Enabling Quantitative Analysis of RNA Dynamics through Comparison to Structural Data

Statement of Problem

The ultimate goal of computational modeling is to provide an all-atom look at how biomolecular structures evolve over time. This is severely limited when potentials describing molecules fail in reproducing experimental observables, as is particularly found for ribonucleic acid (RNA) molecules. To address this problem I propose tuning molecular mechanics force fields to better reproduce the conformational dynamics of RNA in their native environment by using very well-refined solution state experimental datasets, including small- and wide-angle X-ray scattering (SAXS and WAXS), solution neutron scattering (SANS), and NMR, derived from model RNA systems.

Background and Significance

The biomolecular function of RNA is intimately related to its structure and dynamics.[1] RNA adopts an enormous range of secondary and tertiary structures despite the minimum variability of nucleic acid bases from which it is polymerized. Sequence-based structure prediction for RNA remains challenging, but is an ongoing area of research since understanding structure-function relationships could help access a wealth of novel, potentially druggable targets.[2,3] Molecular dynamics (MD) simulations provide a way to look at the motions of these molecules across multiple timescales at an atomistic level, directly linking structure, dynamics and function.[4] MD simulations are also an important complement to experimental data, which can lack dynamics (such as in X-ray crystallography), or average out relevant motions to yield a consensus structure (as in NMR). MD can be used to determine relative drug binding affinities for RNA, and explain discrepancies seen in NMR refinement of hairpin and loop structures.[5] With recent advancements in ion models, the solvent dependent structure shift of ribozyme stem loops can be modeled.[6,7] All-atom MD can reversibly sample folding pathways of tetraloop structures, providing critical information about how these smaller motifs fold, giving us insight into the role they play as pervasive motifs in the ribosome.[8] MD has become critical to end-stage structure refinement, and new force fields and solvent models are playing increasingly large roles in refining biomolecular structure and function.

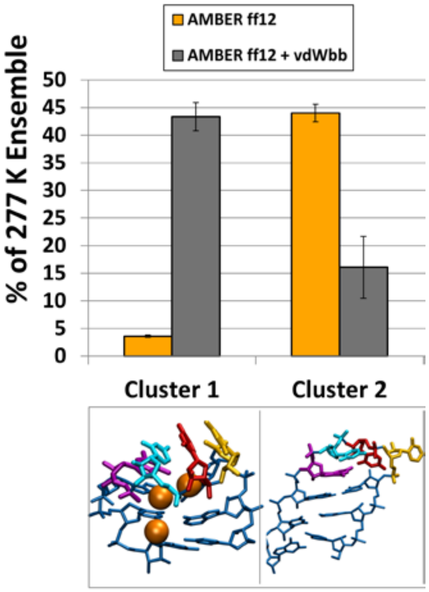
The ability of MD simulations to predict structure-function relationships is limited by two main components: the accuracy of the underlying potential function (or force field), which describes the energy of the system, and the overwhelming sampling required to reach biological timescales.[9,10] It is almost impossible with short MD simulations to determine whether a problem is caused by a flawed force field or by the limited nature of sampling. This leads to inaccurate descriptions of RNA structures, giving little hope of success in identifying potential drug targets. This problem is ubiquitous throughout the community of researchers using MD methods, potentially affecting hundreds of research groups world-wide.

Validating MD force fields against RNA experimental structural studies is currently limited due to the lack of high resolution experimental data. Out of approximately 120,000 experimentally determined macromolecular structures deposited in the Protein Data Bank, less than 3% contain RNA, half of those only as a minor component. Difficulties in crystallization of often flexible RNA architectures and distortions of the native structure due to crystal contacts and tight cation association are common when using X-ray crystallography.[11] On the other hand, derivation of accurate structures of RNA via solution NMR is also complicated by the typically low density of available restraints, which quickly deteriorates further with the increase in the oligonucleotide chain length.[12]

By applying the multi-dimensional replica exchange (M-REMD) enhanced sampling method[13] on large-scale high-performance computing resources, I have previously tested the current nucleic acid force fields by generating highly converged structure populations of several well-studied RNA structures, which can then be directly compared to experimental results, giving us a broad overview of problems within these force fields and a method by which solutions can be tested.[14] M-REMD can be used to efficiently search conformational space in order to generate a well-converged structure ensemble.[15] However, this study showed the limitations in cases that lack highly resolved experimental data, where only qualitative improvement was observed using modified van der Waals (vdW) parameters and newer water models.[16]

The accuracy of the force field parameters is particularly a concern for nucleic acid simulations, where force fields are not as well-tested as they are for proteins, and the high charge and high flexibility of nucleic acid polymers makes generating effective parameters for these systems more challenging.[17] Additionally, the lack of experimental data specifically for RNA limits our ability to generate and validate parameters. RNA tetranucleotides, which are short RNA sequences that approximate hairpin loops and direct recognition of RNA by regulatory proteins[18], have been used as tractable test systems for generating a converged set of conformational substates. The r(GACC) and r(CCCC) tetranucleotide sequences have been refined by solution NMR, providing a data set for comparison to simulation results.[19,20] Results for the r(GACC) tetranucleotide illustrates the ability of independent M-REMD simulations to converge to the same conformational distribution,[15,21] though the conformational distribution does not correctly predict the NMR Major and Minor structures. I extended this study to the UUCG tetraloop, a well resolved loop structure prevalent in the ribosome[22], which has been a notorious problem for MD simulations.[8] For the unmodified Amber12 force field there is an imbalance between dihedral parameters and non-bonded parameters, causing artificial hydrogen bonding between the sugar hydroxyl and the backbone bridging phosphate oxygens (Figure 1, Cluster 2). For

Figure 1. Amber force field structure predictions for UUCG. Orange spheres represent K$^+$ ions.



modified force fields (specifically vdW parameter modifications), an imbalance between non-bonded interactions and the ion model causes artificially high occupancies for K^+ ions in the major groove, disrupting the loop structure (Figure 1, Cluster 1). The results suggest two areas for further exploration: (1) Correcting the balance between chi dihedral and oxygen vdW parameters, and (2) Correcting the balance between solute-solute and solute-solvent non-bonded interactions, including ions.

To improve the force field description quantitatively, very complete structural data is required for a number of systems. Solution X-ray and neutron scattering could become valuable sources of structural restraints due to the high precision and rich information content of these datasets, and these approaches can provide structural data over broad length scales. Application of solution scattering techniques is hindered, however, by the need to accurately describe RNA conformational dynamics and the structure of the solvation shell in order to accurately model the observed scattering profiles.[23] Both of these issues are largely unexplored due to severe scarcity of systems for which both a high density of solution-state experimental (NMR) restraints and high-quality wide-angle solution scattering data sets are available. I propose investigating RNA structure and conformational dynamics for several systems that exhibit highest available densities of solution NMR restraints, including residual dipolar coupling (RDC) restraints which yield precision descriptors of bond vectors relative to a common molecular axis, or a large number of independently determined crystallographic models. The optimized force fields and the resulting model of RNA conformational dynamics will then be applied to investigate structure, conformational dynamics, and effects of ligand binding in larger RNA systems.

Methodology

Refining MD force fields using high level experimental data is an iterative process, and I expect to perform parts of each Specific Aim concurrently.

Specific Aim 1: RNA force field optimization using parameter scanning and experimental restraints. Preliminary results suggest that similar to current ion models, an RNA force field will need to be parametrized with respect to a specific solvent model/ion model combination. For the smaller RNA systems described previously, supplemented by the larger RNA systems described in Specific Aim 2 as data is collected, I propose M-REMD simulations with simultaneous scaling of the vdW radii of non-bridging phosphate oxygens and the oxygen atom of the ribose sugar 2' hydroxyl, as well as the bonded dihedral terms for the glycosidic torsion and sugar pucker, allowing us to select optimal parameters that reproduce experimental conformational distributions. A single simulation having multiple dimensions can be converged in all dimensions, yielding useful data at several different combinations of parameters. This refitting will start from the Amber RNA force field[24–27], using the TIP3P[28] and OPC[29] water models and their respective monovalent ion parameters[30] to generate solvent-model-specific parameter sets.

For larger RNAs, simulations will be performed on a 24-nucleotide stem-loop containing ribosomal helix 35ψ from *E. coli* with approximately 15 RDCs per nucleotide. The conformation of the helical part of the stem-loop was determined via NMR using a single-conformer fit (PDB ID: 2GBH[31]), but the structure of the more dynamic 8-nucleotide loop was undetermined, even though the same density of experimental RDCs is available. The fit, employing the optimized force field, will be performed against experimental RDCs and newly acquired SAXS and WAXS data in H_2O, and SANS data in D_2O for both the 16-nucleotide duplex-only A-form stem construct and the complete 24-nuclotide stem-loop sequence. Both X-ray and neutron scattering data will serve as main sources of experimental restraints for specification of the hydration and counter-ion layer surrounding the oligonucleotide. Neutron scattering data will be useful for systems that exhibit high propensity for radiation damage upon X-ray exposure. Empirical corrections to the force field will be explored to reproduce the NMR and solution scattering data during unrestrained simulations.

Additional systems containing a high density of RDC restraints for which similar analyses will be carried out include 16-nucleotide A-form RNA duplex (PDB ID: 2KYD[32]) and the full length 14-nucleotide cUUCGg tetraloop hairpin (PDB ID: 2KOC[33]). Several other systems with a high number of independently determined crystallographic and NMR models will be explored as well, including the 14-nucleotide acceptor stem microhelix of tRNA^{Ala} from *E. coli*[34,35], the 27-nucleotide sarcin/ricin domain from *E. coli* 23S ribosomal RNA[36], and the 46-nucleotide kissing loop dimer for the HIV-1 type F dimerization initiation site (DIS) RNA.[37,38]

Specific Aim 2: Collection and structural analysis of experimental X-ray and neutron solution scattering data for the selected test RNA systems. Initial efforts in generating very well resolved data will focus on short RNA constructs exhibiting highest densities of NMR restraints, or largest numbers of independently determined crystal structures that could also reveal some aspects of their conformational dynamics and the impact of crystal packing.

The samples containing shorter RNA sequences – truncated helix 35ψ stem and stem-loop (PDB IDs 2GBH[31] & 2LV0[39]), A-form duplex, and the tRNA^{Ala} acceptor stem micro-helix (PDB ID 1IKD[40]), and the sarcin/ricin domain (PDB ID 3DVZ[36]) will be prepared commercially, followed by standard hybridization, annealing, and HPLC purification protocols. HIV-1 DIS kissing complex (PDB ID 1ZCI[37]) is also available commercially and will be purified according to previously published procedures. Solution X-ray and neutron scattering data will be collected at the Advanced Photon Source (Argonne, IL) and NIST Center for Neutron Research (Gaithersburg, MD), respectively.

Expected Results and Significance

To evaluate force field performance, and tune it to quantitatively reproduce experimental data, simulations with different experimentally derived restraints, including NOE (distance) and dihedral angle restraints, J-couplings, and RDCs, can be linked together in single M-REMD runs. The benefit of performing this type of simulation is that multiple dimensions can be added to the same simulation, as computational resources allow. In performing these simulations, the structure ensemble generated by a new force field should fit the experimentally determined restraints. However, if a systematic bias exists, indicating a discrepancy between experimental restraints and the simulated ensemble, these simulations will give insight into the sources of the bias and lead to empirical corrections.

Resulting converged ensembles will be evaluated by cluster analysis of all structures to pinpoint the most populated conformations in each ensemble. The most populated conformation should identify the native structure as a balance between the RNA parameters and solvent/ion parameters is struck across scaled replicas. One possible issue is that performing simulations in multiple dimensions is limited by the available computational resources. This can be overcome by optimizing replica overlap for each parameter scaled, and can go further by breaking the M-REMD simulation down into smaller runs with fewer dimensions as computational resources allow. Simulations can be run on either CPUs or GPUs, using computational resources at the Institute for Bioscience and Biotechnology Research collaboration between NIST and the University of Maryland.

The timeline for completing this research is as follows:

1st year: Aims 1 and 2 can begin concurrently during this time. Optimizing M-REMD simulations with respect to replica spacing and parameter scaling will take place within 6 months. After simulations are performed on small tetranucleotide test systems and the UUCG tetraloop, finalist parameters can be used for larger systems. Additionally, during this year RNA sample preparation, collection of the experimental SAXS/WAXS and SANS data and their analyses for Aim 2 will be completed. Aim 1 simulations on larger RNA systems can begin as data is collected, as described in Aim 2.

2^{nd} **year:** Aim 1 will be completed in this year. SAXS and SANS data analysis of the underlying structural distributions will also be completed within this time frame.

Application

RNA is an essential molecule for all forms of life, playing key roles in protein synthesis, gene transcription, expression, regulation, and viral replication.[41] Because of the ubiquitous nature of RNA, it is turning into a popular cellular target for natural product drugs, as well as rationally designed compounds.[42] Recent research has postulated targeting RNA for overcoming antibacterial resistance, selectively inhibiting specific protein functions, and regulating gene expression.[43–46] The success of these novel methods depends, in part, on a deep understanding of the structure-activity relationship in RNA. Our understanding of proteins' relationship between structure, dynamics, and function is relatively robust, and multiple areas of research address proteins as drug targets. RNA, with its incredible flexibility and the many roles it plays in biology is even less studied than DNA, and our understanding of the balance between its structure and function is still under constant investigation and revision.

Preliminary work has demonstrated a basic failure in most current MD force fields to identify the most energetically favorable structure. Although these force fields are widely used throughout the research community, many of them produce incorrect results. This impacts hundreds of research groups worldwide that rely on these force fields to obtain insight into RNA function and dynamics. I propose methodical, validated ways of searching conformational space for several test systems which are model RNA motifs. Modeling results will be supplemented by the experimental data from a broad array of techniques that characterize RNA at physiological conditions, including solution NMR and X-ray and neutron scattering, for which access to NIST Center for Neutron Research would be critical. This proposal to create a reliable experimental data set for building and validating an RNA force field is complementary to the long standing NIST effort to create measurement standards. Nucleic acid force field development can be advanced by proposing and evaluating modifications which bring force fields in better alignment with experimental data, leading to a better understanding of the interplay between RNA structure, function, and dynamics.

References

(1) Meister, G. *RNA biology : an introduction*; Wiley-VCH: Weinheim, 2011.
(2) Zuker, M. *Nucleic Acids Res.* **2003**, *31*, 3406–3415.
(3) Turner, D. H.; Sugimoto, N.; Freier, S. M. In *Annual review of biophysics and biophysical chemistry*; 1988; pp. 167–192.
(4) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham 3rd, T. E.; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6*, 3836–3849.
(5) Henriksen, N. M.; Davis, D. R.; Cheatham 3rd, T. E. *J. Biomol. NMR* **2012**, *53*, 321–339.
(6) Bergonzo, C.; Hall, K. B.; Cheatham 3rd, T. E. *J. Phys. Chem. B* **2015**, *119*, 12355–12364.
(7) Bergonzo, C.; Hall, K. B.; Cheatham 3rd, T. E. *J. Chem. Theory Comput.* **2016**, *12*, 3382–3389.
(8) Kührová, P.; Banáš, P.; Best, R. B.; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2013**, *9*, 2115–2125.
(9) Bergonzo, C.; Galindo-Murillo, R.; Cheatham 3rd, T. E. In *Current protocols in nucleic acid chemistry*; 2013; Vol. 54, pp. 7.8.1–7.8.21.
(10) Galindo-Murillo, R.; Bergonzo, C.; Cheatham 3rd, T. E. In *Current protocols in nucleic acid chemistry*; 2013; Vol. 54, pp. 7.5.1–7.5.13.
(11) Ke, A.; Doudna, J. A. *Methods* **2004**, *34*, 408–414.
(12) Lipsitz, R. S.; Tjandra, N. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 387–413.
(13) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
(14) Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Cheatham 3rd, T. E. *RNA* **2015**, *21*, 1578–1590.
(15) Roe, D. R.; Bergonzo, C.; Cheatham 3rd, T. E. *J. Phys. Chem. B* **2014**, *118*, 3543–3552.
(16) Bergonzo, C.; Cheatham 3rd, T. E. *J. Chem. Theory Comput.* **2015**, *11*, 3969–3972.
(17) Bergonzo, C.; Galindo-Murillo, R.; Cheatham 3rd, T. E. In *Current protocols in nucleic acid chemistry*; 2013; Vol. 55, pp. 7.9.1–7.9.27.
(18) Jensen, K. B.; Musunuru, K.; Lewis, H. A.; Burley, S. K.; Darnell, R. B. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 5740–5745.
(19) Yildirim, I.; Stern, H. A.; Tubbs, J. D.; Kennedy, S. D.; Turner, D. H. *J. Phys. Chem. B* **2011**, *115*, 9261–9270.
(20) Tubbs, J. D.; Condon, D. E.; Kennedy, S. D.; Hauser, M.; Bevilacqua, P. C.; Turner, D. H. *Biochemistry* **2013**, *52*, 996–1010.
(21) Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Swails, J. M.; Roitberg, A. E.; Cheatham 3rd, T. E. *J. Chem. Theory Comput.* **2014**, *10*, 492–499.
(22) Woese, C. R.; Winkers, S.; Gutell, R. R. *Proc. Natl. Acad. Sci.* **1990**, *87*, 8467–8471.
(23) Koch, M. H. J.; Vachette, P.; Svergun, D. I.; Koch, M.; Svergun, D. *Q. Rev. Biophys.* **2003**, *36*, 147–227.
(24) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
(25) Cheatham 3rd, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.
(26) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham 3rd, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.
(27) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham 3rd, T. E.; Jurečka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

(28) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(29) Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. *J Phys Chem Lett* **2014**, *5*, 3863–3871.

(30) Joung, I. S.; Cheatham 3rd, T. E. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(31) O'Neil-Cabello, E.; Bryce, D. L.; Nikonowicz, E. P.; Bax, A. *J. Am. Chem. Soc.* **2004**, *126*, 66–67.

(32) Tolbert, B. S.; Miyazaki, Y.; Barton, S.; Kinde, B.; Starck, P.; Singh, R.; Bax, A.; Case, D. A.; Summers, M. F. *J. Biomol. NMR* **2010**, *47*, 205–219.

(33) Nozinovic, S.; Fürtig, B.; Jonker, H. R. A.; Richter, C.; Schwalbe, H. *Nucleic Acids Res.* **2010**, *38*, 683–694.

(34) Mueller, U.; Schübel, H.; Sprinzl, M.; Heinemann, U. *RNA* **1999**, *5*, 670–677.

(35) Mueller, U.; Muller, Y. A.; Herbst-Irmer, R.; Sprinzl, M.; Heinemann, U. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1999**, *55*, 1405–1413.

(36) Olieric, V.; Rieder, U.; Lang, K.; Serganov, A.; Schulze-Briese, C.; Micura, R.; Dumas, P.; Ennifar, E. *RNA* **2009**, *15*, 707–715.

(37) Ennifar, E.; Dumas, P. *J. Mol. Biol.* **2006**, *356*, 771–782.

(38) Ennifar, E.; Paillart, J. C.; Bodlenner, A.; Walter, P.; Weibel, J. M.; Aubertin, A. M.; Pale, P.; Dumas, P.; Marquet, R. *Nucleic Acids Res.* **2006**, *34*, 2328–2339.

(39) Nikonowicz, E. P.; Wang, J.; Moran, S.; Donarski, J. Solution Structure of Helix-35 Stem-loop from E. coli 23S rRNA. *To be Published*, 2013.

(40) Ramos, A.; Varani, G. *Nucleic Acids Res.* **1997**, *25*, 2083–2090.

(41) Hermann, T.; Westhof, E. *Curr. Opin. Biotechnol.* **1998**, *9*, 66–73.

(42) Detering, C.; Varani, G. *J. Med. Chem.* **2004**, *47*, 4188–4201.

(43) Blount, K. F.; Breaker, R. R. *Nat. Biotechnol.* **2006**, *24*, 1558–1564.

(44) Milhavet, O.; Gary, D. S.; Mattson, M. P. **2003**, *55*, 629–648.

(45) Amort, T.; Soulière, M. F.; Wille, A.; Jia, X.-Y.; Fiegl, H.; Wörle, H.; Micura, R.; Lusser, A. *RNA Biol.* **2013**, *10*, 1003–1008.

(46) Wahlestedt, C. *Nat. Rev. Drug Discov.* **2013**, *12*, 433–446.