849.

61. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *J. Comput. Chem. 25*, 1605-1612.

62. Tubbs, J. D., Condon, D. E., Kennedy, S. D., Hauser, M., Bevilacqua, P. C., and Turner, D. H. (2013) The nuclear magnetic resonance of CCCC RNA reveals a right-handed helix, and revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics, *Biochemistry 52*, 996-1010.

63. Rosta, E., and Hummer, G. (2009) Error and efficiency of replica exchange molecular dynamics simulations, *J. Chem. Phys. 131*, 165102.

CHAPTER 5

CONCLUSION

In this chapter, we summarize in a two-part fashion the research presented in Chapters 2 – 4. First, the significance of the research is evaluated in the context of current methods. Second, we discuss ongoing projects and consider the long term directions of RNA simulation research.

5.1 Significance

In Chapter 2, we have demonstrated that simulations can provide a robust tool for evaluating structural ensembles from NMR refinements. By studying two domain 5 variants of the group II intron, we revealed that traditional structure refinement procedures can mask restraint violations and structural artifacts. By performing a simulation, rather than selecting structures with the fewest restraint violations, a researcher can monitor the cumulative effect of the restraints within the context of a fully solvated energy potential. In the case of the ai5γ domain 5, three restraints were clearly responsible for creating an unphysical kink in the RNA structure. By monitoring such problematic restraints over the time course of the simulation, a dynamic analysis of the restraints can be considered. This type of end-stage refinement

of NMR ensembles, which includes sophisticated energetic analysis in the context of explicit solvent, will likely expand the range of molecules for which useful structural data can be obtained.

We were able to make similar observations regarding the inhibitor bound HCV IRES using unrestrained simulations. Experience has taught us that the current AMBER RNA force field models with reasonable accuracy local conformational minima, but the relative weighting of each minima might be incorrect. The observation that nearly all of the simulations which started in the NMR conformation became unstable within just a few nanoseconds suggests that the NMR conformation is not near an energy minima. In contrast, the crystal conformation is clearly in an energy minima, as only very subtle changes in its structure occur during simulation. Additionally, we have shown that binding energy estimates based on implicit solvent energy terms yield incorrect results. As an alternative, the use of binding enthalpy, including explicit solvent, appears to be a better estimate of the total free energy. Neglecting entropy appears to be reasonable in this case due to the similarity of the ligands investigated. These results are some of the first for an RNA/small molecule complex and will be instructive for future research as the field grows.

Despite the aforementioned utility, the AMBER RNA force field has additional room to improve. As mentioned earlier, RNA conformations may be sampled at improper frequencies. This is likely due to subtle force field errors. Unfortunately, determining the source of such problems has been difficult due to the extreme time and resource investment required to exhaustively sample

the available conformational space. Our investigation of the tetranucleotide rGACC appears to be ideal for this purpose. Its small size reduces the computational burden of advanced sampling techniques, yet it populates a diverse set of conformations for which many appear to be incorrectly favored by the force field. This combination of size and conformational richness allows us to evaluate emerging algorithms for simulation as well as develop and test new RNA force fields.

5.2 Future Directions

Based on the research presented here, a variety of near term follow-up projects can be considered. In addition to presenting these, many of which are already underway, we also speculate on the longer term direction of RNA MD simulations.

The re-refinement research presented in Chapter 2 suggests that many published NMR ensembles could likely be improved by the use of restrained MD simulation in explicit solvent. This would be unnecessary if the NMR data defining positional and orientational restraints were dense enough that solvent representation and advanced force fields were unnecessary. Unfortunately it is often the case, especially for noncanonical regions of RNA, that the NMR data are somewhat underdetermined. In these situations, traditional NMR refinement techniques can produce unusual and pathological conformations.

Of course, given the known limitations of RNA force fields, it is also true that if the NMR data are too sparse, even restrained simulations will give

anomalous results. Thus it would be of interest to study a well-determined RNA structure which maintains the correct geometry during restrained simulation but deviates significantly during unrestrained simulation. By selectively removing more and more restraints, thereby determining the "transition point" between well-behaved and un-behaved simulations, insight into force field flaws can be uncovered. Additionally, it is also likely the use of time-averaged distance restraints might produce a more realistic ensemble, especially in the case of highly flexible RNA. Such an approach has been used with small furanose models *(1)* and nucleic acid helices *(2, 3)*. However, recent advances in simulation timescales would again make this approach interesting for systems such as an RNA tetranucleotide or small hairpin. The inherent flexibility in these latter systems would provide a more robust test of whether time-averaged restraints produce a more accurate ensemble than traditional refinement approaches. As force fields improve, significantly fewer experimental restraints will be required to generate an accurate conformational ensemble. Eventually, we envision the generation of refined models from simple two-dimensional structure maps of RNA. Explicitly solvated simulations would likely provide the final step in such a structure determination process.

The study on inhibitor binding to the HCV IRES element in Chapter 3 has a variety of additional work already in progress. For example, despite the well-behaved nature of the X-ray structure in simulation, certain aspects of that conformation are not reconcilable with the NMR data. A number of possible

explanations could account for this discrepancy, including multiple binding modes and alternate RNA conformations, and we are actively testing these hypotheses. Additionally, we are studying a variety of novel inhibitor structures in order to identify scaffolds which offer tighter binding than existing inhibitors. These studies represent some of the first steps in computer-aided drug design on RNA targets. Given the growing appreciation for the multi-faceted roles of RNA in controlling cell processes, we anticipate growing interest in small molecule drugs targeting RNA.

Finally, our work on enhanced sampling of a small RNA oligonucleotide in Chapter 4 has enabled us to quickly and quantitatively test new force field modifications. Quantitative results are critical because individual trajectories from conventional MD at laboratory temperatures require extraordinary simulation times to converge and can lead to misleading interpretations of the force field. We are currently testing a re-parameterization of the ribose sugar pucker based on comparison to reference quantum mechanical calculations. To build on this work, we will also investigate a variety of other tetranucleotide sequences and build a collaborative effort with NMR experimentalists in order to identify and correct simulation weaknesses. Additionally, we hope to use this sampling method with larger, more structured RNA. In order to avoid complete melting of the structure at high reservoir temperatures, base pair restraints will be used to limit unfolding. This approach will allow the detailed study of flexible loop and bulge regions without the associated cost of complete unfolding. Finally, given our thorough conformational analysis of the

rGACC tetranucleotide, we have begun testing emerging sampling methods including Hamiltonian replica exchange and multidimensional replica exchange. Through comparison with the data in Chapter 4, we can assess the suitability (and potential liabilities) of new sampling methods.

MD simulations of RNA have improved in accuracy while advances in computational hardware are opening new timescales for active research. However, further room for improvement is still evident. In order to maximize the utility of these methods, deep collaborations need to be built between experimental and theoretical approaches. The cross-talk between both approaches will yield lasting insights into biomolecular structure and function.

5.3 References

1. Hendrickx, P. M. S., Corzana, F., Depraetere, S., Tourwé, D. A., Augustyns, K., and Martins, J. C. (2010) The use of time-averaged 3JHH restrained molecular dynamics (tar-MD) simulations for the conformational analysis of five-membered ring systems: methodology and applications, *J. Comput. Chem. 31*, 561-572.

2. Gonzalez, C., Stec, W., Reynolds, M. A., and James, T. L. (1995) Structure and dynamics of a DNA/RNA hybrid duplex with a chiral phosphorothioate moiety: NMR and molecular dynamics with conventional and time-averaged restraints, *Biochemistry 34*, 4969-4982.

3. Isaacs, R. J., and Spielmann, H. P. (2003) Insight into G−T mismatch recognition using molecular dynamics with time-averaged restraints derived from NMR spectroscopy, *J. Am. Chem. Soc. 126*, 583-590.